CVA Hedging with Reinforcement Learning

Roberto Daluiso
Intesa Sanpaolo
roberto.daluiso@intesasanpaolo.com

Marco Pinciroli
Intesa Sanpaolo
marco.pinciroli@intesasanpaolo.com

Michele Trapletti
Intesa Sanpaolo
michele.trapletti@intesasanpaolo.com

Edoardo Vittori
Intesa Sanpaolo
edoardo.vittori@intesasanpaolo.com

ABSTRACT

This work considers the problem of a trader who must manage the Credit Valuation Adjustment (CVA) of a derivative, defined as the risk-neutral expectation of losses incurred if the counterparty of the derivative defaults. CVA can be regarded as a hybrid product, one of the most complex actively managed by a trading desk. Standard delta hedging based on sensitivities is not completely satisfactory for this product, because it ignores trading costs and jump-to-default risk while introducing unavoidable simplifications in the pricing model. In this paper we use risk-averse Reinforcement Learning to learn a superior hedging strategy compared to the standard delta hedging approach. Specifically, we generalize risk-averse Reinforcement Learning to stochastic horizons, to be compatible with counterparty defaults, and we introduce a realistic framework for the mechanics of the hedger's portfolio in which the data generating process of the underlying risk drivers can be inconsistent with the risk-neutral laws used to price the CVA and the hedging instruments. The potential of the proposed approach is investigated empirically by numerical examples on hedging the CVA of a forex forward.

CCS CONCEPTS

• **Theory of computation** → **Sequential decision making**; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

deep hedging, risk aversion, reinforcement learning, credit valuation adjustment, foreign exchange, transaction costs, model misspecification

ACM Reference Format:
Roberto Daluiso, Marco Pinciroli, Michele Trapletti, and Edoardo Vittori. 2023. CVA Hedging with Reinforcement Learning. In *4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3604237.3626852

ICAIF '23, November 27–29, 2023, Brooklyn, NY, USA

ACM ISBN 979-8-4007-0240-2/23/11…$15.00
https://doi.org/10.1145/3604237.3626852

1 INTRODUCTION

Given a financial derivative product Π in place between two parties I and C, the Credit Valuation Adjustment (CVA) experienced by I is a measure of the average loss incurred by I due to the fact that C may default when the value of Π is in favour of I. A financial institution is due to adjust Π's book value by the CVA, which enters I's balance sheet, and, since its value depends on the market conditions, contributes to its volatility and has to be actively managed. In this sense, the CVA can be seen as another derivative, built on Π, but it is typically more complex, since its value depends both on the value of Π (which, as a derivative, depends on one or more risk drivers), and on the default probability of C.

CVA risks are usually managed by buying and selling both the underlying risk drivers and Credit Default Swaps (CDS) on the counterparty. The CVA risk can be measured and monitored using its first (and eventually higher) order derivatives with respect to the risk drivers, but this standard practitioner "delta hedge" may be highly suboptimal for a number of misalignments between the risk-neutral CVA pricing setup and reality, including trading costs and model misspecification. Moreover, sensitivity-based hedging addresses by definition only continuous movements of the risk factors, while CVA jumps at C's default time.

In this work we propose to use risk-averse Reinforcement Learning (RL) to learn the CVA hedging strategy and to overcome the mentioned issues.

The theoretical benefits of our approach are manifold:

(1) We describe the hedger's portfolio in a general and realistic way including market frictions, time-discretization and practical details of the market instruments. This is because RL algorithms are data driven and the realism can simply be modelled in the environment.
(2) We decouple the generating process of the underlying risk drivers from the risk-neutral laws used to price the CVA and the hedging instruments, which may even change along the life of the deal (e.g., because of recalibration). This enhances our description of the actual profit-and-loss (PnL), which in practice is calculated according to rules which are not under the control of the trader. Such flexibility enables for instance a numerical investigation of the quality of the RL-optimized hedging strategy in presence of hedging costs and correlations even when the book value of CVA ignores them.
(3) We use a realistic objective function. Indeed, the adopted Trust Region Volatility Optimization (TRVO) [4] algorithm is tailored to the way in which the performance of traders is typically monitored; in particular, it optimizes a trade-off of return and risk in which the latter is defined by reward

volatility: a *pathwise* measure of PnL variability, instead of a theoretical variance or risk-measure abstractly defined in population sense at an arbitrarily fixed time horizon, as in standard risk-averse control or RL.

(4) We explicitly address the stochasticity of the optimization horizon by an appropriate modification of the TRVO algorithm.

Finally, the potential of our approach is investigated empirically by numerical examples with the objective of hedging the CVA of a forex (FX) forward contract.

1.1 Related literature

Hedging of derivative products has been tackled without machine learning in countless studies. Particularly relevant to our setting are those considering transaction costs or correlation among risk drivers.

As for transaction costs, some authors just postulate the rules to build and dynamically adjust the hedging portfolio, and then concentrate on quantifying the impact on pricing: the earliest attempt in this direction was probably Leland [27], followed by Dewynne et al. [18] and many others up to the recent Burnett [7]. Other authors try to optimize the hedging strategy by stochastic control tools, getting either numerical or asymptotic solutions: e.g. Davis et al. [17], Hodges and Neuberger [22], Whalley and Wilmott [42], Zakamouline [43] just to mention a few. The objective is usually utility maximization, often defined on terminal wealth, although some authors try definitions based on the local PnL [e.g. 25] which are closer in spirit to our approach.

As for the impact of correlation, the greatest attention was drawn by the specific "shadow-delta" case of co-movements of an asset and its volatility [1–3, 14, 23, 39]. A generalization to a generic set of factors with any cardinality was described in Daluiso and Morini [15], which is unusually close to our research in the choice of the numerical example: indeed, the case study is CVA hedging, although for an interest rate swap.

Turning to machine learning, hedging has been constantly mentioned since the earliest applications to finance [24, 28], but it was not until recently that Reinforcement Learning was attempted with hedging as the primary focus. There is a major distinction between those optimizing a risk measure defined on final performance, and those which use a pricing model to define a daily performance. The first family includes the "deep hedging" series: Buehler et al. [6], Mikkilä and Kanniainen [30], Murray et al. [32]. The present work belongs to the second family and is particularly aligned with the point of view of Vittori et al. [41] and Mandelli et al. [29]. It has also a relevant aspect in common with Cao et al. [9, 10] in that they discuss the possibility to price with a model and simulate with a different one. Further papers which work on the the daily PnL include Cannelli et al. [8], Du et al. [19], Halperin [20, 21], Kolm and Ritter [26], Vittori et al. [40].

2 REINFORCEMENT LEARNING AND ITS APPLICATION TO HEDGING

Reinforcement Learning [36] is a machine learning framework for sequential decision-making processes, and as such we deem it suitable for our CVA hedging problem.

2.1 Reinforcement learning definitions

The basic building block to apply RL algorithms to a problem is a description of the latter as a Markov Decision Process (MDP) [34].

Definition 2.1 (Markov Decision Process). *A discrete-time MDP is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{P}(\cdot|s,a)$ is a Markovian transition model that assigns to each state-action pair (s,a) the probability of reaching the next state s', $\mathcal{R}(s,a)$ is a bounded reward function, $\gamma \in [0,1)$ is the discount factor, and μ is the distribution of the initial state. The policy of an agent is characterized by $\pi(\cdot|s)$, which defines for each state s an action with a probability distribution over the action space.*

We consider finite horizon problems in which future rewards are exponentially discounted with γ. Let us define a trajectory as a sequence of states, actions, and rewards, up to a stopping time ε:

$$(s_0, a_0, r_1, s_1, a_1, r_2, ..., s_{\varepsilon-1}, a_{\varepsilon-1}, r_\varepsilon).$$

Remark 2.1 (Termination time). *The episode termination timestep ε can be modelled without loss of generality as the first time at which the state s_ε would enter an absorbing termination region $\mathcal{T} \subset \mathcal{S}$, so that its law is included in the definition of $\mathcal{P}$. We suppose that $\varepsilon > 0$; moreover, if $\gamma = 1$, we require that ε is almost surely finite for every choice of π.*[1]

We then define the discounted sum of the rewards of a trajectory:

$$\mathcal{G} = \sum_{i=1}^{\varepsilon} \gamma^{i-1} r_i.$$

The objective in *risk-neutral* RL is the maximization of the expected return, given an initial state distribution:

$$\hat{J}_\pi \coloneqq \mathop{\mathbb{E}_\pi}_{s_0 \sim \mu} [\mathcal{G}].$$

We also introduce its normalized version as:

$$J_\pi \coloneqq \mathop{\mathbb{E}_\pi}_{s_0 \sim \mu} \left[\Gamma^{-1} \mathcal{G}\right], \qquad \Gamma = \sum_{i=1}^{\varepsilon} \gamma^{i-1}.$$

Remark 2.2 (Differences from infinite horizon). *The normalization factor Γ is chosen so that the weighting $\left(\gamma^{i-1}\Gamma^{-1}\right)_{i=1,\dots,\varepsilon}$ is a probability measure on the set of time steps $\{1,\dots,\varepsilon\}$. The original definition was written for an infinite horizon problem, hence it required $\gamma < 1$ and used $\Gamma = (1-\gamma)^{-1}$; the above is adapted to a finite random horizon, allowing for $\gamma \leq 1$ and giving*

$$\Gamma = \frac{1-\gamma^\varepsilon}{1-\gamma} \text{ if } \gamma < 1, \qquad \Gamma = \varepsilon \text{ if } \gamma = 1.$$

2.2 Risk-aversion in reinforcement learning

A number of modified risk-aware objectives have been studied, for example introducing a trade-off with the minimization of variance of the returns, in a mean-variance [33, 38] or Sharpe ratio [31] fashion. Others, such as Tamar et al. [37], have studied the minimization of CVaR or more generally of a coherent risk measure.

Nevertheless, all these approaches consider only the minimization of the long-term risk, while in financial trading interim results are also fundamental, and keeping a low-varying intermediate PnL

[1] For $\gamma < 1$ one can be easily weaken the requirement thanks to the exponential decay of γ^N, but this is irrelevant for our purposes.

becomes crucial. Moreover, the analytical intractability of these formulations does not allow the related algorithms to perform (in terms of learning improvements) as the state-of-the-art algorithms in the risk-neutral RL framework, such as Trust Region Policy Optimization (TRPO) [35]. For these reasons, Bisi et al. [4] introduced a new risk measure, which takes into account the variance of the reward at each time-step with respect to state visitation probabilities:

Definition 2.2 (Unnormalized reward volatility). *The unnormalized reward volatility a.k.a. unnormalized reward variance is expressed as:*

$$\hat{v}_\pi^2 = \mathop{\mathbb{E}_\pi}_{s_0\sim\mu}\left[\sum_{i=1}^{\varepsilon}\gamma^{i-1}\left(r_i - J_\pi\right)^2\right].$$

Remark 2.3 (Differences from infinite horizon). *A factor Γ^{-1} inside the expected value would make $\hat{v}_\pi^2$ a true variance across population and time, giving the (normalized) reward volatility v_π^2 defined in the original infinite-horizon formulation. Here we remove that factor because we believe that with stochastic episode lengths, an underweighting of returns belonging to longer episodes would be financially inappropriate. Note that when Γ is deterministic as in the original paper, putting or removing a multiplier in the definition of v_π^2 is just a matter of notation, since it can be absorbed in the risk-aversion coefficient β defined here below.*

In most trading and even hedging applications, achieving a profit is at least as relevant as being risk-averse thus, we decide to consider an objective that handles the risk-return trade-off through a risk-aversion coefficient, the parameter β. The objective related to the policy π can be defined as:

$$\hat{\eta}_\pi := \hat{J}_\pi - \beta\hat{v}_\pi^2,$$

called *mean-volatility* hereafter, where $\beta \geq 0$ allows to trade-off expected return maximization with risk minimization. Once more, the original normalized goal $\eta_\pi := J_\pi - \beta v_\pi^2$ [4] is just a multiple of ours in deterministic horizon settings. On the other hand, the standard mean-variance objective is $\hat{J}_\pi - \beta\sigma_\pi^2$, where σ_π is the *return* variance as defined in Tamar and Mannor [38]:

$$\sigma_\pi^2 := \mathop{\mathbb{E}_\pi}_{s_0\sim\mu}\left[\left(\mathcal{G} - \hat{J}_\pi\right)^2\right].$$

Remark 2.4 (Time inconsistency). *The optimal policy for the mean-volatility objective may be time-inconsistent, i.e. it is not guaranteed that it is also optimal for the* conditional *mean-volatility as seen from future states. While non-standard objectives have been proposed to fix this seemingly paradoxical effect [see e.g. 12], we point out that time-inconsistency is also generated by more traditional objectives based on return variance or risk measures, and has not prevented their adoption in practice.*

It is possible to generalize the TRVO algorithm and proofs to finite and stochastic horizons. We have used the generalized version for the experiments in this paper, but have not included the theory in the interest of space.

3 FINANCIAL ENVIRONMENT

3.1 Mathematical setting

The financial universe is described by the following objects:

- $\mathscr{H}$: the set of assets that can appear in the trading book; in general, it includes the CVA that must be hedged and a finite collection of hedging and funding instruments.
- $\mathscr{C}$: the set of currencies which appear in the book, where $\bar{c}$ is the main currency used to measure the performance of the trader.
- $\bar{t}$: optimization horizon.

The book dynamics are fully specified by any model for the following processes:

- $\phi_t^{c_1c_2}$: fair price in currency c_2 of a unit of currency c_1 for any $c_1, c_2 \in \mathscr{C}$.
- X_t^h: book valuation (price) of asset h at time t, expressed in an asset specific currency c_h.
- $Y_t^{h,c}$: cumulative cash flows of asset h in currency c which are paid up to time t inclusive.
- N_t^h: units of asset h held at time t.
- $T_t(\boldsymbol{N})$: cumulative costs of trading strategy $\boldsymbol{N}$ up to time t exclusive, expressed in the evaluation currency $\bar{c}$.

Indeed, given the above definitions, one can define the cumulative gains of the book as

$$G_t := \underbrace{\sum_{h\in\mathscr{H}}\int_0^t N_s^h d\bar{X}_s^h}_{\text{capital gains}} + \underbrace{\sum_{h\in\mathscr{H}}\int_0^t N_s^h d\bar{Y}_s^h}_{\text{cash flow gains}} \underbrace{- T_t(\boldsymbol{N})}_{\text{trans. costs}}, \qquad (1)$$

where we introduced the conversion of X_t^h and $(Y_t^{h,c})_{c\in\mathscr{C}}$ into the main evaluation currency $\bar{c}$:

$$\bar{X}_t^h := \phi_t^{c_h\bar{c}} X_t^h,$$
$$\bar{Y}_t^h := \sum_{c\in\mathscr{C}}\int_0^t \phi_s^{c\bar{c}} dY_s^{h,c} = \sum_{c\in\mathscr{C}}\int_0^t \left(\phi_{s-}^{c\bar{c}} dY_s^{h,c} + d[\phi^{c\bar{c}}, Y^{h,c}]_s\right).$$

3.1.1 Constrained and free variables. The set of admissible allocations $\boldsymbol{N}$ is subject to a set of financial constraints listed below:

(1) no CVA transfer: the CVA remains in the book for the whole optimization horizon. This translates into the constraint $N_t^{\text{CVA}} \equiv 1$.
(2) Self financing: no cash is injected or withdrawn. In particular the variations in the bank account $\bar{h}$ in currency $\bar{c}$ equal the gains or losses deriving from the other assets.

We conclude that the free variables for optimization are the quantity of the hedging instruments to be held:

$$\boldsymbol{N}_t^{\mathscr{F}} := (N_t^h)_{h\in\mathscr{F}}, \quad \text{where} \quad \mathscr{F} := \mathscr{H} \setminus \{\text{CVA}, \bar{h}\}.$$

3.2 Financial instruments: price and dividend processes

This section specifies the book value processes X^h, the dividend processes $Y^{h,c}$, and the cost process $T = T(\boldsymbol{N})$.

3.2.1 Credit Valuation Adjustment. We identify by $h = \text{CVA}$ the book item representing the CVA to be hedged.

The price process X_t^{CVA} is the risk-neutral t-conditional expectation $\mathbb{E}_t^{\mathbb{Q}_t}$ of the loss $\text{LGD}_\tau < 0$ recorded at the counterparty default time τ [11]:

$$X_t^{\text{CVA}} = \mathbb{1}_{\tau>t}\mathbb{E}_t^{\mathbb{Q}_t}\left[D(t,\tau)\text{LGD}_\tau\right]$$

computed with some pricing model $\mathbb{Q}_t$ and discount factor $D(t,\tau)$; note that X_t^{CVA} is by convention a non positive quantity. We allow the model $\mathbb{Q}_t$ to depend on time inconsistently with the real-world dynamics of LGD_τ, e.g. because of periodic model recalibration, since pricing rules for a complex object like CVA typically give up maximum realism due to model risk and computational feasibility concerns.

The dividend process Y^{CVA} should be a jump process with a single random negative jump at default time τ equal to the loss given default LGD_τ.

In Section 5, we consider the CVA of an uncollateralized portfolio consisting of a single FX forward, namely an agreement to receive at a future time $t' > t$ an amount N^{c_1} in a currency c_1 paying an amount $N^{\bar{c}}$ in currency $\bar{c}$. We suppose that the loss at default LGD_τ is a fixed fraction $(1 - \text{Rec})$ of the positive value $\max(E_\tau, 0)$, of the contract, where $E_t = \phi_t^{c_1\bar{c}} P_t^{c_1}(t') N^{c_1} - P_t^{\bar{c}}(t') N^{\bar{c}}$ for suitable deterministic interest rate curves P^{c_1}, $P^{\bar{c}}$. The pricing rule $\mathbb{Q}_t$ assumes an instantaneous default probability process λ, independent of $\phi^{c_1\bar{c}}$.

3.2.2 Credit Default Swaps. We suppose that $\mathscr{H}$ can include Credit Default Swaps (CDS) on the derivative counterparty. By convention, we decide that $N_t^h > 0$ for such instruments indicates that the agent has a long position with respect to the risk (he sold protection). The price process X_t^h is the upfront price. Denoting by τ the default time of the counterparty, the dividend process is a step function with positive jumps equal to the quarterly coupons C on a schedule T_i^h contingent to counterparty survival, and a negative jump equal to the protection flow $(1 - \text{Rec})$ at default τ.

3.2.3 Cash accounts. We call "cash account" any element $b \in \mathscr{H}$ such that $X_t^b \equiv 1$ yielding a continuous flow $dY_t^b = r_t^b dt$. We suppose that $\mathscr{H}$ includes a funding account $f(c) \in \mathscr{H}$ for each currency $c \in \mathscr{C}$. Recall that the home funding $f(\bar{c})$ was already introduced in section 3.1.1 with the notation $\bar{h}$.

3.2.4 Rebalancing costs. To model transaction costs, we suppose that the following market operations are used to rebalance the portfolio:

- non-cash assets ($h \neq f(c)$ for all $c \in \mathscr{C}$) can be exchanged for cash in the asset currency c_h, paying a unit cost γ_t^h. The cumulated $\bar{c}$-converted cost of such operations is:
$$\bar{T}_t^h := \int_0^t \phi_s^{c_h\bar{c}} dT_s^h = \int_0^t \phi_s^{c_h\bar{c}} \gamma_s^h |dN_s^h|.$$
- Foreign funding accounts $f(c)$ for $c \neq \bar{c}$ can be converted to domestic cash $\bar{h}$ with spot operations on the foreign exchange market, paying a unit cost γ_t^c. The cumulated notional F_t^c in c currency of such operations up to time t excluded must be computed net of all other c denominated cash flows, so it satisfies by definition
$$N_{t-}^{f(c)} = F_t^c - \sum_{\substack{h \in \mathscr{H} \setminus \{f(c)\} \\ c_h = c}} T_t^h + \sum_{h \in \mathscr{H}} \int_0^{t-} N_s^h dY_s^{h,c},$$
and generates a cost
$$\bar{T}_t^{f(c)} = \int_0^t \gamma_s^c |dF_s^c|.$$

Finally, we can compute the total transaction costs as
$$T_t(\boldsymbol{N}) := \sum_{h \in \mathscr{H}} \bar{T}_t^h.$$

4 IMPLEMENTATION

This section describes a set of modelling choices we made to test numerically the above approach.

4.1 Financial setting

In this subsection we specify the financial environment with the general notation of Section 3.

4.1.1 Assets and currencies. The hedged CVA is due to a single EURUSD FX forward. The set of currencies which appear in the book are $\mathscr{C} = \{\text{EUR}, \text{USD}\}$ and EUR is considered to be the main evaluation currency, i.e. $\bar{c} = \text{EUR}$.

We suppose that $\mathscr{H}$ includes a CDS on the CVA counterparty with a fixed maturity date, which we identify with the notation $h = \text{CDS}$.[2] The set of assets is therefore
$$\mathscr{H} = \{\text{CVA}, \text{CDS}, f(\text{EUR}), f(\text{USD})\}.$$
Under these assumptions the free set is $\mathscr{F} = \{\text{CDS}, f(\text{USD})\}$, which we call informally "hedging assets". We also assume that interest rates are zero for all cash accounts, and so are their dividend processes.

4.1.2 Data generation. The risk drivers used to generate the dataset are the FX rate $\phi_t := \phi_t^{\text{USDEUR}}$ and the default intensity λ_t.

We simulate the FX rate via a Geometric Brownian Motion (GBM)
$$\frac{d\phi_t}{\phi_t} = \sigma^\phi \, dW_t^\phi, \tag{2}$$
with null drift, volatility σ^ϕ, and Wiener process W_t^ϕ.

We simulate the default intensity via the Cox Ingersoll Ross (CIR) model [13]
$$d\lambda_t = k^\lambda(\theta^\lambda - \lambda_t)\, dt + \sigma^\lambda \sqrt{\lambda_t}\, dW_t^\lambda, \tag{3}$$
where k^λ is the mean reversion speed, θ^λ is the long term intensity, σ^λ the volatility, and W_t^λ another Wiener process.

An instantaneous correlation $\rho_{\lambda\phi}^{\mathbb{P}}$ between the stochastic terms dW_t^ϕ and dW_t^λ can be naturally introduced, which affects only mildly the level of the FX rate at the default time, but is sufficient to generate a correlation between the dynamics of the FX rate and the credit spread, whose effects on the hedging problem are among the main topics of this paper.

The CVA pricing model $\mathbb{Q}_t$ assumes the same dynamics (2)-(3), but allows for a different correlation $\rho_{\lambda\phi}^{\mathbb{Q}}$ between the Brownian motions, e.g. $\rho_{\lambda\phi}^{\mathbb{Q}} = 0 \neq \rho_{\lambda\phi}^{\mathbb{P}}$.

FX trading costs are modelled by a constant $\gamma_t^{\text{USD}} = \gamma^{\text{USD}}$. Analogously, once λ_t is simulated, we generate bid and ask default intensities $\lambda^\pm$ by applying a semi-spread γ^λ, i.e. $\lambda_t^\pm = \lambda_t \pm \gamma^\lambda$. Eventually, standard pricing formulas [e.g. 5] map the simulated $\lambda_t^\pm$ to bid and ask prices $X_t^{\text{CDS},\pm}$; we take their midpoint for the book value X_t^{CDS}, and their half-difference for the infinitesimal transaction cost γ_t^{CDS}.

[2] For long trading horizons, one may want to consider a synthetic rolling instrument, to use always the most liquid on-the-run maturity for hedging. In such case, the dividend process should be carefully defined to include the roll costs.

4.2 Reinforcement learning setting

In this subsection we specify the MDP with the general notation of Section 2.1.

4.2.1 State. The state vector includes:

(1) Time t to CVA maturity, in days.
(2) Value of the risk drivers λ_t and ϕ_t.
(3) Current allocation in the hedging assets, expressed without loss of generality by the first order sensitivity of the hedging book to λ_t and ϕ_t.
(4) First order sensitivities of the CVA w.r.t. λ_t and ϕ_t;
(5) First order sensitivity of the CDS w.r.t. λ_t.

Item 1 is included because the hedged object is non-stationary due to its fixed maturity. Items 2 and 3 are a complete status vector for the Markovian dynamics. Items 4 and 5 are functions of (t, λ_t, ϕ_t) which the algorithm may learn by itself, but every modern front office system already calculates first order derivatives of all book items, so there is no reason not to give them to the RL agent as useful pre-engineered features. A similar consideration applies to item 3: one may naively put into the state the notionals N^{CDS} and $N^{f(\text{USD})}$, but sensitivities are universally considered by practitioners as a better representation of risk when trading, so we let the AI trader start from information in this form.

4.2.2 Action. The action vector $\boldsymbol{A}_{t_i}$ should describe the allocations in the interval $(t_i, t_{i+1}]$. For the same reasons as in Section 4.2.1, we express them as the sensitivity of the allocation to λ_t and to ϕ_t.

4.2.3 Reward. Following the definition of the gain process in (1), the signed increment $G_t - G_s$ represents the performance over the time period $[s, t)$ of the hedging strategy, usually referred to as PnL. Traders aim at maximizing such increment and at controlling its variability, not only in distributional sense as the possibility of a large negative PnL over the full trading period $[t_0, \bar{t})$, but also in the time direction: e.g., they cannot accept recording a large loss $G_{\bar{t}/2} \ll 0$, even if it leads eventually to a positive gain $G_{\bar{t}} > 0$ with high probability. All of this suggests that a good description of the trader's objective should consider a *set* of increments over a time grid $t_0 < t_1 < \cdots < t_N = \bar{t}$: $R_{t_{i+1}} = G_{t_{i+1}} - G_{t_i}$, to be maximized in volatility-averse sense. The reward is thus defined as:

$$R_{t_{i+1}} = \sum_{h \in \mathscr{H}} N^h_{t_{i+1}} (\bar{X}^h_{t_{i+1}} - \bar{X}^h_{t_i} + \bar{Y}^h_{t_{i+1}} - \bar{Y}^h_{t_i}) - T_{t_{i+1}}(\boldsymbol{N}) + T_{t_i}(\boldsymbol{N}).$$

Remark 4.1 (Role of the pricing model). *The return of an episode $\mathcal{G} = \sum_{i=1}^{t_N} \gamma^i R_{t_i}$ collapses telescopically to $G_{\bar{t}}$ if $\gamma = 1$. Its maximization is equivalent to the maximization of the eventual profit without risk-aversion; moreover, if the optimization horizon $\bar{t}$ is the maturity of the portfolio, then there is no dependence of the objective on the pricing model $\mathbb{Q}_t$. The latter plays a role only when either $\gamma < 1$ (encoding a preference in PnL timing, be it real or a bias introduced for better algorithm convergence), or when the time distribution of the PnL is part of the risk-aversion as in TRVO.*

4.2.4 Episode termination time. ε corresponds to the earliest between the trading horizon $\bar{t}$ and default time τ (discretized on the time grid).

5 NUMERICAL RESULTS

We collect here empirical evidence on the behaviour of the algorithm and of the optimized policy with different choices of model parameters and algorithmic hyper-parameters.

5.1 Common parameters and assumptions

Unless otherwise specifically stated, we consider the same maturity date for the FX forward and the CDS and we set it equal to 5 years. Furthermore, we choose $N^{\text{USD}} = 1.1$, $N^{\text{EUR}} = 1$, mid FX rate at the pricing date equal to 1, $\sigma^{\phi} = 10\%$. Concerning the CIR model, we calibrate it to so to fit a flat credit curve with fixed spread equal to 1% (100 bps) and an at-the-money payer credit swaption with 1-year expiry and 5-year final maturity with a 50% implied volatility[3], obtaining $\lambda_{t_0} = 1.66\%$, $k^{\lambda} = 0.3769$, $\theta^{\lambda} = 1.87\%$ and $\sigma^{\lambda} = 19.22\%$.

The time grid $t_0 < t_1 < \cdots < t_N$ spans 90 trading days and considers a 2-hours spacing within each trading day, for a total of 5 timesteps per day; note that this realistically implies a non-uniform spacing in calendar time, with larger steps (and market movements) across the nights, and even larger no-action gaps across weekends. The actor and critic in the TRVO algorithm are represented by a neural network with two hidden layers of 10 units each and hyperbolic tangent activation function, and trained with batches of 500 episodes.

Performance metrics are computed with $\gamma = 1$ even though $\gamma < 1$ is used in training to ease convergence.

5.2 Baseline

We consider as baseline an environment with null transaction costs and null correlation between the two Brownian drivers: $\rho^{\mathbb{P}}_{\lambda\phi} = \rho^{\mathbb{Q}}_{\lambda\phi} = 0$.

We neglect the possibility of defaults, which should be very unlikely for an IG CDS considering the episode length. Note that the no-default assumption could in principle induce a bias, since selling (buying) protection on CDS means experiencing a benefit (cost) from the premium leg without a counterbalancing cost (benefit) from the protection leg, thus, inducing an "aggressive" agent to hold a long credit outright position. However, our results show that this bias is essentially immaterial for an IG CDS curve.

With the above assumptions, any trading strategy in the hedging instruments has zero expected value, because we are simulating such assets in a risk-neutral model with null interest rates and transaction costs. Therefore, for any positive level of risk-aversion coefficient β and any value of the discount factor γ, we expect the optimal strategy to just minimize the reward volatility. Given the high frequency of rebalancing, such optimum should be very close to the classical hedge matching first order sensitivities of the CVA.

Therefore, we performed 1500 optimization steps with three different values of β, and with γ fixed to 0.99. The out-of-sample statistics show that the optimized policies perform similar to the delta hedge: the return average and standard deviation differ by max 2%, the reward standard deviation differs by max 1%. The actions are in perfect overlap both in CDS and FX space.

[3]These values are compatible with the market values observable between mid 2022 and mid 2023 for the EURUSD and for a generic investment-grade (IG) CDS curve.

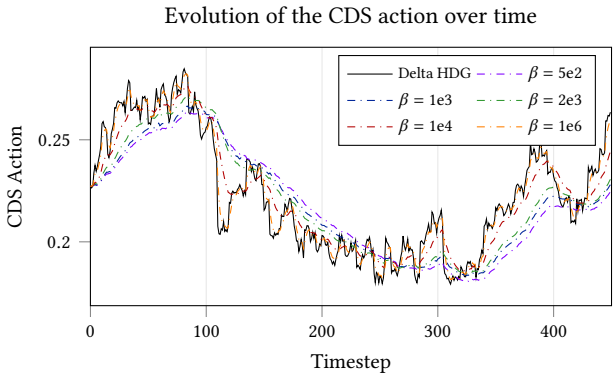

Figure 1: The plot represents the CDS action for an out-of-sample episode, expressed as the sensitivity to λ_t of the hedging portfolio, chosen by the delta hedging strategy and by agents trained at different values of the risk-aversion coefficient β, in an environment that includes transaction costs. In this simulation $\gamma^{\lambda} = 1.66\text{e}{-3}$ (i.e., bid-ask semi-spread of about 10 bps), $\gamma^{\text{USD}} = 5\text{e}{-5}$, and $\rho^{\mathbb{P}}_{\lambda\phi} = 0$.

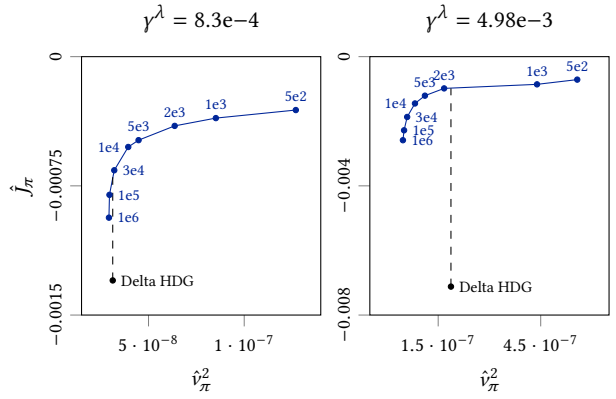

Figure 2: Each dot represents the average performance of an agent over 2000 out-of-sample episodes in terms of return and unnormalized reward volatility, depending on β (annotated next to each dot) and γ^{λ}. The CDS bid-ask semi-spread is about 5 bps in the left plot and 30 bps in the right plot; in each exercise $\gamma^{\text{USD}} = 5\text{e}{-5}$, $\rho^{\mathbb{P}}_{\lambda\phi} = 0$ and CVA at inception equals -0.0712 Eur.

5.3 With transaction costs

In this section, we change the baseline setting of Section 5.2 by including transaction costs.

In this situation the agent improves on delta hedging by trading less frequently, to avoid frequent small movements of the allocation in opposite directions, which would be costly without significant risk reductions. This results in a smoothing of the hedging position, as shown in Figure 1, for a specific testing episode. The smoothing is more pronounced if the risk-aversion is lower, while at high risk-aversion the agent strategy and delta hedging coincide, as previously shown by [41] in an equity environment. Similarly,

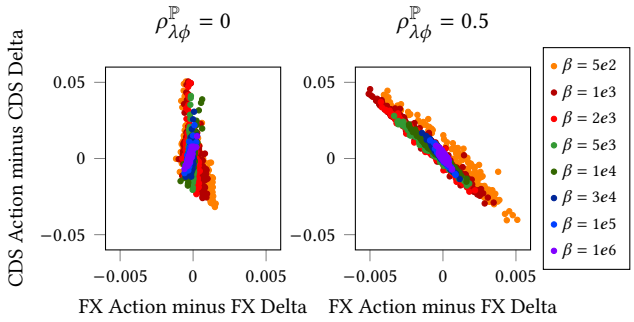

Figure 3: Each dot represents the difference between agent and delta hedging on a simulated market in terms of FX action and CDS action, expressed respectively as the sensitivity to ϕ_t and λ_t of the hedging portfolio, depending on β and $\rho^{\mathbb{P}}_{\lambda\phi}$. In each exercise $\gamma^{\lambda} = 8.3\text{e}{-4}$ (i.e., bid-ask semi-spread of about 5 bps), and $\gamma^{\text{USD}} = 5\text{e}{-5}$.

smoothing is more pronounced if transaction costs are higher; for FX, where realistic costs are extremely low, the agent's action overlaps delta hedging, hence its plot is omitted. The economic effects of the smoothing are summarized in Figure 2: at lower risk-aversions the agent's return is higher, while the reward volatility increases. The obtained frontiers dominate delta hedging already at relatively low costs (see the l.h.s. of Figure 2), at higher costs dominance increases (plot on the r.h.s.).

5.4 With correlation (and costs)

In this section, we consider correlation between the Brownian motions driving λ_t and ϕ_t in the evolution of the risk drivers but not in the CVA pricing formula, as anticipated in Section 4.1.2.

Without transaction costs the first order sensitivities are still close to optimal, so in the figures we represent only results with costs. As we introduce transaction costs, the agent implements the smoothing strategy observed in Section 5.3, but slightly modifies the policy to exploit the expected co-movements of the risk drivers and save on costly rebalances. This can be seen in Figure 3: the figure on the left represents how the actions of the agent differ from the delta hedge with no correlation between the risk drivers and we can observe that there is no linear relationship, thus we obtain a symmetric shape with more dispersion in the CDS direction, which is caused by the fact that the smoothing on the credit actions is more evident due to higher costs. If instead we look at the figure on the right, where $\rho^{\mathbb{P}}_{\lambda\phi} = 0.5$, we can clearly notice a linear relationship, specifically when the agent is over-hedging compared to the delta hedging strategy in terms of credit, then it is under-hedging (w.r.t. delta hedge) on the FX side. This is because, since the two risk drivers are correlated, it is possible to hedge some of the credit risk using FX, and since it is less expensive to trade FX, the agent exploits this correlation to save on trading costs. This improvement is obtained when the CVA pricing is misspecified as a zero-correlation pricer, this is a strong indication that a complete coherence between $\mathbb{P}$ and $\mathbb{Q}$ dynamics is not strictly necessary for the agent to learn an efficient policy.

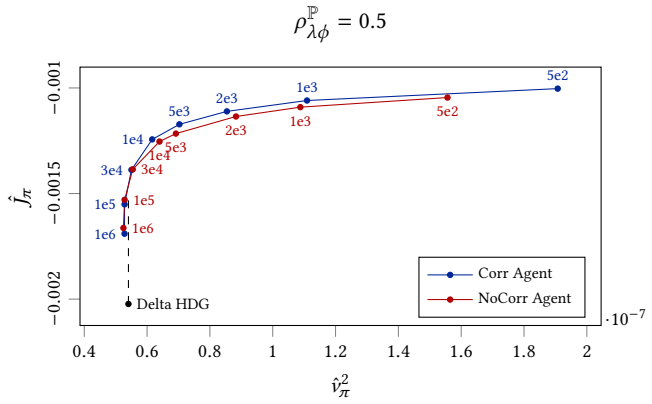

Figure 4: Each dot represents the average performance of an agent over 2000 out-of-sample episodes in terms of return and unnormalized reward volatility, depending on β (annotated next to each dot). In each exercise $\gamma^{\text{USD}} = 5\text{e}{-5}$, $\gamma^{\lambda} = 8.3\text{e}{-4}$ (i.e., bid-ask semi-spread of about 5 bps), $\rho^{\mathbb{P}}_{\lambda\phi} = 0.5$ and CVA at inception equals -0.0712 Eur. The blue dots are built from a policy trained on data with the same correlation $\rho^{\mathbb{P}}_{\lambda\phi} = 0.5$ and red dots from a policy trained on data with $\rho^{\mathbb{P}}_{\lambda\phi} = 0$.

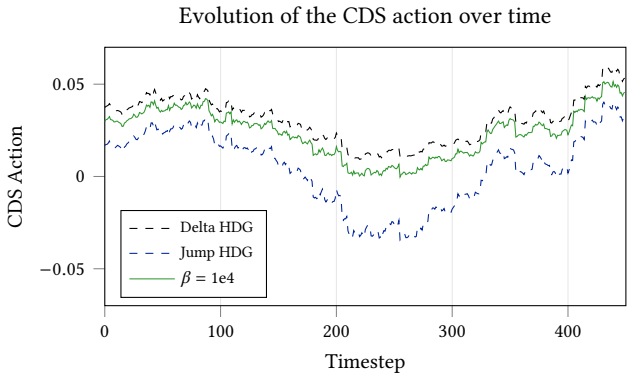

Figure 5: The plots shows the CDS action chosen by the agent and the benchmarks delta hedging and jump hedging, all expressed as the sensitivity to λ_t of the hedging portfolio, for an out-of-sample episode in a zero-cost environment. The initial value for the credit process is $\lambda_0 = 0.22412$, which corresponds to a CDS spread of about 1500 bps.

Figure 4 shows the average results of two different policies with different risk-aversion coefficients (similarly to Figure 2). The environment in which we are testing considers a 50% correlation. The red policy has been trained with no correlation, while the blue policy has been trained with correlation. What we can see is that, as expected, the red policy performs better. On the other hand, this superior performance is only marginal, indicating how the agent is robust to changes in the behavior of the environment and thus of the risk drivers.

5.5 With default

In this section we go back to the zero-cost case and we face counteparties with credit standing spreading between High Yield and distressed. We considered credit curves with fixed spread equal to 500 bps, 1000 bps, 1500 bps, and 3000 bps and we calibrate a CIR model for each of them. These levels are incompatible with a no-default framework, so we adopt the developments of [16] to accommodate a stochastic termination, both in the RL framework (see Section 4.2.4) and in the simulations.

As in Section 5.2, any hedging strategies with the above assumptions have zero expected value. Therefore, for any positive level of risk-aversion coefficient β we expect the optimal strategy to just minimize the reward volatility. In this section we only consider $\beta = 10^4$, which obtained a good convergence in the training phase, but other cases have also been tested with comparable results.

We consider an additional benchmark hedging strategy, defined so that the notional of hedging CDS at a certain time is chosen to perfectly offset the possible loss the FX forward could incur if the counterparty defaults at that time. We call this strategy "jump hedge", since it aims at ensuring that, at default, the loss due to the counterparty risk is perfectly balanced by the CDS protection.

We expect the agent to select a trade-off between the delta hedging and jump hedging strategies since it is only allowed to trade one CDS expiry. In fact, neither of the two strategies is optimal: the former allows for a going concern hedge (i.e., when $\tau > \bar{t}$); the latter gives a perfect hedge only at default time (i.e., if $\tau \leq \bar{t}$). This trade-off is indeed what the trained policy does in terms of CDS actions, as can be observed in Figure 5 for a specific testing episode. The plot of FX actions is instead omitted since it would have shown a perfect overlap between agent's and benchmarks' actions.

Table 1 summarizes the performance of the agent and the benchmarks. It shows that the agent achieves better results in terms of reward volatility. Average return is also shown, even if it should be statistically insignificant and hence irrelevant for the performance measurement.

6 CONCLUSION

Reinforcement Learning is a promising tool to optimize hedging strategies under a realistic financial description, without constraints on analytical tractability and dimensionality typical of classical control approaches. In this paper, we showed how RL can indeed be used for hedging of a CVA, a hybrid product which can depend on several risk factors, including the non-diffusive default indicator.

With this aim, we started by describing a concept of risk-aversion that considers the interest of traders in reducing profit and loss swings in the time direction.

Secondly, we introduced a flexible setup for the description of the CVA hedging mechanics, with almost no assumptions on the underlying stochastic processes. We used this flexibility to vary the environment in the numerical experiments, and study how different elements of realism affect the optimized policy. The tests show that the algorithm converges to theoretical optima when available, while it finds nontrivial improvements to the optima in presence of transaction costs, correlation among the risk drivers, or non negligible probabilities of counterparty default.

Table 1: Performance metrics of the trained agents, delta hedging and jump hedging, in the setting of Section 5.5, computed over 2000 out-of-sample episodes for different levels of the flat CDS spread curve to which the CIR model is calibrated.

Policy	CDS Spread	$\hat{J}_\pi$	$\hat{v}_\pi^2$
TRVO $\beta = 10^4$	500	3.17×10^{-5}	2.78×10^{-6}
Delta hedge	500	3.49×10^{-5}	5.27×10^{-6}
Jump hedge	500	4.56×10^{-5}	3.36×10^{-6}
TRVO $\beta = 10^4$	1000	-0.59×10^{-5}	3.73×10^{-6}
Delta hedge	1000	-1.71×10^{-5}	5.79×10^{-6}
Jump hedge	1000	3.40×10^{-5}	8.00×10^{-6}
TRVO $\beta = 10^4$	1500	-1.37×10^{-5}	4.78×10^{-6}
Delta hedge	1500	-1.77×10^{-5}	6.90×10^{-6}
Jump hedge	1500	0.50×10^{-5}	10.24×10^{-6}
TRVO $\beta = 10^4$	3000	-0.49×10^{-5}	1.40×10^{-6}
Delta hedge	3000	0.22×10^{-5}	1.73×10^{-6}
Jump hedge	3000	2.14×10^{-5}	2.05×10^{-6}

DISCLAIMER

The opinions expressed in this document are solely those of the authors and do not represent in any way those of their present and past employers.

REFERENCES

[1] Carol Alexander and Leonardo M. Nogueira. 2007. Model-free hedge ratios and scale-invariant models. *Journal of Banking & Finance* 31, 6 (2007), 1839–1861. https://doi.org/10.1016/j.jbankfin.2006.11.011

[2] Carol Alexander, Alexander Rubinov, Markus Kalepky, and Stamatis Leontsinis. 2012. Regime-Dependent Smile-Adjusted Delta Hedging. *Journal of Futures Markets* 32, 3 (2012), 203–229. https://doi.org/10.1002/fut.20517

[3] Bruce Bartlett. 2006. Hedging under SABR model. *Wilmott Magazine* 2023, 22 (2 2006), 1–4.

[4] Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, Matteo Papini, and Marcello Restelli. 2020. Risk-Averse Trust Region Optimization for Reward-Volatility Reduction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Virtual Event) *(IJCAI '20)*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 4583–4589. https://doi.org/10.24963/ijcai.2020/632

[5] Damiano Brigo and Fabio Mercurio. 2013. *Interest Rate Models - Theory and Practice*. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-34604-3

[6] Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. 2019. Deep Hedging. *Quantitative Finance* 19, 8 (2019), 1271–1291. https://doi.org/10.1080/14697688.2019.1571683

[7] Benedict Burnett. 2021. Hedging valuation adjustment: fact and friction. *Risk Magazine* (2 2021).

[8] Loris Cannelli, Giuseppe Nuti, Marzio Sala, and Oleg Szehr. 2023. Hedging using reinforcement learning: Contextual k-armed bandit versus Q-learning. *The Journal of Finance and Data Science* 9 (2023), 100101. https://doi.org/10.1016/j.jfds.2023.100101

[9] Jay Cao, Jacky Chen, Soroush Farghadani, John Hull, Zissis Poulos, Zeyu Wang, and Jun Yuan. 2023. Gamma and Vega Hedging using Deep Distributional Reinforcement Learning. *Frontiers in Artificial Intelligence* 6, 1129370 (2023), 1–11. https://doi.org/10.3389/frai.2023.1129370

[10] Jay Cao, Jacky Chen, John Hull, and Zissis Poulos. 2021. Deep Hedging of Derivatives Using Reinforcement Learning. *The Journal of Financial Data Science* 3, 1 (2021), 10–27. https://doi.org/10.3905/jfds.2020.1.052

[11] Giovanni Cesari, John Aquilina, Niels Charpillon, Zlatko Filipović, Gordon Lee, and Ion Manda. 2009. *Modelling, Pricing, and Hedging Counterparty Credit Exposure*. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04454-0

[12] Anthony Coache and Sebastian Jaimungal. 2023. Reinforcement Learning with Dynamic Convex Risk Measures. *Mathematical Finance* (2023), 1–31. https://doi.org/10.1111/mafi.12388

[13] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. 1985. A Theory of the Term Structure of Interest Rates. *Econometrica* 53, 2 (1985), 385–407. https://doi.org/10.2307/1911242

[14] Stéphane Crépey. 2004. Delta-Hedging Vega Risk? *Quantitative Finance* 4, 5 (2004), 559–579. https://www.tandfonline.com/doi/pdf/10.1080/14697680400008718

[15] Roberto Daluiso and Massimo Morini. 2017. Hedging Efficiently under Correlation. *Quantitative Finance* 17, 10 (2017), 1535–1547. https://doi.org/10.1080/14697688.2017.1299201

[16] Roberto Daluiso, Marco Pinciroli, Michele Trapletti, and Edoardo Vittori. 2023. CVA Hedging by Risk-Averse Stochastic-Horizon Reinforcement Learning. (2023). In preparation.

[17] Mark H. A. Davis, Vassilios G. Panas, and Thaleia Zariphopoulou. 1993. European Option Pricing with Transaction Costs. *SIAM Journal on Control and Optimization* 31, 2 (1993), 470–493. https://doi.org/10.1137/0331022

[18] J. N. Dewynne, A. E. Whalley, and P. Wilmott. 1994. Path-Dependent Options and Transaction Costs. *Philosophical Transactions: Physical Sciences and Engineering* 347, 1684 (1994), 517–529. http://www.jstor.org/stable/54362

[19] Jiayi Du, Muyang Jin, Petter N. Kolm, Gordon Ritter, Yixuan Wang, and Bofei Zhang. 2020. Deep Reinforcement Learning for Option Replication and Hedging. *The Journal of Financial Data Science* 2, 4 (2020), 44–57. https://doi.org/10.3905/jfds.2020.1.045

[20] Igor Halperin. 2019. The QLBS Q-Learner Goes NuQLear: Fitted Q Iteration, Inverse RL, and Option Portfolios. *Quantitative Finance* 19, 9 (2019), 1543–1553. https://doi.org/10.1080/14697688.2019.1622302

[21] Igor Halperin. 2020. QLBS: Q-Learner in the Black-Scholes(-Merton) Worlds. *The Journal of Derivatives* 28, 1 (2020), 99–122. https://doi.org/10.3905/jod.2020.1.108

[22] Stewart D. Hodges and Anthony Neuberger. 1989. Optimal Replication of Contingent Claims under Transaction Costs. *The Review of Futures Markets* 8, 2 (11 1989), 222–239.

[23] John Hull and Alan White. 2017. Optimal Delta Hedging for Options. *Journal of Banking & Finance* 82, C (2017), 180–190. https://doi.org/10.1016/j.jbankfin.2017.05.006

[24] James M. Hutchinson, Andrew W. Lo, and Tomaso Poggio. 1994. A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks. *The Journal of Finance* 49, 3 (1994), 851–889. https://doi.org/10.1111/j.1540-6261.1994.tb00081.x

[25] Jan Kallsen. 1999. A Utility Maximization Approach to Hedging in Incomplete Markets. *Mathematical Methods of Operations Research* 50 (10 1999), 321––338. https://doi.org/10.1007/s001860050100

[26] Petter N. Kolm and Gordon Ritter. 2019. Dynamic Replication and Hedging: A Reinforcement Learning Approach. *The Journal of Financial Data Science* 1, 1 (2019), 159–171. https://doi.org/10.3905/jfds.2019.1.1.159

[27] Hayne E. Leland. 1985. Option Pricing and Replication with Transactions Costs. *The Journal of Finance* 40, 5 (12 1985), 1283–1301. https://doi.org/10.1111/j.1540-6261.1985.tb02383.x

[28] Mary Malliaris and Linda Salchenberger. 1993. A Neural Network Model for Estimating Option Prices. *Journal of Applied Intelligence* 3 (9 1993), 193–206. https://doi.org/10.1007/BF00871937

[29] Francesco Mandelli, Marco Pinciroli, Michele Trapletti, and Edoardo Vittori. 2023. *Reinforcement Learning for Credit Index Option Hedging*. arXiv:2307.09844

[30] Oskari Mikkilä and Juho Kanniainen. 2023. Empirical Deep Hedging. *Quantitative Finance* 23, 1 (2023), 111–122. https://doi.org/10.1080/14697688.2022.2136037

[31] John Moody and Matthew Saffell. 2001. Learning to Trade via Direct Reinforcement. *IEEE Transactions on Neural Networks* 12, 4 (2001), 875–889. https://doi.org/10.1109/72.935097

[32] Phillip Murray, Ben Wood, Hans Buehler, Magnus Wiese, and Mikko Pakkanen. 2022. Deep Hedging: Continuous Reinforcement Learning for Hedging of General Portfolios across Multiple Risk Aversions. In *Proceedings of the Third ACM International Conference on AI in Finance* (New York, NY, USA) *(ICAIF '22)*. Association for Computing Machinery, New York, NY, USA, 361–368. https://doi.org/10.1145/3533271.3561731

[33] L. A. Prashanth and Mohammad Ghavamzadeh. 2013. Actor-Critic Algorithms for Risk-Sensitive MDPs. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (Lake Tahoe, Nevada) *(NIPS '13, Vol. 1)*. Curran Associates, Inc., Red Hook, NY, USA, 252—-260. https://dl.acm.org/doi/10.5555/2999611.2999640

[34] Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316887

[35] John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. 2015. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML '15, Vol. 37)*. JMLR.org, Lille, France, 1889–1897. https://dl.acm.org/doi/10.5555/3045118.3045319

[36] Richard S. Sutton. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine learning* 3, 1 (8 1988), 9–44. https://doi.org/10.1007/BF00115009

[37] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. 2017. Sequential Decision Making With Coherent Risk. *IEEE Transactions on Automatic Control* 62, 7 (7 2017), 3323–3338. https://doi.org/10.1109/TAC.2016.2644871

[38] Aviv Tamar and Shie Mannor. 2013. *Variance Adjusted Actor Critic Algorithms.* arXiv:1310.3697

[39] Sami Vähämaa. 2004. Delta hedging with the Smile. *Financial Markets and Portfolio Management* 18, 3 (2004), 241––255. https://doi.org/10.1007/s11408-004-0302-y

[40] Edoardo Vittori, Amarildo Likmeta, and Marcello Restelli. 2021. Monte Carlo Tree Search for Trading and Hedging. In *Proceedings of the Second ACM International Conference on AI in Finance* (Virtual Event) *(ICAIF '21)*. Association for Computing Machinery, New York, NY, USA, Article 37, 9 pages. https://doi.org/10.1145/3490354.3494402

[41] Edoardo Vittori, Michele Trapletti, and Marcello Restelli. 2020. Option Hedging with Risk Averse Reinforcement Learning. In *Proceedings of the First ACM International Conference on AI in Finance* (New York, NY, USA) *(ICAIF '20)*. Association for Computing Machinery, New York, NY, USA, Article 27, 8 pages. https://doi.org/10.1145/3383455.3422532

[42] A. E. Whalley and P. Wilmott. 1997. An Asymptotic Analysis of an Optimal Hedging Model for Option Pricing with Transaction Costs. *Mathematical Finance* 7, 3 (1997), 307–324. https://doi.org/10.1111/1467-9965.00034

[43] Valeri I. Zakamouline. 2005. Optimal Hedging of Options with Transaction Costs. *Wilmott Magazine* 2005, 18 (7 2005), 70–82.