DEEP REINFORCEMENT LEARNING FOR OPTION PRICING AND HEDGING UNDER DYNAMIC EXPECTILE RISK MEASURES

A PREPRINT

Saeed Marzban
HEC Montréal, Montréal,
H3T 2A7, Canada
saeed.marzban@hec.ca

Erick Delage
HEC Montréal, Montréal,
H3T 2A7, Canada
erick.delage@hec.ca

Jonathan Yumeng Li
Telfer School of Management, University of Ottawa,
Ottawa, Ontario K1N 6N5, Canada
jonathan.li@telfer.uottawa.ca

March 27th, 2023

ABSTRACT

Recently equal risk pricing, a framework for fair derivative pricing, was extended to consider dynamic risk measures. However, all current implementations either employ a static risk measure that violates time consistency, or are based on traditional dynamic programming solution schemes that are impracticable in problems with a large number of underlying assets (due to the curse of dimensionality) or with incomplete asset dynamics information. In this paper, we extend for the first time a famous off-policy deterministic actor-critic deep reinforcement learning (ACRL) algorithm to the problem of solving a risk averse Markov decision process that models risk using a time consistent recursive expectile risk measure. This new ACRL algorithm allows us to identify high quality time consistent hedging policies (and equal risk prices) for options, such as basket options, that cannot be handled using traditional methods, or in context where only historical trajectories of the underlying assets are available. Our numerical experiments, which involve both a simple vanilla option and a more exotic basket option, confirm that the new ACRL algorithm can produce 1) in simple environments, nearly optimal hedging policies, and highly accurate prices, simultaneously for a range of maturities 2) in complex environments, good quality policies and prices using reasonable amount of computing resources; and 3) overall, hedging strategies that actually outperform the strategies produced using static risk measures when the risk is evaluated at later points of time.

1 Introduction

Derivative pricing remains to be a challenging problem in finance when the markets are incomplete and the derivatives are dependent on multiple underlying assets. The incompleteness of a market implies that the price of a derivative cannot be uniquely determined by the standard replication argument, as in such a market no self-financing hedging strategy exists that can perfectly replicate the payoffs of the derivative. Most of existing literature tackle this problem by using two different mechanisms. The first one is to exploit a fixed risk-neutral martingale measure, including for example Hull & White (1987); Heston (1993); Amin (1993); Delbaen & Schachermayer (1995), and Brennan (1979), whereas the second set of approaches rely on identifying the indifference price of a risk-averse hedging problem, including for example Jaschke & Küchler (2001); Carr et al. (2001); Föllmer et al. (1985); Schweizer (1996); Gourieroux et al. (1998), and Bertsimas et al. (2001). Nevertheless, most of these approaches are developed from the perspective of a single trader. Unfortunately, a price that is set only according to one party's interest, e.g. a super-replication price that a seller may wish to charge, may not be acceptable to the buyer and thus does not represent a plausible transaction price. Recently, a new pricing scheme, known as Equal Risk Pricing (ERP), was proposed by Guo & Zhu (2017) and further adapted to convex risk measures in the work of Marzban et al. (2022).

The scheme of ERP is built upon the idea of modelling separately the risk exposure of the buyer and the seller of a derivative, and seeking a price that ensures that the risk exposure of both parties is the same under their respective optimal self-financing hedging strategy. The price generated from ERP thus has the merit of fairness to both parties. While ERP has its conceptual appeal, there remains a gap between its general construct and the actual implementation.

In particular, as shown in Marzban et al. (2022), great care must be taken to define properly how risk should be measured in a dynamic hedging setting in order to obtain hedging problems that are operationally meaningful and computationally solvable. The work of Marzban et al. (2022) provides necessary analysis for solving the equal risk pricing and hedging problem based on dynamic programming (DP). It is known however that DP suffers from the issue of the curse of dimensionality, which restricts the applicability of the results in Marzban et al. (2022). In addition, DP assumes the knowledge of a stochastic model that precisely captures the dynamics of the markets, which may not be available in practice.

In the past decade, Deep Reinforcement Learning (DRL) has proven to be a powerful tool for solving dynamic optimization problems when the number of state variables is large and/or when no stochastic model is known for the underlying system dynamics. Traditional DRL methods have been used to find option hedging strategies that minimize the expected replication error using symmetric and asymmetric loss functions (see Fecamp et al. (2021) and Carbonneau (2021)), or that minimize the expected sum of utility of each profit-and-loss increment (see Kolm & Ritter (2019) and Mikkilä & Kanniainen (2023)). Buehler et al. (2019) was the first to propose a DRL approach for hedging to minimize a convex risk measure applied on the terminal wealth. This DRL approach is used in Carbonneau & Godin (2021a) and Carbonneau & Godin (2021b) who are the first that apply DRL to solve ERP problems to price a broad range of over-the-counter options such as basket options. Unfortunately, the DRL approaches described in Buehler et al. (2019) can only be used in settings where the risk is measured according to a static risk measure. This raises the serious issue that the hedging problem exploited by the ERP could be time inconsistent, i.e. the hedging decisions planned for future state of the world may not be considered optimal anymore once the state is visited. The violation of time consistency implies that equal risk prices calculated based on static risk measures will assume a hedging policy that cannot be implemented in practice, and thus are optimistically biased. From a numerical perspective, employing a static risk mesure in ERP also limits the type of DRL algorithms that can be used to solve each party's hedging problem. Specifically, methods based on Buehler et al. (2019) employ a policy optimization scheme, a.k.a. Actor-Only RL (AORL) algorithm (see Williams (1992) as an example of this method), while other approaches such as critic-only or actor-critic algorithms (such as Mnih et al. (2015) and Silver et al. (2014) respectively) that rely on an equivalent DP formulation remain out of reach. Cao et al. (2021) tackles the latter issue by formulating a hedging problem where the portfolio at each period minimizes long term mean-variance of profit-and-loss performance. Unfortunately, such an approach continues to suffer from a form of time inconsistency given that delayed actions are not required to be optimal with respect to earlier views of long term risk averse performance.[1]

In this paper, we seek to develop a DRL approach for solving a class of time-consistent ERP problems under a convex risk measure.[2] It is known that to ensure time consistency, a dynamic risk measure should be employed to measure risk in a recursive fashion. In particular, motivated by the theory of coherent risk measures, which identifies expectile risk measures as the only elicitable coherent risk measures, we propose in this paper the use of dynamic expectile risk measures to formulate time consistent ERP problems. The dynamic nature of risk measures suggests the consideration of an Actor-Critic RL (ACRL) algorithm for solving the hedging problem. It turns out that the elicitability property of expectile risk measures facilitates greatly the design of a model-free ACRL algorithm. The convergence of this algorithm is also greatly improved due to the translation invariance property of the risk measures.

Overall, we may summarize the contribution of this paper as follows:

> We present the first model-free DRL based algorithm for computing equal risk prices that rely on option hedging strategies that are time-consistent. To reinforce the importance of this contribution, we in fact demonstrate using a simple single asset two-period horizon option pricing problem how equal risk prices might suffer from an optimistic bias when static risk measures are used (as in Buehler et al. (2019), Carbonneau & Godin (2021b) and Carbonneau & Godin (2021a)). A side benefit from pricing an option with maturity T using dynamic risk measures will be that we will easily obtain equal risk prices for any other maturity $T^0 < T$.
>
> The ACRL algorithm that we propose is the first model-free DRL algorithm to naturally extend the famous off-policy deterministic actor-critic method presented in Silver et al. (2014) to the risk averse setting. Unlike the ACRL proposed in Tamar et al. (2015) and Huang et al. (2021) for risk-averse DRL, which can employ up to five neural networks, our algorithm will only require two deep neural networks: a policy network (actor) and a Q network (critic). While our policy network will be trained following a stochastic gradient procedure similar to Silver et al. (2014), to the best of our knowledge we are the first to leverage the elicitability property (i.e. existence of a scoring function) of expectile risk measures and to propose a procedure for training the "risk-to-go" Q network that is also based on stochastic gradient descent.

[1] We also note that unlike stated by the authors a weighted sum of mean and standard deviation of asset value does not constitute a coherent risk measure. In particular, the measure used in Cao et al. (2021) does not satisfy the fundamental monotonicity axiom.

[2] Note that in Carbonneau & Godin (2021c), which was made public shortly before this one, the authors deployed a policy optimization algorithm for ERP under an expected utility framework that is time consistent yet violates translation invariance.

We perform a comprehensive evaluation of the training efficiency, quality of option hedging strategies, and quality of equal risk prices obtained using our ACRL algorithm on a synthetic multi-asset geometric Brownian motion market model. In the simple case of vanilla option pricing, we provide empirical evidence that ACRL provides nearly optimal hedging policies, and highly accurate prices, simultaneously for a range of maturities. The latter is in sharp contrast with approaches, such as in Carbonneau & Godin (2021a), that employ time inconsistent risk measures and produce investment strategies that are visibly outperformed by the ACRL strategy in terms of the risk measured as time to maturity reduces. This phenomenon is also observed, although less prominently, in the context of a basket option over 5 underlying assets, where good quality policies and prices are obtained using our ACRL algorithm using a reasonable amount of computing resources.

Remark 1. *While our work was the first that showed how the elicitability property of risk measures can be exploited to design a deep model-free ACRL algorithm for solving time-consistent dynamic problems, two other related works have appeared since its initial public release (see Marzban et al. (2021)). First, Coache & Jaimungal (2022) generalizes the work of Tamar et al. (2015) and Huang et al. (2021) to dynamic **convex** risk measures (instead of coherent ones) yet suffers from the similar issues as Tamar et al. (2015), namely the need for a simulator that can be reinitialized at any given state, which makes it "computationally expensive" or even "impracticable" in real world applications (see Coache et al. (2022)). Second, following a key insight of our work, Coache et al. (2022) addresses this issue by exploiting the conditional elicitability property of **spectral risk measures**. Their ACRL algorithm however differs significantly from ours in two ways: it employs on-policy (rather than off-policy) learning and optimizes a stochastic (instead of deterministic) policy. Both of these algorithmic choices are known to lead to a poor usage of sample data (obtained either from a simulator or actual real world interactions). Indeed, with off-policy learning, samples produced at each epochs of training can be reused (unlike with on-policy) through a replay buffer in all following epochs when computing policy and value estimation updates thus reducing the need for new samples and improving the usage of existing ones (see Mnih et al. (2015)). Silver et al. (2014) further argues that stochastic policy gradient algorithms are data inefficient for tasks with high-dimensional action spaces. Overall, this suggests that our ACRL algorithm has potentially more practical value than those of Coache & Jaimungal (2022) and Coache et al. (2022), albeit being limited in its current form to dynamic expectile risk measures. We leave as future work a formal comparison of the data efficiency of both family of methods.*

The rest of this paper is organized as follows. Section 2 introduces equal risk pricing and illustrates using a simple two-period pricing problem the practical issues related to using static risk measures for option hedging and pricing. Section 3 adapts the ERP framework to the case of a dynamic expectile risk measure and proposes the new ACRL algorithm. Finally, Section 4 presents and discusses our numerical experiments.

2 Equal risk pricing and hedging under coherent risk measures

In this section, we provide a brief overview of ERP under coherent risk measures based on the recent work of Marzban et al. (2022). We pay particular attention to the issue of time (in)consistency by presenting an example that demonstrates numerically that employing a time-inconsistent static risk measure can lead to an under-evaluation of the risk to which each party are actually exposed in practice.

2.1 ERP under coherent risk measures

The problem of ERP can be formalized as follows. Consider a frictionless market, i.e. no transaction cost, tax, etc, that contains m risky assets, and a risk-free bank account with zero interest rate. Let $\mathbf{S}_t : \quad ! \ \mathsf{R}^m$ denote the values of the risky assets adapted to a filtered probability space $(\ , F, \mathsf{F} := fF_t g_{t=0}^T, \mathsf{P})$, i.e. each $\mathbf{S}_t$ is F_t measurable. It is assumed that $\mathbf{S}_t$ is a locally bounded real-valued semi-martingale process and that the set of equivalent local martingale measures is non-empty (i.e. no arbitrage opportunity). The set of all admissible self-financing hedging strategies with the initial capital $p_0 \ 2 \ \mathsf{R}$ is shown by $X(p_0)$:

$$X(p_0) = \left\{ X : \quad ! \ \mathsf{R}^T \,\middle|\, 9f\xi_t g_{t=0}^{T \ 1}, \quad X_t = p_0 + \sum_{t^0=0}^{t \ 1} \xi_{t^0}^{>} \quad \mathbf{S}_{t^0+1}, \quad 8t = 1, \ldots, T \right\},$$

where $\mathbf{S}_{t+1} := \mathbf{S}_{t+1} \quad \mathbf{S}_t$, the hedging strategy $\xi_t \ 2 \ \mathsf{R}^m$ is a vector of random variables adapted to the filtration F and captures the number of shares of each of the risky assets held in the portfolio during the period $[t, \ t+1]$, $\xi_t^{>} \quad \mathbf{S}_{t^0+1}$ is the inner product of the two random vectors, and X_t is the accumulated wealth.

Let $F(f\mathbf{S}_t g_{t=1}^T)$ denote the payoff of a derivative. Throughout this paper, we assume $F(f\mathbf{S}_t g_{t=1}^T)$ admits the formulation of $F(\mathbf{S}_T, \mathbf{Y}_T)$ where $\mathbf{Y}_t$ is an auxiliary fixed-dimensional stochastic process that is F_t-measurable. This class of payoff functions is common in the literature, (see for example Bertsimas et al. (2001) and Marzban et al. (2022)). The

problem of ERP is defined based on the following two hedging problems that seek to minimize the risk of hedging strategies, one is for the writer and the other is for the buyer of the derivative:

$$\text{(Writer)} \qquad \varrho^w(p_0) = \inf_{X \in \mathcal{X}(p_0)} \rho^w(F(\mathbf{S}_T, \mathbf{Y}_T) \quad X_T) \tag{1}$$

$$\text{(Buyer)} \qquad \varrho^b(p_0) = \inf_{X \in \mathcal{X}(\ p_0)} \rho^b(\ F(\mathbf{S}_T, \mathbf{Y}_T) \quad X_T)\,, \tag{2}$$

where ρ^w and ρ^b are two risk measures that capture respectively the writer and the buyer's risk aversion. In words, equation (1) describes a writer that is receiving p_0 as the initial payment and implements an optimal hedging strategy for the liability captured by $F(\mathbf{S}_T, \mathbf{Y}_T)$. On the other hand, in (2) the buyer is assumed to borrow p_0 in order to pay for the option and then to manage a portfolio that will minimize the risks associated to his final wealth $F(\mathbf{S}_T, \mathbf{Y}_T) + X_T$. With equations (1) and (2), ERP defines a fair price p_0 as the value of an initial capital that leads to the same risk exposure to both parties, i.e.

$$\rho^w(p_0) = \rho^b(p_0).$$

Motivated by the theory of coherent risk measures (Artzner et al. (1999)), Marzban et al. (2022) study the ERP problem by imposing the property of coherency to the risk measures ρ^w and ρ^b. Namely, a risk measure is said to be coherent if it satisfies the following five conditions:

Monotonicity: if $X \quad Z \quad a.s.$ then $\rho(X) \quad \rho(Z)$

Subadditivity: $\rho(X + Z) \quad \rho(X) + \rho(Z)$

Positive homogeneity: If $\lambda \quad 0$, then $\rho(\lambda X) = \lambda\rho(X)$

Translation invariance: If $m \in \mathbb{R}$, then $\rho(X + m) = \rho(X) + m$

Normalized risk: $\rho(0) = 0$.

It is well known that Value-at-Risk (VaR), a risk measure commonly applied in financial risk management, is not coherent, whereas its convex counterpart, namely Conditional Value-at-Risk (CVaR) is coherent. The application of CVaR in ERP can be found for example in Carbonneau & Godin (2021b). As one of the key results in ERP, Marzban et al. (2022) establishes that an equal risk price p_0 can actually be found by solving the writer and buyer's hedging problem with no initial payment, i.e. (1) and (2), separately. Namely, it can be calculated by the following result.

Theorem 1 (Proposition 2.1 in Marzban et al. (2022))**.** *Let ρ^w and ρ^b be two coherent risk measures. In the case where the equal risk price p_0 exists, it can be calculated by*

$$p_0 = (\varrho^w(0) \quad \varrho^b(0))/2\,,$$

when $1 > \varrho^w(0) \quad \varrho^b(0) > \quad 1$.

2.2 The issue of time inconsistency

As briefly mentioned in the introduction, measuring risk in a dynamic setting requires additional care. The use of a coherent risk measure, without any further adaptation to a dynamic setting, can lead to solutions that suffer from the issue of time inconsistency. The goal of this section is to carefully demonstrate this point by presenting a numerical example that quantifies the impact of time inconsistency. Our demonstration is inspired by the work of Rudloff et al. (2014), where the impact of time inconsistency is discussed in a portfolio management problem. Here, we present an example based on a vanilla option hedging problem.

In this example, we consider a stock price process modelled by a simple two-stage trinomial tree. Specifically, the horizon spans $t \in \{0, 1, 2\}$ and the probability space $(\ , \mathcal{F}, \mathbb{P})$ is such that $\ = \{\omega_i\}_{i=1}^{9}$, $\mathcal{F}_1 := \sigma(\{\{\omega_i\}_{i=1}^{3}, \{\omega_i\}_{i=4}^{6}, \{\omega_i\}_{i=7}^{9}\})$, and all outcomes are equiprobable. The market contains a risk-free asset (with a risk-free rate of zero) and a risky asset S which are used to hedge a vanilla at-the-money call option on S_2 with strike price $K := S_0$. The details of the price process is shown in Table 1. For simplicity, we set the initial capital for hedging to zero and employ a $\text{CVaR}_{60\%}$ risk measure for hedging.

When hedging the call-option using a static CVaR measure, the writer of the option solves the following two-period optimization model:

$$\min_{\xi_0, \xi_1} \qquad \text{CVaR}_{60\%}((S_2(\omega) \quad K)^+ \quad (S_1(\omega) \quad S_0)\xi_0 \quad (S_2(\omega) \quad S_1(\omega))\xi_1(\omega)) \tag{3}$$

where $(y)^+ := \max(y, 0)$ and $K := S_0$. The optimal solution of this problem will prescribe purchasing 0.93 shares of the risky asset at time 0, i.e. $\xi_0 = 0.9341$, using money borrowed at the risk-free rate (see Table 1 for the optimal

shares to hold at $t = 1$). The resulting $\text{CVaR}_{60\%}$ is 26.36, implying that if the writer charges the buyer with a price above 26.36, the writer would consider the price being sufficient to cover the hedged risk of this call option.

Note that in the risk averse hedging problem (3), it is not clear what motivates the writer of the option to implement the prescribed hedging strategy once new information is revealed at time $t = 1$. In particular, he/she might be curious to compare the prescribed strategy with the strategy that minimizes the CVaR from the new perspective at $t = 1$, i.e., the following hedging problem:

$$\min_{\xi_1} \quad \text{CVaR}_{\alpha_1}((S_2(\omega) \quad K)^+ \quad (S_1(\omega) \quad S_0)\xi_0 \quad (S_2(\omega) \quad S_1(\omega))\xi_1(\omega) j F_1), \tag{4}$$

where $\alpha_1 := 60\%$ and where $\xi_0 = 0.9341$, i.e. the optimal first stage solution in (3).

Table 1 presents the optimal conditional hedging strategy ξ_1 as a function of the information revealed by F_1. While it does appear that ξ_1 agrees with ξ_1 when $\omega \in f\omega_i g_{i=1}^3$, the investment in the risky asset ends up significantly reduced in the other two sets of outcomes. More importantly, we established that in order to motivate the prescribed hedging strategy ξ_1, the risk aversion level used in problem (4) would need to be in the range of $[0.4580,\ 0.4585]$, when $\omega \in f\omega_i g_{i=4}^6$, or $[0.1992,\ 0.2]$, when $\omega \in f\omega_i g_{i=7}^9$. This confirms that ξ is likely to be perceived as overly risky given the information revealed at time $t = 1$. Ultimately, in the likely case where the writer decides to replace ξ_1 with ξ_1, one can establish that the overall exposition to risk from the perspective of $t = 0$ should have rather been estimated to 27.94 instead of 26.36. This implies that employing a static risk measure here underestimated the necessary coverage capital by 6%.

While this issue of time consistency has been discussed significantly in the recent years, a common approach to overcome it is to employ a so-called dynamic risk measure as will be done in the following section. In the context of this example, this would reduce to replacing problem (3) with:

$$\min_{\xi_0,\xi_1} \quad \text{CVaR}_\alpha(\ \text{CVaR}_\alpha((S_2(\omega) \quad K)^+ \quad (S_2(\omega) \quad S_1(\omega))\xi_1(\omega) \quad (S_1(\omega) \quad S_0)\xi_0 j F_1(\omega))\ ), \tag{5}$$

where α can be chosen to characterize the right level of risk aversion for the "dynamic conditional value-at-risk measure". This formulation ensures that the prescribed policy at time $t = 1$ remains optimal (according to problem (4)) at the moment where it is actually implemented thus preventing the necessary coverage capital from being under estimated.

Table 1: Example of a time inconsitent hedging strategy obtained from employing a static risk measure. ξ is obtained by solving problem (3), α_1 is the risk aversion level that motivates ξ_1 at $t = 1$, ξ_1 is the actual investment prescribed by $\text{CVaR}_{60\%}$ at $t = 1$.

Atoms of F_1	Price process			Time inconsistent hedging strategy			Optimal conditional hedging strategy
	$S_0(\omega)$	$S_1(\omega)$	$S_2(\omega)$	ξ_0	$\xi_1(\omega)$	$\alpha_1(\xi\ )$	$\xi_1(\omega)$
$\omega \in f\omega_i g_{i=1}^3$	100	150	$f270,150,75g$	0.9341	0.8718	[0.4580,1.0000]	0.8718
$\omega \in f\omega_i g_{i=4}^6$	100	100	$f180,100,50g$	0.9341	0.7665	[0.4580,0.4585]	0.6154
$\omega \in f\omega_i g_{i=7}^9$	100	80	$f120,80,64g$	0.9341	0.5000	[0.1992,0.2000]	0.3571

3 ERP under dynamic expectile risk measure and an actor-critic algorithm

While time consistent ERP problems can be formulated by employing dynamic risk measures and be calculated, in principle, by solving a set of dynamic programming (DP) equations (Marzban et al. (2022)), there remains the challenge of determining which dynamic risk measure one should employ and how these equations might be solved in high dimension, i.e. multiple underlying assets. In this section, we address the two issues by first motivating the use of dynamic expectile risk measures to formulate time consistent ERP hedging problems and then presenting a Deep Reinforcement Learning approach (DRL) for approximately solving this problems.

3.1 Dynamic expectile risk measures and DP equations

Expectile has been proposed in the recent literature (see Bellini & Bignozzi (2015)) as a replacement of VaR and CVaR given that it is not only coherent but also elicitable. It is known that VaR is not coherent but is elicitable, whereas CVaR is coherent but is not elicitable. A risk measure is said to be elicitable if it can be expressed as the minimizer of a certain scoring function, and this property is found to be critical in practice due to the need of backtesting (Chen, 2018). In fact, expectile is the only elicitable coherent risk measure. Recall the following definition of expectile.

Definition 1 (Bellini & Bernardino (2017))**.** *The τ-expectile of a random liability X is defined as:*

$$\rho(X) := \arg\min_q (1 \quad \tau)\mathsf{E}\left[(q \quad X)_+^2\right] + \tau\mathsf{E}\left[(q \quad X)^2\right],$$

with τ 2 [0.5, 1).

Like CVaR, expectile covers at one extreme the case of risk-neutrality, i.e. with $\tau = 1/2$, and at the other extreme the case of converging towards the worst-case risk, i.e. as $\tau \; ! \; 1$. Thus, expectile also allows for modelling a wide spectrum of risk aversion. Using expectile as the basis, we define its dynamic version as follows (see Ruszczyński & Shapiro (2006) and Pichler et al. (2022) for general definitions of dynamic recursive risk models).

Definition 2. *A dynamic recursive expectile risk measure takes the form:*

$$\rho(X) := \rho_0(\rho_1(\ldots \rho_{T \;\; 1}(X))),$$

where each $\rho(\;)$ is an expectile risk measure that employs the conditional distribution based on F_t. Namely,

$$\rho_t(X_{t+1}) := \arg\min_q (1 \quad \tau)\mathsf{E}\left[(q \quad X_{t+1})_+^2 j\mathsf{F}_t\right] + \tau\mathsf{E}\left[(q \quad X_{t+1})^2 j\mathsf{F}_t\right]$$

where X_{t+1} a random liability measureable on F_{t+1}.

We apply dynamic expecile risk measures to formulate the two hedging problems in ERP. By further imposing the following assumption that there exists a sufficient statistic process ψ_t such that $f(\mathbf{S}_t, \mathbf{Y}_t, \psi_t)g_{t=0}^T$ satisfies the Markov property, we can obtain compact dynamic equations for them.

Assumption 1. *[Markov property] There exists a sufficient statistic process ψ_t adapted to F such that $f(\mathbf{S}_t, \mathbf{Y}_t, \psi_t)g_{t=0}^T$ is a Markov process relative to the filtration F. Namely, $\mathsf{P}((\mathbf{S}_{t+s}, \mathbf{Y}_{t+s}, \psi_{t+s})\; 2\; Aj\mathsf{F}_t) = \mathsf{P}((\mathbf{S}_{t+s}, \mathbf{Y}_{t+s}, \psi_{t+s})\; 2\; Aj\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$ for all t, for all $s \quad 0$, and all sets A.*

In particular, based on Proposition 3.1 and the examples presented in section 3.3 of Marzban et al. (2022), together with the fact that both ρ^w and ρ^b are dynamic recursive expectile risk measures, the Markovian assumption allows us to conclude that the ERP can be calculated using two sets of dynamic programming equations. Specifically, on the writer side, we can define

$$V_T^w(\mathbf{S}_T, \mathbf{Y}_T, \psi_T) := F(\mathbf{S}_T, \mathbf{Y}_T),$$

and recursively

$$V_t^w(\mathbf{S}_t, \mathbf{Y}_t, \psi_t) := \inf_t \rho(\;\; \xi_t^> \;\; \mathbf{S}_{t+1} + V_{t+1}^w(\mathbf{S}_t + \;\; \mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1})j\mathbf{S}_t, \mathbf{Y}_t, \psi_t),$$

where $\mathbf{S}_{t+1} := \mathbf{S}_{t+1} \quad \mathbf{S}_t$ and where $\rho(\; j\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$ is the expectile risk measure that uses $\mathsf{P}(\; j\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$. This leads to considering $\varrho^w(0) = V_0^w(\mathbf{S}_0, \mathbf{Y}_0, \psi_0)$. On the other hand, for the buyer we similarly define:

$$V_T^b(\mathbf{S}_T, \mathbf{Y}_T, \psi_T) := \quad F(\mathbf{S}_T, \mathbf{Y}_T),$$

and

$$V_t^b(\mathbf{S}_t, \mathbf{Y}_t, \psi_t) := \inf_t \rho(\;\; \xi_t^> \;\; \mathbf{S}_{t+1} + V_{t+1}^b(\mathbf{S}_t + \;\; \mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1})j\mathbf{S}_t, \mathbf{Y}_t, \psi_t),$$

with $\varrho^b(0) = V_0^b(\mathbf{S}_0, \mathbf{Y}_0, \psi_0)$. The following lemma summarizes how DP can be used to compute ERP.

Lemma 2. *Under Assumption 1, the ERP that employs dynamic recursive expectile riks measure can be computed as: $p_0 = (V_0^w(\mathbf{S}_0, \mathbf{Y}_0, \psi_0) \quad V_0^b(\mathbf{S}_0, \mathbf{Y}_0, \psi_0))/2$.*

3.2 A novel Expectile-based actor-critic algorithm for ERP

In this section, we formulate each option hedging problem as a finite horizon Markov Decision Process (MDP) described with (S, A, r, P). In this regard, the agent (i.e. the writer or buyer) interacts with a stochastic environment by taking an action $a_t \quad \xi_t \; 2 \; A := [\;\; 1, 1]^m$ after observing the state $s_t \; 2 \; S$, which includes $\mathbf{S}_t$, $\mathbf{Y}_t$, and ψ_t. Note that to simplify exposition, in this section we drop the reference to the specific identity (i.e. w or b) of the agent in our notation. The action taken at each time t results in the immediate stochastic reward that takes the shape of the immediate hedging portfolio return, i.e. $r_t(s_t, a_t, s_{t+1}) := \xi_t^> \;\; \mathbf{S}_{t+1}$ when $t < T$ and otherwise of the option liability/payout $r_T(s_T, a_T, s_{T+1}) := F(\mathbf{S}_T, \mathbf{Y}_T)(1 \quad 2 \;\; \mathbf{1}_{f\text{agent}=\text{writer}g})$, which is insensitive to s_{T+1}. Finally, the Markovian exogeneous dynamics described in Assumption 1 are modeled using P as $P(s_{t+1}js_t, a_t) = \mathsf{P}(\mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1}j\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$. Overall, each of the two dynamic derivative hedging problems presented in Section 3.1 reduce to a version of the following general risk averse reinforcement learning problem:

$$\varrho(0) = V_0(\mathbf{S}_0, \mathbf{Y}_0, \psi_0) = \min_\pi Q_0^\pi(s_0, \pi_0(s_0)),$$

where $\pi : S \quad f0, \dots, Tg \; ! \; A$ is a policy and $s_0 := (\mathbf{S}_0, \mathbf{Y}_0, \psi_0)$ is the initial state in which the option is priced while

$$Q_t^\pi(s_t, a_t) := \rho(\quad r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^\pi(s_{t+1}, \pi_{t+1}(s_{t+1})) j s_t) \,,$$

$Q_T^\pi(s_T, a_T) := \quad r_T(s_T, a_T, s_T)$, and where ρ is an expectile risk measure. Equipped with these definitions, we can now motivate our proposed extension of the model-free off-policy deterministic ACRL algorithm to the general finite horizon risk-averse MDP setting. In doing so, we start with a proposition that will provide the motivation for a stochastic gradient scheme to optimize a policy network, while the optimization of a risk-to-go network will follow from the elicitability property of the expectile risk measure.

Proposition 3. *Let π be an arbitrary reference policy and μ an arbitrary distribution over the initial state s_0, such that there is a strictly positive probability on all of A for each state, and has a strictly positive probability of reaching all of S for all $t \quad 1$ when starting from $s_0 \quad \mu$.*[3] *For any π that satisfies*

$$\pi \quad 2 \arg\min_\pi \mathsf{E}_{\substack{t \quad f0,\dots,Tg,\, s_0 \quad \mu, \\ s_{t+1} \quad P(\, j s_t, \pi_t(s_t))}} [Q_t^\pi(s_t, \pi_t(s_t))] \tag{6}$$

where t is uniformly drawn, we necessarily have that $\pi \quad 2 \arg\min_\pi Q_0^\pi(s_0, \pi_0(s_0))$ hence $\varrho(0) = Q_0^\pi \; (s_0, \pi_0(s_0))$.

Proof. We start by proving first that given any π that satisfies (6), it must also satisfy

$$\pi \quad 2 \arg\min_\pi \mathsf{E}_{(t,s) \quad \beta}[Q_t^\pi \; (s, \pi_t(s))] \,, \tag{7}$$

where β captures the distribution of (t, s_t) used in (6). We do so by contradiction. Let's assume that there exists a $\hat{\pi}$ such that

$$\mathsf{E}_{(t,s) \quad \beta}[Q_t^\pi \; (s, \hat{\pi}_t(s))] < \mathsf{E}_{(t,s) \quad \beta}[Q_t^\pi \; (s, \pi_t(s))] \,.$$

Then, one can design the following policy:

$$\hat{\pi}_t(s) := \begin{cases} \hat{\pi}_t(s) & \text{if } Q_t^\pi \; (s, \hat{\pi}_t(s)) < Q_t^\pi \; (s, \pi_t(s)) \\ \pi_t(s) & \text{otherwise.} \end{cases}$$

Using a recursive argument, one can show that $Q_t^{\hat{\pi}} \; (s_t, a_t) \quad Q_t^\pi \; (s_t, a_t)$ for all t and (s_t, a_t) pair. In this recursion, we start with:

$$Q_T^{\hat{\pi}} \; (s_T, a_T) = \quad r_T(s_T, a_T, s_T) = Q_T^\pi \; (s_T, a_T) \,.$$

Moreover, for all $t < T$, and (s_t, a_t) pairs, we have that:

$$\begin{aligned} Q_t^{\hat{\pi}} \; (s_t, a_t) = \rho(\quad & r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\hat{\pi}}(s_{t+1}, \hat{\pi} \; (s_{t+1})) j s_t) \\ & \rho(\quad r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^\pi(s_{t+1}, \hat{\pi} \; (s_{t+1})) j s_t) \\ & \rho(\quad r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^\pi(s_{t+1}, \pi \; (s_{t+1})) j s_t) = Q_t^\pi \; (s_t, a_t), \end{aligned}$$

where we exploited the monotonicity of expectile risk measures, together with $Q_{t+1}^{\hat{\pi}}(s_t, a_t) \quad Q_{t+1}^\pi(s_t, a_t)$ and the definition of $\hat{\pi}_t$ for the first and second inequality respectively. With this result in hand we can obtain that for all t and s_t

$$Q_t^{\hat{\pi}} \; (s_t, \hat{\pi}_t(s_t)) \quad Q_t^\pi \; (s_t, \hat{\pi}_t(s_t)) \quad Q_t^\pi \; (s_t, \hat{\pi}_t(s_t)),$$

where we again used the definition of $\hat{\pi}$. Finally, we must therefore have that:

$$\mathsf{E}_{(s,t) \quad \beta}[Q_t^{\hat{\pi}} \; (s, \hat{\pi}_t(s))] \quad \mathsf{E}_{(s,t) \quad \beta}[Q_t^\pi \; (s, \hat{\pi}_t(s))] < \mathsf{E}_{(s,t) \quad \beta}[Q_t^\pi \; (s, \pi_t(s))]$$

which leads to a contradiction, hence (7) must hold.

Next, applying the interchangeability property (see Proposition 2.2 of Shapiro (2017)) to equation (7) and using the fact that the β distribution puts positive probability on all time periods and all sub-regions of $S \quad A$, we know that the following necessarily hold:

$$\pi_t(s) \; 2 \arg\min_a Q_t^\pi \; (s, a), \; 8s \; 2 \; S, 8t \; 2 \; f0, \dots, Tg \,.$$

Our last step involves using recursion to show that $\pi \quad 2 \arg\min_\pi Q_t^\pi(s_t, \pi_t(s_t))$ for all t and all s_t. We start once more at $t = T$ where for all s_T:

$$Q_T^\pi \; (s_T, \pi_T(s)) = \min_{a_T} Q_T^\pi \; (s_T, a_T) = \min_{a_T} \quad r_T(s_T, a_T, s_T) \quad Q_T^\pi(s_T, \pi_T(s_T)), \quad 8\pi.$$

[3] In our option hedging problem, given that s_t is entirely exogenous, the distribution of s_{t+1} is unaffected by π, which can therefore be chosen arbitrarily. In more general settings, more care must be put in ensuring that (μ, π) satisfy the minimum state-action visit probability condition.

And then recursively for all $t < T$ and all s_t,,

$$\begin{aligned} Q_t^{\pi}(s_t, \pi_t(s_t)) &= \min_{a_t} Q_t^{\pi}(s_t, a_t) \\ &= \min_{a_t} \rho(r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1})) | s_t) \\ &\quad \min_{a_t} \rho(r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1})) | s_t) \; \forall \pi \\ &\quad \rho(r_t(s_t, \pi_t(s_t), s_{t+1}) + Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1})) | s_t) \; \forall \pi \\ &\quad \min_{\pi} Q_t^{\pi}(s_t, \pi_t(s_t)). \end{aligned}$$

□

In the context of a deep reinforcement learning approach, we can employ a procedure based on off-policy deterministic policy gradient (Silver et al., 2014) to optimize (6). Specifically, given a policy network π^{θ}, we wish to optimize:

$$\min_{\theta} \mathbb{E}_{\substack{t \sim \{0,\dots,T-1\}, s_0 \sim \mu, \\ s_{t+1} \sim P(\cdot | s_t, \pi_t(s_t))}} \left[Q_t^{\pi^{\theta}}(s_t, \pi_t^{\theta}(s_t)) \right],$$

using a stochastic gradient algorithm after assuming that the MDP and π^{θ} satisfy the regularity properties needed for a gradient in θ to exist.[4]

In doing so, we rely on the fact that:

$$\begin{aligned} &\nabla_{\theta} \mathbb{E}_{\substack{t \sim \{0,\dots,T-1\}, s_0 \sim \mu, \\ s_{t+1} \sim P(\cdot | s_t, \pi_t(s_t))}} \left[Q_t^{\pi^{\theta}}(s_t, \pi^{\theta}(s_t)) \right] \\ &\quad = \mathbb{E}_{\substack{t \sim \{0,\dots,T-1\}, s_0 \sim \mu, \\ s_{t+1} \sim P(\cdot | s_t, \pi_t(s_t))}} \left[\nabla_{\theta} Q_t^{\pi^{\theta}}(s_t, a) \Big|_{a = \pi_t^{\theta}(s_t)} + \nabla_a Q_t^{\pi^{\theta}}(s_t, a) \nabla_{\theta} \pi_t^{\theta}(s_t) \Big|_{a = \pi_t^{\theta}(s_t)} \right] \\ &\qquad \mathbb{E}_{\substack{t \sim \{0,\dots,T-1\}, s_0 \sim \mu, \\ s_{t+1} \sim P(\cdot | s_t, \pi_t(s_t))}} \left[\nabla_a Q_t^{\pi^{\theta}}(s_t, a) \nabla_{\theta} \pi_t^{\theta}(s_t) \Big|_{a = \pi_t^{\theta}(s_t)} \right]. \end{aligned}$$

Note that in the above equation, we have dropped the the term that depends on $\nabla_{\theta} Q_t^{\pi^{\theta}}$ as is commonly done in off-policy deterministic gradient methods and usually motivated by a result of Degris et al. (2012), who argue that this approximation preserves the set of local optima in a risk neutral setting, i.e. $\rho(\cdot) := \mathbb{E}[\cdot]$. While we do consider as an important subject of future research to extend this motivation to general recursive risk measures, our numerical experiments (see Section 4.3) will confirm empirically that the quality of this approximation permits the identification of nearly optimal hedging policies.

Given that we do not have access to an exact expression for $Q_t^{\pi^{\theta}}(s_t, a)$, this operator needs to be estimated directly from the training data. Exploiting the fact that ρ is a utility-based shortfall risk measure, we get that:

$$Q_t^{\pi}(s_t, a_t) \in \arg\min_{q} \mathbb{E}_{s_{t+1} \sim P(\cdot | s_t, a_t)} \left[\ell \left(q + r(s_t, a_t, s_{t+1}) - Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1})) \right) \right]$$

where $\ell(y) := ((1-\tau)\mathbf{1}\{y > 0\} + \tau \mathbf{1}\{y \leq 0\}) y^2$ is the score function associated to the τ-expectile risk measure (see Definition 1). As explained in Shen et al. (2014), in a tabular MDP environment one can apply the following stochastic gradient step:

$$\hat{Q}_t(s_t, a_t) \leftarrow \hat{Q}_t(s_t, a_t) - \alpha \partial \ell(\hat{Q}_t(s_t, a_t) + r_t(s_t, a_t, s_{t+1}) - \hat{Q}_{t+1}(s_{t+1}, \pi_{t+1}(s_{t+1}))),$$

where $\partial \ell(y) := 2((1-\tau) \max(0, y) - \tau \max(0, -y))$ is the derivative of $\ell(y)$, within a properly designed Q-learning algorithm and have the guarantee that $\hat{Q}_t(s_t, a_t)$ will almost surely converge to $Q_t^{\pi}(s_t, a_t)$ for all t, s_t, and a_t (see Theorem 2 in Shen et al. (2014)).

In the non-tabular setting, we replace $\hat{Q}_t^{\pi}(s_t, a_t)$ with two estimators: i.e. the "main" network $Q_t^{\pi}(s_t, a_t | \theta^Q)$ for the immediate conditional risk and the "target" network $Q_t^{\pi}(s_t, a_t | \theta^{Q'})$ for the next period's conditional risk. The procedure consists in iterating between a step that attempts to make the main network $Q_t^{\pi}(s_t, a_t | \theta^Q)$ a good estimator of

[4] In fact, given that the expectile risk measure is only sub-differentiable with respect to the random variable, there might exist some values of θ for which only directional derivatives exist. For those values, one should interpret ∇ as Clarke's generalized gradient (see Clarke (1990)). In practice however, this is not an issue as we will replace $Q_t^{\pi^{\theta}}(s, a)$ with a neural network approximation $Q_t(s, a | \theta^Q)$, which can be differentiable by design.

$\rho(\quad r(s_t, a_t, s_{t+1}) + Q^{\pi}_{t+1}(s_{t+1}, a_{t+1} j \theta^{Q'}))$ and a step that replaces the target network $Q^{\pi}_t(s_t, a_t j \theta^{Q'})$ with a network more similar to the main one $Q^{\pi}_t(s_t, a_t j \theta^{Q})$. The former is achieved, similarly as with the policy network, by searching for the optimal θ^Q according to:

$$\min_{\theta^Q} \mathsf{E}_{\substack{t \quad f0,\dots,Tg, s_0 \quad \mu, \\ s_{t+1} \quad P(\, j s_t, \pi_t(s_t))}} [\ell(Q^{\pi}_t(s_t, \pi_t(s_t) j \theta^Q) + r(s_t, \pi_t(s_t), s_{t+1}) \quad Q^{\pi}_{t+1}(s_{t+1}, \pi_{t+1}(s_{t+1}) j \theta^{Q'}))], \tag{8}$$

which suggests a stochastic gradient update of the form:[5]

$$\theta^Q \quad \theta^Q \quad \alpha \partial \ell(Q^{\pi}_t(s_t, \pi_t(s_t) j \theta^Q) + r(s_t, \pi_t(s_t), s_{t+1}) \quad Q^{\pi}_{t+1}(s_{t+1}, \pi_{t+1}(s_{t+1}) j \theta^{Q'})) \, r_{\theta^Q} Q^{\pi}_t(s_t, \pi_t(s_t) j \theta^Q). \tag{9}$$

We recall that the existence of a stochastic gradient update for risk estimation is only possible for elicitable risk measures. Moreover, although all convex utility-based shortfall risk measures satisfy this property, expectiles are known to be the only elicitable one-dimensional **coherent** risk measures (see Bellini & Bignozzi (2015)).

These two types of updates are integrated in our proposed expectile-based actor-critic deep RL (a.k.a. ACRL) algorithm. A first version, Algorithm 1, is designed for a model-based environment where one can simulate trajectories according to $P(\, j s_t, a_t)$ from state s_0. One may note that in each episode, the reference policy π_t is updated to be a perturbed version of the main policy network in order to focus the accuracy of the main critic network $Q(s, a j \theta^Q)$ value and derivatives on actions that are more likely to be produced by the main policy network. We also choose to update the target networks using convex combinations operations as is done in Lillicrap et al. (2015) in order to improve stability of learning. A second more general version of ACRL, designed for data-driven environments with possibly action-dependant state dynamics, is presented in Appendix A and mimics the original DDPG by generating minibatches using a replay buffer. We note that, while the practical value of Algorithm 1 will be demonstrated in Section 4 for the ERP problem, we will leave as future work the question of validating the performance of the more general algorithm presented in the Appendix.

Remark 2. *We note that in our problem, $P(s_{t+1} j s_t, a_t) = P(s_{t+1} j s_t, a^0_t) = \mathsf{P}(\mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1} j \mathbf{S}_t, \mathbf{Y}_t, \psi_t)$, meaning that the action is not affecting the distribution of state in the next period. This is a direct consequence of using a translation invariant risk measure, which eliminates the need to keep track of the accumulated wealth in the set of state variables as explained in Marzban et al. (2022) (see Remark 1 and proof of Proposition 3.1) and allows the reward function to provide an immediate signal regarding the quality of implemented actions. In the context of our deep reinforcement learning approach, we observed in our experiments that empirical convergence speed is improved in training due to this property. Furthermore, the fact that this property makes the dynamics exogenous lifts the need for keeping a replay buffer, which is also known to affect negatively convergence speed.*

Remark 3. *It is worth noting that while there has been a large number of DRL approaches recently proposed to address risk averse MDP using coherent risk measures, to the best of our knowledge all of those that are model-free, except for two exceptions, consider a law invariant risk measure (i.e. a static risk measure) applied on the discounted sum of total rewards (see Castro et al. (2019); Singh et al. (2020); Urpí et al. (2021); Bisi et al. (103765, 2022)). Such methods therefore suffer from the issues identified in Section 2.2. The two exceptions consist of Tamar et al. (2015) and Huang et al. (2021) who propose ACRL algorithms to deal with general dynamic law-invariant coherent risk measures. While being applicable to a wider range of dynamic risk measures, the two algorithms either assume that it is possible to generate samples from a perturbed version of the underlying dynamics, or rely on training three additional neural networks (namely a state distribution reweighting network, a transition perturbation network, and a Lagrangean penalisation network) concurrently with the actor and critic networks. Furthermore, only Huang et al. (2021) was to this date implemented yet only tested on toy tabular problem involving 12 states and 4 actions where it produced questionable performances[6]. While our approach can only be used with the dynamic expectile risk measure, it offers a much simpler implementation that naturally extends DDPG to the risk averse setting. Section 4 will present a real application of this approach on an option hedging problem involving a portfolio of 6 different assets.*

4 Experimental results

In this section we provide two different sets of experiments that are run over one vanilla and one basket option. We will compare both algorithmic efficiency and quality, in terms of pricing and hedging strategies, of the dynamic risk

[5] This update rule is analogous to the update proposed in Shen et al. (2014) when using the "risk-adjusted reward" measure $\varrho(X) := \sup f m j \mathsf{E}[u(X \quad s)] \quad x_0 g$ with $x_0 := 0$ and $u(y) := \min(\tau y, (1 \quad \tau) y) = (1 \quad \tau) \max(0, y) \quad \tau \max(0, \quad y) = (1/2) \partial \ell(y)$.

[6] At the time of writing this paper, the risk averse implementation of this algorithm was unable to recommend an optimal risk neutral policy in a deterministic setting, while the risk neutral implementation produced policies that were outperformed by risk averse ones in a stochastic setting.

Algorithm 1 Actor-critic RL algorithm for the dynamic recursive expectile option hedging problem with known dynamic model (ACRL)

Inputs: number of episodes J, learning rates $\{\alpha_j^{\pi}\}_{j=1}^{J}$, $\{\alpha_j^{Q}\}_{j=1}^{J}$, and γ, mini-batch size N, Expiration time of the option T

Randomly initialize the main actor and critic networks' parameters θ^{π} and θ^{Q}

Initialize the target actor and critic networks: $\theta^{\pi^0} \leftarrow \theta^{\pi}$, $\theta^{Q^0} \leftarrow \theta^{Q}$

for $j = 1 : J$ **do**

Randomly select $t \in \{0, 1, ..., T-1\}$

Sample a minibatch of N triplets $\{(s_t^i, a_t^i, s_{t+1}^i)\}_{i=1}^{N}$ from $P(\cdot|s_t, \pi_t(s_t))$, where

$$\pi_t(s_t) := \pi_t(s_t|\theta^{\pi}) + \mathcal{N}(0, \sigma)$$

Set the realized losses y_t^i as

$$r_t(s_t^i, a_t^i, s_{t+1}^i) + Q_{t+1}(s_{t+1}^i, \pi_{t+1}(s_{t+1}^i|\theta^{\pi^0})|\theta^{Q^0})$$

Update the main critic network θ^Q as:

$$\theta^{Q} \leftarrow \alpha_j^{Q} \frac{1}{N} \sum_{i=1}^{N} \partial \ell(Q_t(s_t^i, a_t^i|\theta^{Q}) - y_t^i) \nabla_{\theta^{Q}} Q_t(s_t^i, a_t^i|\theta^{Q})$$

Update the main actor network θ^{π} as:

$$\theta^{\pi} \leftarrow \alpha_j^{\pi} \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q_t(s_t^i, a|\theta^{Q})|_{a=\pi_t(s_t^i|\theta^{\pi})} \nabla_{\theta^{\pi}} \pi_t(s_t^i|\theta^{\pi})$$

Update the target networks:

$$\begin{aligned} \theta^{Q^0} &\leftarrow \gamma\theta^{Q} + (1-\gamma)\theta^{Q^0}, \\ \theta^{\pi^0} &\leftarrow \gamma\theta^{\pi} + (1-\gamma)\theta^{\pi^0} \end{aligned} \tag{10}$$

end for

model (DRM), which employs a dynamic expectile risk measure and is solved using our new ACRL algorithm, and the static risk model (SRM), which employs a static expectile measure and is solved using an AORL algorithm similar to Carbonneau & Godin (2021a). All experiments are done using simulated price processes of five risky assets: AAPL, AMZN, FB, JPM, GOOGL. The price paths are simulated using correlated Brownian motions considering the empirical mean, variance, and the correlation matrix of five reference stocks (APPL, AMZN, FB, KPM, and GOOGL) over the period that spans from January 2019 to January 2021. The vanilla option will be over AAPL while the basket option will contain all five stocks. In both cases, the maturity of the option will be one year and the hedging portfolios will be rebalanced on a monthly basis. Table 2 provides the descriptive statistics of our underlying stochastic process.[7]

Table 2: Stock data including the mean, standard deviation, and the correlation matrix

	AAPL	AMZN	FB	JPM	GOOGL
S_0	78.81	1877.94	221.77	137.25	1450.16
μ	-0.0015	-0.0017	-0.0001	0.0006	-0.0004
σ	0.0298	0.0243	0.0295	0.0345	0.0246
AAPL	1.0000	0.7133	0.7744	0.5383	0.7680
AMZN	0.7133	1.0000	0.6903	0.2685	0.6837
FB	0.7744	0.6903	1.0000	0.4807	0.8054
JPM	0.5383	0.2685	0.4807	1.0000	0.6060
GOOGL	0.7680	0.6837	0.8054	0.6060	1.0000

In what follows, we first explain the network architecture of our ACRL model, which is composed of an actor and a critic network. Then, the training procedure of the network under the conditional risk measurement using unconditional

[7]Note that all numbers were rounded to four decimal places while exact values can be found on the github repository: https://anonymous.4open.science/r/ERP-Dynamic-Expectile-RM-805B.

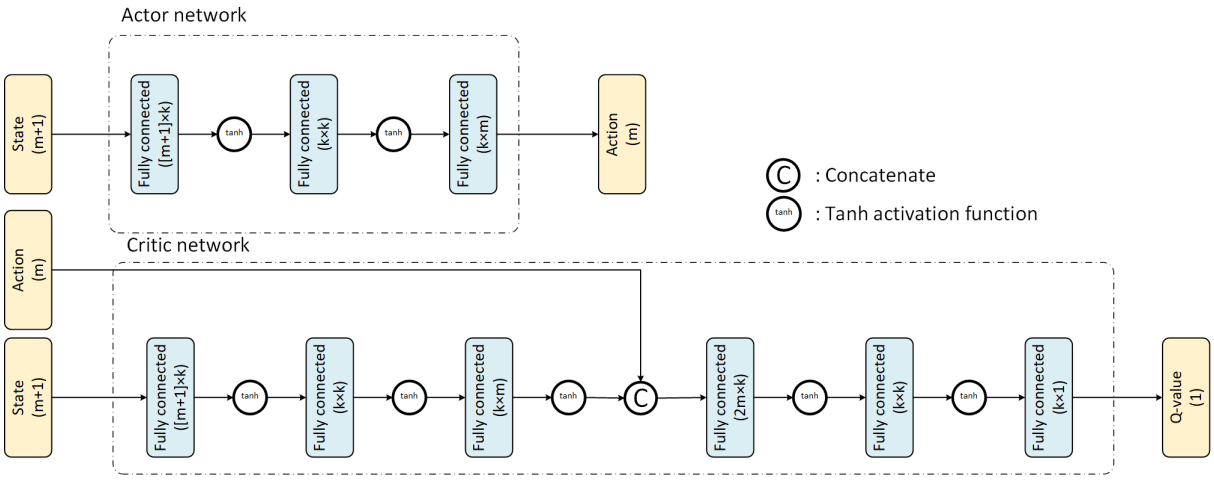


Figure 1: The architecture of the actor and critic networks in ACRL algorithm.

assessment of risk is elaborated. We also numerically demonstrate the benefit of exploiting translation invariance in an option hedging problem using RL, which is for a different purpose than what is previously shown by Marzban et al. (2022) in a DP setting. Finally, the main numerical results of the paper is presented for pricing and hedging a vanilla and a basket option, where the advantages of having a time consistent risk measurement compared to time inconsistent approach is illustrated. In particular, we first focus on the vanilla option to show the precision of our approach by bench-marking its results against a discretized DP model and then extend the results to the case of basket options. All codes are available on the following anonymous Github repository: https://anonymous.4open.science/r/ERP-Dynamic-Expectile-RM-805B.

4.1 Actor and critic network architecture

Our implementation of the ACRL algorithm involves two networks, one for the actor and one for the critic, both of which are presented in Figure 1. Since the numerical experiments assume that the underlying assets of the options follow a Brownian motion process, the model only needs to consider the most recent price for each asset to make investment decisions and the time to maturity. Consequently, the input state to each of the actor and critic networks includes the logarithm of each asset's cumulative return, and the time remaining until maturity, which together correspond to an input vector of dimension $m + 1$.

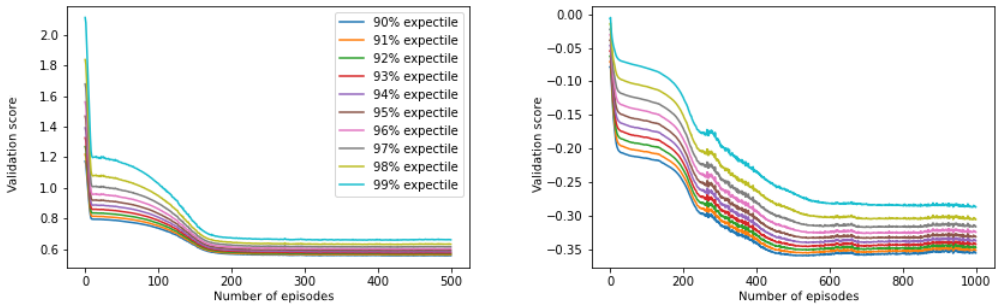
The actor network is composed of three fully connected layers where the number of neurons are considered to be $k = 32$ in the first two layers and then maps back to the number of assets in the last layer so that the model generates the investment policy accordingly for each asset. The activation functions in our networks are considered to be *tanh* functions. In the last layer, this implies that the actions will lie in $[-1, 1]^m$.

The critic network is operating on the same state information, while the m dimensional action information vector is only concatenated to the output of the third layer. The first three layers of the critic network follow the same structure as the actor network in terms of the number of neurons, then after concatenating the action into the network, the two fully connected layers following the concatenation maps the number of neurons again to $k = 32$. Finally, the last layer is a fully connected layer with one neuron to make sure that the output is a scalar representing the approximated Q value function.

4.2 ACRL training procedure for DRM and the role of translation invariance

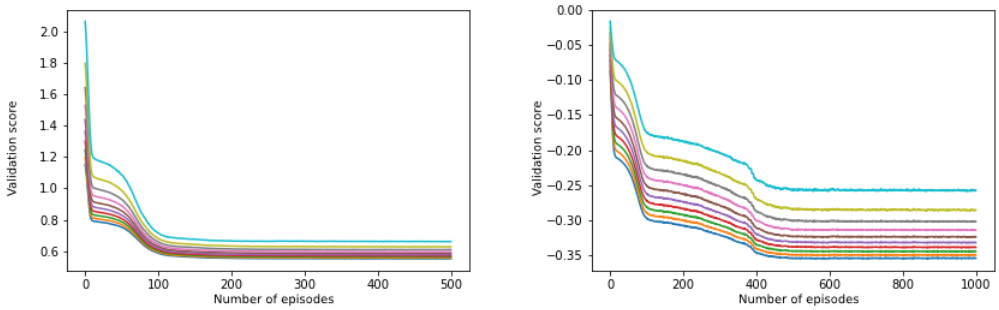
We now explain the training procedure employed for the actor and critic networks in the DRM. Recall that in an SRM setting, overfitting of any DRL algorithm can be controlled by measuring the performance of the trained policy on a validation data set using an empirical estimate of the risk-averse objective as validation score. Unfortunately, this is no longer possible in the case of DRMs since the risk measure relies on conditional risk measurements of the trajectories produced by our policy. In theory, estimates of such conditional measurements could be obtained by training a new critic network using the validation set (while maintaining the policy fixed to the trained one). In practice, this is highly computationally demanding to perform in the training stage and raises a new issue of how to control overfitting of the validation score estimate. Our solution for this problem is to rely on using a static risk measure as validation score.

Given that it is unclear how to best replace a dynamic expectile risk measure with a static one, we choose to compute a set of validation scores that report the performance for a set of static expectiles at risk levels that are larger or equal to the risk level of the DRM. Figure 2(a) and (b) show examples of learning curves for the validation performance of a DRM when trained to hedge the writer and buyer positions of a vanilla option at a risk level of $\tau = 90\%$. In this experiment, it appears that convergence roughly happens at all levels of $\tau \quad 90\%$. This approach is applied in all of our experiments for choosing the optimal number of episodes. We also note that both our training and validation sets included 1000 trajectories from the underlying geometric Brownian motion process. This implies that the training procedure used in these experiments can naturally extend to settings where only historical data is available. The initial learning rates in these experiments were chosen using a cross-validation process, which pointed to $1e^{-3}$ and $5e^{-6}$ respectively for the main critic and actor networks in ACRL, while AORL was trained with an initial rate of $5e^{-5}$. All training employed an exponential decay rate of 0.9999 in order to smooth out the learning process. Finally, a rate of $\gamma = 0.1$ was used to update the target networks.



(a) ACRL for DRM's writer

(b) ACRL for DRM's buyer

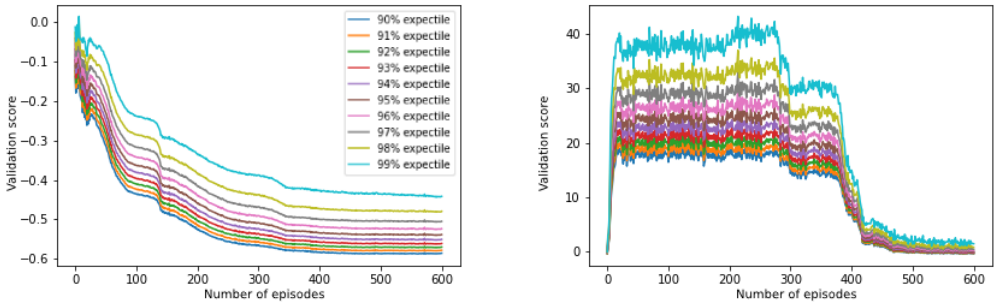


(c) AORL for SRM's writer

(d) AORL for SRM's buyer

Figure 2: Learning curves of the DRM and SRM for an at-the-money vanilla call option on AAPL when a 90% expectile measure is used. The graphs show the validation scores for a range of static expectile measures with risk level ranging from 90% to 99%.

We close this section with a short discussion about the role of the translation invariance property of dynamic risk measures. In particular, the work of Marzban et al. (2022) explains how without this property, the dynamic programming equations need to keep track of the wealth accumulated since $t = 0$ using an additional state variable that gets only employed at $t = T$. More importantly, without translation invariance, the MDP representation ends up only having a reward at $t = T$ thus preventing the ACRL algorithm from receiving quick feedback about the quality of the actions that it is proposing. To illustrate the effect of this property, we compared the convergence of the training process for the ACRL algorithm under both form of DP representation of the buyer's DRM. Namely, Figure 3 presents the learning curves of ACRL with immediate rewards as described in Section 3.2, while (b) presents the learning curves for an implementation in which all the rewards are delayed (using an additional state variable) until $t = T$. These figures clearly show that the MDP with immediate rewards is much easier to train than the delayed rewards MDP. In particular, not only does this model converge in less number of episodes, it also ends up converging to a better solution: the immediate rewards MDP converges to a risk of -0.59 for the buyer (0.91 for the writer), while with delayed rewards it converges to -0.41 (1.01).



(a) With immediate rewards

(b) With delayed rewards

Figure 3: Learning curves of the ACRL algorithm for the buyer's DRM when using (a) the immediate rewards versus (b) delayed rewards in the hedging of a vanilla call at-the-money option.

4.3 Vanilla call option pricing and hedging

In our first set of experiments, we consider pricing and hedging an at-the-money vanilla call option on AAPL. We should first note that solving a hedging problem, e.g. DRM, for a vanilla option is not particularly difficult since the number of state variables in this case is small. It is possible to obtain (approximately) optimal solutions by dynamic programming (Marzban et al. (2022)). Our purpose of considering the case of vanilla option is twofold. First, it provides a useful basis for checking the accuracy of solutions obtained from our deep reinforcement learning (DRL) methods against the "true" optimal solutions, namely by comparing against the DP solutions. Such an accuracy check would be useful for justifying our use of DRL later in this paper as a general means to evaluate hedging performance and calculate the equal risk price (which becomes necessary for problems that cannot be solved by DP such as the case of basket options discussed in the next section). Second, the setting of a vanilla option also allows us to provide a more accurate comparison between DRM and SRM and demonstrate the advantage of the former, i.e. the benefit of time-consistent hedging policies, particularly when options with different time to maturity need to be considered.

To proceed, we first detail how the experiments are conducted. First, the initial price of the underlying stock AAPL is always set to be 78.81, and the hedging portfolio is rebalanced on a monthly basis. Options with different time to maturity are considered, ranging from one month to one year. We generate from a Brownian motion three sets of price trajectories with one year time window, one for training, one for validation, and one for testing, and each consists of 1000 trajectories. In the training phase, we solve both DRM and SRM for the writer and buyer's hedging problems using the longest maturity time, i.e. one year, as the hedging horizon. In solving the DRM, a policy and a critic network are trained using ACRL, whereas in solving the SRM, only a policy network is trained using AORL. See also Section 4.2 regarding how the validation is done to guide the training. Figure 2 presents the learning curves for the training of the hedging policies of the DRM and the SRM with a risk level of $\tau = 90\%$. SRM appears to have a faster rate of convergence than DRM, which might not be surprising given that the architecture of SRM is simpler [8]. It is however worth noting that the issue of time inconsistency for SRM implies that it can potentially produce poor quality policies and prices when the maturity of the option is modified unless it is completely retrained for each type of maturity. This is not the case for DRM and will be further discussed below.

With the trained DRM and SRM policy networks for a fixed 1 year maturity and risk aversion level $\tau \in \{75\%, 90\%, 95\%\}$, we can evaluate the writer and the buyer's (out-of-sample) risk exposure over a pre-specified time horizon so as to calculate the corresponding ERP. We consider the following three metrics for measuring the realized risk under different hedging policy and explain the methods used for calculating the metrics:

> *Out-of-sample static expectile risk*: Given a trained policy network, use the testing data to calculate the static expectile risk obtained when hedging the option using this policy. This is the metric that should be minimized by the SRM.
>
> *RL based out-of-sample dynamic expectile risk estimation*: Use the testing data to train only the critic network in ACRL for evaluating the out-of-sample dynamic risk. In particular, by fixing the policy network in ACRL to a trained policy network, the critic network trained based on testing data provides an estimate of the

[8] The policy network at SRM model is exactly the actor network of DRM, while the quality of actions are directly evaluated in the absence of a trained critic network.

out-of-sample dynamic expectile risk. To speed up the training of the critic network, one may initialize the critic network using the network trained previously with the training data. This is an estimate of the metric minimized by the DRM.

DP based out-of-sample dynamic expectile risk estimation: Given a trained policy network, evaluate the "true" dynamic expectile risk by solving the dynamic programming equations, under the fixed policy, using a high precision discretization of the states, actions, and transitions. Note that this metric is available neither for the case of basket option nor in a data-driven environment where the stochastic process is unknown.

We note that our RL based estimate of out-of-sample dynamic risk is a novel concept, which refers to the calculation of dynamic risk based on testing data. This is possible, as explained above, by training only the critic network using ACRL on the test data. This metric is especially relevant given that classical methods for calculating dynamic risk, such as our DP based estimate, assume full knowledge of the stochastic model that captures the dynamics of an underlying system, i.e. stock price, and require the resolution of dynamic programming equations, which is known to suffer from the curse of dimensionality. Consequently, such methods can no longer be used when the DP equations require a large state space, as can be the case with basket options, or when the description of the underlying stochastic process is unknown.

In our experiments, we apply the second and third metric to the trained DRM policies and the first metric to both the trained DRM and SRM policies. In the former case, we are interested in demonstrating that the RL based out-of-sample expectile risk estimate is an accurate metric. Namely, we will compare the RL based estimate against the "true" DP based estimate. In the latter case, we will shed light on how the DRM policy performs when evaluated according to other metrics that are also of interest to practitioners. In particular, the static expectile risk measure, despite its issue of time inconsistency, can still have its intuitive appeal as a metric, and one may be interested in knowing how a DRM policy performs against this metric as compared to an SRM policy.

Figure 4 summarizes the evaluations of out-of-sample dynamic risk for DRM policies trained for 1 year maturity then applied to options of different maturities ranging from 12 months to 2 months. One can observe that the risk of the writer decreases monotonically for options of shorter maturities, whereas the risk of the buyer increases monotonically. This is consistent with the fact that there is less uncertainty for a shorter hedging horizon, which favors the writer's risk exposure more than the buyer's when considering an at-the-money option. This also provides the evidence that the DRM policies, albeit only trained based on the longest time to maturity, i.e. one year, can be well applied to hedge options with shorter time to maturity and be used to draw consistent conclusion. The observation that the DRM policies remain good policies for problems with shorter time to maturity testifies of the value of using a time consistent hedging model. Another important observation one can make is that the RL based out-of-sample dynamic risk estimate is generally very close to the DP based estimate across all conditions. The difference between the two appears to be more noticeable for the case of high risk aversion, i.e. $\tau = 95\%$ and long time to maturity, but the difference remains minor overall. This observation allows us to confirm the accuracy of our RL based out-of-sample dynamic risk estimation procedure as a replacement for the DP based estimation in settings where the latter cannot be used.

Figure 5 reports the out-of-sample static risk for both SRM policies and DRM policies. The results are interesting and perhaps surprising. First, unlike the consistent behavior observed in the case of dynamic risk, i.e. Figure 4, the static risk of SRM policies for the seller (resp. buyer) may increase (resp. decrease) when hedging an option with shorter maturity. The possibility that a seller's policy may actually increase risk when applied to an option with shorter maturity is clearly problematic when the underlying asset follows a geometric brownian motion with positive drift, as it is inconsistent with the fact that there is less uncertainty (and lower expected value) regarding the payout of such options. This inconsistency occurs because the SRM policies are only trained based on the longest time to maturity, i.e. one year, and they cannot be well applied, unlike for the case of DRM policies, to problems with shorter time to maturity due to the violation of the time consistency property. It is clear from the figures that the SRM policies can be far from the optimal policies when applied to a shorter time to maturity. On the other hand, the DRM policies can actually be found not only to outperform SRM policies in terms of static risk exposure but also to generate consistent results across time, i.e. risk decreases (resp. increases) for the seller (resp. buyer) as the time to maturity decreases. This can be somewhat surprising, as the DRM policies are optimized based on dynamic risk measures rather than the static ones, but the policies can still perform well when evaluated according to static risk measures. Overall, the results presented in Figure 5 best showcase the strength of time consistent policies and why such policies are important to consider in settings where risk needs to be re-evaluated across different time points or maturity dates.[9] We suspect that the possibility that SRM policies may not account properly for risk aversion at some future time point or for other range of option maturities should seriously hinder their use in practice.

[9]Indeed, recall that the example in Section 2.2 demonstrated that the fact that SRM was time inconsistent implied that its policy might not remain a reasonable risk averse policy at future time points. This phenomenon is implicitly observed in Figure 5 given that the MDP is stationary so that the risk measured for a maturity t is exactly equal to the risk measured at time $T \quad t$ when $\mathbf{S}_t = \mathbf{S}_0$.

In order to be more precise about results presented in figures 4 and 5, we detail in Table 3 all the numerical results for the case of high risk aversion, i.e. $\tau = 90\%$, along with the equal risk prices calculated based on RL based out-of-sample dynamic risk estimate and based on the discretized DP (referred as True ERP).[10] One first confirm that the RL based estimateof ERP is a high quality approximation of the true ERP in this vanilla option pricing setting, with a maximum approximation error of 0.01 over all maturity dates. Moreover, we can see that the prices for the SRM polices are generally higher than the prices for the DRM polices. The observation is that while DRM policies are less risky than SRM policies across different time to maturity, it is the writer that benefits more from the use of DRM than the buyer. This could be related to the fact that the writer's loss due to the option payout is unbounded while the option protects the buyer from losses. This in turns implies that the writer's risk exposure is larger in this transaction. Thus, the choice of a policy can be more critical to the writer than the buyer. As the risk exposure of the writer decreases more than for the buyer, this leads to lower ERP price for DRM policies.

Table 3: The out-of-sample dynamic and static 90%-expectile risk imposed to the two sides of vanilla at-the-money call options over AAPL, with maturities ranging from 12 to 0 months, when hedged using the DRM and the SRM policies trained at risk level $\tau = 90\%$ and for a 12 months maturity. Associated ERPs under the DRM are also compared to the "true" ERP measured using a discretized MDP.

		Time to maturity											
Policy	Est.[y]	12	11	10	9	8	7	6	5	4	3	2	1
Dynamic 90%-expectile risk													
Writer's DRM	RL	0.77	0.73	0.69	0.65	0.62	0.58	0.53	0.48	0.45	0.38	0.29	0.23
	DP	0.75	0.71	0.68	0.65	0.61	0.57	0.53	0.49	0.43	0.38	0.31	0.23
Buyer's DRM	RL	-0.22	-0.21	-0.20	-0.19	-0.18	-0.16	-0.15	-0.13	-0.11	-0.09	-0.07	-0.05
	DP	-0.23	-0.22	-0.21	-0.20	-0.18	-0.17	-0.16	-0.14	-0.12	-0.11	-0.08	-0.06
Static 90%-expectile risk													
Writer's SRM	ED	0.55	0.54	0.54	0.53	0.53	0.53	0.52	0.50	0.48	0.46	0.41	0.31
Writer's DRM	ED	0.56	0.54	0.52	0.50	0.47	0.44	0.42	0.39	0.36	0.33	0.29	0.24
Buyer's SRM	ED	-0.35	-0.33	-0.30	-0.27	-0.23	-0.20	-0.17	-0.13	-0.09	-0.07	-0.07	-0.06
Buyer's DRM	ED	-0.36	-0.34	-0.32	-0.30	-0.28	-0.26	-0.24	-0.21	-0.18	-0.14	-0.11	-0.06
Equal risk prices with DRM													
True ERP		0.49	0.47	0.45	0.42	0.40	0.37	0.34	0.31	0.28	0.24	0.19	0.14
DRM	RL	0.50	0.47	0.45	0.42	0.40	0.37	0.34	0.31	0.28	0.24	0.18	0.14
SRM	RL	0.49	0.46	0.44	0.43	0.40	0.38	0.35	0.33	0.30	0.27	0.24	0.22

[y] Estimation (Est.) is either made based on reinforcement learning (RL), discretized dynamic programming (DP), or with the empirical distribution (ED).

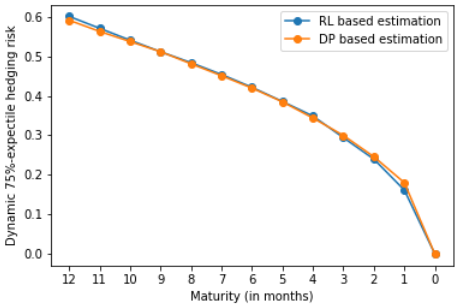
Finally, Figure 6 presents the optimal policies of the two models (i.e., DRM and SRM), together with the actual optimal policy of DRM, obtained using a high precision dynamic program (referred as DP-DRM). Each subfigure shows the policy as a function of current price (x-axis) and time period (colors). The figure further confirms that the policies of both DRM and SRM follow a similar pattern as DP-DRM, which ensures the quality of implementation of both AORL for SRM and ACRL for DRM.

4.4 Basket options

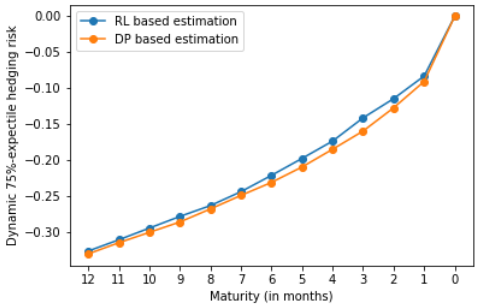
In our second set of experiments, we extend the application of ERP pricing framework to the case of basket options where traditionnal DP solution schemes are not computationally tractable. In particular, we consider an at-the-money basket option with the strike price of 753$ on five underlying assets: AAPL, AMZN, FB, JPM, and GOOGL, where the option payoff is determined by the average price of the underlyings. Similarly to the case of the vanilla option, the rebalancing of the portfolio is happening once per month, options with different maturities from one month to twelve months are considered, and three sets of price trajectories are used for training, validation, and testing the models. We train the ACRL and AORL networks for a one year basket option and then use the same policy network for hedging options with shorter time to maturity.

Our first observation in this set of experiments relates to the training time of the model for the basket option with five assets. Figure 7 presents the convergence of the training of the ACRL model under $\tau = 90\%$. When comparing to the case of the vanilla option, the convergence rate appears to have a similar behavior, i.e., the number of episodes and the time spent on each episode is similar for both the case of the writer and the buyer. This is important as it indicates that the training time might not be very sensitive to the number of assets, while traditional DP approaches are known to become intractable when the option is written on multiple assets.

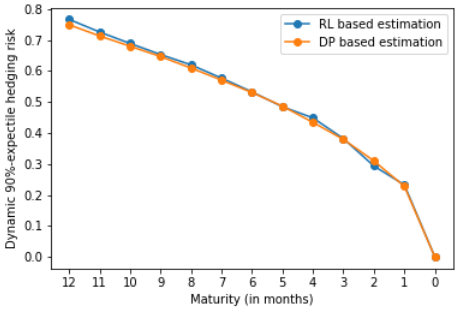
[10] Note that in a purely data-driven setting, the ERP could either be estimated using the in-sample trained critic network, or by calculating our RL based estimate using some freshly reserved data to reduce overfitting biases.



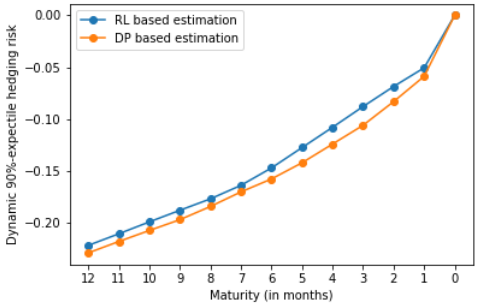
(a) Writer, $\tau = 75\%$



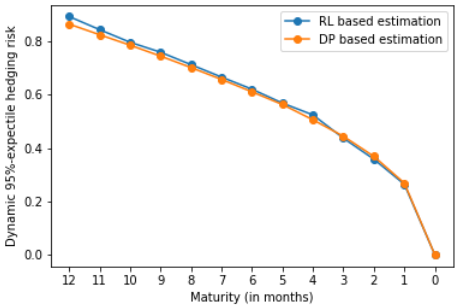
(b) Buyer, $\tau = 75\%$



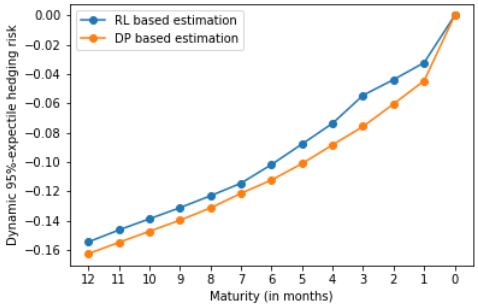
(c) Writer, $\tau = 90\%$



(d) Buyer, $\tau = 90\%$



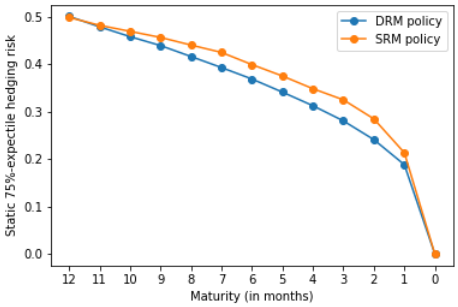
(e) Writer, $\tau = 95\%$



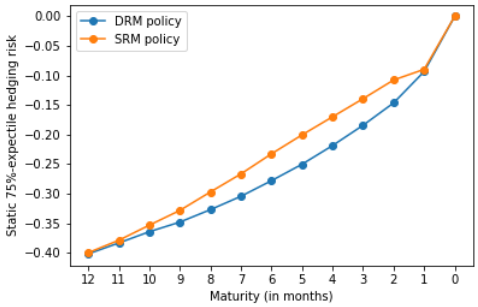
(f) Buyer, $\tau = 95\%$

Figure 4: The out-of-sample dynamic risk imposed to the two sides of a vanilla at-the-money call option over AAPL (with maturity ranging from 12 months to 0 months) under the DRM policy trained for a 12 months maturity and at different risk levels $\tau \in \{75\%, 90\%, 95\%\}$.

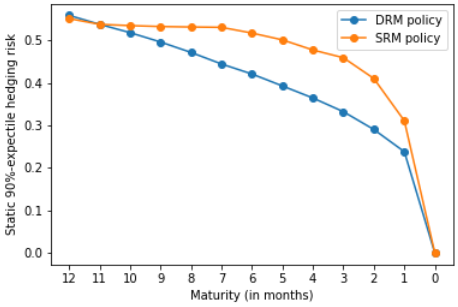
In this section, dynamic risk is estimated using the RL based estimator described in Section 4.3 given that the DP estimator requires too much computations and that the RL based one was shown to provide a relatively high precision estimation of the "true" dynamic risk. Following this, in Figure 8 (a) and (b) we present the dynamic risk obtained from applying the DRM policy on the test data when the model is trained for a one year maturity option. Hedging risk using the same trained policy is presented for 12 different options with maturity ranging from 0 to 12 months. Similar to the vanilla option case, the dynamic risk of the writer is monotonically decreasing as we get closer to the maturity of the option, which can be attributed to the reduced probability that the average price of the assets significantly diverges from the initial average (i.e., the strike price of the option). On the other side, i.e. for the buyer of the option, although overall the risk is increasing to zero as the maturity gets closer to zero, for longer time to maturities we observe some degradation of risk. We attribute this behavior to the estimation error of the RL based dynamic risk estimator.



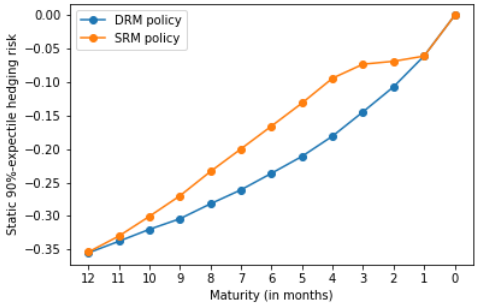
(a) Writer, $\tau = 75\%$



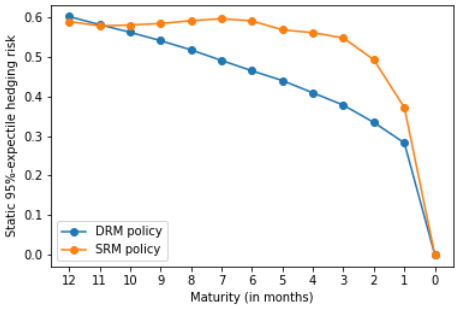
(b) Buyer, $\tau = 75\%$



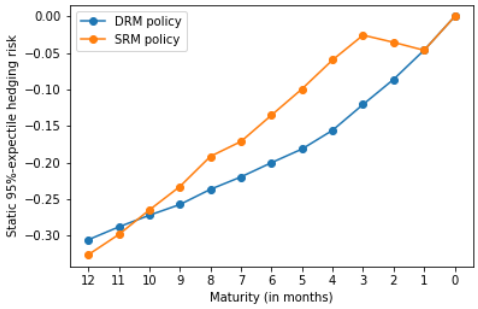
(c) Writer, $\tau = 90\%$



(d) Buyer, $\tau = 90\%$



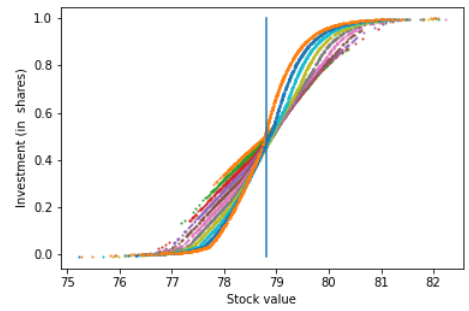
(e) Writer, $\tau = 95\%$



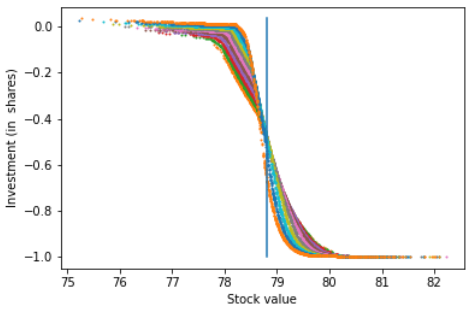
(f) Buyer, $\tau = 95\%$

Figure 5: The out-of-sample static risk imposed to the two sides of a vanilla at-the-money call option over AAPL (with maturity ranging from 12 months to 2 months) under the DRM and SRM policies trained for a 12 months maturity and at different risk levels $\tau \in \{75\%, 90\%, 95\%\}$.

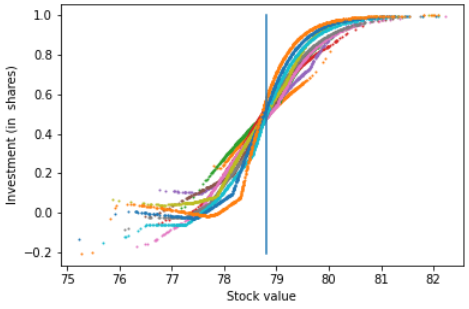
In order to have a view of risk that is not perturbed by estimation errors, we also compare the static risk under DRM and SRM as we did for vanilla options. Figure 9 (a) and (b) shows the static risk under $\tau = 90\%$. One can first recognize the same monotone convergence to zero of the two sides of the options. However, contrary to the case of the vanilla option, the difference between the static risk performance of DRM and SRM policies are rather similar for all maturity times. It therefore appears that in these experiments with a basket option, both SRM and DRM produce similar polices. One possible reason could be that the range of "optimal" risk averse investment plans, whether using DRM or SRM, is more limited. Indeed, while for the vanilla option, we observed that the optimal policies generated investments in the range [0, 1] and [-1, 0] for the writer and the buyer respectively, for the basket option we observed wealth allocations that are more concentrated around 0.20 (i.e. the uniform portfolio known for its risk hedging properties) and -0.20 for each of the 5 assets asset respectively. Finally, similar to the vanilla option case, Table 4 presents more details on the results used to produce figures 8 and 9, along with the equal risk prices computed based on our RL based out-of-sample



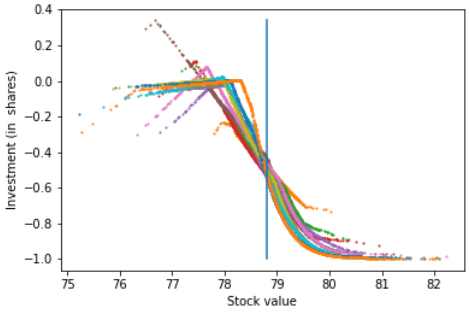
(a) DRM's writer



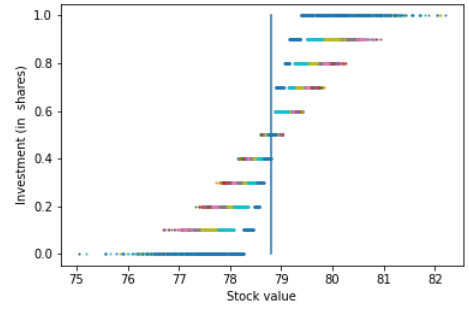
(b) DRM's buyer



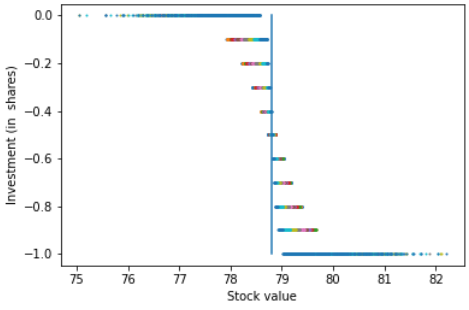
(c) SRM's writer



(d) SRM's buyer



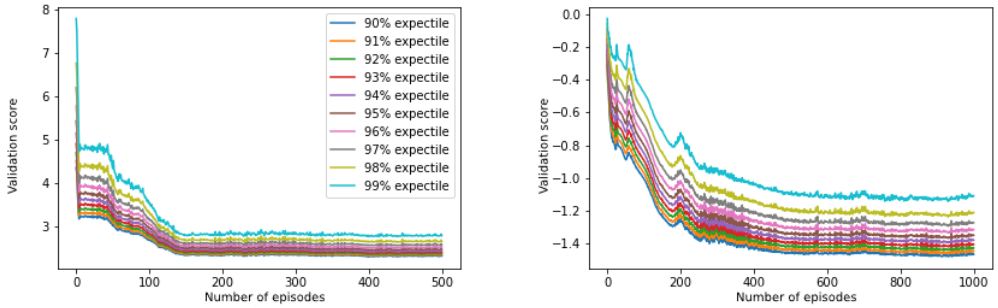
(e) DP-DRM's writer



(f) DP-DRM's buyer

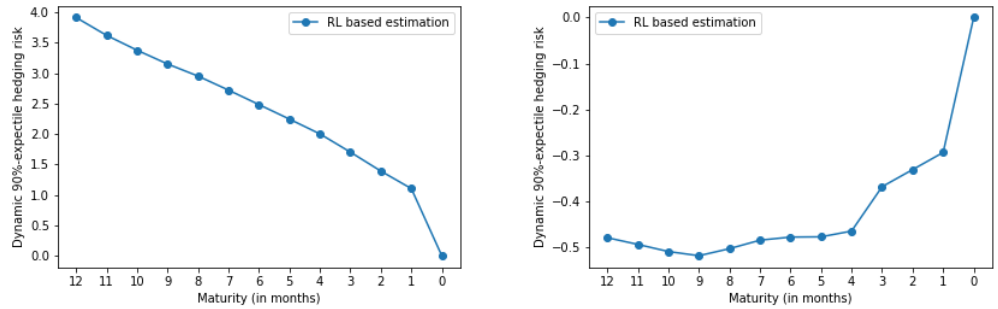
Figure 6: Comparison of the optimal DRL policies obtained for DRM and SRM (with 90% expectile measures) to the discretized DP solution (DP-DRM) for an at-the-money vanilla call option on AAPL with a one year maturity. Each figure presents the sampled actions in our simulated trajectories as a function of the AAPL stock value. The strike price is marked at 78.81.

dynamic risk estimator. The higher ERP price for the SRM policy is an obvious observation in this table, which again can be attributed to the better performing (in terms of dynamic risk) hedging policy produced by ACRL for the DRM, compared to the policy produced by AORL for the SRM.



(a) Writer (b) Buyer

Figure 7: Learning curves of the ACRL algorithm for the writer and buyer's DRM for a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$. The graphs show the validation scores for a range of static expectile measures with risk level ranging from 90% to 99%.

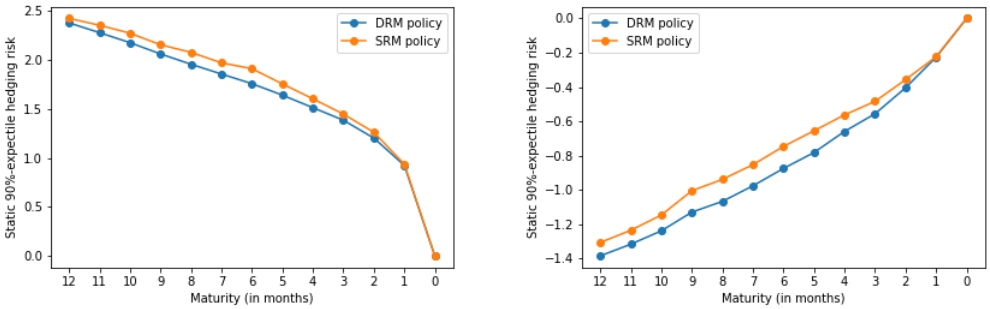


(a) Writer (b) Buyer

Figure 8: The out-of-sample dynamic risk imposed to the two sides of a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$ (as maturity ranges from 12 to 0 months) under a DRM policy trained for a 12 months maturity.

5 Conclusion

In this paper, we developed and implemented the first deep reinforcement learning algorithm for calculating equal risk prices under time consistent dynamic risk measures. This algorithm exploits the elicitability property of the expectile risk measure to extend in a natural way the famous off-policy deterministic actor-critic method presented in Silver et al. (2014) to the risk averse setting. Our numerical experiments confirmed that it can identify risk averse hedging strategies of good quality and be used to estimate the ERP, simultaneously for a range of maturities, using a reasonable amount of computational resources in conditions where traditional DP methods are impracticable. We also demonstrated important issues regarding the implementability of hedging strategies that are based on static (time inconsistent) risk measures. Namely, both our illustrative example and two numerical experiments demonstrated how the time consistent policy produced using the DRM might in fact appear preferable to the investor (from the point of view of the time inconsistent static risk measure) as the risk is measured at later points of time, i.e. with shorter maturity. Overall, as the first paper that is considering option pricing under ERP using time consistent dynamic risk measures, we only evaluated the performance of our model in a synthetic environment using simple neural network architectures. It would be interesting to further examine the performance of Algorithm 2 using real market data with the objective of producing a purely data-driven option pricing scheme. We also consider an important direction of future work to deploy and evaluate the proposed ACRL algorithms in other fields of application.



(a) Static risk, writer (b) Static risk, buyer

Figure 9: The out-of-sample static risk imposed to the two sides of a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$ (as maturity ranges from 12 to 0 months) under the DRM and SRM policies trained for a 12 months maturity.

Table 4: The out-of-sample dynamic and static 90%-expectile risk imposed to the two sides of basket at-the-money call options over AAPL, AMZN, FB, JPM, and GOOGL, with maturities ranging from 12 to 0 months, when hedged using the DRM and the SRM policies trained at risk level $\tau = 90\%$ and for a 12 month maturity. Associated ERPs under the DRM are also compared.

		Time to maturity											
Policy	Est.[y]	12	11	10	9	8	7	6	5	4	3	2	1
Dynamic 90%-expectile risk													
Writer's DRM	RL	3.92	3.62	3.38	3.15	2.95	2.72	2.48	2.25	2.00	1.70	1.39	1.10
Buyer's DRM	RL	-0.48	-0.49	-0.51	-0.52	-0.50	-0.49	-0.48	-0.48	-0.47	-0.37	-0.33	-0.29
Static 90%-expectile risk													
Writer's SRM	ED	2.43	2.36	2.28	2.16	2.08	1.97	1.91	1.76	1.61	1.45	1.26	0.94
Writer's DRM	ED	2.38	2.28	2.18	2.06	1.96	1.86	1.76	1.64	1.51	1.39	1.20	0.92
Buyer's SRM	ED	-1.31	-1.24	-1.15	-1.01	-0.94	-0.85	-0.75	-0.66	-0.56	-0.48	-0.36	-0.22
Buyer's DRM	ED	-1.39	-1.32	-1.24	-1.13	-1.07	-0.98	-0.88	-0.78	-0.66	-0.56	-0.40	-0.23
Equal risk prices with DRM													
DRM	RL	2.20	2.06	1.95	1.84	1.73	1.61	1.48	1.37	1.24	1.04	0.86	0.70
SRM	RL	2.23	2.10	2.01	1.91	1.79	1.65	1.52	1.39	1.21	1.03	0.92	0.82

[y] Estimation (Est.) is either made based on reinforcement learning (RL), discretized dynamic programming (DP), or with the empirical distribution (ED).

Acknowledgement

The authors are thankful to Marc Bellemare for valuable discussions. The authors also gratefully acknowledge the financial support from the Canadian Natural Sciences and Engineering Research Council [Grants RGPIN-2016-05208 and RGPIN-2014-05602], Compute Canada, and the Canada Research Chair program.

A Adapting DDPG to handle dynamic expectile risk measures

We include below the extension of deep deterministic policy gradient (DDPG) algorithm to a risk averse MDP that employs a dynamic expectile risk measure. In **bold** we highlight the modification to DDPG that is due to the use of a dynamic expectile risk measure. Note that after assuming that the information about t is included in the state, we drop the subscript t notation to increase similarity with Lillicrap et al. (2015). For completeness, we make precise that the original DDPG uses $\partial\ell(y) := y$ while this risk averse DDPG, with risk level τ, uses $\partial\ell(y) := 2((1 \quad \tau)\max(0, y) \quad \tau\max(0, \quad y))$. Note that the two approaches are equivalent (up to a scaling of the gradient) when $\tau := 0.5$.

The algorithm assumes an online environment in which the initial state s_0 is drawn from some distribution μ (satisfying the assumption of Proposition 3) and a policy can be deployed over a horizon of T periods. In our option hedging problem, this can be achieved using historical data of the stochastic process $\mathbf{S}_t$ since the latter is assumed exogenous. Namely, one can first randomly sample a starting point t_0 in the historical data and compute

Algorithm 2 General risk averse deep deterministic policy gradient

Randomly initialize the main actor and critic networks' parameters θ^{π} and θ^{Q}
Initialize the target actor, θ^{π^0} θ^{π}, and critic, θ^{Q^0} θ^{Q}, networks
Initialize replay buffers R
for $j = 1 : \#Episodes$ **do**
  Initialize a random process N for action exploration;
  Receive initial observation state s_0
  for $t = 0 : T \quad 1$ **do**
    Select action $a_t = \pi_t(s_t | \theta^{\pi}) + N_t$
    Execute a_t and observe reward r_t and new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in R
    Sample a minibatch of N transitions $f(s^i_{t_i}, a^i_{t_i}, r^i_{t_i}, s^i_{t_i+1})g^N_{i=1}$ in R
    Set the realized losses $y^i := \quad r^i_{t_i} + Q(s^i_{t_i+1}, \pi(s^i_{t_i+1} | \theta^{\pi^0}) | \theta^{Q^0})$
    Update the main critic network:

$$\boldsymbol{\theta^Q \leftarrow \theta^Q - \alpha \frac{1}{N} \sum_{i=1}^{N} \partial \ell(Q(s^i_{t_i}, a^i_{t_i} | \theta^Q) - y^i) \nabla_{\theta^Q} Q(s^i_{t_i}, a^i_{t_i} | \theta^Q)}$$

    where $\partial \ell(y) := 2((1 \quad \tau) \max(0, y) \quad \tau \max(0, \quad y))$
    Update the main actor network: $\theta^{\pi} \quad \theta^{\pi} \quad \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q(s^i_{t_i}, a | \theta^Q)|_{a=\pi(s^i_{t_i} | \theta^{\pi})} \nabla_{\theta^{\pi}} \pi(s^i_{t_i} | \theta^{\pi})$
    Update the target networks: $\theta^{Q^0} \quad \alpha\theta^{Q} + (1 \quad \alpha)\theta^{Q^0}, \quad \theta^{\pi^0} \quad \alpha\theta^{\pi} + (1 \quad \alpha)\theta^{\pi^0}$
  end for
end for

the trajectory $s_0, a_0, r_0, s_1, a_1, r_1, , \ldots, s_T$ obtained on an arbitrary policy based on the observed historical evolution $\mathbf{S}_{t_0}, \mathbf{S}_{t_0+1}, \ldots, \mathbf{S}_{t_0+T}$. Due to the use of a replay buffer R, the algorithm can also be used in context where the state transitions are affected by the policy, namely in hedging problems with transactions cost where the state space must include information about the state of the portfolio.

References

Kaushik I Amin. Jump diffusion option valuation in discrete time. *The Journal of Finance*, 48(5):1833–1863, 1993.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Fabio Bellini and Elena Di Bernardino. Risk management with expectiles. *The European Journal of Finance*, 23(6): 487–506, 2017.

Fabio Bellini and Valeria Bignozzi. On elicitable risk measures. *Quantitative Finance*, 15(5):725–733, 2015.

Dimitris Bertsimas, Leonid Kogan, and Andrew W Lo. Hedging derivative securities and incomplete markets: an ϵ-arbitrage approach. *Operations Research*, 49(3):372–397, 2001.

Lorenzo Bisi, Davide Santambrogio, Federico Sandrelli, Andrea Tirinzoni, Brian D. Ziebart, and Marcello Restelli. Risk-averse policy optimization via risk-neutral policy optimization. *Artificial Intelligence*, 103765, 2022.

Michael J. Brennan. The pricing of contingent claims in discrete time models. *The Journal of Finance*, 34(1):53–68, 1979.

Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019.

Jay Cao, Jacky Chen, John Hull, and Zissis Poulos. Deep hedging of derivatives using reinforcement learning. *The Journal of Financial Data Science*, 3(1):10–27, 2021.

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

Alexandre Carbonneau and Frédéric Godin. Deep equal risk pricing of financial derivatives with multiple hedging instruments. *arXiv preprint arXiv:2102.12694*, 2021a.

Alexandre Carbonneau and Frédéric Godin. Equal risk pricing of derivatives with deep hedging. *Quantitative Finance*, 21(4):593–608, 2021b.

Alexandre Carbonneau and Frédéric Godin. Deep equal risk pricing of financial derivatives with non-translation invariant risk measures. *arXiv preprint arXiv:2107.11340*, 2021c.

Peter Carr, Helyette Geman, and Dilip B Madan. Pricing and hedging in incomplete markets. *Journal of Financial Economics*, 62(1):131 – 167, 2001.

Dotan Di Castro, J. Oren, and Shie Mannor. Practical risk measures in reinforcement learning. *arXiv preprint arXiv:1908.08379*, 2019.

James Ming Chen. On exactitude in financial regulation: Value-at-risk, expected shortfall, and expectiles. *Risks*, 6(2): 61, 2018.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

Anthony Coache and Sebastian Jaimungal. Reinforcement learning with dynamic convex risk measures. *arXiv preprint arXiv:2112.13414*, 2022.

Anthony Coache, Sebastian Jaimungal, and Álvaro Cartea. Conditionally elicitable dynamic risk measures for deep reinforcement learning. *arXiv preprint arXiv:2206.14666*, 2022.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, pp. 179–186, Madison, WI, USA, 2012. Omnipress.

Freddy Delbaen and Walter Schachermayer. The variance-optimal martingale measure for continuous processes. *Bernoulli*, 2:81–105, 1995.

Simon Fecamp, Joseph Mikael, and Xavier Warin. Deep learning for discrete-time hedging in incomplete markets. *Journal of Computational Finance*, 25:51–85, 2021.

Hans Föllmer, Hans Sondermann, and Dieter Sondermann. *Hedging of non-redundant contingent claims*, pp. 205 – 223. 06 1985. doi: 10.13140/RG.2.1.3298.8322.

Christian Gourieroux, Jean Paul Laurent, and Huyên Pham. Mean-variance hedging and numéraire. *Mathematical Finance*, 8(3):179–200, 1998.

Ivan Guo and Song-Ping Zhu. Equal risk pricing under convex trading constraints. *Journal of Economic Dynamics and Control*, 76:136–151, 2017.

Steven L Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343, 1993.

Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. On the convergence and optimality of policy gradient for Markov coherent risk. *arXiv preprint arXiv:2103.02827*, 2021.

John Hull and Alan White. The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2): 281–300, 1987.

Stefan Jaschke and Uwe Küchler. Coherent risk measures and good-deal bounds. *Finance and Stochastics*, 5(2): 181–200, 2001.

Petter N. Kolm and Gordon Ritter. Dynamic replication and hedging: A reinforcement learning approach. *The Journal of Financial Data Science*, 1(1):159–171, 2019.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Saeed Marzban, Erick Delage, and Jonathan Yumeng Li. Deep reinforcement learning for equal risk pricing and hedging under dynamic expectile risk measures. *arXiv preprint arXiv:2109.04001*, 2021. URL `https://arxiv.org/abs/2109.04001`.

Saeed Marzban, Erick Delage, and Jonathan Yu-Meng Li. Equal risk pricing and hedging of financial derivatives with convex risk measures. *Quantitative Finance*, 22(1):47–73, 2022.

Oskari Mikkilä and Juho Kanniainen. Empirical deep hedging. *Quantitative Finance*, 23(1):111–122, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Alois Pichler, Rui Peng Liu, and Alexander Shapiro. Risk-averse stochastic programming: Time consistency and optimal stopping. *Operations Research*, 70(4):2439–2455, 2022.

Birgit Rudloff, Alexandre Street, and Davi M Valladão. Time consistency and risk averse dynamic decision models: Definition, interpretation and practical consequences. *European Journal of Operational Research*, 234(3):743–750, 2014.

Andrzej Ruszczyński and Alexander Shapiro. Conditional risk mappings. *Mathematics of Operations Research*, 31(3): 544–561, 2006.

Martin Schweizer. Approximation pricing and the variance-optimal martingale measure. *The Annals of Probability*, 24 (1):206–236, 1996.

Alexander Shapiro. Interchangeability principle and dynamic equations in risk averse stochastic programming. *Operations Research Letters*, 45(4):377–381, 2017.

Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, 2014.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pp. 387–395. PMLR, 2014.

Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger (eds.), *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 958–968, 2020.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.