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

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently equal risk pricing, a framework for fair derivative pricing, was extended to consider coherent risk measures. However, all current implementations either employ a static risk measure or are based on traditional dynamic programming solution schemes that are impracticable in realistic settings: when the number of underlying assets is large or only historical trajectories are available. This paper extends for the first time the deep deterministic policy gradient algorithm to the problem of solving a risk averse Markov decision process that models risk using a time consistent dynamic expectile risk measure. Our numerical experiments, which involve both a simple vanilla option and a more exotic basket option, confirm that the new ACRL algorithm can produce high quality hedging strategies that produce accurate prices in simple settings, and 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. Recently, the scheme of equal risk pricing (ERP), proposed by Guo & Zhu (2017) and extended to coherent risk measures in Marzban et al. (2020), was 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. Having at its core a financial hedging problem, the practical implementation of ERP faces two important challenges. First, while the work of Marzban et al. (2020) provides sufficient conditions for making the equal risk pricing and hedging problem amenable to dynamic programming (DP), such formulation is necessarily susceptible to the curse of dimensionality. In addition, traditional solution schemes rely on 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. In particular, the recent works of Carbonneau & Godin (2020) and Carbonneau & Godin (2021a) successfully applied DRL to solve ERP's hedging problems, where risk aversion is modeled using coherent risk measures, and they demonstrate the possibility of pricing a broad range of over-the-counter options such as asian options. Unfortunately, the DRL approaches proposed in the work of Carbonneau & Godin can only be used in settings where the risk is measured according to a static risk measure (i.e. that disregards the temporal structure of the random variable). 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. In other words, equal risk prices calculated based on static risk measures assume a hedging policy that can-

not be implemented in practice, and thus are biased.[1] From a numerical perspective, employing a static risk measure in ERP also limits the type of DRL algorithms that can be used to solve each party's hedging problem. Specifically, Carbonneau & Godin employ a policy optimization scheme, a.k.a. Actor-Only RL (AORL) algorithm (see Williams (1992)), while other critic-only or actor-critic algorithms (such as Mnih et al. (2015) and Lillicrap et al. (2015) respectively) remain out of reach.

In this paper, we develop a new DRL algorithm for solving a general class of risk averse MDP in order to solve a time-consistent ERP hedging problems. It is known that to ensure time consistency, a dynamic risk measure should be employed. In particular, motivated by the theory of coherent risk measures, which identifies expectile risk measures as the only elicitable coherent risk measures, we propose the use of dynamic expectile risk measures and exploit elicitability in the design of our model-free Actor-Critic RL (ACRL) algorithm.

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

- We present the first model-free DRL based algorithm for identifying optimal risk averse option hedging strategies that are time-consistent with respect to a dynamic convex risk measure, and for computing their associated equal risk prices. A side benefit from pricing/hedging an option with maturity T using dynamic risk measures will be that we will easily obtain equal risk prices and hedging strategies for any other maturity $T' < T$. We believe that ERP constitutes an original and fertile application in which to develop and test new risk averse deep RL methods.
- The ACRL algorithm that we propose is the first to naturally extend the popular deep deterministic policy gradient algorithm (DDPG) (see Lillicrap et al. (2015)) to a risk averse setting where a time consistent coherent risk measure is used. 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 an actor and a critic network. 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 of expectile risk measures and to propose a procedure for training the "risk-to-go" deep Q-network that is also based on stochastic gradient descent.
- We evaluate 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.

The rest of this paper is organized as follows. Section 2 introduces equal risk pricing and its associated DP equations. Section 3 proposes the new ACRL algorithm for risk averse MDP with expectile measures. Finally, Section 4 presents and discusses our numerical experiments.

2 Equal Risk Pricing and Hedging under Dynamic Expectile Risk Measures

As described in Marzban et al. (2020), 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 $\mathbf{S}_t : \Omega \to \mathbb{R}^m$ denote the values of the risky assets adapted to a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} := \{\mathcal{F}_t\}_{t=0}^T, \mathbb{P})$, i.e. each $\mathbf{S}_t$ is $\mathcal{F}_t$ 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 $p_0 \in \mathbb{R}$ is shown by $\mathcal{X}(p_0)$:

$$\mathcal{X}(p_0) = \left\{ X : \Omega \to \mathbb{R}^T \,\middle|\, \exists \{\boldsymbol{\xi}_t\}_{t=0}^{T-1}, \quad X_t = p_0 + \sum_{t'=0}^{t-1} \boldsymbol{\xi}_{t'}^\top \Delta \mathbf{S}_{t'+1}, \quad \forall t = 1, \ldots, T \right\},$$

[1] Note that in a very recent paper of Carbonneau & Godin (2021b), the authors deployed a policy optimization algorithm for ERP under an expected utility framework that is time consistent yet violates an important axiom of monetary risk measures Föllmer & Schied (2016).

where $\Delta \mathbf{S}_{t+1} := \mathbf{S}_{t+1} - \mathbf{S}_t$, the hedging strategy $\boldsymbol{\xi}_t \in \mathbb{R}^m$ is a vector of random variables adapted to the filtration $\mathbb{F}$ and captures the number of shares of each risky asset held in the portfolio during the period $[t,\ t+1]$, and X_t is the accumulated wealth.

Let $F(\{\mathbf{S}_t\}_{t=1}^T)$ denote the payoff of a derivative. Throughout this paper, we assume $F(\{\mathbf{S}_t\}_{t=1}^T)$ admits the formulation of $F(\mathbf{S}_T, \mathbf{Y}_T)$ where $\mathbf{Y}_t$ is an auxiliary fixed-dimensional stochastic process that is $\mathcal{F}_t$-measurable. This class of payoff functions is common in the literature, (see for example Bertsimas et al. (2001) and Marzban et al. (2020)). 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:

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

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

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 $F(\mathbf{S}_T, \mathbf{Y}_T)$. 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. 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. $\varrho^w(p_0^*) = \varrho^b(p_0^*)$.

In particular, based on Proposition 3.1 and the examples presented in section 3.3 of Marzban et al. (2020), together with the fact that both ρ^w and ρ^b are dynamic recursive law invariant risk measures, a Markovian assumption allows us to conclude that the ERP can be calculated using two sets of dynamic programming equations.

Assumption 2.1. *[Markov property] There exists a sufficient statistic process ψ_t adapted to $\mathbb{F}$ such that $\{(\mathbf{S}_t, \mathbf{Y}_t, \psi_t)\}_{t=0}^T$ is a Markov process relative to the filtration $\mathbb{F}$. Namely, $\mathbb{P}((\mathbf{S}_{t+s}, \mathbf{Y}_{t+s}, \psi_{t+s}) \in \mathcal{A}|\mathcal{F}_t) = \mathbb{P}((\mathbf{S}_{t+s}, \mathbf{Y}_{t+s}, \psi_{t+s}) \in \mathcal{A}|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$ for all t, for all $s \geq 0$, and all sets $\mathcal{A}$.*

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

$$V_t^w(\mathbf{S}_t, \mathbf{Y}_t, \psi_t) := \inf_{\boldsymbol{\xi}_t} \bar{\rho}(-\boldsymbol{\xi}_t^\top \Delta \mathbf{S}_{t+1} + V_{t+1}^w(S_t + \Delta \mathbf{S}_{t+1}, \mathbf{Y}_t + \Delta Y_{t+1}, \psi_{t+1})|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)\,,$$

where $\bar{\rho}(\cdot|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$ is a law invariant risk measure that uses $\mathbb{P}(\cdot|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$. 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: $V_T^b(\mathbf{S}_T, \mathbf{Y}_T, \psi_T) := -F(\mathbf{S}_T, \mathbf{Y}_T)$ and

$$V_t^b(\mathbf{S}_t, \mathbf{Y}_t, \psi_t) := \inf_{\boldsymbol{\xi}_t} \bar{\rho}(-\boldsymbol{\xi}_t^\top \Delta \mathbf{S}_{t+1} + V_{t+1}^b(\mathbf{S}_t + \Delta \mathbf{S}_{t+1}, \mathbf{Y}_t + \Delta \mathbf{Y}_{t+1}, \psi_{t+1})|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)\,,$$

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 the ERP.

Lemma 2.1 (Marzban et al. (2020)). *Under Assumption 2.1, the ERP that employs dynamic expectile risk measure can be computed as: $p_0^* = (V_0^w(\mathbf{S}_0, \mathbf{Y}_0, \psi_0) - V_0^b(\mathbf{S}_0, \mathbf{Y}_0, \psi_0))/2$.*

In what follows, we will further assume that the risk measure is a dynamic expectile risk measure.

Definition 2.1. *A dynamic expectile risk measure takes the form: $\rho(X) := \bar{\rho}_0(\bar{\rho}_1(\ldots \bar{\rho}_{T-1}(X)))$ where each $\bar{\rho}(\cdot)$ is an expectile risk measure that employs the conditional distribution based on $\mathcal{F}_t$. Namely, $\bar{\rho}_t(X_{t+1}) := \arg\min_q \tau \mathbb{E}\left[(q - X_{t+1})_+^2 | \mathcal{F}_t\right] + (1-\tau)\mathbb{E}\left[(q - X_{t+1})_-^2 | \mathcal{F}_t\right]$ where X_{t+1} is a random liability measureable on $\mathcal{F}_{t+1}$.*

Like conditional value at risk, the expectile measure (see Bellini & Bignozzi (2015)) covers the range of risk attitudes from risk neutrality, when $\tau = 1/2$, to worst-case risk, when $\tau \to 1$.

3 A NOVEL ACTOR-CRITIC ALGORITHM FOR RISK AVERSE MDP UNDER A DYNAMIC EXPECTILE RISK MEASURE

With the dynamic programming equations in hand, it now becomes apparent that each option hedging problem in ERP can be formulated as a Markov Decision Process (MDP) described with

$(\mathcal{S}, \mathcal{A}, r, P)$. In this regard, the agent (i.e. the writer or buyer) interacts with a stochastic environment by taking an action $a_t \equiv \boldsymbol{\xi}_t \in [-1, 1]^m$ after observing the state $s_t \in \mathcal{S}$, 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. $r_t(s_t, a_t, s_{t+1}) := \boldsymbol{\xi}_t^\top \Delta \mathbf{S}_{t+1}$ when $t < T$ and otherwise of the option liability/payout $r_T(s_T, a_T, s_{T+1}) := F(\mathbf{S}_T, \mathbf{Y}_T)(1 - 2 \cdot \mathbf{1}\{\text{agent=writer}\})$, which is insensitive to s_{T+1}. Finally, the Markovian exogeneous dynamics described in Assumption 2.1 are modeled using P as $P(s_{t+1}|s_t, a_t) = \mathbb{P}(\mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1}|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$. Overall, each of the two dynamic derivative hedging problems presented in Section 2 reduce to a version of the following general risk averse reinforcement learning problem:

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

where $\bar{s}_0 := (\mathbf{S}_0, \mathbf{Y}_0, \psi_0)$ is the initial state in which the option is priced while $Q_t^\pi(s_t, a_t) := \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^\pi(s_{t+1}, \pi_{t+1}(s_{t+1}))|s_t)$, $Q_T^\pi(s_T, a_T) := r_T(s_T, a_T, s_T)$, and where $\bar{\rho}$ is an expectile risk measure, i.e. $\bar{\rho}(X) := \arg\min_q \mathbb{E}\left[\tau(q - X)_+^2 + (1 - \tau)(q - X)_-^2\right]$. Equipped with these definitions, we can now motivate our proposed extension of the model-free off-policy deterministic ACRL algorithm to the risk-averse MDP setting.

First, by the nature of the MDP, which exploits the interchangeability principle (see Pichler et al. (2021)), we have that $\varrho(0) = Q_0^{\pi^*}(\bar{s}_0, \pi_0^*(\bar{s}_0))$, where

$$\pi_t^*(s_t) \in \arg\min_{a_t} \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi^*}(s_{t+1}, \pi_{t+1}^*(s_{t+1}))|s_t) \ = \ \arg\min_{a_t} Q_t^{\pi^*}(s_t, a_t) \ .$$

Hence, π^* must also minimize:

$$\pi^* \in \arg\min_\pi \mathbb{E}_{\substack{\tilde{t} \sim \{0, \ldots, T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} [Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \pi_{\tilde{t}}(s_{\tilde{t}}))] \tag{3}$$

where $\tilde{t}$ is uniformly drawn from the range $\{0, \ldots, T-1\}$, s_0 is drawn from some reference initial distribution (e.g. $P(s_0 = \bar{s}_0) = 1$), and where $\bar{\pi}$ is an arbitrary reference policy[2] that gives non-zero probability of occurring to all states in all time periods.

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 (3). Specifically, given a policy network π^θ, we wish to optimize:

$$\min_\theta \mathbb{E}_{\substack{\tilde{t} \sim \{0, \ldots, T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} [Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, \pi_{\tilde{t}}^\theta(s_{\tilde{t}}))] \ ,$$

using a stochastic gradient algorithm. In doing so, we rely on the fact that:

$$\begin{aligned}
&\nabla_\theta \mathbb{E}_{\substack{\tilde{t} \sim \{0, \ldots, T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} [Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, \pi^\theta(s_{\tilde{t}}))] \\
&= \mathbb{E}_{\substack{\tilde{t} \sim \{0, \ldots, T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} \left[\nabla_\theta Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a)\Big|_{a=\pi_{\tilde{t}}^\theta(s_{\tilde{t}})} + \nabla_a Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a) \nabla_\theta \pi_{\tilde{t}}^\theta(s_{\tilde{t}})\Big|_{a=\pi_{\tilde{t}}^\theta(s_{\tilde{t}})} \right] \\
&\approx \mathbb{E}_{\substack{\tilde{t} \sim \{0, \ldots, T-1\} \\ s_{t+1} \sim P(\cdot|s_t, \bar{\pi}_t(s_t))}} \left[\nabla_a Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a) \nabla_\theta \pi_{\tilde{t}}^\theta(s_{\tilde{t}})\Big|_{a=\pi_{\tilde{t}}^\theta(s_{\tilde{t}})} \right] \ .
\end{aligned}$$

Note that in the above equation, we have dropped the term that depends on $\nabla_\theta Q_{\tilde{t}}^{\pi^\theta}$ 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. $\bar{\rho}(\cdot) := \mathbb{E}[\cdot]$. While we do consider as an important subject of future research to extend this motivation to more general 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.

[2] In fact, given that s_t is entirely exogeneous, the distribution of s_{t+1} is unaffected by $\bar{\pi}$ in our option hedging problem.

Given that we do not have access to an exact expression for $Q_{\tilde{t}}^{\pi^\theta}(s_{\tilde{t}}, a)$, 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:

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

where $\ell(y) := (\tau\mathbf{1}\{y > 0\} - (1-\tau)\mathbf{1}\{y \leq 0\})y^2$ is the score function associated to the τ-expectile risk measure. As explained in Shen et al. (2014) for the case of a tabular MDP, this suggests using the following stochastic gradient step to improve each expectile estimators:

$$Q_t^\pi(s_t, a_t) \leftarrow Q_t^\pi(s_t, a_t) - \alpha\partial\ell(Q_t^\pi(s_t, a_t) + r_t(s_t, a_t, s_{t+1}) - Q_{t+1}^\pi(s_{t+1}, \pi_{t+1}(s_{t+1}))\,,$$

where $\partial\ell(y) := \tau\max(0, y) - (1-\tau)\max(0, -y)$ is the subdifferential of $\ell(y)$.

In the non-tabular setting, we replace $Q_t^\pi(s_t, a_t)$ 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 $\rho(-r(s_t, a_t, s_{t+1}) + Q_{t+1}^\pi(s_{t+1}, a_{t+1}|\theta^{Q'}))$ and a step that replaces the target network $Q_t^\pi(s_t, a_t|\theta^{Q'})$ with a network more similar to the main one $Q_t^\pi(s_t, a_t|\theta^Q)$. The former is achieved, similarly as with the policy network, by searching for the optimal θ^Q according to:

$$\min_{\theta^Q} \mathbb{E}_{\substack{\tilde{t}\sim\{0,\dots,T-1\} \\ s_{t+1}\sim P(\cdot|s_t,\bar{\pi}_t(s_t))}} [\ell(Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}})|\theta^Q) + r_t(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}}), s_{\tilde{t}+1}) - Q_{\tilde{t}+1}^\pi(s_{\tilde{t}+1}, \pi_{\tilde{t}+1}(s_{\tilde{t}+1})|\theta^{Q'}))]\,,$$

which suggests a stochastic gradient update of the form $\theta^Q \leftarrow \theta^Q - \alpha\Delta$, where Δ is

$$\partial\ell(Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}})|\theta^Q) + r_t(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}}), s_{\tilde{t}+1}) - Q_{\tilde{t}+1}^\pi(s_{\tilde{t}+1}, \pi_{\tilde{t}+1}(s_{\tilde{t}+1})|\theta^{Q'}))\nabla_{\theta^Q} Q_{\tilde{t}}^\pi(s_{\tilde{t}}, \bar{\pi}_{\tilde{t}}(s_{\tilde{t}})|\theta^Q)\,.$$

These two types of updates are integrated in our proposed expectile-based actor-critic deep RL (a.k.a. ACRL) algorithm described in Algorithm 1. One may note that in each episode, the reference policy $\bar{\pi}_t$ is updated to be a perturbed version of the main policy network in order to focus the accuracy of the main critic network's 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. For completeness, a version of ACRL that mimics the way that the original DDPG generates mini-batch using a replay buffer can also be found in Appendix A.

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 static risk measure applied on the discounted sum of total rewards (see Prashanth & Ghavamzadeh (2013); Chow & Ghavamzadeh (2014); Castro et al. (2019); Singh et al. (2020); Urpí et al. (2021)). Such methods therefore suffer from time-inconsistency. The two exceptions consist of Tamar et al. (2015) and Huang et al. (2021) who propose ACRL algorithms to deal with dynamic law-invariant coherent risk measures. While applicable to a wide range of dynamic risk measures, the two algorithms either assume that it is possible to generate samples from a perturbed version of the 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[3]. While our approach can only be used with the dynamic expectile risk measure, it offers a much simpler implementation that naturally extends the DDPG algorithm 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.

We finally note that in our problem, $P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, a_t') = \mathbb{P}(\mathbf{S}_{t+1}, \mathbf{Y}_{t+1}, \psi_{t+1}|\mathbf{S}_t, \mathbf{Y}_t, \psi_t)$, 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,

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

Algorithm 1: The actor-critic RL algorithm for the dynamic recursive expectile option hedging problem (ACRL)

Randomly initialize the main actor and critic networks' parameters θ^{π} and θ^{Q};
Initialize the target actor, $\theta^{\pi'} \leftarrow \theta^{\pi}$, and critic, $\theta^{Q'} \leftarrow \theta^{Q}$, networks;
for $j = 1 : \#Episodes$ **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 $P(\cdot|s_t, \bar{\pi}_t(s_t))$, where $\bar{\pi}_t(s_t) := \pi_t(s_t|\theta^{\pi}) + \mathcal{N}(0, \sigma)$;
  Set the realized losses $y_t^i := -r_t(s_t^i, a_t^i, s_{t+1}^i) + Q_{t+1}(s_{t+1}^i, \pi_{t+1}(s_{t+1}^i|\theta^{\pi'})|\theta^{Q'})$;
  Update the main critic network:

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

  Update the main actor network:

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

  Update the target networks:

$$\theta^{Q'} \leftarrow \alpha\theta^{Q} + (1-\alpha)\theta^{Q'}, \qquad \theta^{\pi'} \leftarrow \alpha\theta^{\pi} + (1-\alpha)\theta^{\pi'}\,; \tag{4}$$

end

which eliminates the need to keep track of the accumulated wealth in the set of state variables as explained in Marzban et al. (2020) 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 that convergence speed is significantly improved in training due to this property (see Figure 4 in Appendix).

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 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: APPL, AMZN, FB, JPM, and 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. In both experiments, the maturity of the option will be one year and the hedging portfolios will be rebalanced on a monthly basis. Table 3 in the appendix provides the descriptive statistics of our underlying hidden stochastic process.

In what follows, we first explain the architectures of our ACRL model. Then, the training procedure of the networks under the dynamic risk measurement is elaborated. Finally, the main numerical results of the paper are presented for pricing and hedging a vanilla, where the precision of our approach will be empirically demonstrated, and a basket option. All codes are available at https://anonymous.4open.science/r/ERP-Dynamic-Expectile-RM-4BEA.

4.1 Actor and Critic Network Architecture

Our implementation of the ACRL algorithm involves two simple networks presented in Figure 1. Since the underlying assets follow a Brownian motion, the actor and critic networks can define

the input state as the logarithm of the cumulative returns of each asset and the time remaining to maturity (i.e. 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 to generate 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 only concatenates the m dimensional action information vector after its third layer. In the case of SRM, only the actor network is used.

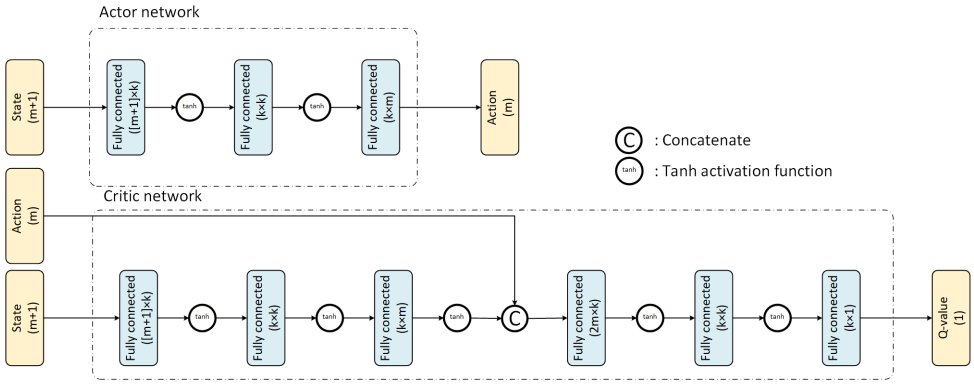

Figure 1: The architecture of the actor and critic networks in ACRL algorithm.

4.2 Training procedure and learning curves

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 static risk measures as validation score, namely a set of static expectiles at risk levels that are larger or equal to the risk level of the DRM. Figures 2 and 3 in the appendix show examples of learning curves for the validation performance of a DRM and SRM approaches on vanilla and basket options at a risk level of $\tau = 90\%$. SRM appeared to have a faster rate of convergence than DRM, due to its simpler architecture. Being a time-inconsistent model, SRM must however be retrained whenever the maturity of the option is modified. When comparing convergence rates between vanilla and basket options, we observed similar behavior, which indicates that the training time might not be very sensitive to the number of assets, thus suffering less from the curse of dimensionality. We finally note that both our training and validation sets included 1000 trajectories from the underlying geometric Brownian motion process, implying that the procedure can be applied in settings where only historical data is available.

4.3 Vanilla option hedging and pricing

In our first set of experiments, we consider pricing and hedging an at-the-money vanilla call option on APPL. In this setting, it is possible to obtain (approximately) optimal solutions by dynamic programming via discretization of the state space. The initial price of APPL is set to 78.81 and options with time to maturity ranging from one month to one year are considered. Both DRM and SRM are trained using a one year maturity/horizon.

With the trained DRM and SRM policy networks, 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, use the test data to calculate the static expectile risk.
- *RL based out-of-sample dynamic expectile risk estimation*: Given the trained policy, use the test data to only train a critic network using ACRL to produce an estimate of out-of-sample dynamic expectile risk.
- *DP based out-of-sample dynamic expectile risk estimation*: Given a trained policy, evaluate the "true" dynamic expectile risk by solving the dynamic programming equations using a high precision discretization of the states, actions, and transitions.[4]

We note that our RL based estimate of out-of-sample dynamic risk is a novel approach, which tackles the important challenge of policy evaluation in RL with dynamic risk measures.

Table 1: The out-of-sample dynamic and static 90%-expectile risk imposed to the two sides of vanilla at-the-money call options over APPL.

		Time to maturity									
Policy	Est.†	12	11	10	9	8	$\cdots$	4	3	2	1
Dynamic 90%-expectile risk											
Writer's DRM	RL	0.77	0.73	0.69	0.65	0.62	$\cdots$	0.45	0.38	0.29	0.23
	DP	0.75	0.71	0.68	0.65	0.61	$\cdots$	0.43	0.38	0.31	0.23
Buyer's DRM	RL	-0.22	-0.21	-0.20	-0.19	-0.18	$\cdots$	-0.11	-0.09	-0.07	-0.05
	DP	-0.23	-0.22	-0.21	-0.20	-0.18	$\cdots$	-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	$\cdots$	0.48	0.46	0.41	0.31
Writer's DRM	ED	0.56	0.54	0.52	0.50	0.47	$\cdots$	0.36	0.33	0.29	0.24
Buyer's SRM	ED	-0.35	-0.33	-0.30	-0.27	-0.23	$\cdots$	-0.09	-0.07	-0.07	-0.06
Buyer's SRM	ED	-0.36	-0.34	-0.32	-0.30	-0.28	$\cdots$	-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	$\cdots$	0.28	0.24	0.19	0.14
DRM	RL	0.50	0.47	0.45	0.42	0.40	$\cdots$	0.28	0.24	0.18	0.14
SRM	RL	0.49	0.46	0.44	0.43	0.40	$\cdots$	0.30	0.27	0.24	0.22

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

Table 1 summarizes the evaluations of out-of-sample dynamic risk for DRM policies trained for 1 year maturity at risk level $\tau = 90\%$ then applied to options of different maturities ranging from 12 months to 1 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. 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.

Table 1 also reports the out-of-sample static risk for both SRM policies and DRM policies. The results are interesting and perhaps surprising. First, unlike with DRM, the static risk of SRM policies for the seller (resp. buyer) can 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 here 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 SRM violates the time consistency property. On the other hand, the DRM policies outperform SRM policies in terms of static risk exposure and also generate consistent policies across time, even though they were trained using a different risk measure. 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.

[4]Note that this metric is available neither for the case of basket option nor in a data-driven environment.

Finally, Table 1 reports 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).[5] One first confirms that the RL based estimate is of high quality, with a maximum approximation error of 0.01 over all maturities. Moreover, we can see that the prices for the SRM polices are generally higher than the prices for the DRM polices, perhaps due to the fact that it is the writer that benefits most from the improved DRM policy than the buyer, as he is more exposed to tail risks in this transaction. We further refer the reader to section B.2 of the appendix for additional results regarding the performance of SRM and DRM in this vanilla option setting.

4.4 Basket option Hedging and Pricing

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: APPL, AMZN, FB, JPM, and GOOGL, where the option payoff is determined by the average price of the underlyings. In this section, dynamic risk is only estimated using the RL based estimator defined in Section 4.3, given that exact DP resolution has become intractable.

Table 2: The out-of-sample dynamic and static 90%-expectile risk imposed to the two sides of basket at-the-money call options. Associated ERPs under the DRM are also compared.

		Time to maturity									
Policy	Est.[†]	12	11	10	9	8	$\cdots$	4	3	2	1
Dynamic 90%-expectile risk											
Writer's DRM	RL	3.92	3.62	3.38	3.15	2.95	$\cdots$	2.00	1.70	1.39	1.10
Buyer's DRM	RL	-0.48	-0.49	-0.51	-0.52	-0.50	$\cdots$	-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	$\cdots$	1.61	1.45	1.26	0.94
Writer's DRM	ED	2.38	2.28	2.18	2.06	1.96	$\cdots$	1.51	1.39	1.20	0.92
Buyer's SRM	ED	-1.31	-1.24	-1.15	-1.01	-0.94	$\cdots$	-0.56	-0.48	-0.36	-0.22
Buyer's SRM	ED	-1.39	-1.32	-1.24	-1.13	-1.07	$\cdots$	-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	$\cdots$	1.24	1.04	0.86	0.70
SRM	RL	2.23	2.10	2.01	1.91	1.79	$\cdots$	1.21	1.03	0.92	0.82

[†] Estimation (Est.) is either made based on reinforcement learning (RL), or the empirical distribution (ED).

Table 2 presents the dynamic risk obtained from training the DRM policy for a one year maturity option and applying it on the test data for maturity ranging from 1 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, while for the writer the monotonic behavior seems to be slightly perturbed by estimation error. The table also compares the static risk under DRM and SRM. 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 more 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, Table 2 presents the equal risk prices computed based on our RL based out-of-sample dynamic risk estimator. Once again, the higher ERP price for the SRM policy are notable, 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. Further details are presented in section B.3 of the Appendix.

[5] Note that in a real 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 statistical biases.

REFERENCES

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.

Alexandre Carbonneau and Frédéric Godin. Equal risk pricing of derivatives with deep hedging. *Quantitative Finance*, pp. 1–16, 2020.

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. Deep equal risk pricing of financial derivatives with non-translation invariant risk measures, 2021b.

Dotan Di Castro, J. Oren, and Shie Mannor. Practical risk measures in reinforcement learning. *ArXiv*, abs/1908.08379, 2019.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in MDPs. *Advances in neural infor- mation processing systems*, abs/1406.3339:3509–3517, 2014.

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

Hans Föllmer and Alexander Schied. *Stochastic finance*. de Gruyter, 2016.

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

Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. On the convergence and optimality of policy gradient for markov coherent risk, 2021.

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. Equal risk pricing and hedging of financial derivatives with convex risk measures. *arXiv preprint arXiv:2002.02876*, 2020.

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*, 2021.

LA Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. *Advances in neural infor- mation processing systems*, abs/1406.3339:252–260, 2013.

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.