OptionRL: Estimating with Differential Equations (Draft ver.)

Dongsheng Hou*
Department of Computer Science and Engineering
Southern University of Science and Technology
12410421@mail.sustech.edu.cn

Yanqiao Chen*
Department of Computer Science and Engineering
Southern University of Science and Technology
12412115@mail.sustech.edu.cn

February 4, 2026

Contents

1 Introduction

2 Related Works

3 OptionRL

We proposed OptionRL, a novel framework that integrates the concept of options into reinforcement learning to enhance decision-making processes. OptionRL leverages differential equations to model the dynamics of options, allowing for more efficient learning and execution of complex tasks.

Usually in RL tasks, we encountered environments with sparse rewards, which makes it difficult for agents to learn optimal policies. To address this challenge, we introduce the concept of options pricing, which allow agents to refine their policies with both present values and future expected rewards. By incorporating options, agents can make more informed decisions, leading to improved performance in environments with sparse rewards.

3.1 Levy Pricing

Like classical RL algorithms, OptionRL also relies on the Bellman equation to estimate the value functions. However, we extend the traditional Bellman equation by incorporating differential equations to model the evolution of options over time. This allows us to capture the dynamics of options more accurately, leading to better value function estimates.

In the OptionRL framework, we apply the two essential assumptions:

Assumption 3.1 (Levy Process Assumption)**.** *The noise term in the environment follows a Levy process.*

The another one is the classical assumption of Black-Scholes-Merton model:

Assumption 3.2 (Neutral Risk Assumption)**.** *The expected return of the option is the risk-free interest rate.*

Under the following assumptions, we can derive the differential equations that govern the evolution of options in the OptionRL framework. These equations allow us to estimate the value functions more accurately, leading to improved decision-making capabilities for agents.

Definition 3.1 (Levy Process)**.** *A stochastic process $X = \{X_t, t \geq 0\}$ is called a Levy process if it satisfies the following properties:*

1. $X_0 = 0$ *almost surely.*
2. X_t *has independent increments: for any* $0 \leq t_0 < t_1 < \ldots < t_n$*, the random variables* $X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}$ *are independent.*
3. X_t *has stationary increments: for any* $s, t \geq 0$*, the distribution of* $X_{t+s} - X_s$ *depends only on* t*.*
4. X_t *is stochastically continuous: for any* $t \geq 0$ *and* $\epsilon > 0$*,* $\lim_{h \to 0} P(|X_{t+h} - X_t| > \epsilon) = 0$*.*

Such process can be rewritten as in differential form:

$$\mathrm{d}X_t = \mu(X_{t-})\mathrm{d}t + \sigma(X_{t-})\mathrm{d}W_t + \int_{\mathbb{R}\setminus\{0\}} \gamma(X_{t-}, z)\tilde{N}(\mathrm{d}t, \mathrm{d}z)$$

Applying the Ito formula for Levy processes, we can derive the following differential equation for the option price $V(t, S_t)$.

Theorem 3.3 (Levy Option Pricing Equation)**.**

$$\frac{\partial V}{\partial t} + rS_t\frac{\partial V}{\partial S_t} + \frac{1}{2}\sigma^2 S_t^2\frac{\partial^2 V}{\partial S_t^2} + \int_{\mathbb{R}\setminus\{0\}} \Big[V(t, S_t + \gamma(S_t, z)) - V(t, S_t) - \gamma(S_t, z)\frac{\partial V}{\partial S_t}\,\mathbf{1}_{\{|z|<1\}}\Big]\nu(\mathrm{d}z) - rV = 0$$

Where r is the risk-free interest rate, σ is the volatility of the underlying asset, and ν is the Levy measure associated with the jump component of the process.

Under the neutral risk assumption, we can solve the above differential equation to obtain the option price $V(t, S_t)$. This price can then be used to refine the value function estimates in the OptionRL framework, leading to improved decision-making capabilities for agents.

Theorem 3.4 (OptionRL Pricing Equation)**.** *The price of the value of a agent is given by the following differential equation:*

$$C_t = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[R_T|\mathcal{F}_t]$$

Where C_t is the option price at time t, R_T is the reward at terminal time T, and $\mathbb{Q}$ is the risk-neutral measure.

3.2 OptionRL Framework

We treat the state value of an agent as an option, which can be priced using the Levy option pricing equation. By incorporating the option price into the value function estimates, we can refine the agent's policy and improve its decision-making capabilities.

This means that, the agent will be rewarded not only based on the immediate rewards it receives from the environment, but also based on the expected future rewards, discounted by the risk-free interest rate. This allows the agent to make more informed decisions with prior knowledge, leading to improved performance in environments with sparse rewards.

The idea of OptionRL can be described in a single equation:

$$C_t = \mathrm{LevyPrice}(S_t, K, T, r, \sigma, \nu)$$

And the action value function can be defined as:

$$\begin{aligned} Q^{\pi}(s, a) &= r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a)C_{t+1}(s') \\ &= r(s, a) + \gamma\mathbb{E}_{s'\sim\mathbb{P}}\left[e^{-r(T-(t+1))}\mathbb{E}^{\mathbb{Q}}[R_T|\mathcal{F}_{t+1}]\right] \end{aligned}$$

We introduce the measure transformation technique via the Radon–Nikodym derivative.

Theorem 3.5 (Radon–Nikodym Derivative)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let $\mathbb{Q} \ll \mathbb{P}$ with Radon–Nikodym density $\Lambda := \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$. Then for any integrable random variable X and any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, we have*

$$\mathbb{E}^{\mathbb{Q}}[X] = \mathbb{E}^{\mathbb{P}}[\Lambda X], \qquad \mathbb{E}^{\mathbb{Q}}[X \mid \mathcal{G}] = \frac{\mathbb{E}^{\mathbb{P}}[\Lambda X \mid \mathcal{G}]}{\mathbb{E}^{\mathbb{P}}[\Lambda \mid \mathcal{G}]}$$

whenever the denominators are almost surely positive.

Applying this to $X = R_T$ and $\mathcal{G} = \mathcal{F}_{t+1}$ with $\Lambda_T := \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\big|_{\mathcal{F}_T}$, the action-value function can be rewritten purely under $\mathbb{P}$ as

$$\begin{aligned} Q^{\pi}(s,a) &= r(s,a) + \gamma\, \mathbb{E}_{s' \sim \mathbb{P}}\Big[e^{-r(T-(t+1))}\, \mathbb{E}^{\mathbb{Q}}[R_T \mid \mathcal{F}_{t+1}]\Big] \\ &= r(s,a) + \gamma e^{-r(T-(t+1))}\, \mathbb{E}_{s' \sim \mathbb{P}}\left[\frac{\mathbb{E}^{\mathbb{P}}[\Lambda_T R_T \mid \mathcal{F}_{t+1}]}{\mathbb{E}^{\mathbb{P}}[\Lambda_T \mid \mathcal{F}_{t+1}]}\right]. \end{aligned}$$

4 Theoretical Analysis

4.1 MDP Formulation for OptionRL

4.2 Convergence Analysis

4.3 Variance Analysis

A Proofs

A.1 Derivation of Levy Option Pricing Equation (Theorem 3.2)

To derive the partial integro-differential equation (PIDE) for option pricing under Levy processes, we employ the principle of no-arbitrage and Ito's Lemma for semimartingales with jumps.

Step 1: Dynamics of the Asset Price Assume the underlying asset price S_t follows a geometric process driven by a Levy process. Under the risk-neutral measure $\mathbb{Q}$, the dynamics of S_t (assuming it pays no dividends) are governed by:

$$\mathrm{d}S_t = S_{t-}\left(r\mathrm{d}t + \sigma\mathrm{d}W_t + \int_{\mathbb{R}}(e^z - 1)\tilde{N}(\mathrm{d}t, \mathrm{d}z)\right)$$

where r is the risk-free rate, W_t is a standard Brownian motion under $\mathbb{Q}$, and $\tilde{N}(\mathrm{d}t, \mathrm{d}z) = N(\mathrm{d}t, \mathrm{d}z) - \nu(\mathrm{d}z)\mathrm{d}t$ is the compensated Poisson random measure with Levy measure ν. The jump size is represented by $e^z - 1$, so that $S_t = S_{t-}e^z$ after a jump of size z.

For a general jump function $\gamma(S_t, z)$, the SDE is:

$$\mathrm{d}S_t = rS_t\mathrm{d}t + \sigma S_t\mathrm{d}W_t + \int_{\mathbb{R}}\gamma(S_{t-}, z)\tilde{N}(\mathrm{d}t, \mathrm{d}z)$$

Step 2: Ito's Lemma for Jump-Diffusion Let $V(t, S_t)$ be the price of the option at time t. By Ito's Lemma for semimartingales, the differential $\mathrm{d}V(t, S_t)$ is given by:

$$\begin{aligned}\mathrm{d}V = &\frac{\partial V}{\partial t}\mathrm{d}t + \frac{\partial V}{\partial S}\mathrm{d}S_t + \frac{1}{2}\frac{\partial^2 V}{\partial S^2}\mathrm{d}\langle S^c\rangle_t \\ &+ \int_{\mathbb{R}\setminus\{0\}}\Big(V(t, S_{t-} + \gamma(S_{t-}, z)) - V(t, S_{t-})\Big)\tilde{N}(\mathrm{d}t, \mathrm{d}z) \\ &+ \int_{\mathbb{R}\setminus\{0\}}\Big(V(t, S_{t-} + \gamma(S_{t-}, z)) - V(t, S_{t-}) - \frac{\partial V}{\partial S}(t, S_{t-})\,\gamma(S_{t-}, z)\,\mathbf{1}_{\{|z|<1\}}\Big)\,\nu(\mathrm{d}z)\,\mathrm{d}t\end{aligned}$$

Here S^c denotes the continuous martingale part of S, so $\mathrm{d}S^c_t = \sigma S_t\,\mathrm{d}W_t$ and $\mathrm{d}\langle S^c\rangle_t = \sigma^2 S_t^2\,\mathrm{d}t$; the jump size is $\Delta S_t = \gamma(S_{t-}, z)$. (The truncation $\mathbf{1}_{\{|z|<1\}}$ is the standard Levy–Ito convention to ensure the integral is well-defined.) Actually, it is more convenient to work with the discounted price process $\tilde{V}_t = e^{-rt}V(t, S_t)$. For $\tilde{V}_t$ to be a martingale under $\mathbb{Q}$, the drift term of $\mathrm{d}(e^{-rt}V(t, S_t))$ must be zero.

Applying Ito's product rule:

$$\mathrm{d}(e^{-rt}V) = -re^{-rt}V\mathrm{d}t + e^{-rt}\mathrm{d}V$$

Expanding $\mathrm{d}V$ and collecting the $\mathrm{d}t$ terms (drift): The drift comes from:

- Time decay: $\frac{\partial V}{\partial t}$
- Continuous drift of S: Since $\mathbb{E}^{\mathbb{Q}}[\mathrm{d}S_t] = rS_t\mathrm{d}t$, the effective drift contributing to the Ito term involves the compensator for the jump.
- Mean magnitude of jumps: $\int_{\mathbb{R}}[V(t, S+\gamma(S,z)) - V(t,S) - \frac{\partial V}{\partial S}\gamma(S,z)]\nu(\mathrm{d}z)$

Specifically, we use the property that $\tilde{N}(\mathrm{d}t, \mathrm{d}z) = N(\mathrm{d}t, \mathrm{d}z) - \nu(\mathrm{d}z)\mathrm{d}t$. The stochastic integral with respect to $\tilde{N}$ is a martingale. The remaining $\mathrm{d}t$ terms from the jump part form the integral operator.

The condition that the drift is zero yields:

$$-rV+\frac{\partial V}{\partial t}+rS_t\frac{\partial V}{\partial S}+\frac{1}{2}\sigma^2 S_t^2\frac{\partial^2 V}{\partial S^2}+\int_{\mathbb{R}\setminus\{0\}}\Big[V(t,S+\gamma(S,z))-V(t,S)-\gamma(S,z)\frac{\partial V}{\partial S}(t,S)\,\mathbf{1}_{\{|z|<1\}}\Big]\nu(\mathrm{d}z)=0$$

This concludes the derivation of the PIDE in Theorem 3.2.

A.2 Proof for theorem 3.4

Proof. The proof relies on the Fundamental Theorem of Asset Pricing, which states that under the assumption of no arbitrage, there exists an equivalent martingale measure (risk-neutral measure) $\mathbb{Q}$ such that the discounted price process of any tradable asset is a martingale.

Assume $\mathbb{E}^{\mathbb{Q}}[|R_T|] < \infty$ so the conditional expectations below are well-defined.

Let C_t be the price of the value option at time t. By Feynman-Kac formula, we construct the discounted price process $M_t = e^{-rt}C_t$. Under the risk-neutral measure $\mathbb{Q}$, M_t satisfies the martingale property:

$$M_t = \mathbb{E}^{\mathbb{Q}}[M_T|\mathcal{F}_t]$$

where $\mathcal{F}_t$ represents the filtration (information history) up to time t.

Substituting the definition of M_t back into the equation:

$$e^{-rt}C_t = \mathbb{E}^{\mathbb{Q}}[e^{-rT}C_T|\mathcal{F}_t]$$

At the terminal time T, the option expires and its value is determined entirely by the final payoff. In our RL context, this payoff is the terminal reward R_T:

$$C_T = R_T$$

Substituting C_T into the conditional expectation:

$$e^{-rt}C_t = \mathbb{E}^{\mathbb{Q}}[e^{-rT}R_T|\mathcal{F}_t]$$

Since r and T are constants relative to the expectation at time T, we can factor out the discounting term on the Right Hand Side (RHS), but usually, we just move the exponential term from LHS to RHS:

$$C_t = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[R_T|\mathcal{F}_t]$$

This confirms that the current option price is the risk-neutral expected present value of the future terminal reward. □

B Experiment Details

B.1 Practical Implementation Algorithm

While the theoretical derivation involves solving complex Partial Integro-Differential Equations (PIDE) as shown in Theorem 3.2, solving these equations numerically at each time step during training is computationally prohibitive.

To bridge the gap between theory and practice, we implement OptionRL using **Potential-Based Reward Shaping (PBRS)**. Instead of solving the PIDE directly, we utilize the analytical solution (or a robust approximation) of the Merton Jump Diffusion model as a potential function $\Phi(s)$. This potential function captures the "theoretical value" of a state under risk-neutral measure, which is then used to reshape the sparse reward signal, accelerating the convergence of the model-free Q-learning agent.

The implementation details are described in Algorithm 1.

Algorithm 1 OptionRL via Merton Potential Shaping

Require: State space $\mathcal{S}$, Action space $\mathcal{A}$, Goal Threshold K
Require: Hyperparameters: Volatility σ, Jump Intensity λ, Risk-free rate r
Ensure: Optimal Policy π^*

Initialize $Q(s, a)$ arbitrarily for all $s \in \mathcal{S}, a \in \mathcal{A}$
Initialize Potential $\Phi(s) = 0$
function GetMertonPotential(s)
    Map state s to proxy asset price S_t (e.g., semantic proximity)
    Map state s to time-to-go T
    $d_1 \leftarrow \frac{\ln(S_t/K)+(r+\frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}$
    $P_{BS} \leftarrow S_t\Phi_{norm}(d_1) - Ke^{-rT}\Phi_{norm}(d_1 - \sigma\sqrt{T})$ ▷ Vanilla Black-Scholes
    $P_{Merton} \leftarrow P_{BS} \cdot (1 + \lambda T)$ ▷ Approx. Jump Premium
    return P_{Merton}
end function
for each episode $1 \dots M$ **do**
    Initialize state s
    repeat
        Choose action a from s using ϵ-greedy policy derived from Q
        Take action a, observe reward r_{env} and next state s'
        $\Phi_t \leftarrow$ GetMertonPotential(s)
        $\Phi_{t+1} \leftarrow$ GetMertonPotential(s')
        $F_t \leftarrow \gamma\Phi_{t+1} - \Phi_t$ ▷ Calculate Shaping Reward
        $r_{total} \leftarrow r_{env} + F_t$
        $Q(s, a) \leftarrow Q(s, a) + \alpha[r_{total} + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
        $s \leftarrow s'$
    until s is terminal
end for