

# 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 3, 2026

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Related Works</b>	<b>2</b>
<b>3</b>	<b>OptionRL</b>	<b>2</b>
3.1	Levy Pricing . . . . .	2
3.2	OptionRL Framework . . . . .	3
<b>4</b>	<b>Theoretical Analysis</b>	<b>5</b>
4.1	MDP Formulation for OptionRL . . . . .	5
4.2	Convergence Analysis . . . . .	5
4.3	Variance Analysis . . . . .	5
<b>A</b>	<b>Proofs</b>	<b>5</b>
A.1	Derivation of Levy Option Pricing Equation (Theorem 3.2) . . . . .	5
A.2	Proof for theorem 3.4 . . . . .	6
<b>B</b>	<b>Experiment Details</b>	<b>6</b>

# 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  $0 \leq t_0 < t_1 < \dots < t_n$ , the random variables  $X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}}$  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  $\epsilon > 0$ ,  $\lim_{h \rightarrow 0} P(|X_{t+h} - X_t| > \epsilon) = 0$ .

Such process can be rewritten as in differential form:

$$dX_t = \mu(X_{t-})dt + \sigma(X_{t-})dW_t + \int_{\mathbb{R} \setminus \{0\}} \gamma(X_{t-}, z)\tilde{N}(dt, dz)$$

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).

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

Where  $r$  is the risk-free interest rate,  $\sigma$  is the volatility of the underlying asset, and  $\nu$  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:

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

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 observe a double expectation structure involving both the physical measure  $\mathbb{P}$  (environment transition) and the risk-neutral measure  $\mathbb{Q}$  (pricing). To simplify, we adopt the **Risk-Neutral RL Assumption**, approximating the environmental dynamics  $\mathbb{P}$  with  $\mathbb{Q}$ . By applying the **Tower**

**Property** (Law of Iterated Expectations) of conditional expectations, i.e.,  $\mathbb{E}_t[\mathbb{E}_{t+1}[X]] = \mathbb{E}_t[X]$ , and assuming the discount factor satisfies  $\gamma = e^{-r}$ , the double expectation collapses:

$$\mathbb{E}^{\mathbb{Q}}[C_{t+1} | \mathcal{F}_t] = e^{-r(T-(t+1))} \mathbb{E}^{\mathbb{Q}}[\mathbb{E}^{\mathbb{Q}}[R_T | \mathcal{F}_{t+1}] | \mathcal{F}_t] = e^{-r(T-(t+1))} \mathbb{E}^{\mathbb{Q}}[R_T | \mathcal{F}_t] = e^r C_t.$$

Therefore, with  $\gamma = e^{-r}$ , we obtain  $\gamma \mathbb{E}^{\mathbb{Q}}[C_{t+1} | \mathcal{F}_t] = C_t$ .

Therefore, under this option-pricing interpretation, and with the approximation  $\mathbb{P} \approx \mathbb{Q}$  together with  $\gamma = e^{-r}$ , the Bellman update reduces to a simple one-step form in which the option price  $C_t$  can be used directly as the bootstrapped value:

$$Q^{\pi}(s, a) \approx r(s, a) + C_t(s) \tag{1}$$

This suggests that, rather than learning the value function purely from data, one can use the Levy option price as a model-informed proxy for the state value.

## 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:

$$dS_t = S_{t-} \left( rdt + \sigma dW_t + \int_{\mathbb{R}} (e^z - 1) \tilde{N}(dt, dz) \right)$$

where  $r$  is the risk-free rate,  $W_t$  is a standard Brownian motion under  $\mathbb{Q}$ , and  $\tilde{N}(dt, dz) = N(dt, dz) - \nu(dz)dt$  is the compensated Poisson random measure with Levy measure  $\nu$ . 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:

$$dS_t = rS_t dt + \sigma S_t dW_t + \int_{\mathbb{R}} \gamma(S_{t-}, z) \tilde{N}(dt, dz)$$

**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  $dV(t, S_t)$  is given by:

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

Here  $S^c$  denotes the continuous martingale part of  $S$ , so  $dS_t^c = \sigma S_t dW_t$  and  $d\langle S^c \rangle_t = \sigma^2 S_t^2 dt$ ; 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  $d(e^{-rt} V(t, S_t))$  must be zero.

Applying Ito's product rule:

$$d(e^{-rt} V) = -re^{-rt} V dt + e^{-rt} dV$$

Expanding  $dV$  and collecting the  $dt$  terms (drift): The drift comes from:

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

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

The condition that the drift is zero yields:

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

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.  $\square$

## B Experiment Details