

Reinforcement Learning for Jump-Diffusions, with Financial Applications

Xuefeng Gao* Lingfei Li† Xun Yu Zhou‡

August 22, 2025

Abstract

We study continuous-time reinforcement learning (RL) for stochastic control in which system dynamics are governed by jump-diffusion processes. We formulate an entropy-regularized exploratory control problem with stochastic policies to capture the exploration–exploitation balance essential for RL. Unlike the pure diffusion case initially studied by Wang et al. (2020), the derivation of the exploratory dynamics under jump-diffusions calls for a careful formulation of the jump part. Through a theoretical analysis, we find that one can simply use the same policy evaluation and q-learning algorithms in Jia and Zhou (2022a, 2023), originally developed for controlled diffusions, without needing to check a priori whether the underlying data come from a pure diffusion or a jump-diffusion. However, we show that the presence of jumps ought to affect parameterizations of actors and critics in general. We investigate as an application the mean–variance portfolio selection problem with stock price modelled as a jump-diffusion, and show that both RL algorithms and parameterizations are invariant with respect to jumps. Finally, we present a detailed study on applying the general theory to option hedging.

Keywords. Reinforcement learning, continuous time, jump-diffusions, exploratory formulation, well-posedness, Hamiltonian, martingale, q-learning.

1 Introduction

Recently there is an upsurge of interest in continuous-time reinforcement learning (RL) with continuous state spaces and possibly continuous action spaces. Continuous RL problems are important because: 1) many if not most practical problems are naturally continuous in time (and in space), such as autonomous driving, robot navigation, video game play and high frequency trading; 2) while one can discretize time upfront and turn a continuous-time problem into a discrete-time MDP, it has been known, indeed shown experimentally in e.g., Munos (2006), Tallec et al. (2019) and Park et al. (2021), that this approach is very sensitive to time discretization and performs poorly with small time steps; 3) there are more analytical tools available for the continuous setting that

*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China. E-mail: xfgao@se.cuhk.edu.hk

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China. E-mail: lfi@se.cuhk.edu.hk

‡Department of Industrial Engineering and Operations Research and The Data Science Institute, Columbia University, New York, NY 10027, USA. Email: xz2574@columbia.edu

enable a rigorous and thorough analysis leading to interpretable (instead of black-box) and general (instead of ad hoc) RL algorithms.

Compared with the vast literature of RL for MDPs, continuous-time RL research is still in its infancy with the latest study focusing on establishing a rigorous mathematical theory and devising resulting RL algorithms. This strand of research starts with Wang et al. (2020) that introduces a mathematical formulation to capture the essence of RL – the exploration–exploitation tradeoff – in the continuous setting, followed by a “trilogy” (Jia and Zhou 2022a,b, 2023) that develops intertwining theories on policy evaluation, policy gradient and q-learning, respectively. The common underpinning of the entire theory is the martingale property of certain stochastic processes, the enforcement of which naturally leads to various temporal difference algorithms to train and learn q-functions, value functions and optimal (stochastic) policies. The research is characterized by carrying out all the analysis in the continuous setting, and discretizing time only at the final, implementation stage for approximating the integrated rewards and the temporal difference. The theory has been adapted and extended in different directions; see e.g. Reisinger and Zhang (2021), Guo et al. (2022), Dai et al. (2023), as well as employed for applications; see e.g. Wang and Zhou (2020), Huang et al. (2022), Gao et al. (2022), Wang et al. (2023), and Wu and Li (2024).

The study so far has been predominantly on pure diffusion processes, namely the state processes are governed by controlled stochastic differential equations (SDEs) with a drift part and a diffusion one. While it is reasonable to model the underlying data generating processes as diffusions within a short period of time, sudden and drastic changes can and do happen over time. An example is a stock price process: Although it is approximately a diffusion over a sufficiently short period, it may respond dramatically to a surprisingly good or bad earning report. Other examples include neuron dynamics (Giraud and Sacerdote 1997), stochastic resonance (Gammaitoni et al. 1998) and climate data (Goswami et al. 2018). It is therefore natural and necessary to extend the continuous RL theory and algorithms to the case when jumps are present. This is particularly important for decision makings in financial markets, where it has been well recognized that using jumps to capture large sudden movements provides a more realistic way to model market dynamics; see the discussions in Chapter 1 of Cont and Tankov (2004). The financial modeling literature with jumps dates back to the seminal work of Merton (1976), who extends the classical Black–Scholes model by introducing a compound Poisson process with normally distributed jumps in the log returns. Since then alternative jump size distributions have been proposed in e.g. Kou (2002) and Cai and Kou (2011). Empirical success of jump-diffusion models has been documented for many asset classes; see Bates (1991), Andersen et al. (2002), and Ait-Sahalia et al. (2012) for stocks and stock indices, Bates (1996) for exchange rates, Das (2002) for interest rates, and Li and Linetsky (2014) for commodities, among many others.

This paper makes two major contributions. The first is several mathematical results that form the foundation of the RL theory for jump-diffusions, including a suitable exploratory formulation, a well-posedness proof of the resulting exploratory SDE, and the convergence of the value functions of the grid sample state processes to the exploratory value function. Wang et al. (2020) apply the classical stochastic relaxed control to model the exploration or randomization prevalent in RL, and derive an exploratory state equation that dictates the dynamics of the “average” of infinitely many state processes generated by repeatedly sampling from the same exploratory, stochastic policy. The drift and variance coefficients of the exploratory SDE are the means of those coefficients against the given stochastic policy (which is a probability distribution) respectively. The derivation therein

is based on a law of large number (LLN) argument to the first two moments of the diffusion process. That argument fails for jump-diffusions which are not uniquely determined by the first two moments. We overcome this difficulty by analyzing instead the infinitesimal behavior of the grid sample state process, which is obtained by sampling the stochastic policy on a discrete time grid. Based on this analysis, we identify the dynamic of the exploratory state process. Inspired by Kushner (2000) who studies relaxed control for jump-diffusions, we formulate the exploratory SDE by extending the original Poisson random measures for jumps to capture the effect of random exploration. It should be noted that, like almost all the earlier works on relaxed control, Kushner (2000) is motivated by answering the theoretical question of whether an optimal control exists, as randomization convexifies the universe of control strategies. In comparison, our formulation is guided by the practical motivation of exploration for learning. There is also another subtle but important difference. We consider stochastic *feedback* policies while Kushner (2000) does not. This in turn creates technical issues in studying the well-posedness of the exploratory SDE in our framework. Although the exploratory state process plays an important role in theoretical analysis, it is unobservable and hence cannot be used in the implementation of any RL algorithm. Instead, we consider the grid sample state processes, which is what one obtains from the environment by following the stochastic policy on a discrete time grid. We prove that the value functions of grid sample state processes converge to that of the exploratory state process at first order as the grid size shrinks to zero, which extends a major result in Jia et al. (2025) from diffusions to jump-diffusions.

The second main contribution is several implications regarding the impact of jumps on RL algorithm design. Thanks to the established exploratory formulation, we can define the Hamiltonian that, compared with the pure diffusion counterpart, has to include an additional term corresponding to the jumps. The resulting HJB equation – called the exploratory HJB – is now a partial integro-differential equation (PIDE) instead of a PDE due to that additional term. However, when expressed *in terms of the Hamiltonian*, the exploratory HJB equation has exactly the same form as that in the diffusion case. This leads to several completely identical *statements* of important results, including the optimality of the Gibbs exploration, definition of a q-function, and martingale characterizations of value functions and q-functions. Here by “identical” we mean in terms of the Hamiltonian; in other words, these statements differ between diffusions and jump-diffusions entirely because the Hamiltonian is defined differently (which also causes some differences in the proofs of the results concerned). Most important of all, in the resulting RL algorithms, the Hamiltonian (or equivalently the q-function) can be computed using temporal difference of the value function by virtue of the Itô lemma; as a result the algorithms are completely identical no matter whether or not there are jumps. This has a significant practical implication: we can just use the same RL algorithms without the need of checking in advance whether the underlying data come from a pure diffusion or a jump-diffusion. It is significant for the following reason. In practice, data are always observed or sampled at discrete times, no matter how frequent they arrive. Thus, we encounter successive discontinuities along the sample trajectory even when the data actually come from a diffusion process. There are some criteria that can be used to check whether the underlying process is a diffusion or a jump-diffusion, e.g. Aït-Sahalia et al. (2012), Wang and Zheng (2022). But these methods typically require data with very high frequency to be effective, which may not always be available. In addition, noises must be properly handled for them to work.

Even though we can apply the same RL algorithms irrespective of the presence of jumps, the *parametrization* of the policy and value function may still depend on it, if we try to exploit certain special structure of the problem instead of using general neural networks for parameterization. In-

deed, we give an example in which the optimal exploratory policy is Gaussian when there are no jumps, whereas an optimal policy either does not exist or becomes non-Gaussian when there are jumps. However, in the mean–variance (MV) portfolio selection problem we present as a concrete application, the optimal Gibbs exploration measure again reduces to Gaussian and the value function is quadratic as in Wang and Zhou (2020), both owing to the inherent linear–quadratic (LQ) structure of the problem. Hence in this particular case jumps do not even affect the parametrization of the policy and value function/q-function for learning.

We also consider MV hedging of options as another application. This is a non-LQ problem and hence more difficult to solve than MV portfolio selection. The MV hedging problem has been studied in various early works, such as Schweizer (1996, 2001) and Lim (2005). Here, we introduce the entropy-regularized MV hedging objective for an asset following a jump-diffusion and derive analytical representations for the optimal stochastic policy, which is again Gaussian, as well as the optimal value function. We use these representations to devise an actor–critic algorithm to learn the optimal hedging policy from data.

We compare our work with five recent related papers. (1) Bender and Thuan (2023) consider the continuous-time MV portfolio selection problem with exploration under a jump-diffusion setting. Our paper differs from theirs in two main aspects. First, they consider the specific MV problem only while we study RL for general controlled jump-diffusions with MV as an instance of application. Second, they do not develop algorithms based on their solution of the exploratory MV portfolio selection, which we do in this paper. (2) During the revision process of our paper, we became aware of the concurrent study Bender and Thuan (2024), which independently investigates RL for jump-diffusions using a random measure approach. Specifically, they derive a grid sample state process as an equation driven by suitable random measures and establish a limit theorem for these measures as the sampling grid size approaches zero. By contrast, our work focuses on developing q-learning theory and algorithms for jump-diffusions, with a particular emphasis on applying this general framework to two financial applications. (3) Bo et al. (2024) develop q-learning for jump-diffusions by using Tsallis’ entropy for regularization instead of Shannon’s entropy considered in our paper and Jia and Zhou (2022b, 2023).¹ While this entropy presents an interesting alternative for developing RL algorithms, it may make the exploratory control problem less tractable to solve and lead to policy distributions that are inefficient to sample for exploration in certain applications. (4) Guo et al. (2023) consider continuous-time RL for linear–convex models with jumps. The scope and motivation are different from ours: They focus on the Lipschitz stability of feedback controls for this special class of control problems where the diffusion and jump terms are not controlled, and propose a least-square model-based algorithm and obtain sublinear regret guarantees in the episodic setting. By contrast, we consider RL for general jump-diffusions and develop model-free algorithms without considering regret bounds. (5) Denkert et al. (2024) aim to unify certain types of stochastic control problems by considering the so-called randomized control formulation which leads to the same optimal value functions as those of the original problems. They develop a policy gradient representation and actor–critic algorithms for RL. The randomized control formulation is fundamentally different from the framework we are considering: therein the control is applied at a set of random time points generated by a random point process.

The remainder of the paper is organized as follows. In Section 2, we discuss the formulation of the control problem and the exploratory state process. In Section 3, we present the theory

¹The paper Bo et al. (2024) came to our attention after a previous version of our paper was completed and posted.

of q-learning for jump-diffusions, followed by the discussion of q-learning algorithms in Section 4. In Section 5, we apply the general theory and algorithms to a mean-variance portfolio selection problem, and discuss the impact of jumps. Section 6 presents the application to a mean-variance option hedging problem. Finally, Section 7 concludes. The appendix contains the proofs and some results regarding the convergence of the value functions of the grid sample state processes to the value function of the exploratory state process.

2 Problem Formulation and Preliminaries

For readers' convenience, we first recall some basic concepts for one-dimensional (1D) Lévy processes, which can be found in standard references such as Sato (1999) and Applebaum (2009). A 1D process $L = \{L_t : t \geq 0\}$ is a Lévy process if it is continuous in probability, has stationary and independent increments, and $L_0 = 0$ almost surely. Denote the jump of L at time t by $\Delta L_t = L_t - L_{t-}$, and let \mathbf{B}_0 be the collection of Borel sets of \mathbb{R} whose closure does not contain 0. The Poisson random measure (or jump measure) of L is defined as

$$N(t, B) = \sum_{s:0 < s \leq t} 1_B(\Delta L_s), \quad B \in \mathbf{B}_0,$$

which gives the number of jumps up to time t with jump size in a Borel set B away from 0. The Lévy measure ν of L is defined by $\nu(B) = \mathbb{E}[N(1, B)]$ for $B \in \mathbf{B}_0$, which shows the expected number of jumps in B in unit time, and $\nu(B)$ is finite. For any $B \in \mathbf{B}_0$, $\{N(t, B) : t \geq 0\}$ is a Poisson process with intensity given by $\nu(B)$. The differential forms of these two measures are written as $N(dt, dz)$ and $\nu(dz)$, respectively. If ν is absolutely continuous, we write $\nu(dz) = \nu(z)dz$ by using the same letter for the measure and its density function. The Lévy measure ν must satisfy the integrability condition

$$\int_{\mathbb{R}} \min\{z^2, 1\} \nu(dz) < \infty. \quad (1)$$

However, it is not necessarily a finite measure on $\mathbb{R}/\{0\}$ but always a σ -finite measure. The Lévy process L is said to have jumps of finite (infinite activity) if $\int_{\mathbb{R}} \nu(dz) < \infty$ ($= \infty$). The number of jumps on any finite time interval is finite in the former case but infinite in the latter. For any Borel set B with its closure including 0, $\nu(B)$ is finite in the finite activity case but infinite otherwise. Finally, the compensated Poisson random measure is defined as $\tilde{N}(dt, dz) = N(dt, dz) - \nu(dz)dt$. For any $B \in \mathbf{B}_0$, the process $\{\tilde{N}(t, B) : t \geq 0\}$ is a martingale.

Throughout the paper, we use the following notations. By convention, all vectors are column vectors unless otherwise specified. We use \mathbb{R}^k and $\mathbb{R}^{k \times \ell}$ to denote the space of k -dimensional vectors and $k \times \ell$ matrices, respectively. For matrix A , we use A^\top for its transpose, $|A|$ for its Euclidean/Frobenius norm, and write $A^2 := AA^\top$. Given two matrices A and B of the same size, we denote by $A \circ B$ the inner product between A and B , which is given by $\text{tr}(AB^\top)$. For a positive semidefinite matrix A , we write $\sqrt{A} = UD^{1/2}V^\top$, where $A = UDV^\top$ is its singular value decomposition with U, V as two orthogonal matrices and D as a diagonal matrix, and $D^{1/2}$ is the diagonal matrix whose entries are the square root of those of D . We use $f = f(\cdot)$ to denote the function f , and $f(x)$ to denote the function value of f at x . We use $\partial_x f, \partial_x^2 f$ for the first and second partial derivatives of a function f with respect to x . We write the minimum of two values a and b as $a \wedge b$. The notation $\mathcal{U}(B)$ denotes the uniform distribution over set B while $\mathcal{N}(\mu, \Sigma)$ refers to the Gaussian distribution with mean vector μ and covariance matrix Σ .

2.1 Classical stochastic control of jump-diffusions

Consider a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ satisfying the usual hypothesis. Assume that this space is rich enough to support $W = \{W_t : t \geq 0\}$, a standard Brownian motion in \mathbb{R}^m , and ℓ independent one-dimensional (1D) Lévy processes L_1, \dots, L_ℓ , which are also independent of W . Let $N(dt, dz) = (N_1(dt, dz_1), \dots, N_\ell(dt, dz_\ell))^\top$ be the vector of their Poisson random measures, and similarly define $\nu(dz)$ and $\tilde{N}(dt, dz)$. The controlled system dynamics are governed by the following Lévy SDE (Øksendal and Sulem 2007, Chapter 3):

$$dX_s^a = b(s, X_{s-}^a, a_s)ds + \sigma(s, X_{s-}^a, a_s)dW_s + \int_{\mathbb{R}^\ell} \gamma(s, X_{s-}^a, a_s, z)\tilde{N}(ds, dz), \quad s \in [0, T], \quad (2)$$

where

$$b : [0, T] \times \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}^d, \quad \sigma : [0, T] \times \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}^{d \times m} \quad \text{and} \quad \gamma : [0, T] \times \mathbb{R}^d \times \mathcal{A} \times \mathbb{R}^\ell \rightarrow \mathbb{R}^{d \times \ell},$$

a_s is the control or action at time s , $\mathcal{A} \subseteq \mathbb{R}^n$ is the control space, and $a = \{a_s : s \in [0, T]\}$ is the control process assumed to be predictable with respect to $\{\mathcal{F}_s : s \in [0, T]\}$. We denote the k -th column of the matrix γ by γ_k . The goal of stochastic control is, for each initial time-state pair (t, x) of (2), to find the optimal control process a that maximizes the expected total reward:

$$\mathbb{E} \left[\int_t^T e^{-\beta(s-t)} r(s, X_s^a, a_s) ds + e^{-\beta(T-t)} h(X_T^a) \middle| X_t^a = x \right], \quad (3)$$

where $\beta \geq 0$ is a discount factor that measures the time value of the payoff.

The stochastic control problem (2)–(3) is very general; in particular, control processes affect the drift, diffusion and jump coefficients. We now make the following assumption to ensure well-posedness of the problem. Define $\mathbb{R}_K^d := \{x \in \mathbb{R}^d : |x| \leq K\}$.

Assumption 1. *Suppose the following conditions are satisfied by the state dynamics and reward functions:*

- (i) b, σ, γ, r, h are all continuous functions in their respective arguments;
- (ii) (local Lipschitz continuity) for any $K > 0$ and any $p \geq 2$, there exist positive constants C_K and $C_{K,p}$ such that $\forall (t, a) \in [0, T] \times \mathcal{A}, (x, x') \in \mathbb{R}_K^d$,

$$\begin{aligned} |b(t, x, a) - b(t, x', a)| + |\sigma(t, x, a) - \sigma(t, x', a)| &\leq C_K |x - x'|, \\ \sum_{k=1}^{\ell} \int_{\mathbb{R}} |\gamma_k(t, x, a, z_k) - \gamma_k(t, x', a, z_k)|^p \nu_k(dz) &\leq C_{K,p} |x - x'|^p; \end{aligned}$$

- (iii) (linear growth in x) for any $p \geq 1$, there exist positive constants C and C_p such that $\forall (t, x, a) \in [0, T] \times \mathbb{R}^d \times \mathcal{A}$,

$$\begin{aligned} |b(t, x, a)| + |\sigma(t, x, a)| &\leq C(1 + |x| + |a|), \\ \sum_{k=1}^{\ell} \int_{\mathbb{R}} |\gamma_k(t, x, a, z)|^p \nu_k(dz) &\leq C_p(1 + |x|^p + |a|^p); \end{aligned}$$

(iv) there exists a constant $C > 0$ such that

$$|r(t, x, a)| \leq C(1 + |x|^p + |a|^q), \quad |h(x)| \leq C(1 + |x|^p), \quad \forall (t, x, a) \in [0, T] \times \mathbb{R}^d \times \mathcal{A}$$

for some $p \geq 2$ and some $q \geq 1$.

A control process a is admissible if $\mathbb{E}[\int_0^T |a_s|^p ds]$ is finite for any $p \geq 1$. Conditions (i)-(iii) guarantee the existence of a unique strong solution to the controlled Lévy SDE (2) with initial condition $X_t^a = x \in \mathbb{R}^d$ for any admissible control process. Furthermore, for any $p \geq 2$, there exists a constant $C_p > 0$ such that

$$\mathbb{E}_{t,x} \left[\sup_{t \leq s \leq T} |X_s^a|^p \right] \leq C_p(1 + |x|^p); \quad (4)$$

see (Kunita 2004, Theorem 3.2) and (Situ 2006, Theorem 119). With the moment estimate (4), it follows from condition (iv) and the admissibility condition that the expected value in (3) is finite.

Let \mathcal{L}^a be the infinitesimal generator associated with the Lévy SDE (2). Under condition (iii), we have $\int_{\mathbb{R}} |\gamma_k(t, x, a, z)| \nu_k(dz) < \infty$ for $k = 1, \dots, \ell$. Thus, we can write \mathcal{L}^a in the following form:

$$\begin{aligned} \mathcal{L}^a f(t, x) &= \partial_t f(t, x) + b(t, x, a) \circ \partial_x f(t, x) + \frac{1}{2} \sigma^2(t, x, a) \circ \partial_x^2 f(t, x) \\ &\quad + \sum_{k=1}^{\ell} \int_{\mathbb{R}} (f(t, x + \gamma_k(t, x, a, z)) - f(t, x) - \gamma_k(t, x, a, z) \circ \partial_x f(t, x)) \nu_k(dz), \end{aligned} \quad (5)$$

where $\partial_x f \in \mathbb{R}^d$ is the gradient and $\partial_x^2 f \in \mathbb{R}^{d \times d}$ is the Hessian matrix.

We recall Itô's formula, which will be frequently used in our analysis; see e.g. (Øksendal and Sulem 2007, Theorem 1.16). Let X^a be the unique strong solution to (2). For any $f \in C^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d)$, we have

$$\begin{aligned} df(t, X_t^a) &= \partial_t f(t, X_{t-}^a) dt + b(t, X_{t-}^a, a_t) \circ \partial_x f(t, X_{t-}^a) dt + \frac{1}{2} \sigma^2(t, X_{t-}^a, a_t) \circ \partial_x^2 f(t, X_{t-}^a) dt \\ &\quad + \sum_{k=1}^{\ell} \int_{\mathbb{R}} (f(t, X_{t-}^a + \gamma_k(t, X_{t-}^a, a_t, z)) - f(t, X_{t-}^a) - \gamma_k(t, X_{t-}^a, a_t, z) \circ \partial_x f(t, X_{t-}^a)) \nu_k(dz) dt \\ &\quad + \partial_x f(t, X_{t-}^a) \circ \sigma(t, X_{t-}^a, a_t) dW_t + \sum_{k=1}^{\ell} \int_{\mathbb{R}} (f(t, X_{t-}^a + \gamma_k(t, X_{t-}^a, a_t, z)) - f(t, X_{t-}^a)) \tilde{N}_k(dt, dz). \end{aligned} \quad (6)$$

It is known that the Hamilton–Jacobi–Bellman (HJB) equation for the control problem (2)–(3) is given by

$$\sup_{a \in \mathcal{A}} \{r(t, x, a) + \mathcal{L}^a V(t, x)\} - \beta V(t, x) = 0, \quad (t, x) \in [0, T] \times \mathbb{R}^d, \quad (7)$$

$$V(T, x) = h(x),$$

where \mathcal{L}^a is given in (5). Under proper conditions, the solution to the above equation is the optimal value function V^* for control problem (3). Moreover, the following function, which maps a time–state pair to an action:

$$a^*(t, x) = \arg \max_{a \in \mathcal{A}} \{r(t, x, a) + \mathcal{L}^a V^*(t, x)\} \quad (8)$$

is the optimal feedback control policy of the problem, a result called the verification theorem.

Given a smooth function $V(t, x) \in C^{1,2}([0, T] \times \mathbb{R}^d)$, we define the Hamiltonian H by

$$\begin{aligned} H(t, x, a, \partial_x V, \partial_x^2 V, V) &= r(t, x, a) + b(t, x, a) \circ \partial_x V(t, x) + \frac{1}{2} \sigma^2(t, x, a) \circ \partial_x^2 V(t, x) \\ &+ \sum_{k=1}^{\ell} \int_{\mathbb{R}^d} (V(t, x + \gamma_k(t, x, a, z)) - V(t, x) - \gamma_k(t, x, a, z) \circ \partial_x V(t, x)) \nu_k(dz). \end{aligned} \quad (9)$$

Then, the HJB equation (7) can be recast as

$$\begin{aligned} \partial_t V(t, x) + \sup_{a \in \mathcal{A}} \{H(t, x, a, \partial_x V, \partial_x^2 V, V)\} - \beta V(t, x) &= 0, \quad (t, x) \in [0, T] \times \mathbb{R}^d, \\ V(T, x) &= h(x). \end{aligned}$$

What we have just outlined above is the classical *model-based* approach to solving the stochastic control problem (2)–(3), under the assumption that the model primitives b, σ, γ, r, h are known functions. When they are not known, there is no HJB equation (7) to solve and no verification condition (8) to apply, in which case we resort to RL.

2.2 Randomized control and the grid sample state process

A key idea of RL is to explore the unknown environment by randomizing the actions. Let $\pi : (t, x) \in [0, T] \times \mathbb{R}^d \rightarrow \pi(\cdot|t, x) \in \mathcal{P}(\mathcal{A})$ be a given *stochastic* feedback policy, where $\mathcal{P}(\mathcal{A})$ is the set of probability density functions defined on \mathcal{A} . An action $a \in \mathcal{A}$ sampled from $\pi(\cdot|t, x)$ can be expressed as $a = G^\pi(t, x, U)$ for some measurable function G^π and some U following the uniform distribution over $[0, 1]^n$ (n is the dimension of the control space). In the following, we will also interchangeably use the more compact notations $\pi_{t,x}(\cdot)$ for $\pi(\cdot|t, x)$ and $G^{\pi_{t,x}}(\cdot)$ for $G^\pi(t, x, U)$.

When it comes to randomizing actions in *continuous* time, a subtle measurability issue arises from continuum (uncountable in particular) and independent samplings from the stochastic policy; see, e.g. Szpruch et al. (2024) and Bender and Thuan (2024) for detailed discussions. To avoid such issues, we follow Jia et al. (2025) to only consider randomization on a time grid and use piecewise constant actions.

Let $\mathbb{S} = \{0 = t_0 < t_1 < \dots < t_N = T\}$ be a time grid of $[0, T]$ with $N \in \mathbb{N}$ and $|\mathbb{S}| = \max_{i=1, \dots, N} (t_i - t_{i-1})$. Given a stochastic feedback policy π , we follow a *grid sampling scheme* where action randomization occurs at the grid points only. Specifically, to define the state dynamic under such a scheme, we introduce the grid sampling process:

$$U_s^\mathbb{S} = U_1 \cdot 1_{[0, t_1]}(s) + \sum_{i=2}^N U_i \cdot 1_{(t_{i-1}, t_i]}(s), \quad s \in [0, T]. \quad (10)$$

Here (U_i) are i.i.d random vectors uniformly distributed over $[0, 1]^n$, which are independent of W and L_1, \dots, L_ℓ , assuming that the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is rich enough to support these random elements. This process essentially describes sampling from U_i at time t_{i-1} . We now enlarge the original filtration to include the additional randomness from sampling actions. Denote by $\mathcal{G}^\mathbb{S} = (\mathcal{G}_t^\mathbb{S})_{t \in [0, T]}$ the right-continuous, augmented version of the filtration generated by W, L_1, \dots, L_ℓ and the grid sampling process $U^\mathbb{S}$. The new filtered probability space is now $(\Omega, \mathcal{F}, \mathcal{G}^\mathbb{S}, \mathbb{P})$.

Fix an initial system state x_0 at time 0, a time grid \mathbb{S} of $[0, T]$ and a stochastic feedback policy π . Under a grid sampling scheme, the state dynamic $X^{\pi, \mathbb{S}} = \{X_s^{\pi, \mathbb{S}} : 0 \leq s \leq T\}$, which will be henceforth referred to as the *grid sample state process*, satisfies the following SDE: for all $i = 0, \dots, n-1$ and all $s \in (t_i, t_{i+1}]$,

$$\begin{aligned} dX_s^{\pi, \mathbb{S}} &= b(s, X_{s-}^{\pi, \mathbb{S}}, G^\pi(t_i, X_{t_i}^{\pi, \mathbb{S}}, U_s^{\mathbb{S}}))ds + \sigma(s, X_{s-}^{\pi, \mathbb{S}}, G^\pi(t_i, X_{t_i}^{\pi, \mathbb{S}}, U_s^{\mathbb{S}}))dW_s \\ &\quad + \int_{\mathbb{R}^\ell} \gamma(s, X_{s-}^{\pi, \mathbb{S}}, G^\pi(t_i, X_{t_i}^{\pi, \mathbb{S}}, U_s^{\mathbb{S}}), z) \tilde{N}(ds, dz), \end{aligned} \quad (11)$$

where $U_s^{\mathbb{S}} = U_{i+1}$ for $s \in (t_i, t_{i+1}]$. This means a random action $G^\pi(t_i, X_{t_i}^{\pi, \mathbb{S}}, U_{i+1})$ is generated at t_i and remains unchanged over $(t_i, t_{i+1}]$.² Under appropriate technical assumptions on π , we will show later that the SDE (11) admits a unique strong solution that has a desirable moment estimate; see Proposition 2 for details.

2.3 Exploratory state process

For theoretical analysis, we need to formulate the so-called exploratory state process, which represents the averaged controlled dynamic over infinitely many randomized actions. In the case of diffusions, Wang et al. (2020) heuristically formulated such exploratory dynamics by applying an LLN (Law of Large Numbers) argument to the first two moments of the diffusion processes. We will still use the same heuristic approach based on LLN here, but the first two moments cannot determine a jump-diffusion. Instead, we study the infinitesimal behavior of the grid sample state process by averaging randomized actions, based on which we will formulate the exploratory state process.

To this end, let $f \in C_0^{1,2}([0, T] \times \mathbb{R}^d)$ be continuously differentiable in t and twice continuously differentiable in x with compact support. Let $t = t_k$, for some k , be a grid point of \mathbb{S} and fix x . Consider the SDE (11) starting from $X_t^{\pi, \mathbb{S}} = x$. Let $s > 0$ be very small such that $[t, t+s]$ lies within $[t_k, t_{k+1}]$. We apply the random action $G^\pi(t_k, X_{t_k}^{\pi, \mathbb{S}}, U_{k+1})$ throughout $(t, t+s]$. Let a_1, \dots, a_N be N independent samples of this random action. With a slight abuse of notations, denote by $X_{t+s}^{a_i}$ the value of the state process corresponding to a_i at $t+s$. We consider $\mathbb{E}_{t,x}[f(t+s, X_{t+s}^{\pi, \mathbb{S}})]$, where the expectation, conditional on $X_t^{\pi, \mathbb{S}} = x$, is taken over the randomness caused by U_{k+1} as well as W, L_1, \dots, L_ℓ . Then

$$\begin{aligned} &\lim_{s \rightarrow 0} \frac{\mathbb{E}_{t,x}[f(t+s, X_{t+s}^{\pi, \mathbb{S}})] - f(t, x)}{s} \\ &= \lim_{s \rightarrow 0} \frac{\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{t,x}[f(t+s, X_{t+s}^{a_i})] - f(t, x)}{s} \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \lim_{s \rightarrow 0} \frac{\mathbb{E}_{t,x}[f(t+s, X_{t+s}^{a_i})] - f(t, x)}{s} \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (\partial_t f(t, x) + b(t, x, a_i) \circ \partial_x f(t, x) + \frac{1}{2} \sigma^2(t, x, a_i) \circ \partial_x^2 f(t, x)) \end{aligned} \quad (12)$$

²To define the grid sample state process, one can alternatively consider an action process given by $G^\pi(s, X_{s-}^{\pi, \mathbb{S}}, U_s^{\mathbb{S}}) = G^\pi(s, X_{s-}^{\pi, \mathbb{S}}, U_{i+1})$ for $s \in (t_i, t_{i+1}]$ as in Bender and Thuan (2024). In this paper we choose to work with the piecewise constant action process given in (11) because it is simpler and indeed is the actual implemented control process in our RL algorithms.

$$+ \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{\ell} \int_{\mathbb{R}} (f(t, x + \gamma_k(t, x, a_i, z)) - f(t, x) - \gamma_k(t, x, a_i, z) \circ \partial_x f(t, x)) \nu_k(dz). \quad (13)$$

Using LLN, we obtain

$$(12) = \partial_t f(t, x) + \tilde{b}(t, x, \boldsymbol{\pi}_{t,x}) \circ \partial_x f(t, x) + \frac{1}{2} \tilde{\sigma}^2(t, x, \boldsymbol{\pi}_{t,x}) \circ \partial_x^2 f(t, x), \quad (14)$$

where

$$\tilde{b}(t, x, \boldsymbol{\pi}_{t,x}) := \int_{\mathcal{A}} b(t, x, a) \boldsymbol{\pi}(a|t, x) da, \quad \tilde{\sigma}(t, x, \boldsymbol{\pi}_{t,x}) := \left(\int_{\mathcal{A}} \sigma^2(t, x, a) \boldsymbol{\pi}(a|t, x) da \right)^{1/2}. \quad (15)$$

These ‘‘exploratory’’ drift and diffusion coefficients are consistent with those in Wang et al. (2020). It is thus tempting to think the exploratory jump coefficient $\tilde{\gamma}$ is similarly the average of γ with respect to $\boldsymbol{\pi}$; but unfortunately it is generally not true. This in turn is one of the main distinctive features in studying RL for jump-diffusions.

We approach the problem by analyzing the integrals in (13). Using the second-order Taylor expansion, the boundedness of $\partial_x^2 f(t, x)$ for $x \in \mathbb{R}^d$ and condition (iii) of Assumption 1, we obtain that for fixed (t, x) and each k ,

$$\begin{aligned} & \left| \int_{\mathbb{R}} (f(t, x + \gamma_k(t, x, a, z)) - f(t, x) - \gamma_k(t, x, a, z) \circ \partial_x f(t, x)) \nu_k(dz) \right| \\ & \leq C \int_{\mathbb{R}} |\gamma_k(t, x, a, z)|^2 \nu_k(dz) \leq C(1 + |x|^2) \end{aligned}$$

for some constant $C > 0$, which is independent of a . It follows that

$$(13) = \sum_{k=1}^{\ell} \int_{\mathcal{A}} \int_{\mathbb{R}} (f(t, x + \gamma_k(t, x, a, z)) - f(t, x) - \gamma_k(t, x, a, z) \circ \partial_x f(t, x)) \nu_k(dz) \boldsymbol{\pi}(a|t, x) da. \quad (16)$$

Combining (14) and (16), the infinitesimal behavior of the grid sample state process (at grid points) can be characterized by the operator $\mathcal{L}^{\boldsymbol{\pi}}$, which is given by the probability weighted average of the generator \mathcal{L}^a of the classical controlled jump-diffusion, i.e.,

$$\mathcal{L}^{\boldsymbol{\pi}} f(t, x) := \lim_{s \rightarrow 0} \frac{\mathbb{E}_{t,x}[f(t+s, X_{t+s}^{\boldsymbol{\pi}, \mathbb{S}})] - f(t, x)}{s} = \int_{\mathcal{A}} \mathcal{L}^a f(t, x) \boldsymbol{\pi}(a|t, x) da.$$

We now reformulate the integrals in (13) to convert them to the same form as (5). Note that

$$\begin{aligned} & \int_{\mathcal{A}} \int_{\mathbb{R}} (f(t, x + \gamma_k(t, x, a, z)) - f(t, x) - \gamma_k(t, x, a, z) \circ \partial_x f(t, x)) \nu_k(dz) \boldsymbol{\pi}(a|t, x) da \\ & = \int_{\mathbb{R} \times [0,1]^n} (f(t, x + \gamma_k(t, x, G^{\boldsymbol{\pi}_{t,x}}(u), z)) - f(t, x) - \gamma_k(t, x, G^{\boldsymbol{\pi}_{t,x}}(u), z) \circ \partial_x f(t, x)) \nu_k(dz) du. \end{aligned}$$

It follows that $\mathcal{L}^{\boldsymbol{\pi}}$ can be written as

$$\mathcal{L}^{\boldsymbol{\pi}} f(t, x) = \partial_t f(t, x) + \tilde{b}(s, x, \boldsymbol{\pi}_{t,x}) \circ \partial_x f(t, x) + \frac{1}{2} \tilde{\sigma}^2(s, x, \boldsymbol{\pi}_{t,x}) \circ \partial_x^2 f(t, x)$$

$$+ \sum_{k=1}^{\ell} \int_{\mathbb{R} \times [0,1]^n} (f(t, x + \gamma_k(t, x, G^{\pi t, x}(u), z)) - f(t, x) - \gamma_k(t, x, G^{\pi t, x}(u), z) \circ \partial_x f(t, x)) \nu_k(dz) du. \quad (17)$$

For each $k = 1, \dots, \ell$, we can view $\nu_k(dz)du$ as the intensity measure of a new Poisson random measure, denoted by $N'_k(dt, dz, du)$, defined on the product space $[0, T] \times \mathbb{R} \times [0, 1]^n$. We let $N'(dt, dz, du) = (N'_1(dt, dz_1, du), \dots, N'_\ell(dt, dz_\ell, du))^\top$.³

Based on (17), we formulate the exploratory state process as the solution to the following SDE:

$$dX_s^\pi = \tilde{b}(s, X_{s-}^\pi, \pi(\cdot|s, X_{s-}^\pi))ds + \tilde{\sigma}(s, X_{s-}^\pi, \pi(\cdot|s, X_{s-}^\pi))dW_s + \int_{\mathbb{R} \times [0,1]^n} \gamma(s, X_{s-}^\pi, G^\pi(s, X_{s-}^\pi, u), z) \tilde{N}'(ds, dz, du), \quad X_t^\pi = x, \quad s \in [t, T], \quad (18)$$

which we call the *exploratory Lévy SDE*. The solution process, should it exist, is denoted by \tilde{X}^π and called the exploratory (state) process. It is a semimartingale by (18) and determined by three characteristics: the drift, the quadratic variation of the continuous local martingale, and the compensator of the random measure associated with the process's jumps; see [Jacod and Shiryaev \(2013\)](#) for detailed discussions of semimartingales and their characteristics. Given $\tilde{X}_{s-}^\pi = x$, the semimartingale characteristics of \tilde{X}^π over an infinitesimally small time interval $[s, s + ds]$ are given by the triplet $(\tilde{b}(s, x, \pi_{s,x})ds, \tilde{\sigma}^2(s, x, \pi_{s,x})ds, \sum_{k=1}^{\ell} \int_{[0,1]^n} \gamma_k(s, x, G^\pi(s, x, u), z) du \cdot \nu_k(dz)ds)$, where the third characteristic is obtained by calculating $\mathbb{E}[\sum_{k=1}^{\ell} \gamma_k(s, x, G^\pi(s, x, u), z) N_k(ds, dz, du)]$ for Lévy jumps with size from $[z, z + dz]$. Using (15), we have

$$\begin{aligned} \tilde{b}(s, x, \pi_{s,x})ds &= \int_{\mathcal{A}} b(s, x, a) \pi(a|s, x) da ds, \quad \tilde{\sigma}^2(s, x, \pi_{s,x})ds = \int_{\mathcal{A}} \sigma^2(s, x, a) \pi(a|s, x) da ds, \\ \sum_{k=1}^{\ell} \int_{[0,1]^n} \gamma_k(s, x, G^\pi(s, x, u), z) du \cdot \nu_k(dz)ds &= \sum_{k=1}^{\ell} \int_{\mathcal{A}} \gamma_k(s, x, a, z) \nu_k(dz)ds \cdot \pi(a|s, x) da. \end{aligned}$$

Thus, the semimartingale characteristics of the exploratory state process are the averages of those of the original controlled state process (2) over action randomization.

A technical yet foundational question is the well-posedness (i.e. existence and uniqueness of solution) of the exploratory SDE (18), which we now address. For that we first specify the class of admissible strategies, which is the same as that considered in [Jia and Zhou \(2023\)](#) for pure diffusions.

Definition 1. A policy $\pi = \pi(\cdot|\cdot, \cdot)$ is called *admissible*, if

- (i) $\pi(\cdot|t, x) \in \mathcal{P}(\mathcal{A})$, $\text{supp}\pi(\cdot|t, x) = \mathcal{A}$ for every $(t, x) \in [0, T] \times \mathbb{R}^d$ and $\pi(a|t, x) : (t, x, a) \in [0, T] \times \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}$ is measurable;
- (ii) $\pi(a|t, x)$ is continuous in (t, x) , i.e., $\int_{\mathcal{A}} |\pi(a|t, x) - \pi(a|t', x')| da \rightarrow 0$ as $(t', x') \rightarrow (t, x)$. Furthermore, for any $K > 0$, there is a constant $C_K > 0$ independent of (t, a) such that

$$\int_{\mathcal{A}} |\pi(a|t, x) - \pi(a|t, x')| da \leq C_K |x - x'|, \quad \forall x, x' \in \mathbb{R}_K^d;$$

³We assume the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is rich enough to support N' .

(iii) $\forall(t, x), \int_{\mathcal{A}} |\log \pi(a|t, x)| \pi(a|t, x) da \leq C(1 + |x|^p)$ for some $p \geq 2$ and some positive constant C . Moreover, $\int_{\mathcal{A}} |a|^p \pi(a|t, x) da \leq C_p(1 + |x|^p)$ for any $p \geq 1$ and some positive constant C_p .

Next, we establish the well-posedness of (18) under any given admissible policy. The result of the next lemma regarding \tilde{b} and $\tilde{\sigma}$ is provided in the proof of Lemma 2 in Jia and Zhou (2022b), which uses property (ii) of admissibility.

Lemma 1. *Under Assumption 1, for any admissible policy π , the functions $\tilde{b}(t, x, \pi_{t,x})$ and $\tilde{\sigma}(t, x, \pi_{t,x})$ have the following properties:*

(i) (local Lipschitz continuity) for $K > 0$, there exists a constant $C_K > 0$ such that $\forall t \in [0, T], (x, x') \in \mathbb{R}_K^d$,

$$|\tilde{b}(t, x, \pi_{t,x}) - \tilde{b}(t, x', \pi_{t,x'})| + |\tilde{\sigma}(t, x, \pi_{t,x}) - \tilde{\sigma}(t, x', \pi_{t,x'})| \leq C_K |x - x'|;$$

(ii) (linear growth in x) there exists a constant $C > 0$ such that $\forall(t, x) \in [0, T] \times \mathbb{R}^d$,

$$|\tilde{b}(t, x, \pi_{t,x})| + |\tilde{\sigma}(t, x, \pi_{t,x})| \leq C(1 + |x|).$$

We now establish analogous properties for $\gamma(t, x, G^{\pi_{t,x}}(u), z)$ in the following lemmas, whose proofs are relegated to the appendix.

Lemma 2 (linear growth in x). *Under Assumption 1, for any admissible π and any $p \geq 2$, there exists a constant $C_p > 0$ that can depend on p such that $\forall(t, x) \in [0, T] \times \mathbb{R}^d$,*

$$\sum_{k=1}^{\ell} \int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi_{t,x}}(u), z)|^p \nu_k(dz) du \leq C_p(1 + |x|^p).$$

For the local Lipschitz continuity of $\gamma_k(t, x, G^{\pi_{t,x}}(u), z)$, we make an additional assumption.

Assumption 2. *For $k = 1, \dots, \ell$, the following conditions hold.*

(i) *For any $K > 0$ and any $p \geq 2$, there exists a constant $C_{K,p} > 0$ that can depend on K and p such that*

$$\int_{\mathbb{R}} |\gamma_k(t, x, a, z) - \gamma_k(t, x, a', z)|^p \nu_k(dz) \leq C_{K,p} |a - a'|^p, \quad \forall t \in [0, T], a, a' \in \mathcal{A}, x \in \mathbb{R}_K^d, z \in \mathbb{R}.$$

(ii) *For any $K > 0$ and any $p \geq 2$, there exists a constant $C_{K,p} > 0$ that can depend on K and p such that*

$$\int_{[0,1]^n} |G^{\pi}(t, x, u) - G^{\pi}(t, x', u)|^p du \leq C_{K,p} |x - x'|^p, \quad \forall t \in [0, T], x, x' \in \mathbb{R}_K^d.$$

For a stochastic feedback policy $\pi_{t,x} \sim \mathcal{N}(\mu(t, x), A(t, x)A(t, x)^\top)$, we have $G^{\pi}(t, x, u) = \mu(t, x) + A(t, x)\Phi^{-1}(u)$. Clearly, Assumption 2-(ii) holds provided that $\mu(t, x)$ and $A(t, x)$ are locally Lipschitz continuous in x .

Lemma 3 (local Lipschitz continuity). *Under Assumptions 1 and 2, for any admissible policy π , any $K > 0$, and any $p \geq 2$, there exists a constant $C_{K,p} > 0$ that can depend on K and p such that $\forall t \in [0, T], (x, x') \in \mathbb{R}_{K}^d$,*

$$\sum_{k=1}^{\ell} \int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi_{t,x}}(u), z) - \gamma_k(t, x', G^{\pi_{t,x'}}(u), z)|^p \nu_k(dz) du \leq C_{K,p} |x - x'|^p.$$

With Lemmas 1 to 3, we can now apply (Kunita 2004, Theorem 3.2) and (Situ 2006, Theorem 119) to obtain the well-posedness of (18) along with the moment estimate of its solution.

Proposition 1. *Under Assumptions 1 and 2, for any admissible policy π , there exists a unique strong solution $\{\tilde{X}_t^{\pi}, 0 \leq t \leq T\}$ to the exploratory Lévy SDE (18). Furthermore, for any $p \geq 2$, there exists a constant $C_p > 0$ such that*

$$\mathbb{E}_{t,x} \left[\sup_{t \leq s \leq T} |\tilde{X}_s^{\pi}|^p \right] \leq C_p (1 + |x|^p). \quad (19)$$

It should be noted that the conditions imposed in Assumptions 1 and 2 are sufficient but not necessary for obtaining the well-posedness and moment estimate of the exploratory Lévy SDE (18). For a specific problem, weaker conditions may suffice for these results if we exploit special structures of the problem.

Remark 1. *Bender and Thuan (2024) formulate the grid sample state process as an equation driven by certain random measures, and prove a limit theorem for these random measures as the size of the sampling grid goes to zero. This limit theorem in turn motivates a limiting formulation of the grid sample state process, which is referred to as the grid-sampling limit SDE in their paper. Based on the discussion in Section 6.2 of Bender and Thuan (2024), this limiting SDE turns out to have the same probability law as our exploratory SDE, even though it has a different representation.*

2.4 Exploratory HJB equation

Given the exploratory dynamic (18), we now define the (exploratory) value function for any given admissible stochastic policy π . To encourage exploration, we follow Wang et al. (2020) to add an entropy regularizer to the running reward of the original control problem, and set

$$J(t, x; \pi) = \mathbb{E}_{t,x} \left[\int_t^T e^{-\beta(s-t)} \int_{\mathcal{A}} (r(s, \tilde{X}_s^{\pi}, a) - \theta \log \pi(a|s, \tilde{X}_{s-}^{\pi})) \pi(a|s, \tilde{X}_{s-}^{\pi}) da ds + e^{-\beta(T-t)} h(\tilde{X}_T^{\pi}) \right], \quad (20)$$

where the expectation is conditioned on $\tilde{X}_t^{\pi} = x$. Here $\theta > 0$ is the *temperature parameter* that controls the level of exploration. The function $J(\cdot, \cdot; \pi)$ is called the (exploratory) value function of the policy π . The goal of RL is to find the policy that maximizes the value function among admissible policies specified in Definition 1.

Under Assumption 1 and using the admissibility of π and (19), it is easy to see that J has polynomial growth in x . We provide the Feynman–Kac formula for this function in Lemma 4 by

working with the representation (20). In the proof, we consider the finite and infinite jump activity cases separately because special care is needed in the latter. We revise Assumption 1 by adding one more condition for this case.

Assumption 1'. Conditions (i) to (iv) in Assumption 1 hold. We further assume condition (v): if $\int_{\mathbb{R}} \nu_k(dz) = \infty$, then $|\gamma_k(t, x, a, z)|$ is bounded for any $|z| \leq 1$, $t \in [0, T]$, $x \in \mathbb{R}_K^d$ and $a \in \mathcal{A}$.

For Lemma 4, Lemma 5 and Theorem 1, we impose Assumption 1' and assume that the exploratory SDE (18) is well-posed with the moment estimate (19). For simplicity, we do not explicitly mention these assumptions in the statement of the results.

Lemma 4. Given an admissible stochastic policy π , suppose there exists a solution $\phi \in C^{1,2}([0, T] \times \mathbb{R}^d) \cap C([0, T] \times \mathbb{R}^d)$ to the following partial integro-differential equation (PIDE): $\forall (t, x) \in [0, T] \times \mathbb{R}^d$,

$$\partial_t \phi(t, x) + \int_{\mathcal{A}} (H(t, x, a, \partial_x \phi, \partial_x^2 \phi, \phi) - \theta \log \pi(a|t, x)) \pi(a|t, x) da - \beta \phi(t, x) = 0, \quad (21)$$

with terminal condition $\phi(T, x) = h(x)$, $x \in \mathbb{R}^d$. Moreover, for some $p \geq 2$, ϕ satisfies

$$|\phi(t, x)| \leq C(1 + |x|^p), \quad \forall (t, x) \in [0, T] \times \mathbb{R}^d. \quad (22)$$

Then ϕ is the value function of the policy π , i.e. $J(t, x; \pi) = \phi(t, x)$.

For ease of presentation, we henceforth assume the value function $J(\cdot, \cdot; \pi) \in C^{1,2}([0, T] \times \mathbb{R}^d) \cap C([0, T] \times \mathbb{R}^d)$ for any admissible stochastic policy π .

Remark 2. The conclusion in Lemma 4 still holds if we assume $\partial_x \phi(t, x)$ has polynomial growth in x instead of Assumption 1'-(v).

Next, we consider the optimal value function defined by

$$J^*(t, x) = \sup_{\pi \in \Pi} J(t, x; \pi),$$

where Π is the class of admissible strategies. The following result characterizes J^* and the optimal stochastic policy through the so-called exploratory HJB equation.

Lemma 5. Suppose there exists a solution $\psi \in C^{1,2}([0, T] \times \mathbb{R}^d) \cap C([0, T] \times \mathbb{R}^d)$ to the exploratory HJB equation:

$$\partial_t \psi(t, x) + \sup_{\pi \in \mathcal{P}(\mathcal{A})} \int_{\mathcal{A}} (H(t, x, a, \partial_x \psi, \partial_x^2 \psi, \psi) - \theta \log \pi(a|t, x)) \pi(a|t, x) da - \beta \psi(t, x) = 0, \quad (23)$$

with the terminal condition $\psi(T, x) = h(x)$, where H is the Hamiltonian defined in (9). Moreover, for some $p \geq 2$, ψ satisfies

$$|\psi(t, x)| \leq C(1 + |x|^p), \quad \forall (t, x) \in [0, T] \times \mathbb{R}^d,$$

and it holds that

$$\int_{\mathcal{A}} \exp\left(\frac{1}{\theta} H(t, x, a, \partial_x \psi, \partial_x^2 \psi, \psi)\right) da < \infty.$$

Then, the Gibbs measure or Boltzman distribution

$$\pi^*(a|t, x) \propto \exp\left(\frac{1}{\theta} H(t, x, a, \partial_x \psi, \partial_x^2 \psi, \psi)\right) \quad (24)$$

is the optimal stochastic policy and $J^*(t, x) = \psi(t, x)$ provided that π^* is admissible.

Plugging the optimal stochastic policy (22) into Eq. (21) to remove the supremum operator, we obtain the following nonlinear PIDE for the optimal value function J^* : $\forall(t, x) \in [0, T] \times \mathbb{R}^d$,

$$\partial_t J^*(t, x) + \theta \log \left(\int_{\mathcal{A}} \exp \left(\frac{1}{\theta} H(t, x, a, \partial_x J^*, \partial_x^2 J^*, J^*) \right) da \right) - \beta J^*(t, x) = 0, \quad (25)$$

with terminal condition $J^*(T, x) = h(x)$, $x \in \mathbb{R}^d$.

Remark 3. *The optimal value function J^* in (23) for the exploratory control problem is expected to converge to the optimal value function V^* for the classical control problem (3) when $\theta \rightarrow 0$ under certain technical conditions. In the case of pure diffusions without jumps, this has been proved in Tang et al. (2022) for infinite-horizon problems. Establishing such a result for general jump-diffusions necessitates a careful analysis of the dependency of PIDE (23) on θ , which falls beyond the scope of this paper. However, we will see that in each of the two applications considered in this paper (see Sections 5 and 6), the optimal value function J^* of the exploratory problem indeed converges to the classical optimal value function V^* as $\theta \rightarrow 0$, due to the convergence of the optimal stochastic feedback policy to the optimal deterministic feedback policy.*

3 q-Learning Theory

3.1 q-function and policy improvement

We present the q-learning theory for jump-diffusions that includes both policy evaluation and policy improvement, now that the exploratory formulation has been set up. The theory can be developed similarly to Jia and Zhou (2023, 2025); so we will just highlight the main differences in the analysis, skipping the parts that are similar.⁴

Definition 2. *The q-function of the problem (2)–(3) associated with a given policy $\pi \in \Pi$ is defined by*

$$q(t, x, a; \pi) = \partial_t J(t, x; \pi) + H(t, x, a, \partial_x J, \partial_x^2 J, J) - \beta J(t, x; \pi), \quad (t, x, a) \in [0, T] \times \mathbb{R}^d \times \mathcal{A},$$

where J is given in (20) and the Hamiltonian function H is defined in (9).

It is an immediate consequence of Lemma 4 that the q-function satisfies

$$\int_{\mathcal{A}} (q(t, x, a; \pi) - \theta \log \pi(a|t, x)) \pi(a|t, x) da = 0, \quad \forall(t, x) \in [0, T] \times \mathbb{R}^d.$$

The following policy improvement theorem can be proved similarly as in (Jia and Zhou 2023, Theorem 2) by using the arguments in the proof of Lemma 4.

⁴The q-learning theory for pure diffusions is established originally in Jia and Zhou (2023). In the wake of the observations of the aforementioned measurability issue, an erratum that works with grid sample state processes is put forth in Jia and Zhou (2025).

Theorem 1 (Policy Improvement). *For any given $\pi \in \Pi$, define*

$$\pi'(\cdot|t, x) \propto \exp\left(\frac{1}{\gamma}H(t, x, \cdot, \partial_x J(t, x; \pi), \partial_x^2 J(t, x; \pi), J(t, \cdot; \pi))\right) \propto \exp\left(\frac{1}{\gamma}q(t, x, \cdot; \pi)\right).$$

If $\pi' \in \Pi$, then

$$J(t, x, \pi') \geq J(t, x, \pi).$$

Moreover, if the following map

$$\begin{aligned} \mathcal{I}(\pi) &= \frac{\exp(\frac{1}{\theta}H(t, x, \cdot, \partial_x J(t, x; \pi), \partial_x^2 J(t, x; \pi), J(t, \cdot; \pi)))}{\int_{\mathcal{A}} \exp(\frac{1}{\theta}H(t, x, a, \partial_x J(t, x; \pi), \partial_x^2 J(t, x; \pi), J(t, \cdot; \pi))) da}, \quad \pi \in \Pi \\ &= \frac{\exp(\frac{1}{\theta}q(t, x, \cdot; \pi))}{\int_{\mathcal{A}} \exp(\frac{1}{\theta}q(t, x, a; \pi)) da} \end{aligned}$$

has a fixed point π^ , then π^* is an optimal policy.*

3.2 Martingale characterization of the q-function

In this section, we derive the martingale characterization of the q-function associated with a policy $\pi \in \Pi$, which will be used for on-policy and off-policy learning. We cannot work with the exploratory process directly because it is unobservable (i.e. it does not constitute *data*). We consider the grid sample state process instead, which is what one obtains from the environment by following a stochastic policy.

For any $t \in [0, T]$, in the following we consider a grid on $[t, T]$, which will still be denoted by \mathbb{S} for notational simplicity (though it depends on t). That is, $\mathbb{S} = \{t = t_0 < t_1 < \dots < t_N = T\}$ for some N . The grid sampling process on $[t, T]$ can be defined analogously as in (10):

$$U_s^{\mathbb{S}} = U_1 \cdot \mathbf{1}_{[t_0, t_1]}(s) + \sum_{i=2}^N U_i \cdot \mathbf{1}_{(t_{i-1}, t_i]}(s), \quad s \in [t, T],$$

where the first uniform random vector U_1 is sampled at $t = t_0$. The grid sample state process $X^{\pi, \mathbb{S}} = \{X_s^{\pi, \mathbb{S}} : t \leq s \leq T\}$ starts from t and satisfies the SDE (11). The action process $a^{\pi, \mathbb{S}} = \{a_s^{\pi, \mathbb{S}} : t \leq s \leq T\}$ satisfies $a_s^{\pi, \mathbb{S}} = a_{t_n}^{\pi, \mathbb{S}}$ for $s \in (t_n, t_{n+1}]$, where $a_{t_n}^{\pi, \mathbb{S}} := G^{\pi}(t_n, X_{t_n}^{\pi, \mathbb{S}}, U_{n+1})$ is applied over $(t_n, t_{n+1}]$. With a slight abuse of notations, we still denote by $\mathcal{G}^{\mathbb{S}}$ the right-continuous, augmented version of the filtration generated by the Brownian motion W , Lévy processes L_1, \dots, L_ℓ , and the grid sampling process on $[t, T]$.

We need the following result for the grid sample state process $X^{\pi, \mathbb{S}} = \{X_s^{\pi, \mathbb{S}} : t \leq s \leq T\}$.

Proposition 2. *Fix any $(t, x) \in [0, T] \times \mathbb{R}^d$. Under Assumption 1, for any admissible policy π and any time grid \mathbb{S} of $[t, T]$, there exists a unique strong solution $\{X_s^{\pi, \mathbb{S}}, t \leq s \leq T\}$ to the grid sample state SDE (11) with the initial condition $X_t^{\pi, \mathbb{S}} = x$. Furthermore, for any $p \geq 2$, there exists a constant C_p that is independent of the grid \mathbb{S} such that*

$$\mathbb{E}_{t, x} \left[\sup_{t \leq s \leq T} |X_s^{\pi, \mathbb{S}}|^p \right] \leq C_p (1 + |x|^p). \quad (26)$$

We now state in Theorem 2 the martingale characterization of the q-function associated with a policy $\pi \in \Pi$, assuming that its value function has already been learned and known. Note that for Theorems 2–4 below, we impose Assumption 1 without explicitly mentioning it in the theorem statements .

Theorem 2. *Let a policy $\pi \in \Pi$, its value function $J(\cdot, \cdot; \pi) \in C^{1,2}([0, T] \times \mathbb{R}^d) \cap C([0, T] \times \mathbb{R}^d)$ and a continuous function $\hat{q} : [0, T] \times \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}$ be given. Assume that $J(t, x; \pi)$ and $J_x(t, x; \pi)$ both have polynomial growth in x . Then the following results hold.*

- (i) $\hat{q}(t, x, a) = q(t, x, a; \pi)$ for all (t, x, a) if and only if for any (t, x) and any grid \mathbb{S} of $[t, T]$, the following process

$$e^{-\beta s} J(s, X_s^{\pi, \mathbb{S}}; \pi) + \int_t^s e^{-\beta \tau} (r(\tau, X_\tau^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}) - \hat{q}(\tau, X_\tau^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})) d\tau \quad (27)$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale, where $X^{\pi, \mathbb{S}} = \{X_s^{\pi, \mathbb{S}} : t \leq s \leq T\}$ is the grid sample state process defined in (11) with $X_t^{\pi, \mathbb{S}} = x$.

- (ii) If $\hat{q}(t, x, a) = q(t, x, a; \pi)$ for all (t, x, a) , then for any $\pi' \in \Pi$, any (t, x) and any grid \mathbb{S} of $[t, T]$, the following process

$$e^{-\beta s} J(s, X_s^{\pi', \mathbb{S}}; \pi) + \int_t^s e^{-\beta \tau} (r(\tau, X_\tau^{\pi', \mathbb{S}}, a_\tau^{\pi', \mathbb{S}}) - \hat{q}(\tau, X_\tau^{\pi', \mathbb{S}}, a_\tau^{\pi', \mathbb{S}})) d\tau \quad (28)$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale, where $\{X_s^{\pi', \mathbb{S}} : t \leq s \leq T\}$ is the solution to (11) under π' with initial condition $X_t^{\pi', \mathbb{S}} = x$.

- (iii) If there exists $\pi' \in \Pi$ such that for all (t, x) and any grid \mathbb{S} of $[t, T]$, the process (24) is a $\mathcal{G}^{\mathbb{S}}$ -martingale where $X_t^{\pi', \mathbb{S}} = x$, then we have $\hat{q}(t, x, a) = q(t, x, a; \pi)$ for all (t, x, a) .

Moreover, in any of the three cases above, the q-function satisfies

$$\int_{\mathcal{A}} (q(t, x, a; \pi) - \theta \log \pi(a|t, x)) \pi(a|t, x) da = 0, \quad \text{for all } (t, x) \in [0, T] \times \mathbb{R}^d.$$

Remark 4. *Similar to Jia and Zhou (2023, 2025), Theorem 2-(i) facilitates on-policy learning, where learning the q-function of the given target policy π is based on data $\{(s, X_s^{\pi, \mathbb{S}}, a_s^{\pi, \mathbb{S}}), t \leq s \leq T\}$ generated by π . On the other hand, Theorem 2-(ii) and -(iii) are for off-policy learning, where learning the q-function of π is based on data generated by a different, called behavior policy, π' .*

Next, we extend Theorem 7 in Jia and Zhou (2023, 2025) and obtain a martingale characterization of the value function and the q-function simultaneously.

Theorem 3. *Let a policy $\pi \in \Pi$, a function $\hat{J} \in C^{1,2}([0, T] \times \mathbb{R}^d) \cap C([0, T] \times \mathbb{R}^d)$ and a continuous function $\hat{q} : [0, T] \times \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}$ be given satisfying*

$$\hat{J}(T, x) = h(x), \quad \int_{\mathcal{A}} (\hat{q}(t, x, a) - \theta \log \pi(a|t, x)) \pi(a|t, x) da = 0, \quad \forall (t, x) \in [0, T] \times \mathbb{R}^d. \quad (29)$$

Assume that \hat{J} and \hat{J}_x both have polynomial growth in x . Then

- (i) \hat{J} and \hat{q} are respectively the value function and the q-function associated with π if and only if for any $(t, x) \in [0, T] \times \mathbb{R}^d$ and any grid \mathbb{S} of $[t, T]$, the following process

$$e^{-\beta s} \hat{J}(s, X_s^{\pi, \mathbb{S}}) + \int_t^s e^{-\beta \tau} (r(\tau, X_\tau^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}) - \hat{q}(\tau, X_\tau^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})) d\tau$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale, where $X^{\pi, \mathbb{S}} = \{X_s^{\pi, \mathbb{S}} : t \leq s \leq T\}$ satisfies (11) with $X_t^{\pi, \mathbb{S}} = x$.

- (ii) If \hat{J} and \hat{q} are respectively the value function and the q-function associated with π , then for any $\pi' \in \Pi$, any $(t, x) \in [0, T] \times \mathbb{R}^d$ and any grid \mathbb{S} of $[t, T]$, the following process

$$e^{-\beta s} \hat{J}(s, X_s^{\pi', \mathbb{S}}) + \int_t^s e^{-\beta \tau} (r(\tau, X_\tau^{\pi', \mathbb{S}}, a_\tau^{\pi', \mathbb{S}}) - \hat{q}(\tau, X_\tau^{\pi', \mathbb{S}}, a_\tau^{\pi', \mathbb{S}})) d\tau \quad (30)$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale, where $\{X_s^{\pi', \mathbb{S}} : t \leq s \leq T\}$ satisfies (11) with $X_t^{\pi', \mathbb{S}} = x$.

- (iii) If there exists $\pi' \in \Pi$ such that for all (t, x) and any grid \mathbb{S} of $[t, T]$, the process (25) is a $\mathcal{G}^{\mathbb{S}}$ -martingale where $X_t^{\pi', \mathbb{S}} = x$, then we have $\hat{J}(t, x) = J(t, x; \pi)$ and $\hat{q}(t, x, a) = q(t, x, a; \pi)$ for all (t, x, a) .

In any of the three cases above, if it holds that $\pi(a|t, x) = \frac{\exp(\frac{1}{\theta} \hat{q}(t, x, a))}{\int_{\mathcal{A}} \exp(\frac{1}{\theta} \hat{q}(t, x, a)) da}$, then π is the optimal policy and \hat{J} is the optimal value function.

3.3 Optimal q-function

We consider in this section the optimal q-function, i.e., the q-function associated with the optimal policy π^* in (22). Based on Definition 2, we can write it as

$$q^*(t, x, a) = \partial_t J^*(t, x) + H(t, x, a, \partial_x J^*, \partial_x^2 J^*, J^*) - \beta J^*(t, x),$$

where J^* is the optimal value function that solves the exploratory HJB equation in (21).

The following is the martingale condition that characterizes the optimal value function J^* and the optimal q-function.

Theorem 4. Let a function $\hat{J}^* \in C^{1,2}([0, T] \times \mathbb{R}^d) \cap C([0, T] \times \mathbb{R}^d)$ and a continuous function $\hat{q}^* : [0, T] \times \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}$ be given satisfying

$$\hat{J}^*(T, x) = h(x), \quad \int_{\mathcal{A}} \exp\left(\frac{1}{\theta} \hat{q}^*(t, x, a)\right) da = 1, \quad \text{for all } (t, x) \in [0, T] \times \mathbb{R}^d.$$

Assume that $\hat{J}^*(t, x)$ and $\hat{J}_x^*(t, x)$ both have polynomial growth in x . Then

- (i) If \hat{J}^* and \hat{q}^* are respectively the optimal value function and the optimal q-function, then for any $\pi \in \Pi$, for all $(t, x) \in [0, T] \times \mathbb{R}^d$ and any grid \mathbb{S} of $[t, T]$, the following process

$$e^{-\beta s} \hat{J}^*(s, X_s^{\pi, \mathbb{S}}) + \int_t^s e^{-\beta \tau} (r(\tau, X_\tau^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}) - \hat{q}^*(\tau, X_\tau^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})) d\tau \quad (31)$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale, where $X^{\pi, \mathbb{S}} = \{X_s^{\pi, \mathbb{S}} : t \leq s \leq T\}$ satisfies (11) with $X_t^{\pi, \mathbb{S}} = x$. Moreover, in this case, $\hat{\pi}^*(a|t, x) = \exp(\frac{1}{\theta} \hat{q}^*(t, x, a))$ is the optimal stochastic policy.

- (ii) If there exists $\pi \in \Pi$ such that for all (t, x) and any grid \mathbb{S} of $[t, T]$, the process (26) is a $\mathcal{G}^{\mathbb{S}}$ -martingale where $X_t^{\pi, \mathbb{S}} = x$, then \hat{J}^* and \hat{q}^* are respectively the optimal value function and the optimal q-function.

4 q-Learning Algorithms

In this section we present learning algorithms based on the martingale characterization of the q-function discussed in the previous section. We need to distinguish two cases, depending on whether or not the density function of the Gibbs measure generated from the q-function can be computed and integrated explicitly.

We first discuss the case when the normalizing constant in the Gibbs measure can be computed explicitly. We denote by J^ψ and q^ϕ the parameterized function approximators for the optimal value function and optimal q-function, respectively. In view of Theorem 4, these approximators are chosen to satisfy

$$J^\psi(T, x) = h(x), \quad \int_{\mathcal{A}} \exp\left(\frac{1}{\theta} q^\phi(t, x, a)\right) da = 1. \quad (32)$$

We can then update (ψ, ϕ) by enforcing the martingale condition discussed in Theorem 4 and applying the techniques developed in Jia and Zhou (2022a). This procedure has been discussed in details in Section 4.1 of Jia and Zhou (2023), and hence we omit the details.

For reader's convenience, we present Algorithms 1 and 2, which summarize the offline and online q-learning algorithms respectively. In the algorithmic implementation, we execute stochastic policies on a uniform time grid with step size Δt . That is, the grid $\mathbb{S} = \{0 = t_0 < t_1 < \dots < t_K = T\}$, where $t_{i+1} - t_i = \Delta t$ for all i . Algorithms 1 and 2 are based on the so-called martingale orthogonality condition in Jia and Zhou (2022a), where the typical choices of test functions in these algorithms are $\xi_t = \partial_\psi J^\psi(t, X_t^{\pi^\phi})$, and $\zeta_t = \partial_\phi q^\phi(t, X_t^{\pi^\phi}, a_t^{\pi^\phi})$. Here, π^ϕ is the policy generated by q^ϕ , and we use $X_t^{\pi^\phi}$ to denote the grid sample state process $X_t^{\pi^\phi, \mathbb{S}}$ for notational simplicity. Note that these two algorithms are *identical* to Algorithms 2 and 3 in Jia and Zhou (2023).

When the normalizing constant in the Gibbs measure is not available, we take the same approach as in Jia and Zhou (2023) to develop learning algorithms. Specifically, we consider $\{\pi^\phi(\cdot|t, x)\}_{\phi \in \Phi}$, which is a family of density functions of some tractable distributions, e.g. multivariate normal distributions. Starting from a stochastic policy π^ϕ in this family, we update the policy by considering the optimization problem

$$\min_{\phi' \in \Phi} \text{KL}\left(\pi^{\phi'}(\cdot|t, x) \middle| \exp\left(\frac{1}{\theta} q(t, x, \cdot; \pi^\phi)\right)\right).$$

Specifically, using gradient descent, we can update ϕ as in Jia and Zhou (2023), by

$$\phi \leftarrow \phi - \theta \alpha_\phi dt \left(\log \pi^\phi(a_t^{\pi^\phi} | t, X_{t-}^{\pi^\phi}) - \frac{1}{\theta} q(t, X_t^{\pi^\phi}, a_t^{\pi^\phi}; \pi^\phi) \right) \partial_\phi \log \pi^\phi(a_t^{\pi^\phi} | t, X_{t-}^{\pi^\phi}). \quad (33)$$

In the above updating rule, we need only the *values* of the q-function along the trajectory – the “data” – $\{(t, X_t^{\pi^\phi}, a_t^{\pi^\phi}); 0 \leq t \leq T\}$, instead of its full functional form. These values can be learned through the “temporal difference” of the value function along the data. To see this, applying Itô's formula (6) to $J(\cdot, \cdot; \pi^\phi)$, we have

$$\begin{aligned} q(t, X_t^{\pi^\phi}, a_t^{\pi^\phi}; \pi^\phi) dt &= dJ(t, X_t^{\pi^\phi}; \pi^\phi) + (r(t, X_t^{\pi^\phi}, a_t^{\pi^\phi}) - \beta J(t, X_t^{\pi^\phi}; \pi^\phi)) dt \\ &\quad + \partial_x J(t, X_{t-}^{\pi^\phi}; \pi^\phi) \circ \sigma(t, X_{t-}^{\pi^\phi}, a_t^{\pi^\phi}) dW_t \\ &\quad + \int_{\mathbb{R}^d} (J(t, X_{t-}^{\pi^\phi} + \gamma(t, X_{t-}^{\pi^\phi}, a_t^{\pi^\phi}, z)) - J(t, X_{t-}^{\pi^\phi}; \pi^\phi)) \tilde{N}(dt, dz), \end{aligned}$$

Algorithm 1 Offline–Episodic q–Learning Algorithm

Inputs: initial state x_0 , horizon T , time step Δt , number of episodes N , number of mesh grids K , initial learning rates α_ψ, α_ϕ and a learning rate schedule function $l(\cdot)$ (a function of the number of episodes), functional forms of parameterized value function $J^\psi(\cdot, \cdot)$ and q-function $q^\phi(\cdot, \cdot, \cdot)$ satisfying (27), functional forms of test functions $\xi(t, x_{\cdot\Delta t}, a_{\cdot\Delta t})$ and $\zeta(t, x_{\cdot\Delta t}, a_{\cdot\Delta t})$, and temperature parameter θ .

Required program (on-policy): environment simulator $(x', r) = \text{Environment}_{\Delta t}(t, x, a)$ that takes current time–state pair (t, x) and action a as inputs and generates state x' at time $t + \Delta t$ and instantaneous reward r at time t as outputs. Policy $\pi^\phi(a|t, x) = \exp(\frac{1}{\gamma}q^\phi(t, x, a))$.

Required program (off-policy): observations $\{a_{t_k}, r_{t_k}, x_{t_{k+1}}\}_{k=0, \dots, K-1} \cup \{x_{t_K}, h(x_{t_K})\} = \text{Observation}(\Delta t)$ including the observed actions, rewards, and state trajectories under the given behavior policy at the sampling time grid with step size Δt .

Learning procedure:

Initialize ψ, ϕ .

for episode $j = 1$ **to** N **do**

Initialize $k = 0$. Observe initial state x_0 and store $x_{t_k} \leftarrow x_0$.

{**On-policy case**

while $k < K$ **do**

Generate action $a_{t_k} \sim \pi^\phi(\cdot|t_k, x_{t_k})$.

Apply a_{t_k} to environment simulator $(x, r) = \text{Environment}_{\Delta t}(t_k, x_{t_k}, a_{t_k})$, and observe new state x and reward r as outputs. Store $x_{t_{k+1}} \leftarrow x$ and $r_{t_k} \leftarrow r$.

Update $k \leftarrow k + 1$.

end while

}

{**Off-policy case** Obtain one observation $\{a_{t_k}, r_{t_k}, x_{t_{k+1}}\}_{k=0, \dots, K-1} \cup \{x_{t_K}, h(x_{t_K})\} = \text{Observation}(\Delta t)$.

}

For every $k = 0, 1, \dots, K - 1$, compute and store test functions $\xi_{t_k} = \xi(t_k, x_{t_0}, \dots, x_{t_k}, a_{t_0}, \dots, a_{t_k})$, $\zeta_{t_k} = \zeta(t_k, x_{t_0}, \dots, x_{t_k}, a_{t_0}, \dots, a_{t_k})$.

Compute

$$\Delta\psi = \sum_{i=0}^{K-1} \xi_{t_i} (J^\psi(t_{i+1}, x_{t_{i+1}}) - J^\psi(t_i, x_{t_i}) + r_{t_i}\Delta t - q^\phi(t_i, x_{t_i}, a_{t_i})\Delta t - \beta J^\psi(t_i, x_{t_i})\Delta t),$$

$$\Delta\phi = \sum_{i=0}^{K-1} \zeta_{t_i} (J^\psi(t_{i+1}, x_{t_{i+1}}) - J^\psi(t_i, x_{t_i}) + r_{t_i}\Delta t - q^\phi(t_i, x_{t_i}, a_{t_i})\Delta t - \beta J^\psi(t_i, x_{t_i})\Delta t).$$

Update ψ and ϕ by

$$\psi \leftarrow \psi + l(j)\alpha_\psi\Delta\psi.$$

$$\phi \leftarrow \phi + l(j)\alpha_\phi\Delta\phi.$$

end for

Algorithm 2 Online-Incremental q-Learning Algorithm

Inputs: initial state x_0 , horizon T , time step Δt , number of mesh grids K , initial learning rates α_ψ, α_ϕ and learning rate schedule function $l(\cdot)$ (a function of the number of episodes), functional forms of parameterized value function $J^\psi(\cdot, \cdot)$ and q-function $q^\phi(\cdot, \cdot, \cdot)$ satisfying (27), functional forms of test functions $\xi(t, x_{\cdot \wedge t}, a_{\cdot \wedge t})$ and $\zeta(t, x_{\cdot \wedge t}, a_{\cdot \wedge t})$, and temperature parameter θ .

Required program (on-policy): environment simulator $(x', r) = \text{Environment}_{\Delta t}(t, x, a)$ that takes current time–state pair (t, x) and action a as inputs and generates state x' at time $t + \Delta t$ and instantaneous reward r at time t as outputs. Policy $\pi^\phi(a|t, x) = \exp(\frac{1}{\theta}q^\phi(t, x, a))$.

Required program (off-policy): observations $\{a, r, x'\} = \text{Observation}(t, x; \Delta t)$ including the observed actions, rewards, and state when the current time-state pair is (t, x) under the given behavior policy at the sampling time grid with step size Δt .

Learning procedure:

Initialize ψ, ϕ .

for episode $j = 1$ **to** ∞ **do**

Initialize $k = 0$. Observe initial state x_0 and store $x_{t_k} \leftarrow x_0$.

while $k < K$ **do**

{**On-policy case**

Generate action $a_{t_k} \sim \pi^\phi(\cdot|t_k, x_{t_k})$.

Apply a_{t_k} to environment simulator $(x, r) = \text{Environment}_{\Delta t}(t_k, x_{t_k}, a_{t_k})$, and observe new state x and reward r as outputs. Store $x_{t_{k+1}} \leftarrow x$ and $r_{t_k} \leftarrow r$.

}

{**Off-policy case**

Obtain one observation $a_{t_k}, r_{t_k}, x_{t_{k+1}} = \text{Observation}(t_k, x_{t_k}; \Delta t)$.

}

Compute test functions $\xi_{t_k} = \xi(t_k, x_{t_0}, \dots, x_{t_k}, a_{t_0}, \dots, a_{t_k})$, $\zeta_{t_k} = \zeta(t_k, x_{t_0}, \dots, x_{t_k}, a_{t_0}, \dots, a_{t_k})$.

Compute

$$\delta = J^\psi(t_{k+1}, x_{t_{k+1}}) - J^\psi(t_k, x_{t_k}) + r_{t_k} \Delta t - q^\phi(t_k, x_{t_k}, a_{t_k}) \Delta t - \beta J^\psi(t_k, x_{t_k}) \Delta t,$$

$$\Delta\psi = \xi_{t_k} \delta,$$

$$\Delta\phi = \zeta_{t_k} \delta.$$

Update ψ and ϕ by

$$\psi \leftarrow \psi + l(j) \alpha_\psi \Delta\psi.$$

$$\phi \leftarrow \phi + l(j) \alpha_\phi \Delta\phi.$$

Update $k \leftarrow k + 1$

end while

end for

where we also use $a_t^{\pi^\phi}$ to denote the action process $a_t^{\pi^\phi, \mathbb{S}}$ for notational simplicity. We may ignore the dW_t and $\tilde{N}(dt, dz)$ terms which are martingale differences with mean zero, and then the updating rule in (28) becomes

$$\begin{aligned} \phi \leftarrow \phi + \alpha_\phi \left(dJ(t, X_t^{\pi^\phi}; \pi^\phi) + (r(t, X_t^{\pi^\phi}, a_t^{\pi^\phi}) - \beta J(t, X_t^{\pi^\phi}; \pi^\phi)) dt - \theta \log \pi^\phi(a_t^{\pi^\phi} | t, X_{t-}^{\pi^\phi}) dt \right) \\ \cdot \partial_\phi \log \pi^\phi(a_t^{\pi^\phi} | t, X_{t-}^{\pi^\phi}). \end{aligned}$$

Using $J^\psi(\cdot, \cdot)$ as the parameterized function approximator for $J(\cdot, \cdot; \pi^\phi)$, we arrive at the updating rule for the policy parameter ϕ :

$$\begin{aligned} \phi \leftarrow \phi + \alpha_\phi \left(dJ^\psi(t, X_t^{\pi^\phi}) + (r(t, X_t^{\pi^\phi}, a_t^{\pi^\phi}) - \beta J^\psi(t, X_t^{\pi^\phi})) dt - \theta \log \pi^\phi(a_t^{\pi^\phi} | t, X_{t-}^{\pi^\phi}) dt \right) \\ \cdot \partial_\phi \log \pi^\phi(a_t^{\pi^\phi} | t, X_{t-}^{\pi^\phi}). \end{aligned} \quad (34)$$

Therefore, we can update ψ using the PE methods in [Jia and Zhou \(2022a\)](#), and update ϕ using the above rule, leading to actor–critic type of algorithms.

To conclude, we are able to use the *same* RL algorithms to learn the optimal policy and optimal value function, *without having to know a priori whether the unknown environment entails a pure diffusion or a jump-diffusion*. This important conclusion is based on the theoretical analysis carried out in the previous sections.

5 Application: Mean–Variance Portfolio Selection

We now present an applied example of the general theory and algorithms derived. Consider investing in a market where there are a risk-free asset and a risky asset (e.g. a stock or an index). The risk-free rate is given by a constant r_f and the risky asset price process follows

$$dS_t = S_{t-} \left(\mu dt + \sigma dW_t + \int_{\mathbb{R}} (\exp(z) - 1) \tilde{N}(dt, dz) \right). \quad (35)$$

Let X_t be the discounted wealth value at time t , and a_t is the discounted dollar value of the investment in the risky asset. The self-financing discounted wealth process follows

$$dX_t^a = a_t \sigma \rho dt + a_t \sigma dW_t + a_t \int_{\mathbb{R}} (\exp(z) - 1) \tilde{N}(dt, dz), \quad (36)$$

where ρ is the Sharpe ratio of the risky asset and given by

$$\rho = \frac{\mu - r_f}{\sigma}. \quad (37)$$

We assume

$$\int_{|z|>1} \exp(2z) \nu(dz) < \infty. \quad (38)$$

Condition (29) implies that $\mathbb{E}[S_t]$ and $\mathbb{E}[S_t^2]$ are finite for every $t \geq 0$; see ([Cont and Tankov 2004](#), Proposition 3.14). We set

$$\sigma_J^2 := \int_{\mathbb{R}} (\exp(z) - 1)^2 \nu(dz),$$

which is finite by conditions (1) and (29).

Fix the investment horizon as $[0, T]$. The MV portfolio selection problem considers

$$\min_a \text{Var}[X_T^a] \quad \text{subject to } \mathbb{E}[X_T^a] = z.$$

We seek the optimal pre-committed strategy for the MV problem as in Zhou and Li (2000). We can transform the above constrained problem into an unconstrained one by introducing a Lagrange multiplier, which yields

$$\min_a \mathbb{E}[(X_T^a)^2] - z^2 - 2\omega(\mathbb{E}[X_T^a] - z) = \min_a \mathbb{E}[(X_T^a - \omega)^2] - (\omega - z)^2. \quad (39)$$

Note that the optimal solution to the unconstrained minimization problem depends on ω , and we can obtain the optimal multiplier ω^* by solving $\mathbb{E}[X_T^{a^*}(\omega)] = z$, which is given by

$$\omega^* = \left(z \exp\left(\frac{\rho^2 \sigma^2}{\sigma^2 + \sigma_J^2} T\right) - x_0 \right) / \left(\exp\left(\frac{\rho^2 \sigma^2}{\sigma^2 + \sigma_J^2} T\right) - 1 \right), \quad (40)$$

where x_0 is the initial wealth. The optimal deterministic control policy is given by

$$a^*(t, x) = -\frac{\sigma\rho}{\sigma^2 + \sigma_J^2}(x - \omega^*). \quad (41)$$

The exploratory formulation of the control problem is

$$\min_{\pi} \mathbb{E} \left[(\tilde{X}_T^{\pi} - \omega)^2 + \theta \int_0^T \int_{\mathcal{A}} \log \pi(a|t, \tilde{X}_{t-}^{\pi}) \pi(a|t, \tilde{X}_{t-}^{\pi}) da dt \right] - (\omega - z)^2, \quad (42)$$

where the exploratory discounted wealth under a stochastic policy π follows

$$\begin{aligned} d\tilde{X}_t^{\pi} &= \sigma\rho \int_{\mathcal{A}} a\pi(a|t, \tilde{X}_{t-}^{\pi}) da dt + \sigma \sqrt{\int_{\mathcal{A}} a^2 \pi(a|t, \tilde{X}_{t-}^{\pi}) da} dW_t \\ &\quad + \int_{\mathbb{R} \times [0,1]^n} G^{\pi}(t, \tilde{X}_{t-}^{\pi}, u) (\exp(z) - 1) \tilde{N}'(dt, dz, du), \quad t \in [0, T]. \end{aligned} \quad (43)$$

As will be explained shortly, we only need to consider stochastic feedback policies of the form $\mathcal{N}(\cdot|\alpha(t)(x - \beta(t)), \gamma(t))$, where $\alpha(t)$, $\beta(t)$ and $\gamma(t)$ are continuous functions of t . Such a policy π clearly fulfills the requirements in Definition 1 and hence is admissible. It also satisfies condition (ii) in Assumption 2.

We next discuss the well-posedness of the exploratory SDE (33). While Assumptions 1 and 2 are imposed to ensure the moment estimate (19) holds for any $p \geq 2$ in the general setting, we only need such an estimate for the second moment of the exploratory state in the particular MV application. To have such a property, we only need to require the conditions involving p in Assumptions 1 and 2 to hold up to $p = 2$, which are clearly satisfied in the problem under (29). The other requirements in Assumption 1 are also fulfilled. Thus, the exploratory SDE (33) is well-posed for the π mentioned above and Eq. (19) holds for $p = 2$.

5.1 Solution of the exploratory control problem

We consider the HJB equation for problem (32):

$$\partial_t V(t, x) + \inf_{\pi \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} (H(t, x, a, \partial_x V, \partial_x^2 V, V) + \theta \log \pi(a|t, x)) \pi(a|t, x) da = 0, \quad (44)$$

with the terminal condition $V(T, x) = (x - \omega)^2 - (\omega - z)^2$. Note that supremum becomes infimum and the sign before $\theta \log \pi(a|t, x)$ flips compared with (21) because we consider minimization here. The Hamiltonian of the problem is given by

$$\begin{aligned} H(t, x, a, \partial_x V, \partial_x^2 V, V) &= a\sigma\rho\partial_x V(t, x) + \frac{1}{2}a^2\sigma^2\partial_x^2 V(t, x) \\ &\quad + \int_{\mathbb{R}} (V(t, x + \gamma(a, z)) - V(t, x) - \gamma(a, z)\partial_x V(t, x))\nu(dz), \end{aligned}$$

where $\gamma(a, z) = a(e^z - 1)$. We take the following ansatz for the solution of the HJB equation (34):

$$V(t, x) = (x - \omega)^2 f(t) + g(t) - (\omega - z)^2. \quad (45)$$

As $V(t, x)$ is quadratic in x , we can easily calculate the integral term in the Hamiltonian and obtain

$$H(t, x, a, \partial_x V, \partial_x^2 V, V) = a\sigma\rho\partial_x V(t, x) + \frac{1}{2}a^2(\sigma^2 + \sigma_J^2)\partial_x^2 V(t, x). \quad (46)$$

The probability density function that minimizes the integral in (34) is given by

$$\pi_c(\cdot|t, x) \propto \exp\left(-\frac{1}{\theta}H(t, x, a, \partial_x V, \partial_x^2 V, V)\right),$$

which is a candidate for the optimal stochastic policy. From (36), we obtain

$$\pi_c(\cdot|t, x) \sim \mathcal{N}\left(\cdot \mid -\frac{\sigma\rho\partial_x V}{(\sigma^2 + \sigma_J^2)\partial_x^2 V}, \frac{\theta}{(\sigma^2 + \sigma_J^2)\partial_x^2 V}\right).$$

Substituting it back to the HJB equation (34), we obtain a nonlinear PDE as

$$\begin{aligned} \partial_t V - \frac{\rho^2\sigma^2}{2(\sigma^2 + \sigma_J^2)} \frac{(\partial_x V)^2}{\partial_x^2 V} - \frac{\theta}{2} \ln \frac{2\pi\theta}{(\sigma^2 + \sigma_J^2)\partial_x^2 V} &= 0, \quad (t, x) \in [0, T) \times \mathbb{R}, \\ V(T, x) &= (x - \omega)^2 - (\omega - z)^2. \end{aligned}$$

We plug in the ansatz (35) to the above PDE and obtain that $f(t)$ satisfies

$$f'(t) - \frac{\rho^2\sigma^2}{\sigma^2 + \sigma_J^2} f(t) = 0, \quad f(T) = 1,$$

and $g(t)$ satisfies

$$g'(t) - \frac{\theta}{2} \ln \frac{\pi\theta}{(\sigma^2 + \sigma_J^2)f(t)} = 0, \quad g(T) = 0.$$

These two ordinary differential equations can be solved analytically, and we obtain

$$V(t, x) = (x - \omega)^2 \exp\left(-\frac{\rho^2 \sigma^2}{\sigma^2 + \sigma_J^2}(T - t)\right) + \frac{\theta \rho^2 \sigma^2}{4(\sigma^2 + \sigma_J^2)}(T^2 - t^2) - \frac{\theta}{2}\left(\frac{\rho^2 \sigma^2}{\sigma^2 + \sigma_J^2}T + \ln \frac{\pi \theta}{\sigma^2 + \sigma_J^2}\right)(T - t) - (\omega - z)^2.$$

It follows that

$$\pi_c(\cdot | t, x) \sim \mathcal{N}\left(\cdot \mid -\frac{\sigma \rho}{\sigma^2 + \sigma_J^2}(x - \omega), \frac{\theta}{2(\sigma^2 + \sigma_J^2)} \exp\left(\frac{\rho^2 \sigma^2}{\sigma^2 + \sigma_J^2}(T - t)\right)\right), \quad (47)$$

and it is admissible. Furthermore, V solves the HJB equation (34) and shows quadratic growth. Therefore, by Lemma 5, we have the following conclusion.

Proposition 3. *For the unconstrained MV problem (30), the optimal value function $J^*(t, x) = V(t, x)$ and the optimal stochastic policy $\pi^* = \pi_c$.*

When there is no jump, we have $\sigma_J^2 = 0$ and thus recover the expressions of the optimal value function and optimal policy derived in Wang and Zhou (2020) for the unconstrained MV problem in the pure diffusion setting.

5.2 Parametrizations for q-learning

It is important to observe that the optimal value function, optimal policy and the Hamiltonian (36) take the same *structural* forms regardless of the presence of jumps, while the only differences are the constant coefficients in those functions. However, those coefficients are unknown anyway and will be parameterized in the implementation of our RL algorithms. Consequently, we can use the same parameterizations for the optimal value function and optimal q-function for learning as in the diffusion setting of Jia and Zhou (2023). This important insight, concluded only *after* a rigorous theoretical analysis, shows that the continuous-time RL algorithms are robust to the presence of jumps and essentially model-free, at least for the MV portfolio selection problem.

Following Jia and Zhou (2023), we parametrize the value function as

$$J^\psi(t, x; \omega) = (x - \omega)^2 e^{-\psi_3(T-t)} + \psi_2(t^2 - T^2) + \psi_1(t - T) - (\omega - z)^2,$$

and the q-function as

$$q^\phi(t, x, a; \omega) = -\frac{e^{-\phi_2 - \phi_3(T-t)}}{2} (a + \phi_1(x - \omega))^2 - \frac{\theta}{2} (\log(2\pi\theta) + \phi_2 + \phi_3(T - t)).$$

Let $\psi = (\psi_1, \psi_2, \psi_3)^\top$ and $\phi = (\phi_1, \phi_2, \phi_3)^\top$. The policy associated with the parametric q-function is $\pi^\phi(\cdot | t, x; \omega) = \mathcal{N}(-\phi_1(x - \omega), \theta e^{\phi_2 + \phi_3(T-t)})$. In addition to ψ and ϕ , we learn the Lagrange multiplier ω in the same way as in Jia and Zhou (2023) by the stochastic approximation algorithm that updates ω with a learning rate after a fixed number of iterations.

5.3 Simulation study

We assess the effect of jumps on the convergence behavior of our algorithms via a simulation study. We assume the risk-free rate is zero and use the same basic setting as in [Jia and Zhou \(2023\)](#): $x_0 = 1$, $z = 1.4$, $T = 1$ year, $\Delta t = 1/252$ years (corresponding to one trading day), and a chosen temperature parameter $\theta = 0.1$. We consider two market simulators: one is given by the Black–Scholes (BS) model and the other is Merton’s jump-diffusion (MJD) model in which the Lévy density is a scaled Gaussian density, i.e.,

$$\nu(z) = \lambda \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(z-m)^2}{2\delta^2}\right), \quad \lambda > 0, \delta > 0, m \in \mathbb{R}, \quad (48)$$

where λ is the arrival rate of the Poisson jumps. The Gaussian assumption is a common choice in the finance literature for modeling the jump-size distribution (see e.g. [Merton 1976](#), [Bates 1991](#), [Das 2002](#)), partly due to its tractability for statistical estimation and partly because heavy-tailed distributions may not be easily identified from real data when the number of jumps is limited (see [Heyde and Kou 2004](#)).

Under the latter model, we have

$$\sigma_j^2 = \lambda \left(\exp(2m + 2\delta^2) - 2 \exp\left(m + \frac{1}{2}\delta^2\right) + 1 \right).$$

To mimic the real market, we set the parameters of these two simulators by estimating them from the daily data of the S&P 500 index using maximum likelihood estimation (MLE). Our estimation data cover a long period from the beginning of 2000 to the end of 2023. In [Table 1](#), we summarize the estimated parameter values used for the simulators.

Simulator	Parameters
BS	$\mu = 0.0690, \sigma = 0.1965$
MJD	$\mu = 0.0636, \sigma = 0.1347, \lambda = 28.4910, m = -0.0039, \delta = 0.0275$

Table 1: Parameters used in the two simulators

For offline learning ([Algorithm 1](#)), the Lagrange multiplier ω is updated after every 10 iterations, and the parameter vectors ψ and ϕ are initialized as zero vectors. For the learning rates, we set $\alpha_\omega = 0.05$, $\alpha_\psi = 0.001$, $\alpha_\phi = (0.12, 1.5, 0.15)$ for the BS simulator, and $\alpha_\omega = 0.06$, $\alpha_\psi = 0.001$, $\alpha_\phi = (0.06, 1.3, 0.11)$ for the MJD one. We apply a decay rate $l(j) = j^{-0.51}$ to the learning rates, where j is the iteration index. In each iteration, we generate 32 independent T -year trajectories to update the parameters. We train the model for $N = 20000$ iterations.

We also consider online learning ([Algorithm 2](#)) with Δt equal to one trading day. We select a batch size of 128 trading days and update the parameters after having obtained this number of observations. We update ω after collecting 252 trading days of observations and initialize ψ and ϕ as zero vectors. For the learning rates, we set $\alpha_\omega = 0.005$, $\alpha_\psi = 0.001$, $\alpha_\phi = (0.12, 2.9, 0.16)$ for the BS simulator, and $\alpha_\omega = 0.01$, $\alpha_\psi = 0.001$, $\alpha_\phi = (0.027, 1.05, 0.45)$ for the MJD one. We apply the same decay scheme as the offline case. The model is again trained for $N = 20000$ iterations.

We display in [Table 2](#) the learned results for the parameters ϕ_1, ϕ_2 and ϕ_3 as well as the true values of these parameters calculated by plugging the BS/MJD model parameters into [\(37\)](#). For

both offline and online learning, the learned results of the four parameters are close to their true values. In practice, learning ϕ_1 and ω (parameters in the policy’s mean) is more important than ϕ_2 and ϕ_3 (parameters in the policy’s variance). Although we use a stochastic policy to interact with the environment during training to update the policy parameters, for actual execution of portfolio selection we apply a deterministic policy which is the *mean* part of the optimal stochastic policy after it has been learned. This renders an *off-policy* learning. An advantage of doing so among others is to reduce the variance of the final wealth; see [Huang et al. \(2022\)](#) for a discussion. Furthermore, the deterministic policy given by (31) is the one that optimizes the original MV objective (without entropy regularization). Hence, it makes sense to apply the mean part of the learned stochastic policy for execution, which theoretically is identical to the optimal deterministic policy (compare (31) and (37)).

	ϕ_1	ϕ_2	ϕ_3	ω
BS simulator				
True	1.7869	2.5610	0.1233	4.4481
RL-Offline	1.7613	2.5651	0.1241	4.4886
RL-Online	1.7693	2.6158	0.1217	4.3864
MJD simulator				
True	1.5940	2.5282	0.1013	5.1489
RL-Offline	1.5871	2.5451	0.1043	5.1776
RL-Online	1.5873	2.5050	0.0998	5.1840

Table 2: Learned parameters after 20000 iterations from the offline and online q-learning algorithms under two market simulators

Figure 1 plots the learning curves of ϕ_1 for both offline and online learning and under both simulators (one with jumps and one without). All the curves eventually converge, whether jumps are present or not. This demonstrates that convergence of the offline and online q-learning algorithms proposed in [Jia and Zhou \(2023\)](#) under the diffusion setting is robust to the presence of jumps for mean–variance portfolio selection.

5.4 Empirical Study

In this section we report an empirical study on investing in the S&P 500 index for one year. We obtain daily observations of the index value from the beginning of 2000 to the end of 2023, and split them into two periods: the training period (the first 20 years) and the test period (the last 4 years). Recall that the test period includes the U.S. stock market meltdown at the onset of the pandemic in 2020, the 2022 bear period as well as various bull subperiods.

We assume the risk-free rate is zero and perform daily rebalancing. We compare our offline and online RL algorithms with the plug-in approach based on MLE estimates. In the latter, we consider both BS and MJD models and fit them to daily returns of the index in the training period. The MLE estimates are plugged into (5) to obtain the respective optimal Lagrange multipliers and into (31) to obtain the corresponding optimal deterministic policies, which are then applied on the test data. To generate episodes for training the two RL algorithms, we bootstrap index returns in

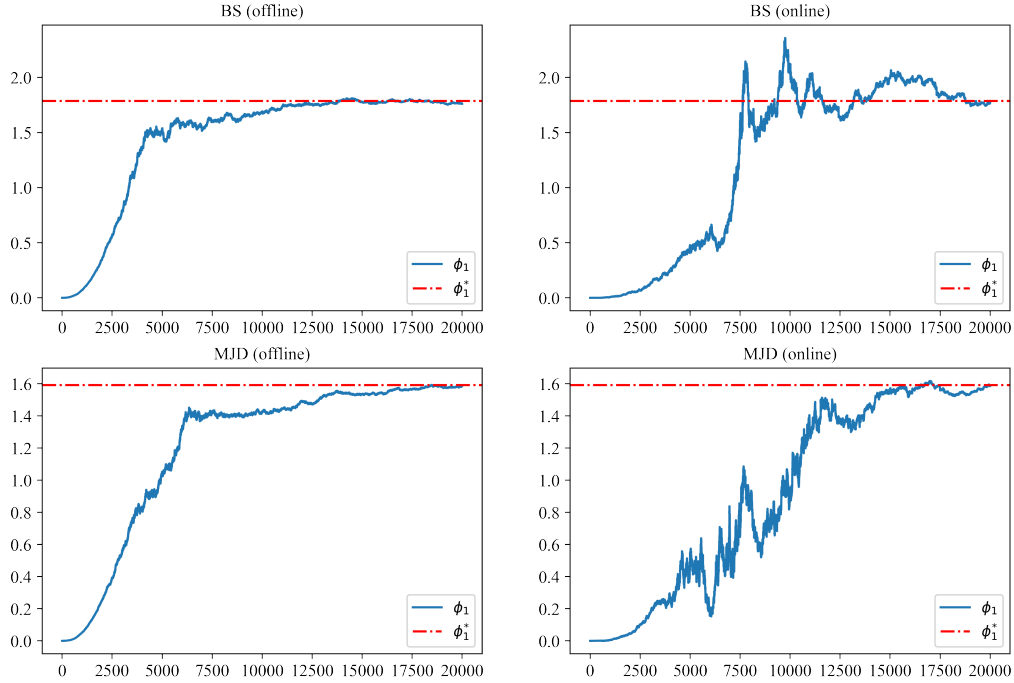


Figure 1: Convergence of the offline and online q-learning algorithms under two market simulators for the policy parameter ϕ_1 (iteration index on the x -axis)

the training period to create one-year episodes of index prices. We initialize the policy parameter ϕ_1 by using the MLE estimates of the MJD model parameters in the formula $\sigma\rho/(\sigma^2 + \sigma_J^2)$. The implementation of RL follows the same practice in our simulation study and the mean part of the learned stochastic policy is applied on the test data. To create the test data, we generate a set of 5000 one-year episodes of index prices by bootstrapping index returns in the test period, which is used for testing the RL algorithms and the plug-in method.

Table 3 compares the out-of-sample test performances of the four methods with different levels of the target wealth. As a natural benchmark, we also report the annualized mean return, volatility and Sharpe ratio of the index in the 4-year test period, which corresponds to a buy-and-hold strategy. It is clear that in terms of the Sharpe ratio, all the methods beat the index yet the two RL algorithms outperform the MLE method significantly. A closer examination shows that this outperformance arises from RL's decisive lower volatilities in all the cases. Indeed, in the case when $z = 1.1$ representing a (reasonably targeted) 10% expected annual return, both RL strategies' volatilities are smaller than that of the index, and the online one is less than half of those of the MLE!

	ϕ_1	ω	MR	Vol	SR
<hr/>					
$z = 1.4$					
RL-Offline	3.0507	2.8560	0.6823	0.7765	0.8788
RL-Online	2.7079	2.5421	0.5062	0.6011	0.8422
MLE-BS	1.6203	5.4748	0.8938	1.2038	0.7425
MLE-MJD	1.3877	6.7992	0.9954	1.3752	0.7238
<hr/>					
$z = 1.3$					
RL-Offline	2.8040	2.5218	0.5165	0.6061	0.8522
RL-Online	2.6593	2.1760	0.3794	0.4532	0.8373
MLE-BS	1.6203	4.3561	0.6704	0.9029	0.7425
MLE-MJD	1.3877	5.3494	0.7466	1.0314	0.7238
<hr/>					
$z = 1.2$					
RL-Offline	3.0599	1.9130	0.3366	0.3826	0.8798
RL-Online	2.4275	1.8672	0.2564	0.3148	0.8143
MLE-BS	1.6203	3.2374	0.4469	0.6019	0.7425
MLE-MJD	1.3877	3.8996	0.4977	0.6876	0.7238
<hr/>					
$z = 1.1$					
RL-Offline	2.9130	1.4729	0.1664	0.1926	0.8637
RL-Online	2.5111	1.4031	0.1231	0.1497	0.8224
MLE-BS	1.6203	2.1187	0.2234	0.3010	0.7425
MLE-MJD	1.3877	2.4498	0.2489	0.3438	0.7238
<hr/>					
Buy-and-hold	N.A.	N.A.	0.0984	0.2312	0.4256
<hr/>					

Table 3: Mean return (MR), volatility (Vol), and Sharpe ratio (SR) of investment portfolios obtained by four methods on 5000 test episodes. We show the learned ϕ_1 and ω under each approach. For MLE-BS and MLE-MJD, ϕ_1 and ω are calculated by plugging the MLE estimates to $\phi_1 = \sigma\rho/(\sigma^2 + \sigma_j^2)$ and (5), respectively.

5.5 Effects of jumps

The theoretical analysis so far in this section shows that, *for the MV portfolio selection problem*, one does not need to know in advance whether or not the stock prices have jumps in order to carry out the RL task, because the optimal stochastic policy is Gaussian and the corresponding value function and q-function have the same structures for parameterization irrespective of the presence of jumps. However, we stress that this is rather an exception than a rule. Here we give a counterexample.

Consider a modification of the MV portfolio selection problem where the controlled state follows

$$dX_t^a = a_t \sigma \rho dt + a_t \sigma dB_t + \int_{\mathbb{R}} \gamma(a_t, z) \tilde{N}(dt, dz), \quad (49)$$

with

$$\gamma(a, z) = a^2. \quad (50)$$

The exploratory value function is given by

$$J^*(t, x; w) = \min_{\pi} \mathbb{E}_{t,x} \left[(\tilde{X}_T^{\pi} - w)^2 + \theta \int_t^T \int_{\mathcal{A}} \log \pi(a|s, \tilde{X}_{s-}^{\pi}) \pi(a|s, \tilde{X}_{s-}^{\pi}) da ds \right] - (\omega - z)^2.$$

Note that this is *not* an MV portfolio selection problem because (5.5) does not correspond to a self-financed wealth equation with a reasonably modelled stock price process.

The Hamiltonian is given by

$$\begin{aligned} H(t, x, a, \partial_x v, \partial_x^2 v, v) &= a \sigma \rho \partial_x v(t, x) + \frac{1}{2} a^2 \sigma^2 \partial_x^2 v(t, x) \\ &+ \int_{\mathbb{R}} (v(t, x + a^2) - v(t, x) - a^2 \cdot \partial_x v(t, x)) \nu(dz). \end{aligned} \quad (51)$$

If an optimal stochastic policy exists, then it must be

$$\pi^*(a|t, x) \propto \exp \left(-\frac{1}{\theta} H(t, x, a, \partial_x J^*, \partial_x^2 J^*, J^*) \right). \quad (52)$$

We show by contradiction that the optimal stochastic policy can not be Gaussian in this case. Note that if there is no optimal stochastic policy, then it would already demonstrate that jumps matter because the optimal stochastic policy for the case of no jumps exists and is Gaussian.

Remark 5. *The existence of optimal stochastic policy in (5.5) is equivalent to the integrability of the quantity $\exp(\frac{1}{\theta} H(t, x, a, \partial_x J^*, \partial_x^2 J^*, J^*))$ over $a \in \mathcal{A} = (-\infty, \infty)$. This integrability depends on the tail behavior of the Hamiltonian and, in particular, the behavior of $J^*(t, x + a^2)$ when a^2 is large.*

Suppose the optimal stochastic policy $\pi^*(\cdot|t, x)$ is Gaussian for all (t, x) , implying that the Hamiltonian $H(t, x, a, \partial_x J^*, \partial_x^2 J^*, J^*)$ is a quadratic function of a . It then follows from (5.5) that there exist functions $h_1(t, x)$ and $h_2(t, x)$ such that

$$J^*(t, x + a^2) - J^*(t, x) - a^2 \partial_x J^*(t, x) = a^2 \cdot h_1(t, x) + a \cdot h_2(t, x), \quad \text{for all } (t, x, a). \quad (53)$$

We do not put a term independent of a on the right-hand side because the left-hand side is zero when $a = 0$. Taking derivative with respect to a , we obtain

$$\partial_x J^*(t, x + a^2) \cdot 2a - 2a \partial_x J^*(t, x) = 2a \cdot h_1(t, x) + h_2(t, x), \quad \text{for all } (t, x, a).$$

Setting $a = 0$, we get $h_2 = 0$. It follows that

$$a \cdot [\partial_x J^*(t, x + a^2) - \partial_x J^*(t, x) - h_1(t, x)] = 0 \quad \text{for all } (t, x, a).$$

Hence, we have $h_1(t, x) = \partial_x J^*(t, x + a^2) - \partial_x J^*(t, x)$ for any $a \neq 0$. Sending a to zero yields $h_1(t, x) = 0$ for all (t, x) . Therefore, we obtain from (5.5) that

$$J^*(t, x + a^2) - J^*(t, x) - a^2 \partial_x J^*(t, x) = 0, \quad \text{for all } (t, x, a).$$

Taking derivative in a in the above we have

$$\partial_x J^*(t, x + a^2) - \partial_x J^*(t, x) = 0, \quad \text{for all } (t, x, a).$$

Thus, $\partial_x J^*$ is constant in x or J^* is affine in x , leading to $J^*(t, x) = g_1(t)x + g_2(t)$ for some functions $g_1(t)$ and $g_2(t)$. The resulting Hamiltonian becomes

$$H(t, x, a, \partial_x J^*, \partial_x^2 J^*, J^*) = a \sigma \rho g_1(t).$$

This is linear in $a \in \mathcal{A} = (-\infty, \infty)$ and hence the integral $\int_{\mathcal{A}} \exp(-\frac{1}{\theta} H(t, x, a, \partial_x J^*, \partial_x^2 J^*, J^*)) da$ does not exist. It follows that $\pi^*(\cdot | t, x)$ does not exist, which is a contradiction. Therefore, we have shown that under (5.5), the optimal stochastic policy either does not exist or is not Gaussian when it exists.

Remark 6. *The argument above works for $\gamma(a, z) = a^m$ for any $m > 1$.*

6 Application: Mean–Variance Hedging of Options

The MV portfolio selection problem considered in Section 5 is an LQ problem. In this section, we present another application that is non-LQ. Consider an option seller who needs to hedge a short position in a European-style option that expires at time T . The option is written on a risky asset whose price process S is described by the SDE (5) with condition (29) satisfied. At the terminal time T , the seller pays $G(S_T)$ to the option holder. We assume that the seller's hedging activity will not affect the risky asset price.

To hedge the random payoff, the seller constructs a portfolio consisting of the underlying risky asset and cash. We consider discounted quantities in the problem: the discounted risky asset price $\hat{S}_t := e^{-r_f t} S_t$ and discounted payoff $\hat{G}(\hat{S}_T) := e^{-r_f T} G(S_T)$, where r_f is again the constant risk-free interest rate. As an example, for a put option with strike price \mathcal{K} , $G(S_T) = (\mathcal{K} - S_T)^+$ ($x^+ := \max(x, 0)$), and $\hat{G}(\hat{S}_T) = (\hat{\mathcal{K}} - \hat{S}_T)^+$, where $\hat{\mathcal{K}} := e^{-r_f T} \mathcal{K}$. Here \hat{S} follows the SDE

$$d\hat{S}_t = \hat{S}_t \left(\rho \sigma dt + \sigma dW_t + \int_{\mathbb{R}} (\exp(z) - 1) \tilde{N}(dt, dz) \right),$$

where ρ is defined in (5).

We denote the discounted dollar value in the risky asset and the discounted value of the hedging portfolio at time t by a_t and X_t , respectively. As in the MV portfolio selection problem, a is the control variable and X follows the SDE (5). The seller seeks a hedging policy to minimize the deviation from the terminal payoff. A popular formulation is mean–variance hedging (also known as quadratic hedging), which considers the objective

$$\min_a \mathbb{E} \left[(X_T^a - \hat{G}(\hat{S}_T))^2 \right]. \quad (54)$$

Note that this expectation is taken under the real-world probability measure, rather than under any martingale measure for option pricing. An advantage of this formulation is that relevant data are observable in the real world (but not necessarily in a risk-neutral world). Although objective (6) looks similar to that of the MV portfolio selection problem, the target the hedging portfolio tries to meet is now random instead of constant. Furthermore, the payoff is generally nonlinear in \hat{S}_T . Thus, MV hedging is not an LQ problem. Also, both \hat{S}_t and X_t^a matter for the hedging decision at time t , whereas \hat{S}_t is irrelevant for decision in the MV portfolio selection problem. This increase in the state dimension makes the hedging problem much more difficult to solve.

We consider the following exploratory formulation to encourage exploration for learning:

$$\min_{\pi} \mathbb{E} \left[(\tilde{X}_T^{\pi} - \hat{G}(\hat{S}_T))^2 + \theta \int_0^T \int_{\mathcal{A}} \log \pi(a|t, X_{t-}^{\pi}) \pi(a|t, X_{t-}^{\pi}) da dt \right], \quad (55)$$

where the discounted value of the hedging portfolio under a stochastic policy π follows (33).

6.1 Solution of the exploratory control problem

We consider the HJB equation for problem (6):

$$\partial_t V + \inf_{\pi \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} (H(t, S, x, a, \partial_S V, \partial_S^2 V, \partial_x V, \partial_x^2 V, \partial_S \partial_x V, V) + \theta \log \pi(a|t, S, x)) \pi(a|t, S, x) da = 0, \quad (56)$$

with terminal condition $V(T, S, x) = (x - \hat{G}(S))^2$. The Hamiltonian of the problem is given by

$$\begin{aligned} & H(t, S, x, a, \partial_S V, \partial_S^2 V, \partial_x V, \partial_x^2 V, \partial_S \partial_x V, V) \\ &= \rho \sigma S \partial_S V + \frac{1}{2} \sigma^2 S^2 \partial_S^2 V + a \rho \sigma \partial_x V + \frac{1}{2} a^2 \sigma^2 \partial_x^2 V + a \sigma^2 S \partial_S \partial_x V \\ &+ \int_{\mathbb{R}} (V(t, S + \tilde{\gamma}(S, z), x + \gamma(a, z)) - V(t, S, x) - \tilde{\gamma}(S, z) \partial_S V(t, S, x) - \gamma(a, z) \partial_x V(t, S, x)) \nu(dz) \end{aligned}$$

with $\gamma(a, z) = a(\exp(z) - 1)$ and $\tilde{\gamma}(S, z) = S(\exp(z) - 1)$.

We make the following ansatz for the solution of the HJB equation (6.1):

$$V(t, S, x) = (x - h(t, S))^2 f(t) + g(t, S). \quad (57)$$

With this, we can simplify the integral term in the Hamiltonian and obtain

$$\begin{aligned} & H(t, S, x, a, \partial_S V, \partial_S^2 V, \partial_x V, \partial_x^2 V, \partial_S \partial_x V, V) \\ &= \rho \sigma S \partial_S V + \frac{1}{2} \sigma^2 S^2 \partial_S^2 V + \frac{1}{2} a^2 (\sigma^2 + \sigma_J^2) \partial_x^2 V \end{aligned}$$

$$\begin{aligned}
& + a \left(\rho\sigma\partial_x V + \sigma^2 S\partial_S\partial_x V - \int_{\mathbb{R}} (\exp(z) - 1)(h(t, S\exp(z)) - h(t, S))\nu(dz) \right) \\
& + \int_{\mathbb{R}} (V(t, S\exp(z), x) - V(t, S, x) - S(\exp(z) - 1)\partial_S V(t, S, x))\nu(dz). \tag{58}
\end{aligned}$$

The probability density function that minimizes the integral in (6.1) is given by

$$\pi_c(\cdot|t, S, x) \propto \exp\left(-\frac{1}{\theta}H(t, S, x, a, \partial_S V, \partial_S^2 V, \partial_x V, \partial_x^2 V, \partial_S\partial_x V, V)\right),$$

which is a candidate for the optimal stochastic policy. From (6.1), we obtain that $\pi_c(\cdot|t, S, x)$ is given by

$$\mathcal{N}\left(\cdot \mid -\frac{\rho\sigma\partial_x V + \sigma^2 S\partial_S\partial_x V - \int_{\mathbb{R}} (\exp(z) - 1)(h(t, S\exp(z)) - h(t, S))\nu(dz)\partial_x^2 V}{(\sigma^2 + \sigma_J^2)\partial_x^2 V}, \frac{\theta}{(\sigma^2 + \sigma_J^2)\partial_x^2 V}\right).$$

Substituting it back to the HJB equation (6.1), we obtain a nonlinear PIDE

$$\begin{aligned}
& \partial_t V + \rho\sigma S\partial_S V + \frac{1}{2}\sigma^2 S^2\partial_S^2 V \\
& + \int_{\mathbb{R}} (V(t, S\exp(z), x) - V(t, S, x) - S(\exp(z) - 1)\partial_S V(t, S, x))\nu(dz) \\
& - \frac{(\rho\sigma\partial_x V + \sigma^2 S\partial_S\partial_x V - \int_{\mathbb{R}} (\exp(z) - 1)(h(t, S\exp(z)) - h(t, S))\nu(dz)\partial_x^2 V)^2}{2(\sigma^2 + \sigma_J^2)\partial_x^2 V} \\
& - \frac{\theta}{2} \ln \frac{2\pi\theta}{(\sigma^2 + \sigma_J^2)\partial_x^2 V} = 0, \quad (t, S, x) \in [0, T] \times \mathbb{R}_+ \times \mathbb{R}, \quad V(T, x) = (x - \hat{G}(S))^2.
\end{aligned}$$

We plug in the ansatz (6.1) to the above PIDE. After some lengthy calculations, we can collect similar terms and obtain the following equations satisfied by f , h and g :

$$\begin{aligned}
& f'(t) - \frac{\rho^2\sigma^2}{\sigma^2 + \sigma_J^2}f(t) = 0, \quad f(T) = 1, \\
& \partial_t h + \int_{\mathbb{R}} (h(t, S\exp(z)) - h(t, S) - S(\exp(z) - 1)\partial_S h(t, S)) \left(1 - \frac{\rho\sigma}{\sigma^2 + \sigma_J^2}(\exp(z) - 1)\right)\nu(dz) \\
& + \frac{1}{2}\sigma^2 S^2\partial_S^2 h = 0, \quad (t, S) \in [0, T] \times \mathbb{R}_+, \quad h(T, S) = \hat{G}(S), \tag{59} \\
& \partial_t g + \rho\sigma S\partial_S g + \frac{1}{2}\sigma^2 S^2\partial_S^2 g + \int_{\mathbb{R}} (g(t, S\exp(z)) - g(t, S) - S(\exp(z) - 1)\partial_S g(t, S))\nu(dz) \\
& + \sigma^2 S^2(\partial_S h)^2 f(t) + \int_{\mathbb{R}} (h(t, S\exp(z)) - h(t, S))^2 f(t)\nu(dz) \\
& - \frac{f(t)}{\sigma^2 + \sigma_J^2} \left(\sigma^2 S\partial_S h + \int_{\mathbb{R}} (\exp(z) - 1)(h(t, S\exp(z)) - h(t, S))\nu(dz) \right)^2 \\
& - \frac{\theta}{2} \ln \frac{\pi\theta}{(\sigma^2 + \sigma_J^2)f(t)} = 0, \quad (t, S) \in [0, T] \times \mathbb{R}_+, \quad g(T, S) = 0. \tag{60}
\end{aligned}$$

The function f is given by

$$f(t) = \exp\left(-\frac{\rho^2\sigma^2}{\sigma^2 + \sigma_J^2}(T - t)\right). \tag{61}$$

However, the two PIDEs cannot be solved in closed form in general. Below we first present some properties of h and g .

Lemma 6. *Assume $\sigma > 0$ and $\hat{G}(S)$ is Lipschitz continuous in S . The PIDE (6.1) has a unique solution in $C^{1,2}([0, T] \times \mathbb{R}_+) \cap C([0, T] \times \mathbb{R}_+)$ that is Lipschitz continuous in S . The PIDE (6.1) also has a unique solution in $C^{1,2}([0, T] \times \mathbb{R}_+) \cap C([0, T] \times \mathbb{R}_+)$. Moreover, there exists a constant $C > 0$ such that*

$$|h(t, S)| \leq C(1 + S), \quad |g(t, S)| \leq C(1 + S^2) \quad \text{for any } (t, S) \in [0, T] \times \mathbb{R}_+.$$

Next, we provide stochastic representations for the solutions to PIDEs (6.1) and (6.1), which will be subsequently exploited for our RL study. Construct a new probability measure \mathbb{Q} from the given probability measure \mathbb{P} with the Radon–Nikodym density process $\Lambda_s := \frac{d\mathbb{Q}}{d\mathbb{P}}|_{\mathcal{F}_s}$ defined by

$$d\Lambda_s = -\frac{\rho\sigma}{\sigma^2 + \sigma_j^2} \Lambda_{s-} \left(\sigma dW_s + \int_{\mathbb{R}} (\exp(z) - 1) \tilde{N}(ds, dz) \right), \quad \Lambda_t = 1.$$

Condition (29) guarantees that $\{\Lambda_s : t \leq s \leq T\}$ is a true martingale with unit expectation. The standard measure change results yield that, under \mathbb{Q} ,

$$W'_s := W_s + \frac{\rho\sigma^2}{\sigma^2 + \sigma_j^2} (s - t), \quad \tilde{N}'(ds, dz) := N(ds, dz) - \left(1 - \frac{\rho\sigma}{\sigma^2 + \sigma_j^2} (\exp(z) - 1) \right) \nu(dz) ds \quad (62)$$

are standard Brownian motion and compensated Poisson random measure, respectively. Using them we can rewrite the SDE for \hat{S} as

$$d\hat{S}_s = \hat{S}_{s-} \left(\sigma dW'_s + \int_{\mathbb{R}} (\exp(z) - 1) \tilde{N}'(ds, dz) \right).$$

Let $Y_t := 1 + \sigma W_t + \int_0^t \int_{\mathbb{R}} (\exp(z) - 1) \tilde{N}'(ds, dz)$. Clearly, it is a Lévy process and also a martingale under condition (29). Process \hat{S} is the stochastic exponential of Y and it follows from (Cont and Tankov 2004, Proposition 8.23) that \hat{S} is also a martingale under \mathbb{Q} . The Feynman–Kac theorem gives the representation of the solution h to PIDE (6.1) as

$$h(t, S) = \mathbb{E}^{\mathbb{Q}}[\hat{G}(\hat{S}_T) | \hat{S}_t = S], \quad t \in [0, T]. \quad (63)$$

We can view \mathbb{Q} as the martingale measure for option valuation and interpret $h(t, S)$ as the option price at time t and underlying price S . For PIDE (6.1), again by the Feynman–Kac theorem we obtain that its solution g is given by

$$\begin{aligned} g(t, S) &= g_e(t, S) - \int_t^T \frac{\theta}{2} \ln \frac{\pi\theta}{(\sigma^2 + \sigma_j^2) f(s)} ds \\ &= g_e(t, S) - \frac{\theta}{2} \left(\ln \left(\frac{\pi\theta}{\sigma^2 + \sigma_j^2} \right) (T - t) + \frac{\rho^2 \sigma^2}{2(\sigma^2 + \sigma_j^2)} (T - t)^2 \right), \end{aligned} \quad (64)$$

where

$$g_e(t, S) := \mathbb{E}^{\mathbb{P}} \left[\int_t^T f(s) \left(\sigma^2 \hat{S}_s^2 (\partial_S h(s, \hat{S}_s))^2 + \int_{\mathbb{R}} (h(s, \hat{S}_s \exp(z)) - h(s, \hat{S}_s))^2 \nu(dz) \right) ds \right]$$

$$- \int_t^T \frac{f(s)}{\sigma^2 + \sigma_J^2} \left(\sigma^2 \hat{S}_s \partial_S h(s, \hat{S}_s) + \int_{\mathbb{R}} (\exp(z) - 1) (h(s, \hat{S}_s \exp(z)) - h(s, \hat{S}_s)) \nu(dz) \right)^2 ds \Big| \hat{S}_t = \mathfrak{S} \Big] \quad (65)$$

Using the expression for $V(t, S, x)$, we obtain $\pi_c(\cdot | t, S, x)$ as

$$\pi_c(\cdot | t, S, x) \sim \mathcal{N} \left(\cdot \mid \mathcal{M}^*(t, S, x), \frac{\theta}{2(\sigma^2 + \sigma_J^2)} \exp \left(\frac{\rho^2 \sigma^2}{\sigma^2 + \sigma_J^2} (T - t) \right) \right), \quad (66)$$

where

$$\mathcal{M}^*(t, S, x) = \frac{\sigma^2 S \partial_S h(t, S) + \int_{\mathbb{R}} (\exp(z) - 1) (h(t, S \exp(z)) - h(t, S)) \nu(dz) - \rho \sigma (x - h(t, S))}{\sigma^2 + \sigma_J^2}. \quad (67)$$

Using Lemma 6, we can directly verify that π_c is admissible. It is also obvious to see that $V(t, S, x)$ given by (6.1) has quadratic growth in S and x . Therefore, by Lemma 5, we have the following conclusion.

Proposition 4. *For the MV hedging problem (6), the optimal value function $J^*(t, S, x) = V(t, S, x)$ and the optimal stochastic policy $\pi^* = \pi_c$ under the assumptions that $\sigma > 0$ and $\hat{G}(S)$ is Lipschitz continuous in S .*

The mean part of π^* , $\mathcal{M}^*(t, S, x)$, comprises three terms. The quantity $\partial_S h(t, S)$ is the delta of the option and hence $S \partial_S h(t, S)$ gives the dollar amount of the risky asset as required by delta hedging, which is used to hedge continuous price movements. The integral term shows the dollar amount of the risky asset that should be held to hedge discontinuous changes. The two hedges are combined using weights $\sigma^2 / (\sigma^2 + \sigma_J^2)$ and $1 / (\sigma^2 + \sigma_J^2)$. The last term reflects the adjustment that needs to be made due to the discrepancy between the hedging portfolio value and option price. It is easy to show that $\mathcal{M}^*(t, S, x)$ is the optimal deterministic policy of problem (6).

The variance of π^* is the same as that in the MV portfolio selection problem (c.f. (37)), which decreases as t approaches the terminal time T . This implies that one gradually reduces exploration over time.

6.2 Parametrizations and actor–critic learning

We use the previously derived solution of the exploratory problem as knowledge for RL. As we will see, the exploratory solution lends itself to natural parametrizations for the policy (“actor”) and value function (“critic”), but less so for parametrizing the q-function.

To simplify the integral in $\mathcal{M}^*(t, S, x)$ defined in (6.1), we first parametrize the Lévy density $\nu(z)$ as in (5.3) where we have three parameters: λ is the jump arrival rate, and m and δ are the mean and standard deviation of the normal distribution for the jump size. We then use the Fourier-cosine method developed in Fang and Oosterlee (2009) to obtain an expression for $h(t, S)$. Let $Y_t := \log(\hat{S}_t / \hat{S}_0)$. Applying Itô’s formula, under \mathbb{P} we obtain

$$d \log \hat{S}_t = \left(\rho \sigma - \frac{1}{2} \sigma^2 - \lambda \left(\exp \left(m + \frac{1}{2} \delta^2 \right) - 1 \right) \right) dt + \sigma dW_t + \int_{\mathbb{R}} z N(dt, dz).$$

Noting (6.1), under \mathbb{Q} we have

$$d \log \hat{S}_t = \left(\rho \sigma \frac{\sigma_J^2}{\sigma^2 + \sigma_J^2} - \frac{1}{2} \sigma^2 - \lambda \left(\exp \left(m + \frac{1}{2} \delta^2 \right) - 1 \right) \right) dt + \sigma dW_t' + \int_{\mathbb{R}} z N(dt, dz),$$

where the Lévy measure of $N(dt, dz)$ becomes

$$\nu'(dz) = \left(1 - \frac{\rho\sigma}{\sigma^2 + \sigma_J^2} (\exp(z) - 1)\right) \frac{\lambda}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(z-m)^2}{2\delta^2}\right) dz.$$

The characteristic function of Y_t under \mathbb{Q} is given by

$$\Phi_{Y_t}^{\mathbb{Q}}(u) = \exp\left(iut \left(\rho\sigma \frac{\sigma_J^2}{\sigma^2 + \sigma_J^2} - \frac{1}{2}\sigma^2 - \lambda \left(\exp\left(m + \frac{1}{2}\delta^2\right) - 1\right)\right) - \frac{1}{2}\sigma^2 u^2 t + tA(u)\right),$$

where

$$\begin{aligned} A(u) &= \int_{\mathbb{R}} (\exp(iuz) - 1) \nu'(dz) \\ &= \lambda \left(\exp\left(ium - \frac{1}{2}\delta^2 u^2\right) - 1 \right) - \frac{\lambda\rho\sigma}{\sigma^2 + \sigma_J^2} \times \\ &\quad \left(\exp\left(m + \frac{1}{2}\delta^2 + iu(m + \delta^2) - \frac{1}{2}\delta^2 u^2\right) - \exp\left(ium - \frac{1}{2}\delta^2 u^2\right) - \exp\left(m + \frac{1}{2}\delta^2\right) + 1 \right). \end{aligned}$$

The Fourier-cosine method then calculates $h(t, S)$ as

$$\hat{h}(t, S) := \sum_{k=0}^{N-1'} \Re\left(\Phi_{Y_t}^{\mathbb{Q}}\left(\frac{k\pi}{b-a}\right) \exp\left(ik\pi \frac{\ln(S/\hat{\mathcal{K}}) - a}{b-a}\right)\right) V_k. \quad (68)$$

Here, $\Re(x)$ denotes the real part of complex number x , Σ' indicates the first term in the sum is halved, $\hat{\mathcal{K}}$ is the discounted strike price of the call or put option under consideration, and $[a, b]$ is the truncation region for Y_T (chosen wide enough so that the probability of Y_T going outside under \mathbb{Q} can be neglected). The expression of V_k is given by Eqs. (24) and (25) in [Fang and Oosterlee \(2009\)](#) for call and put options, respectively, and it only depends on $a, b, \hat{\mathcal{K}}$. We can also calculate $\partial_S h(t, S)$ by differentiating $\hat{h}(t, S)$ w.r.t. S , which yields

$$\partial_S \hat{h}(t, S) := \sum_{k=0}^{N-1'} \Re\left(\Phi_{Y_t}^{\mathbb{Q}}\left(\frac{k\pi}{b-a}\right) \frac{ik\pi}{(b-a)S} \exp\left(ik\pi \frac{\ln(S/\hat{\mathcal{K}}) - a}{b-a}\right)\right) V_k. \quad (69)$$

Actor parametrization. We follow the form of the optimal stochastic policy given in (6.1) to parametrize our policy. To obtain a parsimonious parametrization, we introduce $\phi_1, \phi_2, \phi_3, \phi_4, \phi_5$, which respectively represent the agent's understanding – relative to her learning – of drift μ , volatility σ , jump arrival rate λ , mean m and standard deviation δ of the jump size of the risky asset. We emphasize that in our learning algorithm these parameters are not *estimated* by any statistical method (and hence they do not necessarily correspond to the true dynamics of the risky asset); rather they will be updated based on the hedging experiences (via exploration and exploitation) from the real environment. To implement (6.1), we also need two dependent parameters ϕ_6 and ϕ_7 that are calculated by

$$\phi_6 = \frac{\phi_1 - r_f}{\phi_2}, \quad \phi_7 = \sqrt{\phi_3 \left(\exp(2\phi_4 + 2\phi_5^2) - 2 \exp\left(\phi_4 + \frac{1}{2}\phi_5^2\right) + 1 \right)}.$$

These two parameters show the agent's understanding of ρ and σ_J , and they will be updated in the learning process by the above equations with the update of the independent parameters ϕ_1 to ϕ_5 . Denote $\phi = (\phi_i)_{i=1}^7$.

We use $\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7$ to replace their counterparts $\mu, \sigma, \lambda, m, \delta, \rho, \sigma_J$ in (6.1) to obtain a parametrized policy. For the mean part, we calculate the option price and its delta by plugging ϕ into (6.2) and (6.2), with the results denoted by $\hat{h}^\phi(t, S)$ and $\hat{h}_S^\phi(t, S)$ respectively, and approximate the integral by Gauss-Hermite (G-H) quadrature, which yields

$$\mathcal{M}^\phi(t, S, x) = \frac{\phi_2^2 S \partial_S \hat{h}^\phi(t, S) + \frac{\phi_3}{\sqrt{2\pi}} \sum_{q=1}^Q w_q (\exp(z_q) - 1) \left(\hat{h}^\phi(t, S \exp(z_q)) - \hat{h}^\phi(t, S) \right) - \phi_2 \phi_6 (x - \hat{h}^\phi(t, S))}{\phi_2^2 + \phi_7^2},$$

where $z_q = \phi_4 + \phi_5 u_q$ and $\{(w_q, u_q), q = 1, \dots, Q\}$ are the weights and abscissa of the Q -point G-H rule that are predetermined and independent of ϕ . For the variance part, we use ϕ in (6.1) to obtain $f^\phi(t)$. Thus, our parametrized policy is given by

$$\pi^\phi(\cdot | t, S, x) \sim \mathcal{N}\left(\cdot \mid \mathcal{M}^\phi(t, S, x), \frac{\theta}{2(\phi_2^2 + \phi_7^2) f^\phi(t)}\right).$$

Critic parametrization. For the value function of π^ϕ , following (6.1) we parametrize it as

$$J^{\psi, \phi}(t, S, x) = (x - \hat{h}^\phi(t, S))^2 f^\phi(t) + g_e^\psi(t, S) - \frac{\theta}{2} \left(\ln \left(\frac{\pi \theta}{\phi_2^2 + \phi_7^2} \right) (T - t) + \frac{\phi_2^2 \phi_6^2}{2(\phi_2^2 + \phi_7^2)} (T - t)^2 \right).$$

We do not specify any parametric form for $g_e^\psi(t, S)$, but will learn it using Gaussian process regression (Williams and Rasmussen 2006), where ψ denotes its parameters. We choose Gaussian process for our problem because it is able to generate flexible shapes and can often achieve a reasonably good fit with a moderate amount of data.

Actor-critic learning. Consider a time grid $\{t_k\}_{k=0}^K$ with $t_0 = 0, t_K = T$, and time step Δt . In an iteration, we start with a given policy π^ϕ and use it to collect M episodes from the environment, where the m -th episode is given by $\{(t_k, \hat{S}_{t_k}^{(m)}, X_{t_k}^{\pi^\phi, (m)}, a_{t_k}^{\pi^\phi, (m)})\}_{k=0}^K$. We then perform two steps.

In the critic step, the main task is to learn $g_e^\psi(t_k, S)$ as a function of S for each t_k . The stochastic representation (6.1) shows that g_e can be viewed as the value function of a stream of rewards that depend on time and the risky asset price; so the task is a policy evaluation problem. We first obtain the running reward path in every episode, where the integral involving the Lévy measure is again calculated by the Q -point G-H rule. Specifically, the reward at time t_k for $k = 0, \dots, K-1$ is given by

$$R_{t_k}^g := f^\phi(t_k) \left(\phi_2^2 \hat{S}_{t_k}^2 (\partial_S \hat{h}^\phi(t_k, \hat{S}_{t_k}))^2 + \frac{\phi_3}{\sqrt{2\pi}} \sum_{q=1}^Q w_q (\hat{h}^\phi(t_k, \hat{S}_{t_k} \exp(z_q)) - \hat{h}^\phi(t_k, \hat{S}_{t_k}))^2 \right) \Delta t \\ - \frac{f^\phi(t_k)}{\phi_2^2 + \phi_7^2} \left(\phi_2^2 \hat{S}_{t_k} \partial_S \hat{h}^\phi(t_k, \hat{S}_{t_k}) + \frac{\phi_3}{\sqrt{2\pi}} \sum_{q=1}^Q w_q (\exp(z_q) - 1) (\hat{h}^\phi(t_k, \hat{S}_{t_k} \exp(z_q)) - \hat{h}^\phi(t_k, \hat{S}_{t_k})) \right)^2 \Delta t,$$

where the risky asset price path $\{\{\hat{S}_{t_k}\}_{k=0}^{K-1}\}$ is collected from the real environment. We stress again that the above calculation uses the agent's parameters ϕ and does not involve any parameters

of the true dynamics. We then fit a Gaussian process to the sample of the cumulative rewards $\{\sum_{\ell=k}^{K-1} R_{t_\ell}^{g,(m)}\}_{m=1}^M$ to estimate $g_e^\psi(t_k, S)$. Figure 2 illustrates the fit at three time points with the radial basis kernel. The Gaussian process is able to provide a good fit to the shape of $g_e(t, S)$ as a function of S for each given t .

In the actor step, we update the policy parameter ϕ_i for $i = 1, \dots, 5$ by

$$\begin{aligned} \phi_i \leftarrow \phi_i - \frac{\alpha_{\phi_i}}{M} \sum_{m=1}^M \sum_{k=0}^{K-1} & \left(\theta \log \pi^\phi(a_{t_k}^{\pi^\phi, (m)} | t_k, \hat{S}_{t_k}^{(m)}, X_{t_k}^{\pi^\phi, (m)}) \Delta t - \beta J^{\psi, \phi}(t_k, X_{t_k}^{\pi^\phi, (m)}) \Delta t \right. \\ & \left. + J^{\psi, \phi}(t_{k+1}, \hat{S}_{t_{k+1}}^{(m)}, X_{t_{k+1}}^{\pi^\phi, (m)}) - J^{\psi, \phi}(t_k, \hat{S}_{t_k}^{(m)}, X_{t_k}^{\pi^\phi, (m)}) \right) \\ & \cdot \partial_{\phi_i} \log \pi^\phi(a_{t_k}^{\pi^\phi, (m)} | t_k, \hat{S}_{t_k}^{(m)}, X_{t_k}^{\pi^\phi, (m)}). \end{aligned}$$

This update essentially follows from (4); the difference is that we do batch processing and update the parameters only after observing all the episodes in a batch. Furthermore, we flip the sign before α_{ϕ_i} and θ as we consider minimization here while the problem in the general discussion is for maximization.

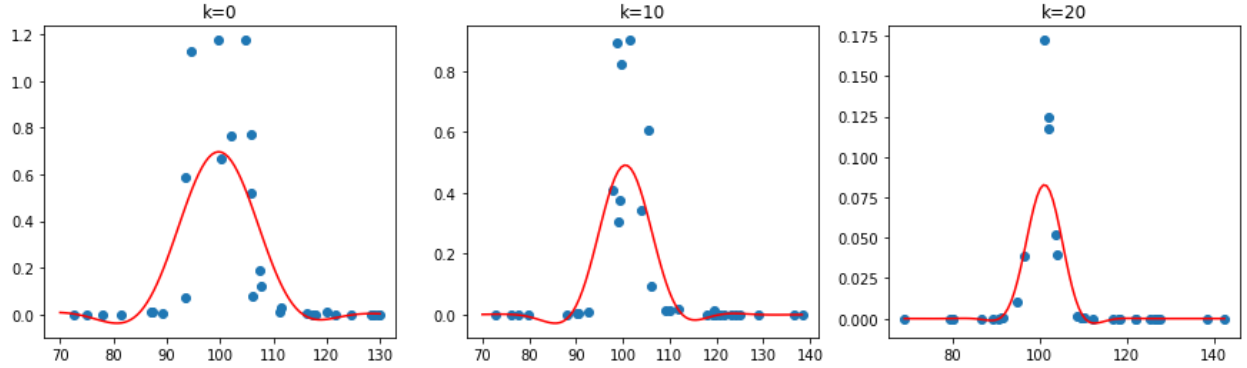


Figure 2: Fit of Gaussian process to the cumulative reward data for g_e at t_k ($k = 0, 10, 20$) for a one-month put option with strike $\mathcal{K} = 100$. The variable S is on the x -axis. We use the MLE estimates of the MJD model shown in Table 4 to calculate the rewards for g_e from the sampled index price paths.

6.3 Simulation study

We check the convergence behavior of our actor-critic algorithm in a simulation study. We assume the environment is described by an MJD model with its true parameters shown in the upper row of Table 4.⁵ We consider hedging a 4-month put option with strike price $\mathcal{K} = 100$ because short-term options with maturity of a few months are typically traded much more actively than longer-term options. We assume the risk-free rate is zero and implement daily rebalancing.

⁵These values are chosen close to the estimates obtained from the market data of S&P 500 in Table 1 to make the MJD model better resemble the reality in the simulation study.

We use 20 quadrature points for the G-H rule, which suffices for high accuracy. In each iteration, we sample $M = 32$ episodes from the environment simulated by the MJD model. The learning rates are chosen to be $\alpha_{\phi_1} = 5e-3$, $\alpha_{\phi_2} = 2e-5$, $\alpha_{\phi_3} = 15$, $\alpha_{\phi_4} = 1e-6$, $\alpha_{\phi_5} = 1e-6$, while the temperature parameter $\theta = 0.1$. We keep the learning rates constant in the first 30 iterations and then they decay at rate $l(j) = j^{-0.5}$, where j is the iteration index. Before our actor-critic learning, we sample a 15-year path from the MJD model as data for MLE and obtain estimates of the parameters shown in Table 4. MLE estimates all the parameters quite well except the drift μ , where the estimate is more than twice the true value.⁶ We then use these estimates as initial values of ϕ_1 to ϕ_5 for learning. Figure 3 clearly shows that our RL algorithm is able to converge near the true values. In particular, the estimate of μ eventually moves closely around the true level despite being distant to it initially.

Parameters	
True	$\mu = 0.06, \sigma = 0.13, \lambda = 28, m = -0.004, \delta = 0.03$
MLE	$\mu = 0.1289, \sigma = 0.1331, \lambda = 26.1091, m = -0.0033, \delta = 0.0317$

Table 4: Parameters of the true MJD model and MLE estimates from 15 years of data

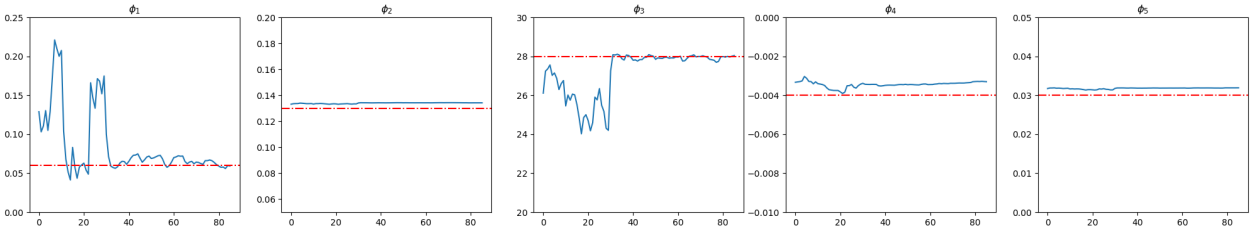


Figure 3: Training paths of five parameters (iteration index on the x -axis)

6.4 Empirical study

In the empirical study we consider hedging a European put option written on the S&P 500 index. This type of option is popular in the financial market as a tool for investors to protect their high beta investment. We assume the risk-free rate is zero, implement daily rebalancing and use the same S&P 500 data as in Section 5.4 with the first 20 years as the training period and the last 4 years as the test period.

We use the MLE estimates from the first 20 years of observations to initialize ϕ_1 to ϕ_5 for learning. To sample an episode from the environment for training our RL algorithm, we bootstrap index returns from the training period. Similarly, we bootstrap index returns from the test period to generate index price paths to test the learned policy. We use the 20-point G-H rule with $M = 32$ episodes in each training iteration. The learning rates are set as $\alpha_{\phi_1} = 5e-3$, $\alpha_{\phi_2} = 5e-4$, $\alpha_{\phi_3} = 50$, $\alpha_{\phi_4} = 5e-6$, $\alpha_{\phi_5} = 1e-5$ with the same learning rate decay schedule as in the simulation study. We set the temperature parameter $\theta = 0.1$.

⁶This is expected due to the well-known mean-blur problem, as 15 years of data are insufficient to obtain a reasonably accurate estimate of the mean using statistical methods.

We consider four maturities ranging from 1 to 4 months for the put option. For a standard S&P500 index option traded on CBOE, in practice its strike is quoted in points commensurate with the scale of the index level and converted to dollar amount by a multiplier of 100. That is, the dollar strike $\mathcal{K} = 100\mathcal{K}_p$, where \mathcal{K}_p is the strike in points.⁷ Let S_t be the index level at time t converted to dollars by the multiplier of 100. We sample S_0 uniformly from the range $[0.7\mathcal{K}, 1.3\mathcal{K}]$ to generate training and test episodes. This allows us to include scenarios where the option is deep out of money or in the money. We train our algorithm for 100 iterations and then test it on 5000 index price episodes. Finally, although we use a stochastic policy to interact with the environment for learning/updating the parameters during training, in the test stage we apply a deterministic policy which is the *mean* of the learned stochastic policy. This can reduce the variance of the final hedging error.

To measure the test performance, we consider the mean squared hedging error (denominated in squared dollars) divided by \mathcal{K}_p^2 . This effectively normalizes the contract’s scale; so we can just set $\mathcal{K}_p = 1$ in our implementation. The test result can be interpreted as the average mean squared hedging error in squared dollars for every point of strike price.

We compare the performances of the policy learned by our RL algorithm and the one obtained by plugging the MLE estimates from the training period data to (6.1). Both policies are tested on the same index price episodes. We check the statistical significance of the difference in the squared hedging errors of the two policies by applying the t-test. Table 5 reports the mean squared hedging error for each policy as well as the p-value of the t-test. Not surprisingly, the hedging error of each policy increases as the maturity becomes longer due to greater uncertainty in the problem. In every case, the RL policy achieves notable reduction in the hedging error compared with the MLE-based policy. The p-value further shows the outperformance of the former is statistically significant.

Maturity	RL policy	MLE policy	p-value
1 month	0.3513 (0.034)	0.4006 (0.043)	3.2e-7
2 months	0.5668 (0.029)	0.6796 (0.050)	1.2e-5
3 months	0.8283 (0.031)	1.0080 (0.060)	1.6e-5
4 months	0.8649 (0.047)	1.1544 (0.068)	4.2e-13

Table 5: Mean squared hedging errors of the two policies in the empirical test with their standard errors shown in parenthesis

7 Conclusions

Fluctuations in data or time series coming from nonlinear, complex dynamic systems are characterized by two types: slow changes and sudden jumps, the latter occurring much rarely than the former. Hence jump-diffusions capture the key *structural* characteristics of many data generating processes in areas such as physics, astrophysics, earth science, engineering, finance, and medicine. As a result, RL for jump-diffusions is important both theoretically and practically. This paper endeavors to lay a theoretical foundation for the study. A key insight from this research is that temporal-difference algorithms designed for diffusions can work seamlessly for jump-diffusions.

⁷For example, for an at-the-money option contract where the current index is 5000 points, we have $\mathcal{K}_p = 5000$. The contract’s notional size is 500000 dollars.

However, unless using general neural networks, policy parameterization does need to respond to the presence of jumps if one is to take advantage of any special structure of an underlying problem.

There are plenty of open questions starting from here, including convergence of the exploratory HJB equation as the temperature parameter decays to zero, convergence of the RL algorithms, regret bounds, decaying rates of the temperature parameters, and learning rates of the gradient decent and/or stochastic approximation procedures involved.

Data Availability Statement

Data for the S&P 500 index were obtained from Yahoo Finance.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Statement

Gao acknowledges support from the Hong Kong Research Grants Council [GRF 14201424, 14212522, 14200123]. Li is supported by the Hong Kong Research Grant Council [GRF 14216322, 14201024]. Zhou is supported by the Nie Center for Intelligent Asset Management at Columbia University. We are grateful to the two reviewers and Associate Editor for their detailed and insightful comments that helped us improve the paper significantly. We also thank Bo Wu for his assistance in the numerical study of the mean–variance portfolio selection application.

References

- Aït-Sahalia, Y., J. Jacod, and J. Li (2012). Testing for jumps in noisy high frequency data. *Journal of Econometrics* 168(2), 207–222.
- Andersen, T. G., L. Benzoni, and J. Lund (2002). An empirical investigation of continuous-time equity return models. *The Journal of Finance* 57(3), 1239–1284.
- Applebaum, D. (2009). *Lévy Processes and Stochastic Calculus*. Cambridge University Press.
- Bates, D. S. (1991). The crash of 87: was it expected? The evidence from options markets. *The Journal of Finance* 46(3), 1009–1044.
- Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *The Review of Financial Studies* 9(1), 69–107.
- Bender, C. and N. T. Thuan (2023). Entropy-regularized mean-variance portfolio optimization with jumps. *arXiv preprint arXiv:2312.13409*.
- Bender, C. and N. T. Thuan (2024). Continuous time reinforcement learning: A random measure approach. *Preprint*.
- Bo, L., Y. Huang, X. Yu, and T. Zhang (2024). Continuous-time q-learning for jump-diffusion models under Tsallis entropy. *arXiv preprint arXiv:2407.03888*.
- Cai, N. and S. G. Kou (2011). Option pricing under a mixed-exponential jump diffusion model. *Management Science* 57(11), 2067–2081.
- Cont, R. and P. Tankov (2004). *Financial Modelling with Jump Processes*. Chapman and Hall/CRC.
- Dai, M., Y. Dong, and Y. Jia (2023). Learning equilibrium mean–variance strategy. *Mathematical Finance* 33(4), 1166–1212.

- Das, S. R. (2002). The surprise element: Jumps in interest rates. *Journal of Econometrics* 106(1), 27–65.
- Denkert, R., H. Pham, and X. Warin (2024). Control randomisation approach for policy gradient and application to reinforcement learning in optimal switching. *arXiv preprint arXiv:2404.17939*.
- Fang, F. and C. W. Oosterlee (2009). A novel pricing method for European options based on Fourier-cosine series expansions. *SIAM Journal on Scientific Computing* 31(2), 826–848.
- Gammaitoni, L., P. Hänggi, P. Jung, and F. Marchesoni (1998). Stochastic resonance. *Reviews of Modern Physics* 70(1), 223–287.
- Gao, X., Z. Q. Xu, and X. Y. Zhou (2022). State-dependent temperature control for langevin diffusions. *SIAM Journal on Control and Optimization* 60(3), 1250–1268.
- Giraud, M. T. and L. Sacerdote (1997). Jump-diffusion processes as models for neuronal activity. *Biosystems* 40(1-2), 75–82.
- Goswami, B., N. Boers, A. Rheinwalt, N. Marwan, J. Heitzig, S. Breitenbach, and J. Kurths (2018). Abrupt transitions in time series with uncertainties. *Nature Communications* 9(1), 48–57.
- Guo, X., A. Hu, and Y. Zhang (2023). Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls. *SIAM Journal on Control and Optimization* 61(2), 755–787.
- Guo, X., R. Xu, and T. Zariphopoulou (2022). Entropy regularization for mean field games with learning. *Mathematics of Operations research* 47(4), 3239–3260.
- Heyde, C. C. and S. G. Kou (2004). On the controversy over tailweight of distributions. *Operations Research Letters* 32(5), 399–408.
- Huang, Y., Y. Jia, and X. Zhou (2022). Achieving mean–variance efficiency by continuous-time reinforcement learning. In *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 377–385.
- Jacod, J. and A. Shiryaev (2013). *Limit Theorems for Stochastic Processes*. Springer.
- Jia, Y., D. Ouyang, and Y. Zhang (2025). Accuracy of discretely sampled stochastic policies in continuous-time reinforcement learning. *arXiv preprint arXiv:2503.09981*.
- Jia, Y. and X. Y. Zhou (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research* 23(154), 1–55.
- Jia, Y. and X. Y. Zhou (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research* 23(154), 1–55.
- Jia, Y. and X. Y. Zhou (2023). q-learning in continuous time. *Journal of Machine Learning Research* 24(161), 1–61.
- Jia, Y. and X. Y. Zhou (2025). Erratum to q-learning in continuous time. *forthcoming in Journal of Machine Learning Research*; available at <https://www.columbia.edu/~xz2574/download/err.pdf>.
- Kou, S. G. (2002). A jump-diffusion model for option pricing. *Management Science* 48(8), 1086–1101.
- Kunita, H. (2004). Stochastic differential equations based on Lévy processes and stochastic flows of diffeomorphisms. In M. M. Rao (Ed.), *Real and Stochastic Analysis: New Perspectives*, pp. 305–373. Birkhuser.
- Kushner, H. J. (2000). Jump-diffusions with controlled jumps: Existence and numerical methods. *Journal of Mathematical Analysis and Applications* 249(1), 179–198.
- Li, L. and V. Linetsky (2014). Time-changed Ornstein–Uhlenbeck processes and their applications in commodity derivative models. *Mathematical Finance* 24(2), 289–330.
- Lim, A. E. (2005). Mean-variance hedging when there are jumps. *SIAM Journal on Control and Optimization* 44(5), 1893–1922.
- Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3(1-2), 125–144.
- Munos, R. (2006). Policy gradient in continuous time. *Journal of Machine Learning Research* 7, 771–791.
- Øksendal, B. K. and A. Sulem (2007). *Applied Stochastic Control of Jump Diffusions*, Volume 498. Springer.
- Park, S., J. Kim, and G. Kim (2021). Time discretization-invariant safe action repetition for policy gradient methods. *Advances in Neural Information Processing Systems* 34, 267–279.
- Reisinger, C. and Y. Zhang (2021). Regularity and stability of feedback relaxed controls. *SIAM Journal on Control and Optimization* 59(5), 3118–3151.
- Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

- Schweizer, M. (1996). Approximation pricing and the variance-optimal martingale measure. *The Annals of Probability* 24(1), 206–236.
- Schweizer, M. (2001). A guided tour through quadratic hedging approaches. In E. Jouini, J. Cvitanic, and M. Musiela (Eds.), *Handbooks in Mathematical Finance: Option Pricing, Interest Rates and Risk Management*, pp. 538–74. Cambridge University Press.
- Situ, R. (2006). *Theory of Stochastic Differential Equations with Jumps and Applications*. Springer.
- Szpruch, L., T. Treetanthiploet, and Y. Zhang (2024). Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning. *SIAM Journal on Control and Optimization* 62(1), 135–166.
- Talleg, C., L. Blier, and Y. Ollivier (2019). Making deep q-learning methods robust to time discretization. *arXiv preprint arXiv:1901.09732*.
- Tang, W., Y. P. Zhang, and X. Y. Zhou (2022). Exploratory HJB equations and their convergence. *SIAM Journal on Control and Optimization* 60(6), 3191–3216.
- Wang, B., X. Gao, and L. Li (2023). Reinforcement learning for continuous-time optimal execution: actor-critic algorithm and error analysis. *Available at SSRN 4378950*.
- Wang, B. and X. Zheng (2022). Testing for the presence of jump components in jump diffusion models. *Journal of Econometrics* 230(2), 483–509.
- Wang, H., T. Zariphopoulou, and X. Y. Zhou (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research* 21(198), 1–34.
- Wang, H. and X. Y. Zhou (2020). Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance* 30(4), 1273–1308.
- Williams, C. K. and C. E. Rasmussen (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Wu, B. and L. Li (2024). Reinforcement learning for continuous-time mean–variance portfolio selection in a regime-switching market. *Journal of Economic Dynamics and Control* 158, 104787.
- Zhou, X. Y. and D. Li (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization* 42(1), 19–33.
- Zhu, C., G. Yin, and N. A. Baran (2015). Feynman–Kac formulas for regime-switching jump diffusions and their applications. *Stochastics* 87(6), 1000–1032.

A Proofs

Proof of Lemma 2. We observe that for each $k = 1, \dots, \ell$,

$$\int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi^{t,x}}(u), z)|^p \nu_k(dz) du = \int_{\mathcal{A}} \int_{\mathbb{R}} |\gamma_k(t, x, a, z)|^p \nu_k(dz) \pi(a|t, x) da.$$

From Assumption 1-(iii), we have $\sum_{k=1}^{\ell} \int_{\mathbb{R}} |\gamma_k(t, x, a, z)|^p \nu_k(dz) \leq C_p(1 + |x|^p + |a|^p)$ for any $(t, x, a) \in [0, T] \times \mathbb{R}^d \times \mathcal{A}$. It follows from condition (iii) of admissibility (see Definition 1) that the desired result holds. \square

Proof of Lemma 3. We consider

$$\int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi^{t,x}}(u), z) - \gamma_k(t, x', G^{\pi^{t,x'}}(u), z)|^p \nu_k(dz) du,$$

which is bounded by

$$C_p \left(\int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi^{t,x}}(u), z) - \gamma_k(t, x, G^{\pi^{t,x'}}(u), z)|^p \nu_k(dz) du \right)$$

$$+ \int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi_{t,x'}}(u), z) - \gamma_k(t, x', G^{\pi_{t,x'}}(u), z)|^p \nu_k(dz) du)$$

for some constant $C_p > 0$.

For the first integral, using Assumption 2 we obtain

$$\begin{aligned} & \int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi_{t,x}}(u), z) - \gamma_k(t, x, G^{\pi_{t,x'}}(u), z)|^p \nu_k(dz) du \\ & \leq C_{K,p} \int_{[0,1]^n} |G^{\pi}(t, x, u) - G^{\pi}(t, x', u)|^p du \leq C'_{K,p} |x - x'|^p. \end{aligned}$$

For the second integral, we have

$$\begin{aligned} & \int_{\mathbb{R} \times [0,1]^n} |\gamma_k(t, x, G^{\pi_{t,x'}}(u), z) - \gamma_k(t, x', G^{\pi_{t,x'}}(u), z)|^p \nu_k(dz) du \\ & = \int_{\mathcal{A}} \int_{\mathbb{R}} |\gamma_k(t, x, a, z) - \gamma_k(t, x', a, z)|^p \nu_k(dz) \pi(a|t, x') da \\ & \leq \int_{\mathcal{A}} C''_{K,p} |x - x'|^p \pi(a|t, x') da \\ & = C''_{K,p} |x - x'|^p, \end{aligned}$$

where we used Assumption 1-(ii) and $C''_{K,p}$ is the constant there. The desired claim is obtained by combining these results. \square

Proof of Lemma 4. Fix $t \in [0, T)$ and suppose $\tilde{X}_t^\pi = x$. Let $a_s^\pi(u) = G^\pi(s, \tilde{X}_{s-}^\pi, u)$. Define a sequence of stopping times $\tau_n = \inf\{s \geq t : |\tilde{X}_s^\pi| \geq n\}$ for $n \in \mathbb{N}$. Applying Itô's formula (6) to the process $e^{-\beta s} \phi(s, \tilde{X}_s^\pi)$, where \tilde{X}_s^π follows the exploratory SDE (18), we obtain $\forall t' \in [t, T]$,

$$\begin{aligned} & e^{-\beta(t' \wedge \tau_n)} \phi(t' \wedge \tau_n, \tilde{X}_{t' \wedge \tau_n}^\pi) - e^{-\beta t} \phi(t, x) \\ & = \int_t^{t' \wedge \tau_n} e^{-\beta s} \left(\mathcal{L}^\pi \phi(s, \tilde{X}_{s-}^\pi) - \beta \phi(s, \tilde{X}_{s-}^\pi) \right) ds \\ & + \int_t^{t' \wedge \tau_n} e^{-\beta s} \partial_x \phi(s, \tilde{X}_{s-}^\pi) \circ \tilde{\sigma}(s, \tilde{X}_{s-}^\pi, \pi(\cdot | s, \tilde{X}_{s-}^\pi)) dW_s \end{aligned} \quad (70)$$

$$+ \int_t^{t' \wedge \tau_n} e^{-\beta s} \sum_{k=1}^{\ell} \int_{\mathbb{R} \times [0,1]^n} \left(\phi(s, \tilde{X}_{s-}^\pi + \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) - \phi(s, \tilde{X}_{s-}^\pi) \right) \tilde{N}'_k(ds, dz, du), \quad (71)$$

where \mathcal{L}^π is given in (17). We next show that the expectations of (A) and (A) are zero.

Note that for $s \in [t, t' \wedge \tau_n]$, $|\tilde{X}_{s-}^\pi| \leq n$. Thus, $\partial_x \phi(s, \tilde{X}_{s-}^\pi)$ is also bounded. Using the linear growth of $\tilde{\sigma}(s, \tilde{X}_{s-}^\pi, \pi(\cdot | s, \tilde{X}_{s-}^\pi))$ in Lemma 1 and the moment estimate (19), we can see that (A) is a square-integrable martingale and hence its expectation is zero.

Next, we analyze the following stochastic integral:

$$\int_t^{t' \wedge \tau_n} e^{-\beta s} \int_{\mathbb{R} \times [0,1]^n} \left(\phi(s, \tilde{X}_{s-}^\pi + \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) - \phi(s, \tilde{X}_{s-}^\pi) \right) \tilde{N}_k(ds, dz, du). \quad (72)$$

We consider the finite and infinite jump activity cases separately.

Case 1: $\int_{\mathbb{R}} \nu_k(dz) < \infty$. In this case, both of the processes

$$\int_t^{t' \wedge \tau_n} e^{-\beta s} \int_{\mathbb{R} \times [0,1]^n} \phi(s, \tilde{X}_{s-}^\pi + \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) \tilde{N}_k(ds, dz, du), \quad (73)$$

$$\int_t^{t' \wedge \tau_n} e^{-\beta s} \int_{\mathbb{R} \times [0,1]^n} \phi(s, \tilde{X}_{s-}^\pi) \tilde{N}_k(ds, dz, du), \quad (74)$$

are square-integrable martingales and hence have zero expectations. We prove this claim for (A) and analyzing (A) is entirely analogous.

Using the polynomial growth of ϕ , Lemma 2, and $|\tilde{X}_{s-}^\pi| \leq n$ for $s \in [t, t' \wedge \tau_n]$, we obtain

$$\begin{aligned} & \mathbb{E}_{t,x} \left[\int_t^{t' \wedge \tau_n} e^{-2\beta s} \int_{\mathbb{R} \times [0,1]^n} \left| \phi(s, \tilde{X}_{s-}^\pi + \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) \right|^2 \nu_k(dz) duds \right] \\ & \leq C_p \cdot \mathbb{E}_{t,x} \left[\int_t^{t' \wedge \tau_n} \int_{\mathbb{R} \times [0,1]^n} \left(1 + |\tilde{X}_{s-}^\pi|^{2p} + |\gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)|^{2p} \right) \nu_k(dz) duds \right] \\ & \leq C'_p \cdot \mathbb{E}_{t,x} \left[\int_t^{t' \wedge \tau_n} (1 + |\tilde{X}_{s-}^\pi|^{2p}) ds \right] < \infty. \end{aligned}$$

This implies the process (A) is a square-integrable martingale; see e.g. (Situ 2006, Section 1.9) for square-integrability of stochastic integrals with respect to compensated Poisson random measures.

Case 2: $\int_{\mathbb{R}} \nu_k(dz) = \infty$. Let $B_1 = \{|z| \leq 1\} \times [0, 1]^n$ and $B_1^c = \{|z| > 1\} \times [0, 1]^n$. The stochastic integral (A) can be written as the sum of two integrals:

$$\int_t^{t' \wedge \tau_n} e^{-\beta s} \int_{B_1} (\phi(s, \tilde{X}_{s-}^\pi + \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) - \phi(s, \tilde{X}_{s-}^\pi)) \tilde{N}_k(ds, dz, du) \quad (75)$$

$$+ \int_t^{t' \wedge \tau_n} e^{-\beta s} \int_{B_1^c} (\phi(s, \tilde{X}_{s-}^\pi + \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) - \phi(s, \tilde{X}_{s-}^\pi)) \tilde{N}_k(ds, dz, du). \quad (76)$$

Using the mean-value theorem, the stochastic integral (A) is equal to

$$\int_t^{t' \wedge \tau_n} e^{-\beta s} \int_{B_1} \partial_x \phi(s, \tilde{X}_{s-}^\pi + \alpha_s \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) \circ \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z) \tilde{N}_k(ds, dz, du)$$

for some $\alpha_s \in [0, 1]$. For $s \in [t, t' \wedge \tau_n]$, $|\tilde{X}_{s-}^\pi| \leq n$. By Assumption 1'-(v), $|\gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)|$ is bounded for any $|z| \leq 1$, $s \in [t, t' \wedge \tau_n]$ and $u \in [0, 1]^n$, which further implies the boundedness of $|\phi_x(s, \tilde{X}_{s-}^\pi + \alpha_s \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z))|$. Now, for each $j = 1, \dots, d$,

$$\int_t^{t' \wedge \tau_n} e^{-\beta s} \int_{B_1} \partial_{x_j} \phi(s, \tilde{X}_{s-}^\pi + \alpha_s \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)) \gamma_{jk}(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z) \tilde{N}_k(ds, dz, du)$$

is a square-integrable martingale because

$$\mathbb{E}_{t,x} \left[\int_t^{t' \wedge \tau_n} \int_{B_1} (\partial_{x_j} \phi(s, \tilde{X}_{s-}^\pi + \alpha_s \gamma_k(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z)))^2 \gamma_{jk}^2(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z) \nu_k(dz) duds \right]$$

$$\begin{aligned}
&\leq C \cdot \mathbb{E}_{t,x} \left[\int_t^{t' \wedge \tau_n} \int_{B_1} \gamma_{jk}^2(s, \tilde{X}_{s-}^\pi, a_s^\pi(u), z) \nu_k(dz) duds \right] \\
&\leq C' \cdot \mathbb{E}_{t,x} \left[\int_t^{t' \wedge \tau_n} (1 + |\tilde{X}_{s-}^\pi|^2) ds \right] < \infty,
\end{aligned}$$

where we used Lemma 2 and boundedness of $|\tilde{X}_{s-}^\pi|$ in the above. Thus, (A) is a square-integrable martingale with mean zero.

For (A), we can use the same arguments as in the finite activity case by observing $\int_{|z|>1} \nu_k(dz) < \infty$ to show that each of the two processes in (A) is a square-integrable martingale with mean zero.

Combining the results above, setting $t' = T$, and taking expectation, we obtain

$$\mathbb{E}_{t,x}[e^{-\beta(T \wedge \tau_n)} \phi(\tilde{X}_{T \wedge \tau_n}^\pi)] - e^{-\beta t} \phi(t, x) = \mathbb{E}_{t,x}^{\mathbb{P}} \left[\int_t^{T \wedge \tau_n} e^{-\beta s} (\mathcal{L}^\pi \phi(s, \tilde{X}_{s-}^\pi) - \beta \phi(s, \tilde{X}_{s-}^\pi)) ds \right].$$

As $\phi(s, x)$ satisfies Eq. (4), it follows from (9) that

$$\begin{aligned}
\phi(t, x) = \mathbb{E}_{t,x} \left[\int_t^{T \wedge \tau_n} e^{-\beta(s-t)} \int_{\mathcal{A}} (r(s, \tilde{X}_{s-}^\pi, a) - \theta \log \pi(a|s, \tilde{X}_{s-}^\pi)) \pi(a|s, \tilde{X}_{s-}^\pi) da ds \right. \\
\left. + e^{-\beta(T \wedge \tau_n - t)} \phi(\tilde{X}_{T \wedge \tau_n}^\pi) \right]. \tag{77}
\end{aligned}$$

It follows from Assumption 1-(iv), Definition 1-(iii), and Eq. (4) that the term on the right-hand side of (A) is dominated by $C(1 + \sup_{t \leq s \leq T} |\tilde{X}_s^\pi|^p)$ for some $p \geq 2$, which has finite expectation from the moment estimate (19). Therefore, sending n to infinity in (A) and applying the dominated convergence theorem, we obtain $\phi(t, x) = J(t, x, \pi)$. \square

Proof of Lemma 5. We first maximize the integral in (21). Applying (Jia and Zhou 2023, Lemma 13), we deduce that π^* given by (22) is the unique maximizer. Next, we show that $\psi(t, x)$ is the optimal value function.

On one hand, given any admissible stochastic policy $\pi \in \Pi$, from (21) we have

$$\partial_t \psi(t, x) + \int_{\mathcal{A}} (H(t, x, a, \partial_x \psi, \partial_x^2 \psi, \psi) - \theta \log \pi(a|t, x)) \pi(a|t, x) da - \beta \psi(t, x) \leq 0.$$

Using similar arguments as in the proof of Lemma 4, we obtain $J(t, x, \pi) \leq \psi(t, x)$ for any $\pi \in \Pi$. Thus, $J^*(t, x) \leq \psi(t, x)$.

On the other hand, Eq. (21) becomes

$$\partial_t \psi(t, x) + \int_{\mathcal{A}} (H(t, x, a, \partial_x \psi, \partial_x^2 \psi, \psi) - \theta \log \pi^*(a)) \pi^*(a) da - \beta \psi(t, x) = 0,$$

with $\psi(T, x) = h(x)$. By Lemma 4, we obtain that $J(t, x, \pi^*) = \psi(t, x)$. It follows that $J^*(t, x) \geq \psi(t, x)$.

Combining these results, we conclude that $J^*(t, x) = \psi(t, x)$ and π^* is the optimal stochastic policy. \square

Proof of Proposition 2. In this proof, C_p is a generic positive constant that depends on p , which may take different values each time it occurs. Consider the grid $\mathbb{S} = \{t = t_0 < t_1 < \dots < t_N = T\}$. We prove the well-posedness by induction. Suppose for $i = 0, \dots, n$, the SDE (11) admits a unique strong solution on $[t_0, t_i]$ satisfying $\mathbb{E}_{t,x}[\sup_{t \leq s \leq t_i} |X_s^{\pi, \mathbb{S}}|^p] < \infty$ for any $p \geq 2$. Now consider the equation on $(t_n, t_{n+1}]$. Note that the random action $a_{t_n}^{\pi, \mathbb{S}} := G^\pi(t_n, X_{t_n}^{\pi, \mathbb{S}}, U_{n+1})$ is applied throughout $(t_n, t_{n+1}]$. For any $p \geq 2$, we have

$$\begin{aligned} \mathbb{E}_{t,x}[|a_{t_n}^{\pi, \mathbb{S}}|^p] &= \mathbb{E}_{t,x}[\mathbb{E}_{t,x}[|G^\pi(t_n, X_{t_n}^{\pi, \mathbb{S}}, U_{n+1})|^p | X_{t_n}^{\pi, \mathbb{S}}]] \\ &= \mathbb{E}_{t,x}\left[\int_{\mathcal{A}} |a|^p \boldsymbol{\pi}(a | t_n, X_{t_n}^{\pi, \mathbb{S}}) da\right] \\ &\leq C_p(1 + \mathbb{E}_{t,x}[|X_{t_n}^{\pi, \mathbb{S}}|^p]), \quad (\text{by Definition 1-(iii)}) \end{aligned} \quad (78)$$

which is finite. This result together with Assumption 1 is sufficient for us to use the arguments in Kunita (2004) and Situ (2006) to show the existence and uniqueness of the strong solution on $(t_n, t_{n+1}]$ that satisfies $\mathbb{E}_{t,x}[\sup_{t_n \leq s \leq t_{n+1}} |X_s^{\pi, \mathbb{S}}|^p] < \infty$ for any $p \geq 2$.

Recall the action process $a^{\pi, \mathbb{S}}$ satisfies $a_\tau^{\pi, \mathbb{S}} = a_{t_n}^{\pi, \mathbb{S}}$ for $\tau \in (t_n, t_{n+1}]$. The SDE can be written as

$$X_s^{\pi, \mathbb{S}} = x + \int_t^s b(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}) d\tau + \int_t^s \sigma(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}) dW_\tau + \int_t^s \int_{\mathbb{R}^\ell} \gamma(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}, z) \tilde{N}(d\tau, dz).$$

Using (Kunita 2004, Theorem 2.11), for any $p \geq 2$ we obtain

$$\begin{aligned} &\mathbb{E}_{t,x}\left[\sup_{t \leq \tau \leq s} |X_\tau^{\pi, \mathbb{S}}|^p\right] \\ &\leq C_p\left(|x|^p + \mathbb{E}\left[\left(\int_t^s |b(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})| d\tau\right)^p\right] + \mathbb{E}\left[\left(\int_t^s |\sigma(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})|^2 d\tau\right)^{p/2}\right]\right. \\ &\quad \left.+ \mathbb{E}\left[\left(\int_t^s \int_{\mathbb{R}^\ell} |\gamma(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}, z)|^2 \nu(dz) d\tau\right)^{p/2}\right] + \mathbb{E}\left[\int_t^s \int_{\mathbb{R}^\ell} |\gamma(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}, z)|^p \nu(dz) d\tau\right]\right). \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}\left[\left(\int_t^s |b(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})| d\tau\right)^p\right] &\leq C_p \mathbb{E}\left[\int_t^s |b(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})|^p d\tau\right] \\ &\leq C_p \int_t^s \mathbb{E}\left[|b(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}})|^p\right] d\tau \\ &\leq C_p \int_t^s (1 + \mathbb{E}[|X_{\tau-}^{\pi, \mathbb{S}}|^p] + \mathbb{E}[|X_{\delta(\tau)}^{\pi, \mathbb{S}}|^p]) d\tau \\ &\leq C_p \int_t^s \left(1 + 2\mathbb{E}\left[\sup_{t \leq u \leq \tau} |X_u^{\pi, \mathbb{S}}|^p\right]\right) d\tau, \end{aligned}$$

where $\delta(\tau) := t_n$ for $\tau \in (t_n, t_{n+1}]$, and we have used the linear growth property in Assumption 1-(iii) and (A). We can also obtain similar estimates for the other three terms. Together, we have

$$\mathbb{E}_{t,x}\left[\sup_{t \leq \tau \leq s} |X_\tau^{\pi, \mathbb{S}}|^p\right] \leq C_p(1 + |x|^p) + C_p \int_t^s \mathbb{E}\left[\sup_{t \leq u \leq \tau} |X_u^{\pi, \mathbb{S}}|^p\right] d\tau.$$

The estimate (2) then follows from Gronwall's inequality. \square

Proof of Theorem 2. In this proof, C is a generic constant whose values may vary from line to line. Consider part (i). Applying Itô's formula to the value function (20) of policy π over the grid sample state process (11) and using the definition of q-function, we obtain that for $0 \leq t < s \leq T$,

$$\begin{aligned} & e^{-\beta s} J(s, X_s^{\pi, \mathbb{S}}; \pi) - e^{-\beta t} J(t, x; \pi) + \int_t^s e^{-\beta \tau} (r(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}) - \hat{q}(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}})) d\tau \\ &= \int_t^s e^{-\beta \tau} (q(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}; \pi) - \hat{q}(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}})) d\tau + \int_t^s e^{-\beta \tau} \partial_x J(\tau, X_{\tau-}^{\pi, \mathbb{S}}; \pi) \circ \sigma(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}) dW_{\tau} \\ & \quad + \sum_{k=1}^{\ell} \int_t^s e^{-\beta \tau} \int_{\mathbb{R}} (J(\tau, X_{\tau-}^{\pi, \mathbb{S}} + \gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)) - J(\tau, X_{\tau-}^{\pi, \mathbb{S}}; \pi)) \tilde{N}_k(d\tau, dz). \end{aligned} \quad (79)$$

Suppose $\hat{q}(t, x, a) = q(t, x, a; \pi)$ for all (t, x, a) . Then the first term on the right-hand side of (A) is zero. We verify the following two conditions:

$$\mathbb{E}_{t,x} \left[\int_t^T e^{-2\beta \tau} |J_x(\tau, X_{\tau-}^{\pi, \mathbb{S}}; \pi) \circ \sigma(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}})|^2 d\tau \right] < \infty, \quad (80)$$

$$\mathbb{E}_{t,x} \left[\int_t^T e^{-2\beta \tau} \int_{\mathbb{R}} |J(\tau, X_{\tau-}^{\pi, \mathbb{S}} + \gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z); \pi) - J(\tau, X_{\tau-}^{\pi, \mathbb{S}}; \pi)|^2 \nu_k(dz) d\tau \right] < \infty. \quad (81)$$

Equality (A) follows from Assumption 1-(iii), the polynomial growth of $\partial_x J$ in x , and the moment estimate (2). For (A), we apply the mean-value theorem to turn the integral to

$$\int_t^T e^{-2\beta \tau} \int_{\mathbb{R}} |\partial_x J(\tau, X_{\tau-}^{\pi, \mathbb{S}} + \alpha_{\tau} \gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z); \pi) \circ \gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^2 \nu_k(dz) d\tau$$

for some $\alpha_{\tau} \in [0, 1]$. Using the polynomial growth of $\partial_x J$ in x , we can bound the integral by

$$\begin{aligned} & \int_t^T e^{-2\beta \tau} \int_{\mathbb{R}} |\partial_x J(\tau, X_{\tau-}^{\pi, \mathbb{S}} + \alpha_{\tau} \gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z); \pi)|^2 \cdot |\gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^2 \nu_k(dz) d\tau \\ & \leq C \int_t^T e^{-2\beta \tau} \int_{\mathbb{R}} (1 + |X_{\tau-}^{\pi, \mathbb{S}} + \alpha_{\tau} \gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^p)^2 \cdot |\gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^2 \nu_k(dz) d\tau \\ & \leq C \int_t^T e^{-2\beta \tau} \int_{\mathbb{R}} (1 + |X_{\tau-}^{\pi, \mathbb{S}}|^p + |\gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^p)^2 \cdot |\gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^2 \nu_k(dz) d\tau \\ & \leq C \int_t^T (1 + |X_{\tau-}^{\pi, \mathbb{S}}|^p)^2 \int_{\mathbb{R}} |\gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^2 \nu_k(dz) d\tau \\ & \quad + C \int_t^T \int_{\mathbb{R}} |\gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^{2p+2} \nu_k(dz) d\tau \\ & \quad + 2C \int_t^T (1 + |X_{\tau-}^{\pi, \mathbb{S}}|^p) \int_{\mathbb{R}} |\gamma_k(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}, z)|^{p+2} \nu_k(dz) d\tau. \end{aligned}$$

Using Assumption 1-(iii) and the moment estimate (2), we obtain (A). It follows that the second and third processes on the right-hand side of (A) are $\mathcal{G}^{\mathbb{S}}$ -martingales. Thus, we have the martingale property of the process given by (2).

Conversely, suppose (2) is a $\mathcal{G}^{\mathbb{S}}$ -martingale for any (t, x) and any grid \mathbb{S} of $[t, T]$. We can infer from (A) that the process

$$\int_t^s e^{-\beta \tau} (q(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}; \pi) - \hat{q}(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}})) d\tau$$

is also a $\mathcal{G}^{\mathbb{S}}$ -martingale. Furthermore, it has continuous sample paths and finite variation and thus is zero almost surely. That is, for any $t < s$,

$$\int_t^s e^{-\beta\tau} f(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}) d\tau = 0, \quad \text{almost surely,} \quad (82)$$

where $f(t, x, a) := q(t, x, a; \pi) - \hat{q}(t, x, a)$, and it is a continuous function by our assumption. We need to show that $f(t, x, a) = 0$ for every (t, x, a) , and we prove this by contradiction. Suppose that there exists some time-state action (t^*, x^*, a^*) such that $f(t^*, x^*, a^*)$ is nonzero. Without loss of generality, we suppose $f(t^*, x^*, a^*) > \epsilon$ for some $\epsilon > 0$. Then there exists some $\delta^* > 0$ such that $f(t, x, a) > \epsilon/2$ for all (u, x', a') with $|u - t^*| \vee |x' - x^*| \vee |a' - a^*| \leq \delta^*$.

Fix $\delta \in (0, \delta^*)$ and choose a grid \mathbb{S} of $[t^*, T]$ such that $[t^*, t^* + \delta]$ is within the first subinterval of the grid. Consider the grid sample state process $X^{\pi, \mathbb{S}}$ starting from (t^*, x^*) , with $X_{t^*}^{\pi, \mathbb{S}} = x^*$. By the definition of the action process, the action at time $s \in [t^*, t^* + \delta]$ is given by $a_s^{\pi, \mathbb{S}} = a_{t^*}^{\pi, \mathbb{S}} = G^{\pi}(t^*, x^*, U)$, where U is a uniform random vector on $[0, 1]^n$. Define

$$T_{\delta, \mathbb{S}} = \inf\{t' \geq t^* : |X_{t'}^{\pi, \mathbb{S}} - x^*| > \delta\} \wedge (t^* + \delta).$$

We can see that $T_{\delta, \mathbb{S}} > t^*$, \mathbb{P} -almost surely. That is, $T_{\delta, \mathbb{S}}(\omega) > t^*$ for all $\omega \in \Omega \setminus \Omega_0$ where $\mathbb{P}(\Omega_0) = 0$. This is because the Lévy processes that drive the controlled grid sample state $X^{\pi, \mathbb{S}}$ are stochastic continuous, i.e., the probability of having a jump at the fixed time t^* is zero. Now for $\omega \in \Omega \setminus \Omega_0$ and $s \in [t^*, T_{\delta, \mathbb{S}}(\omega)] \subset [t^*, t^* + \delta]$, the action is given by $a_s^{\pi, \mathbb{S}}(\omega) = a_{t^*}^{\pi, \mathbb{S}}(\omega) = G^{\pi}(t^*, x^*, U(\omega))$. By Definition 1-(i), if we let $\mathcal{Z} = \{\omega \in \Omega : |a_{t^*}^{\pi, \mathbb{S}}(\omega) - a^*| \leq \delta\}$, then $\mathbb{P}(\mathcal{Z}) > 0$. This implies that for $\omega \in \mathcal{Z} \setminus \Omega_0$ and $\tau \in [t^*, T_{\delta, \mathbb{S}}(\omega)]$, we have $f(\tau, X_{\tau-}^{\pi, \mathbb{S}}(\omega), a_{\tau}^{\pi, \mathbb{S}}(\omega)) > \epsilon/2$. This leads to a contradiction to (A).

Part (ii) directly follows from applying Itô's formula to $J(s, X_s^{\pi', \mathbb{S}}; \pi)$ and together with a similar argument in proving part (i). For part (iii), again a similar argument in proving part (i) deduces that the process

$$\int_t^s e^{-\beta\tau} (q(\tau, X_{\tau-}^{\pi', \mathbb{S}}, a_{\tau}^{\pi', \mathbb{S}}; \pi) - \hat{q}(\tau, X_{\tau-}^{\pi', \mathbb{S}}, a_{\tau}^{\pi', \mathbb{S}})) d\tau$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale. Then the same argument in part (i) applies to obtain the desired conclusion. \square

Proof of Theorem 3. (i) The “if” part directly follows from the argument in the proof of Theorem 2. In the following, we prove the “only if” part.

Assume for any $(t, x) \in [0, T] \times \mathbb{R}^d$ and any grid \mathbb{S} of $[t, T]$, the following process

$$e^{-\beta s} \hat{J}(s, X_s^{\pi, \mathbb{S}}) + \int_t^s e^{-\beta\tau} (r(\tau, X_{\tau}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}) - \hat{q}(\tau, X_{\tau}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}})) d\tau$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale. Applying Itô's formula to \hat{J} over the grid sample state process $(X_s^{\pi, \mathbb{S}})$ similarly as in the proof of part (i) of Theorem 2, and letting $g(t, x, a) := \mathcal{L}^a \hat{J}(t, x) - \beta \hat{J}(t, x)$ where the operator \mathcal{L}^a is given in (5), we obtain

$$e^{-\beta s} \hat{J}(s, X_s^{\pi, \mathbb{S}}) - \int_t^s e^{-\beta\tau} g(\tau, X_{\tau}^{\pi, \mathbb{S}}, a_{\tau}^{\pi, \mathbb{S}}) d\tau$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale. Therefore,

$$\int_t^s e^{-\beta\tau} (r - \hat{q} + g)(\tau, X_\tau^{\pi, \mathbb{S}}, a_\tau^{\pi, \mathbb{S}}) d\tau$$

is a $\mathcal{G}^{\mathbb{S}}$ -martingale, which further implies that $(r - \hat{q} + g)(t, x, a) = 0$ for all (t, x, a) by a similar argument as in the proof of Theorem 2. Hence we have

$$\hat{q}(t, x, a) = r(t, x, a) + \mathcal{L}^a \hat{J}(t, x) - \beta \hat{J}(t, x) = \partial_t \hat{J}(t, x) + H(t, x, a, \partial_x \hat{J}, \partial_x^2 \hat{J}, \hat{J}) - \beta \hat{J}(t, x).$$

The constraint for \hat{q} in (3) then implies that

$$\int_{\mathcal{A}} \left\{ \partial_t \hat{J}(t, x) + H(t, x, a, \partial_x \hat{J}, \partial_x^2 \hat{J}, \hat{J}) - \beta \hat{J}(t, x) - \theta \log \pi(a|t, x) \right\} \pi(a|t, x) da = 0,$$

for all $(t, x) \in [0, T] \times \mathbb{R}^d$. Because $\hat{J}(T, x) = h(x)$, by Lemma 4, we deduce that $\hat{J}(t, x) = J(t, x; \pi)$ for all (t, x) . Then it follows from Theorem 2-(iii) that $\hat{q}(t, x, a) = q(t, x, a; \pi)$ for all (t, x, a) .

Part (ii) follows directly from part (ii) of Theorem 2. Part (iii) follows from a similar argument as in the proof of part (i), by applying Itô's formula to \hat{J} over the grid sample state process $(X_s^{\pi', \mathbb{S}})$ under π' . Finally, if $\pi(a|t, x) = \frac{\exp(\frac{1}{\theta} \hat{q}(t, x, a))}{\int_{\mathcal{A}} \exp(\frac{1}{\theta} \hat{q}(t, x, a)) da}$, then it follows from Theorem 1 that π is the optimal policy and \hat{J} is the optimal value function. \square

Proof of Theorem 4. Part (i) directly follows from Part (ii) of Theorem 3. The proof of part (ii) follows the same argument as in the proof of part (ii) of Theorem 9 in Jia and Zhou (2025). \square

Proof of Lemma 6. (1) Consider the function $h'(t, S)$ defined by the RHS of (6.1). Proposition 12.1 in Cont and Tankov (2004) shows that $h' \in C^{1,2}([0, T] \times \mathbb{R}_+) \cap C([0, T] \times \mathbb{R}_+)$ and it satisfies the PIDE (6.1). Furthermore, $h'(t, S)$ is Lipschitz continuous in S . The Lipschitz continuity of function \hat{G} also implies $|\hat{G}(\hat{S}_T)| \leq C(1 + \hat{S}_T)$ for some constant $C > 0$. It follows that

$$|h'(t, S)| \leq C(1 + \mathbb{E}^{\mathbb{Q}}[\hat{S}_T | \hat{S}_t = S]) \leq C(1 + S),$$

where we used the martingale property of \hat{S} under \mathbb{Q} . The Feynman-Kac Theorem (see Zhu et al. (2015)) implies uniqueness of classical solutions satisfying the linear growth condition.

(2) To study the PIDE (6.1), we consider the function $g'(t, S)$ defined by the RHS of (6.1). Under the assumption of Lemma 6, the arguments of Proposition 12.1 in Cont and Tankov (2004) can be used to show that $g' \in C^{1,2}([0, T] \times \mathbb{R}_+) \cap C([0, T] \times \mathbb{R}_+)$ and it satisfies the PIDE (6.1). The Lipschitz continuity of $h(t, S)$ in S implies that $\partial_S h(t, S)$ is bounded and $|h(t, S \exp(z)) - h(t, S)| \leq C(\exp(z) - 1)$. We also have $\mathbb{E}[\sup_{t \leq s \leq T} \hat{S}_s^2 | \hat{S}_t = S] \leq C(1 + S^2)$ (Kunita 2004, Theorem 3.2). Combining these results, we see that $g'(t, S)$ shows quadratic growth in S . Again the Feynman-Kac Theorem implies uniqueness of classical solutions satisfying the quadratic growth condition. \square

B Convergence of value functions with grid sample state processes

In this section, we show that, under appropriate conditions, the value functions with grid sample state processes converge to the value function of the exploratory state process, and the convergence rate is of the same order in the mesh size of the time grid.

Consider the time grid $\mathbb{S} = \{0 = t_0 < t_1 < \dots < t_N = T\}$ with its mesh size $|\mathbb{S}| := \max_{1 \leq n \leq N} |t_n - t_{n-1}|$. Define

$$g(t, x, a) := r(t, x, a) - \theta \log \pi(a|t, x), \quad \tilde{g}(t, x) := \int_{\mathcal{A}} g(t, x, a) \pi(a|t, x) da.$$

At $t = 0$, the value function associated with the grid sample state process following SDE (11) is

$$J^{\pi, \mathbb{S}}(0, x) := \mathbb{E}_{0, x} \left[\int_0^T e^{-\beta t} g(t, X_{t-}^{\pi, \mathbb{S}}, a_t^{\pi, \mathbb{S}}) dt + e^{-\beta T} h(X_T^{\pi, \mathbb{S}}) \right].$$

For the exploratory state process following SDE (18), the corresponding (initial) value function is

$$J^{\pi}(0, x) := \mathbb{E}_{0, x} \left[\int_0^T e^{-\beta t} \tilde{g}(t, \tilde{X}_{t-}^{\pi}) dt + e^{-\beta T} h(\tilde{X}_T^{\pi}) \right].$$

In the above, $\mathbb{E}_{0, x}$ denotes taking expectation given the initial state is x .

For further analysis, we introduce some terminology and function spaces. When we say a function $f(t, x)$ shows polynomial growth in x , we mean $\sup_{0 \leq t \leq T} |f(t, x)| \leq C(1 + |x|^p)$ for some $p \geq 2$. Additionally, a function $f(t, x, a)$ has polynomial growth in x and a if $\sup_{0 \leq t \leq T} |f(t, x, a)| \leq C(1 + |x|^p + |a|^p)$ for some $p \geq 2$. Let $C_p^k([0, T] \times \mathbb{R}^d)$ be the space of functions $f : [0, T] \times \mathbb{R}^d \mapsto \mathbb{R}$ such that for nonnegative integer i and multi-index j with $2i + |j| \leq k$, $\partial_t^i \partial_x^j f(t, x)$ is continuous and has polynomial growth in x . Let $C_p^{k, 0}([0, T] \times \mathbb{R}^d \times \mathcal{A})$ be the space of functions $f : [0, T] \times \mathbb{R}^d \times \mathcal{A} \mapsto \mathbb{R}$ such that for nonnegative integer i and multi-index j with $2i + |j| \leq k$, $\partial_t^i \partial_x^j f(t, x, a)$ is continuous and shows polynomial growth in x and a . We impose the following assumption.

Assumption 3. *We assume*

(i) $b, \sigma^2 \in C_p^{2, 0}([0, T] \times \mathbb{R}^d \times \mathcal{A});$

(ii) *for any $t \in (0, T]$, the PIDE*

$$\begin{aligned} \mathcal{L}^{\pi} \phi(s, x) &= 0, \quad (s, x) \in [0, t) \times \mathbb{R}^d, \\ \phi(t, x) &= f(x), \quad x \in \mathbb{R}^d, \end{aligned} \tag{83}$$

has a unique solution $\phi \in C_p^4([0, t] \times \mathbb{R}^d)$ for $f = h, \tilde{g}(t, \cdot)$, where \mathcal{L}^{π} is given by (17);

(iii) *for $k = 1, \dots, \ell$ and $p \in \{0, 1\} \cup [2, \infty)$, $\int_{\mathbb{R}} \gamma_k(t, x, a, z)^p \varphi(t, x, a, z) \nu_k(dz)$ has polynomial growth in x and a for $\varphi = \partial_t \gamma_k, \partial_x \gamma_k, (\partial_x \gamma_k)^2, \partial_x^2 \gamma_k$;*

(iv) $g \in C_p^{2, 0}([0, T] \times \mathbb{R}^d \times \mathcal{A})$ and $\tilde{g} \in C_p^2([0, T] \times \mathbb{R}^d)$.

In the two applications considered in Sections 5 and 6, conditions (i) and (iii) clearly hold and condition (iv) can be verified for the Gaussian policies studied. Condition (ii) can be easily verified in the MV portfolio selection problem for the Gaussian policies due to the LQ structure of the problem (the solution is quadratic in x and has smooth dependence on t). It is more difficult to check this condition in the MV hedging problem, but for the optimal stochastic policy given by (6.1) it should be possible to verify the required polynomial growth of the derivatives for calls and

puts based on the representations (6.1) and (6.1) under additional assumptions on the derivatives of the transition density of the Lévy process that models the log stock price.

The following result shows the convergence of the value functions with grid sample state processes to that of the exploratory SDE, which extends Theorem 4.1 and Proposition 5.1 of Jia et al. (2025) from diffusions to jump-diffusions.

Theorem 5. *Suppose Assumptions 1–3 hold. For any admissible stochastic policy π , we have*

$$|J^{\pi, \mathbb{S}}(0, x) - J^{\pi}(0, x)| \leq C|\mathbb{S}|$$

for some positive constant C which may depend on x .

We need the following lemma for proving the theorem.

Lemma 7. *For $\phi \in C_p^4([0, t] \times \mathbb{R}^d)$ and $k = 1, \dots, \ell$, let*

$$I_k^\phi(\tau, x, a) := \int_{\mathbb{R}} (\phi(\tau, x + \gamma_k(\tau, x, a, z)) - \phi(\tau, x) - \gamma_k(\tau, x, a, z) \circ \partial_x \phi(\tau, x)) \nu_k(dz), \quad \tau \in [0, t]. \quad (84)$$

Under Assumption 1-(iii) and Assumption 3-(iii), $\partial_\tau I_k^\phi, \partial_x I_k^\phi, \partial_x^2 I_k^\phi$ have polynomial growth in x and a .

Proof. Without loss of generality we only consider $d = 1$ for notational simplicity. Consider

$$\begin{aligned} \psi(\tau, x, a) &:= \int_{\mathbb{R}} (\partial_\tau \phi(\tau, x + \gamma_k(\tau, x, a, z)) + \partial_x \phi(\tau, x + \gamma_k(\tau, x, a, z)) \partial_\tau \gamma_k(\tau, x, a, z) \\ &\quad - \partial_\tau \phi(\tau, x) - \partial_\tau \gamma_k(\tau, x, a, z) \partial_x \phi(\tau, x) - \gamma_k(\tau, x, a, z) \partial_\tau \partial_x \phi(\tau, x)) \nu_k(dz) \\ &= \int_{\mathbb{R}} (\partial_\tau \partial_x \phi(\tau, x + \alpha \gamma_k(\tau, x, a, z)) \gamma_k(\tau, x, a, z) - \partial_\tau \partial_x \phi(\tau, x) \gamma_k(\tau, x, a, z) \\ &\quad + \partial_x^2 \phi(\tau, x + \alpha' \gamma_k(\tau, x, a, z)) \gamma_k(\tau, x, a, z) \partial_\tau \gamma_k(\tau, x, a, z)) \nu_k(dz) \end{aligned}$$

for some $\alpha, \alpha' \in (0, 1)$ by the mean-value theorem. Using $\phi \in C_p^4([0, t] \times \mathbb{R}^d)$, Assumption 1-(iii) and Assumption 3-(iii), we obtain that $\psi(\tau, x, a)$ has polynomial growth in x and a (and hence has finite values). The interchangeability of the differentiation and integration thus follows leading to $\partial_\tau I_k^\phi = \psi(\tau, x, a)$.

Using the same argument, we obtain

$$\begin{aligned} \partial_x I_k^\phi(\tau, x, a) &= \int_{\mathbb{R}} (\partial_x \phi(\tau, x + \gamma_k(\tau, x, a, z)) (1 + \partial_x \gamma_k(\tau, x, a, z)) - \partial_x \phi(\tau, x) \\ &\quad - \partial_x \gamma_k(\tau, x, a, z) \partial_x \phi(\tau, x) - \gamma_k(\tau, x, a, z) \partial_x^2 \phi(\tau, x)) \nu_k(dz), \\ \partial_x^2 I_k^\phi(\tau, x, a) &= \int_{\mathbb{R}} (\partial_x^2 \phi(\tau, x + \gamma_k(\tau, x, a, z)) (1 + \partial_x \gamma_k(\tau, x, a, z))^2 - \partial_x^2 \phi(\tau, x) \\ &\quad + \partial_x \phi(\tau, x + \gamma_k(\tau, x, a, z)) \partial_x^2 \gamma_k(\tau, x, a, z) - \partial_x \phi(\tau, x) \partial_x^2 \gamma_k(\tau, x, a, z) \\ &\quad - 2 \partial_x \gamma_k(\tau, x, a, z) \partial_x^2 \phi(\tau, x) - \gamma_k(\tau, x, a, z) \partial_x^3 \phi(\tau, x)) \nu_k(dz), \end{aligned}$$

which show polynomial growth in x and a . □

Proof of Theorem 5. In this proof, C is a generic constant independent of n , whose values may vary from line to line. We consider $\beta = 0$ for simplicity, but the proof can be easily adapted to accommodate discounting.

(1) Suppose $t = t_i \in \mathbb{S}$. We first show that for $f = h, \tilde{g}(t, \cdot)$,

$$|\mathbb{E}_{0,x}[f(X_t^{\pi,\mathbb{S}})] - \mathbb{E}_{0,x}[f(\tilde{X}_t^\pi)]| \leq C|\mathbb{S}|.$$

To this end, consider the PIDE (3). By the Feynman–Kac formula, we have $\phi(0, x) = \mathbb{E}_{0,x}[f(\tilde{X}_t^\pi)]$. We also have

$$\begin{aligned} \mathbb{E}_{0,x}[f(X_t^{\pi,\mathbb{S}})] - \mathbb{E}_{0,x}[f(\tilde{X}_t^\pi)] &= \mathbb{E}_{0,x}[\phi(t, X_t^{\pi,\mathbb{S}})] - \phi(0, x) \\ &= \sum_{j=1}^i \mathbb{E}_{0,x}[\phi(t_j, X_{t_j}^{\pi,\mathbb{S}}) - \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}})]. \end{aligned}$$

Itô's formula yields

$$\begin{aligned} &\phi(t_j, X_{t_j}^{\pi,\mathbb{S}}) - \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}}) \\ &= \int_{t_{j-1}}^{t_j} \mathcal{L}^{a_j} \phi(\tau, X_{\tau-}^{\pi,\mathbb{S}}) d\tau + \int_{t_{j-1}}^{t_j} \partial_x \phi(\tau, X_{\tau-}^{\pi,\mathbb{S}}) \circ \sigma(\tau, X_{\tau-}^{\pi,\mathbb{S}}, a_j) dW_\tau \\ &\quad + \sum_{k=1}^{\ell} \int_{t_{j-1}}^{t_j} \int_{\mathbb{R}} (\phi(\tau, X_{\tau-}^{\pi,\mathbb{S}} + \gamma_k(\tau, X_{\tau-}^{\pi,\mathbb{S}}, a_j, z)) - \phi(\tau, X_{\tau-}^{\pi,\mathbb{S}})) \tilde{N}_k(d\tau, dz), \end{aligned}$$

where $a_j = G^\pi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}}, U_j)$ and the operator \mathcal{L}^{a_j} is given by (5). Assumption 1 implies that the two stochastic integrals driven by the Brownian motion and compensated Poisson random measure are martingales. It follows that

$$\mathbb{E}_{0,x}[\phi(t_j, X_{t_j}^{\pi,\mathbb{S}}) - \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}})] = \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} \mathcal{L}^{a_j} \phi(\tau, X_{\tau-}^{\pi,\mathbb{S}}) d\tau \right].$$

As $a_j \sim \pi(\cdot | t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}})$, we have

$$\begin{aligned} \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} \mathcal{L}^{a_j} \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}}) d\tau \right] &= \mathbb{E}_{0,x} \left[\mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} \mathcal{L}^{a_j} \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}}) d\tau \middle| X_{t_{j-1}}^{\pi,\mathbb{S}} \right] \right] \\ &= \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} \mathcal{L}^\pi \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}}) d\tau \right] \\ &= 0, \end{aligned}$$

where the last equality follows from (3). Hence, we can write

$$\mathbb{E}_{0,x}[\phi(t_j, X_{t_j}^{\pi,\mathbb{S}}) - \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}})] = \mathbb{E}_{0,x}[e_j], \quad e_j := \int_{t_{j-1}}^{t_j} (\mathcal{L}^{a_j} \phi(\tau, X_{\tau-}^{\pi,\mathbb{S}}) - \mathcal{L}^{a_j} \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}})) d\tau.$$

Using the expression of \mathcal{L}^{a_j} , we obtain

$$\mathbb{E}_{0,x}[e_j] = \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} (\partial_\tau \phi(\tau, X_{\tau-}^{\pi,\mathbb{S}}) - \partial_\tau \phi(t_{j-1}, X_{t_{j-1}}^{\pi,\mathbb{S}})) d\tau \right]$$

$$\begin{aligned}
& + \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} b(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) \circ (\partial_x \phi(\tau, X_{\tau-}^{\pi, \mathbb{S}}) - \partial_x \phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}})) d\tau \right] \\
& + \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} \frac{1}{2} \sigma^2(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) \circ (\partial_x^2 \phi(\tau, X_{\tau-}^{\pi, \mathbb{S}}) - \partial_x^2 \phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}})) d\tau \right] \\
& + \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} (b(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) - b(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}, a_j)) \circ \partial_x \phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}) d\tau \right] \\
& + \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} \frac{1}{2} (\sigma^2(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) - \sigma^2(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}, a_j)) \circ \partial_x^2 \phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}) d\tau \right] \\
& + \sum_{k=1}^{\ell} \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} (I_k^\phi(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) - I_k^\phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}, a_j)) d\tau \right],
\end{aligned}$$

where I_k^ϕ is defined by (7). The first five terms can be analyzed using the arguments in Jia et al. (2025), which show that their sum is bounded by $C(t_j - t_{j-1})^2$. For the last term, we obtain from Itô's formula that

$$\begin{aligned}
& I_k^\phi(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) - I_k^\phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}, a_j) \\
& = \int_{t_{j-1}}^{\tau} (\partial_s I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}}, a_j) + b(s, X_{s-}^{\pi, \mathbb{S}}, a_j) \circ \partial_x I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}}, a_j)) ds \\
& \quad + \int_{t_{j-1}}^{\tau} \frac{1}{2} \sigma^2(s, X_{s-}^{\pi, \mathbb{S}}, a_j) \circ \partial_x^2 I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}}, a_j) ds \\
& \quad + \int_{t_{j-1}}^{\tau} \int_{\mathbb{R}} (I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}} + \gamma_k(s, X_{s-}^{\pi, \mathbb{S}}, a_j, z), a_j) - I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}}, a_j) \\
& \quad \quad - \gamma_k(s, X_{s-}^{\pi, \mathbb{S}}, a_j, z) \circ \partial_x I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}}, a_j)) \nu_k(dz) ds \\
& \quad + \int_{t_{j-1}}^{\tau} \partial_x I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}}, a_j) \circ \sigma(s, X_{s-}^{\pi, \mathbb{S}}, a_j) dW_s \\
& \quad + \sum_{k=1}^{\ell} \int_{t_{j-1}}^{\tau} \int_{\mathbb{R}} (I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}} + \gamma_k(s, X_{s-}^{\pi, \mathbb{S}}, a_j, z), a_j) - I_k^\phi(s, X_{s-}^{\pi, \mathbb{S}}, a_j)) \tilde{N}_k(ds, dz).
\end{aligned}$$

By Lemma 7 and Proposition 2, the last two terms are martingales and hence their expectations are zero. By Assumptions 1 and 3, the admissibility of π , Lemma 7 and Proposition 2, the expectations of the first three terms are finite and bounded by $C(\tau - t_{j-1})$. It follows that

$$\begin{aligned}
& \left| \mathbb{E}_{0,x} \left[\int_{t_{j-1}}^{t_j} (I_k^\phi(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) - I_k^\phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}, a_j)) d\tau \right] \right| \\
& \leq \int_{t_{j-1}}^{t_j} \mathbb{E}_{0,x} \left[|I_k^\phi(\tau, X_{\tau-}^{\pi, \mathbb{S}}, a_j) - I_k^\phi(t_{j-1}, X_{t_{j-1}}^{\pi, \mathbb{S}}, a_j)| \right] d\tau \\
& \leq C \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{\tau} (\tau - t_{j-1}) ds d\tau \\
& \leq C(t_j - t_{j-1})^2.
\end{aligned}$$

Putting these estimates together, we have $|E[e_j]| \leq C(t_j - t_{j-1})^2$. Finally, we obtain

$$|\mathbb{E}_{0,x}[f(X_T^{\pi,\mathbb{S}})] - \mathbb{E}_{0,x}[f(\tilde{X}_T^\pi)]| \leq \sum_{j=1}^i |\mathbb{E}_{0,x}[e_j]| \leq C|\mathbb{S}|.$$

(2) We estimate the difference between $J^{\pi,\mathbb{S}}(0, x)$ and $J^\pi(0, x)$. Note that

$$J^{\pi,\mathbb{S}}(0, x) - J^\pi(0, x) = (\mathbb{E}_{0,x}[h(X_T^{\pi,\mathbb{S}})] - \mathbb{E}_{0,x}[h(\tilde{X}_T^\pi)]) + \sum_{i=1}^N \mathbb{E} \left[\int_{t_{i-1}}^{t_i} (g(\tau, X_\tau^{\pi,\mathbb{S}}, a_i) - \tilde{g}(\tau, \tilde{X}_\tau^\pi)) d\tau \right],$$

where $a_i = G^\pi(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}, U_i)$. We have already analyzed the first term in part (1), which is bounded by $C|\mathbb{S}|$. For the second term, we have for $\tau \in (t_{i-1}, t_i]$,

$$\begin{aligned} & \mathbb{E}_{0,x}[g(\tau, X_\tau^{\pi,\mathbb{S}}, a_i) - \tilde{g}(\tau, \tilde{X}_\tau^\pi)] \\ &= \mathbb{E}_{0,x}[g(\tau, X_\tau^{\pi,\mathbb{S}}, a_i) - g(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}, a_i)] + \mathbb{E}_{0,x}[g(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}, a_i) - \tilde{g}(\tau, \tilde{X}_\tau^\pi)] \\ &= \mathbb{E}_{0,x}[g(\tau, X_\tau^{\pi,\mathbb{S}}, a_i) - g(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}, a_i)] + \mathbb{E}_{0,x}[\tilde{g}(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}) - \tilde{g}(t_{i-1}, \tilde{X}_{t_{i-1}}^\pi)] \\ & \quad + \mathbb{E}_{0,x}[\tilde{g}(t_{i-1}, \tilde{X}_{t_{i-1}}^\pi) - \tilde{g}(\tau, \tilde{X}_\tau^\pi)] \end{aligned}$$

where we noted $\mathbb{E}_{0,x}[g(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}, a_i)] = \mathbb{E}_{0,x}[\mathbb{E}_{0,x}[g(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}, a_i) | X_{t_{i-1}}^{\pi,\mathbb{S}}]] = \mathbb{E}_{0,x}[\tilde{g}(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}})]$. As $g \in C_p^{2,0}([0, T] \times \mathbb{R}^d \times \mathcal{A})$, it follows from Itô's formula, Assumption 1 and Proposition 2 that

$$|\mathbb{E}_{0,x}[g(\tau, X_\tau^{\pi,\mathbb{S}}, a_i) - g(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}, a_i)]| \leq C|\mathbb{S}|.$$

Part (1) shows

$$|\mathbb{E}_{0,x}[\tilde{g}(t_{i-1}, X_{t_{i-1}}^{\pi,\mathbb{S}}) - \tilde{g}(t_{i-1}, \tilde{X}_{t_{i-1}}^\pi)]| \leq C|\mathbb{S}|.$$

As $\tilde{g} \in C_p^2([0, T] \times \mathbb{R}^d)$, using Itô's formula, Lemma 1, Lemma 2 and Proposition 1 yields

$$|\mathbb{E}_{0,x}[\tilde{g}(t_{i-1}, \tilde{X}_{t_{i-1}}^\pi) - \tilde{g}(\tau, \tilde{X}_\tau^\pi)]| \leq C|\mathbb{S}|.$$

Combining these results, we obtain

$$\left| \sum_{i=1}^N \mathbb{E}_{0,x} \left[\int_{t_{i-1}}^{t_i} (g(\tau, X_\tau^{\pi,\mathbb{S}}, a_i) - \tilde{g}(\tau, \tilde{X}_\tau^\pi)) d\tau \right] \right| \leq C|\mathbb{S}|.$$

The proof of the theorem is now complete. □