Risk-Sensitive Q-Learning in Continuous Time with Application to Dynamic Portfolio Selection

Chuhan Xie
School of Mathematical Sciences
Peking University
Beijing, 100871
ch_xie@pku.edu.cn

Abstract

This paper studies the problem of risk-sensitive reinforcement learning (RSRL) in continuous time, where the environment is characterized by a controllable stochastic differential equation (SDE) and the objective is a potentially nonlinear functional of cumulative rewards. We prove that when the functional is an optimized certainty equivalent (OCE), the optimal policy is Markovian with respect to an augmented environment. We also propose *CT-RS-q*, a risk-sensitive q-learning algorithm based on a novel martingale characterization approach. Finally, we run a simulation study on a dynamic portfolio selection problem and illustrate the effectiveness of our algorithm.

1 Introduction

Traditional reinforcement learning (RL) algorithms typically aim to maximize the expected cumulative reward in a discrete-time setting. However, they often struggle in environments that are inherently continuous, or in applications where the full distributional characteristics of returns are of particular concern. A representative example arises in high-frequency trading: since trades occur irregularly over time, regular discretization of timestamps may cause potential information loss and training instability. Moreover, while achieving higher profits is desirable, it should not come at the cost of excessive volatility or drawdowns. Hence, the variance of the trader's profit, alongside its expectation, becomes a crucial performance criterion.

Existing research has addressed each of these two challenges separately. For the first, Wang et al. (2020); Jia and Zhou (2022a,b, 2023); Zhao et al. (2023); Tang et al. (2022) rigorously extend several RL algorithms to a continuous-time framework. For the second, numerous studies on risk-sensitive learning attempt to optimize risk measures beyond the expectation operator (Mihatsch and Neuneier, 2002; Geibel and Wysotzki, 2005; Shen et al., 2014; Fei et al., 2020; Lütjens et al., 2019; Garcıa and Fernández, 2015). Likewise, distributional reinforcement learning (DRL) focuses on learning the entire return distribution rather than specific moments or functionals (Bellemare et al., 2017; Dabney et al., 2018; Rowland et al., 2019; Bellemare et al., 2023). These approaches have achieved notable success both theoretically and empirically.

However, few studies have explored scenarios where both challenges coexist. This motivates our work, which takes an initial step toward unifying the two perspectives. We aim to bridge their intrinsic connections and establish a rigorous methodological foundation for algorithms that are simultaneously compatible with continuous-time modeling and risk-sensitive objectives. This paper provides a conceptual overview of the core ideas underlying continuous-time risk-sensitive RL, and demonstrates the feasibility of algorithms inspired by these principles.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: GenAI in Finance.

2 Problem Formulation

We first state our formulation of continuous-time risk-sensitive RL. Our problem is to control the state dynamics governed by a stochastic differential equation (SDE), defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_s)_{s\geq 0}, \mathbb{P})$ along with a standard n-dimensional Brownian motion $W = \{W_s, s \geq 0\}$:

$$\mathrm{d}X_s^\pi = \mu(s, X_s^\pi, a_s^\pi)\mathrm{d}s + \sigma(s, X_s^\pi, a_s^\pi)\mathrm{d}W_s, \quad s \in [t, T]. \tag{1}$$

Here, $\mu\colon [0,T] \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ and $\sigma\colon [0,T] \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^{d\times n}$ are given functions, where $\mathcal{A} \subset \mathbb{R}^m$ is the action space. The agent's action a_s^π is sampled independently of the Brownian motion according to a (possibly history-dependent) policy $\pi(\cdot \mid \mathcal{F}_s^X)$, where $\mathcal{F}_s^X$ refers to the natural filtration containing $(X_u^\pi)_{t\leq u\leq s}$ and $(a_u^\pi)_{t\leq u<s}$.

Let $r\colon [0,T] \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ be an instantaneous reward function, $h\colon \mathbb{R}^d \to \mathbb{R}$ be the lump-sum reward function applied at the end of the period, and $\delta > 0$ be a discount factor that measures the time-value of the payoff. By simulating the process (1) from $X_t^\pi = x$, we eventually receive a sum of discounted rewards:

$$Z^\pi(t,x) = \int_t^T e^{-\delta(s-t)} r(s, X_s^\pi, a_s^\pi)\mathrm{d}s + e^{-\delta(T-t)} h(X_T^\pi). \tag{2}$$

Given a risk measure $U\colon L^2(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, our goal is to find the optimal policy π^* that maximizes the risk-sensitive objective with entropy regularization:

$$J_0^\pi(t,x) = U(Z^\pi(t,x)) + \tau\, \mathrm{Ent}(\pi; t, x), \tag{3}$$

where $\mathrm{Ent}(\pi; t, x) = -\mathbb{E}\left[\int_t^T e^{-\delta(s-t)} \log \pi(a_s^\pi \mid \mathcal{F}_s^X)\mathrm{d}s \mid X_t^\pi = x\right]$ is defined as the discounted cumulative entropy of the policy π along the process starting from $X_t^\pi = x$.

3 Resolving Nonlinearity of the Functional Objective

When U is not the expectation operator as in traditional RL (Jia and Zhou, 2022a,b, 2023), typical algorithms based on Bellman equations fail due to its intrinsic nonlinearity. In fact, the optimal policy is generally not Markovian (Wang et al., 2024), which renders it complex to find the optimal policy.

To tackle this issue, we consider a special kind of risk measures called optimized certainty equivalents (OCEs), and show that the optimal policy is Markovian with respect to an augmented SDE in this case. In this way, the original problem breaks down to a conventional policy optimization task followed by a reverse transformation of the learned policy from the augmented SDE to the original one.

3.1 OCE risk measures

Optimized certainty equivalents (OCEs) are special functionals on random variables that admit a variational expression with an expectation operator:

$$U(W) = \mathrm{OCE}_\varphi(W) = \sup_{\eta\in\mathbb{R}} \left\{\eta + \mathbb{E}[\varphi(W - \eta)]\right\}, \tag{4}$$

where $W \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a random variable, and $\varphi\colon \mathbb{R} \to \mathbb{R}$ is a concave utility function.

Table 2 lists a series of common OCE risk measures along with their utility functions, including linear/exponential/logarithm utilities, CVaR risk, etc. The concept generalizes the classical notion of certainty equivalents by incorporating a utility function and an optimization procedure to unify a wider range of risk measures (Ben-Tal and Teboulle, 2007), and has been widely adopted in the optimization and mathematical finance literature.

3.2 Markovian optimality

It can be shown that when the risk measure U is an OCE with respect to the utility function φ, the optimal policy is Markovian with respect to the following augmented SDE:

$$\mathrm{d}\underbrace{\begin{pmatrix} X_s^\pi \\ B_{0,s}^\pi \\ B_{1,s}^\pi \end{pmatrix}}_{X_{\mathrm{aug},s}^\pi} = \underbrace{\begin{pmatrix} \mu(s, X_s^\pi, a_s^\pi) \\ B_{1,s}^\pi r(s, X_s^\pi, a_s^\pi) \\ -\delta B_{1,s}^\pi \end{pmatrix}}_{\mu_{\mathrm{aug}}(s, X_s^\pi, B_{0,s}^\pi, B_{1,s}^\pi, a_s^\pi)} \mathrm{d}s + \underbrace{\begin{pmatrix} \sigma(s, X_s^\pi, a_s^\pi) \\ 0_n^\top \\ 0_n^\top \end{pmatrix}}_{\sigma_{\mathrm{aug}}(s, X_s^\pi, B_{0,s}^\pi, B_{1,s}^\pi, a_s^\pi)} \mathrm{d}W_s, \quad s \in [t, T], \tag{5}$$

with the instantaneous reward function $r_{\text{aug}}(t, x, b_0, b_1, a) \equiv 0$ and the lump-sum reward function $h_{\text{aug}}(x, b_0, b_1) = \varphi(b_0 + b_1 h(x))$. In the sense of conventional RL, the entropy-regularized value function related to (5) is as follows (see Appendix C.1):

$$J^\pi(t, x, b_0, b_1) = \mathbb{E}[\varphi(b_0 + b_1 Z^\pi(t, x))] + \tau b_1 \operatorname{Ent}(\pi; t, x). \tag{6}$$

Compared with (1), the augmented SDE (5) has two additional states $(B^\pi_{0,s}, B^\pi_{1,s})$ that respectively track the cumulative reward so far and the effect of the discount factor (Bäuerle and Ott, 2011; Bäuerle and Glauner, 2021; Wang et al., 2024).

Proposition 3.1. *If the risk measure U is an OCE with respect to the utility function φ, the optimal policy $\pi_0^* = \operatorname{argmax}_\pi J_0^\pi(t, x)$ of the SDE* (1)-(3) *is Markovian with respect to the augmented state $(X_s, B_{0,s}, B_{1,s})$* [1]*, where $B_{0,s} = b_0 + b_1 \int_t^s e^{-\delta(u-t)} r(u, X_u, a_u) \mathrm{d}u$ and $B_{1,s} = b_1 e^{-\delta(s-t)}$ for some $(b_0, b_1) \in \mathbb{R} \times \mathbb{R}_+$; i.e., $\pi_0^*(\cdot \mid \mathcal{F}_s^X) = \pi_0^*(\cdot \mid X_s, B_{0,s}, B_{1,s})$ for any $s \in [t, T]$.*

3.3 Meta-algorithm for policy optimization

In fact, Proposition 3.1 is derived by bridging two value functions, (3) and (6) (see Appendix C.1). Following this idea, we can design a meta-algorithm that first solves policy optimization for the augmented SDE (5)-(6), and then optimizes over a scalar parameter according to the definition of the OCE risk measure U to obtain the optimal value function J_0^* and the optimal policy π_0^*. The meta-algorithm is summarized in Algorithm 1.

4 Risk-Sensitive Q-Learning

In this section, we focus on developing a q-learning algorithm to solve for the optimal value function $J^*(t, x, b_0, b_1)$ and its optimal policy $\pi^*(\cdot \mid t, x, b_0, b_1)$, following the martingale approach proposed in Jia and Zhou (2022a).

4.1 Martingale characterization

Let $\{(M_s^\theta)_{s\in[t,T]}\}_{\theta\in\Theta}$ be a set of parameterized $\mathcal{F}$-adapted stochastic processes, and let $\theta^* \in \Theta$ be the unique parameter such that $(M_s^{\theta^*})_{s\in[t,T]}$ is an $(\mathcal{F}, \mathbb{P})$-martingale. According to Jia and Zhou (2022a,b), $\theta = \theta^*$ if and only if for any $\mathcal{F}$-adapted test process $(\xi_s)_{s\in[t,T]}$, the following *martingale orthogonality condition* holds:

$$\mathbb{E}\left[\int_t^T \xi_s \mathrm{d}M_s^\theta\right] = 0. \tag{7}$$

This motivates us to find a martingale characterization of the value function, so that under proper parameterization, we can obtain a learned optimal value function as long as the condition (7) is satisfied. Theorem 4.1 below provides a rigorous statement of such martingale characterization.

Theorem 4.1. *Let a policy π, a function $\hat{J} \in C^{1,2}([0,T) \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+) \cap C([0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+)$, and a continuous function $\hat{q}\colon [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ \times \mathcal{A} \to \mathbb{R}$ be given such that for any quadruple $(t, x, b_0, b_1) \in [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$,*

$$\hat{J}(T, x, b_0, b_1) = \varphi(b_0 + b_1 h(x)), \quad \int_{\mathcal{A}} \exp\left\{\frac{\hat{q}(t, x, b_0, b_1, a)}{\tau b_1}\right\} \mathrm{d}a = 1. \tag{8}$$

Then, $\hat{J} = J^*$ and $\hat{q} = q^*$ respectively, if and only if for any $(t, x) \in [0,T] \times \mathbb{R}^d$, the following process is an $(\mathcal{F}, \mathbb{P})$-martingale:

$$\hat{J}(s, X_s^\pi, Y_s^\pi, e^{-\delta(s-t)}) - \int_t^s \hat{q}(u, X_u^\pi, Y_u^\pi, e^{-\delta(u-t)}, a_u^\pi)\mathrm{d}u, \quad s \in [t, T]. \tag{9}$$

Here, J^* is the optimal value function, q^* is the optimal q-function as defined in Appendix A.2, and $Y_s^\pi = \int_t^s e^{-\delta(u-t)} r(u, X_u^\pi, a_u^\pi)\mathrm{d}u$ is the discounted cumulative reward up to time s.

[1] We have omitted the superscript π on every state variable for notational simplicity.

4.2 Algorithm

Based on the martingale characterization established in the previous section, we propose an on-policy continuous-time risk-sensitive q-learning (CT-RS-q) algorithm as Algorithm 2. Specifically, we simultaneously parameterize the value function as J^θ and the q-function as q^ψ, and set the test functions ξ_t and ζ_t as their parameter gradients. The parameters θ and ψ are then updated by the average of temporal-difference errors after generation of every whole episode.

5 Application to Dynamic Portfolio Selection

We discuss an application to dynamic portfolio selection to illustrate the effectiveness of our proposed algorithm. Consider a market with two risky assets, whose prices follow the log-normal dynamics:

$$\mathrm{d}S_{1,t} = S_{1,t}(r_1\mathrm{d}t + \sigma_1\mathrm{d}W_{1,t}), \quad \mathrm{d}S_{2,t} = S_{2,t}(r_2\mathrm{d}t + \sigma_2\mathrm{d}W_{2,t}). \tag{10}$$

We would like to invest a \$1 budget on these two assets and are allowed to continuously reallocate the money between assets. Denoting the proportion of money invested on the first asset as a_t, our budget X_t follows another log-normal process:

$$\mathrm{d}X_t = X_t\{a_t r_1 + (1-a_t)r_2\}\mathrm{d}t + X_t\{a_t\sigma_1\mathrm{d}W_{1,t} + (1-a_t)\sigma_2\mathrm{d}W_{2,t}\}. \tag{11}$$

We are concerned with our budget X_T at the end of the whole period. While expecting the mean of X_T to be as large as possible, we would also like to avoid excessive variance; and therefore, we choose the mean-variance risk measure as our objective:

$$\mathrm{MV}(X_T) = \mathbb{E}[X_T] - \frac{\alpha}{2}\mathrm{Var}(X_T). \tag{12}$$

Since $\mathrm{MV}(\cdot)$ is an OCE as shown in Table 2, our algorithm is applicable to the above financial scenario. A rigorous formulation and some detailed discussion are deferred to Appendix B.

Firstly, we examine the convergence of the model parameters as defined in Appendix B.2. Figure 1 illustrates the evolution of eight model parameters, most of which converge to their optimal points. The last two parameters, which belong to the parameterized q-function q^ψ, stay slightly far from their optimal points. We believe that such deviation is caused by the non-shrinking exploration parameter $\tau > 0$ in the training phase.

Secondly, we compare the performance of three policies: (i) Baseline Policy, which always invests a fixed proportion ($a = 0.5$) of the budget between two assets; (ii) CT-RS-q Policy, which is trained according to Algorithm 2; (iii) Optimal Policy, whose analytical formula is given in Appendix B.1. Table 1 lists the cumulative return and the mean-variance objective (12) of the three policies at the end of the whole period, and Figure 2 plots their curves over time.

We find that the trained CT-RS-q policy is close to the optimal policy, both outperforming the baseline policy in the cumulative return and the mean-variance objective. This comes at the cost of a slightly larger volatility, which is further controllable by tuning the regularization parameter $\alpha > 0$ in (12) during the training period.

Table 1: Performance comparison of three policies.

	Cumulative Return ↑ (Std. Dev. ↓)	Mean-Variance Objective ↑
Baseline Policy ($a = 0.5$)	0.2217 (0.0957)	1.2171
CT-RS-q Policy	0.8163 (0.8716)	1.4365
Optimal Policy	0.7128 (0.7205)	1.4532

References

Bäuerle, N. and Glauner, A. (2021). Minimizing spectral risk measures applied to Markov decision processes. *Mathematical Methods of Operations Research*, 94(1):35–69.

Bäuerle, N. and Ott, J. (2011). Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379.

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR.

Bellemare, M. G., Dabney, W., and Rowland, M. (2023). *Distributional reinforcement learning*. MIT Press.

Ben-Tal, A. and Teboulle, M. (2007). An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. (2018). Distributional reinforcement learning with quantile regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):2892–2901.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. (2020). Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395.

Gao, X., Xu, Z. Q., and Zhou, X. Y. (2022). State-dependent temperature control for Langevin diffusions. *SIAM Journal on Control and Optimization*, 60(3):1250–1268.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Geibel, P. and Wysotzki, F. (2005). Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108.

Jia, Y. and Zhou, X. Y. (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 Zhou, X. Y. (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275):1–50.

Jia, Y. and Zhou, X. Y. (2023). q-Learning in continuous time. *Journal of Machine Learning Research*, 24(161):1–61.

Lütjens, B., Everett, M., and How, J. P. (2019). Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE.

Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine learning*, 49(2):267–290.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M. G., and Dabney, W. (2019). Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR.

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. (2014). Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328.

Tang, W., Zhang, Y. P., and Zhou, X. Y. (2022). Exploratory HJB equations and their convergence. *SIAM Journal on Control and Optimization*, 60(6):3191–3216.

Wang, H., Zariphopoulou, T., and Zhou, X. Y. (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34.

Wang, K., Liang, D., Kallus, N., and Sun, W. (2024). A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. *arXiv preprint arXiv:2403.06323*.

Zhao, H., Tang, W., and Yao, D. (2023). Policy optimization for continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 36:13637–13663.

Zhou, X. Y. (2021). Curse of optimality, and how we break it. *Available at SSRN 3845462*.

		Basic Function $u(w)$	Utility Function $\varphi(t)$	OCE Formula $\mathrm{OCE}_{\varphi}(W)$
Risk Aversion	General	$u(w)$ concave, increasing	$\inf_{\delta>0}\left\{\frac{u(\delta+t)-u(\delta)}{u'(\delta)}\right\}$	$u^{-1}(\mathbb{E}[u(W)])$
	Exponential Utility	$-e^{-\alpha w}, \alpha>0$	$\frac{1-e^{-\alpha t}}{\alpha}$	$-\frac{1}{\alpha}\ln\mathbb{E}[e^{-\alpha W}]$
	Power Utility	$\frac{w^{1-\gamma}-1}{1-\gamma}, 0<\gamma<1$	$\inf_{\delta>\max\{0,-t\}}\left\{\delta^{\gamma}\frac{(\delta+t)^{1-\gamma}-\delta^{1-\gamma}}{1-\gamma}\right\}$	$(\mathbb{E}[W^{1-\gamma}])^{\frac{1}{1-\gamma}}$
	Logarithm Utility	$\ln w$	$\inf_{\delta>\max\{0,-t\}}\left\{\delta\ln\left(1+\frac{t}{\delta}\right)\right\}$	$\exp(\mathbb{E}[\ln W])$
Risk Neutral	Linear Utility	w	t	$\mathbb{E}[W]$
Conditional Value-at-Risk (CVaR)		-	$\frac{1}{\beta}\min\{0,t\}$	$\mathbb{E}[W \mid W \le \mathrm{VaR}_{\beta}(W)]$
Mean-Variance (MV)		-	$t-\frac{\beta}{2}t^2$	$\mathbb{E}[W]-\frac{\beta}{2}\mathrm{Var}(W)$
Monotone Mean-Variance (MMV)		-	$\min\{1,t\}-\frac{\beta}{2}(\max\{1,t\})^2$	*no explicit form*

Table 2: Examples of OCE risk measures and their corresponding utility functions.

Algorithm 1: Meta-algorithm

Train based on the augmented SDE:
(i) Optimal value function $\hat{J}^*(t, x, b_0, b_1)$;
(ii) Optimal policy $\hat{\pi}^*(t, x, b_0, b_1)$;
Function `OptimalValueAndPolicy(`t, x, $\hat{J}^*$, $\hat{\pi}^*$`):`
  Optimal initial budget: $\hat{b}^* \leftarrow \operatorname{argmax}_{b\in\mathbb{R}} \{b + \hat{J}^*(t, x, -b, 1)\}$;
  Optimal value function: $\hat{J}_0^*(t, x) \leftarrow \hat{b}^* + \hat{J}^*(t, x, -\hat{b}^*, 1)$;
  Cumulative reward: $Y_s \leftarrow \int_t^s e^{-\delta(u-t)} r(u, X_u, a_u)\mathrm{d}u$;
  Optimal policy: $\hat{\pi}_0^*(\cdot \mid s, X_s, Y_s) \leftarrow \hat{\pi}^*(\cdot \mid s, X_s, Y_s - \hat{b}^*, e^{-\delta(s-t)}), \forall s \in [t, T]$;
  return $\hat{J}_0^*$, $\hat{\pi}_0^*$

Algorithm 2: CT-RS-q: continuous-time risk-sensitive q-learning (on-policy)

Input: initial state x_0, number of episodes N, time horizon T, number of mesh grids K, mesh grids $0 = t_0 < t_1 < \cdots < t_K = T$, learning rates $\{l_j^\theta, l_j^\psi\}_{j=1}^N$, temperature $\tau > 0$, parameterized value function $J^\theta(\cdot,\cdot,\cdot,\cdot)$ and q-function $q^\psi(\cdot,\cdot,\cdot,\cdot,\cdot)$

for *episode* $j = 1$ **to** N **do**
  Observe initial state x_0 and set $(X_{t_0}, B_{0,t_0}, B_{1,t_0}) = (x_0, 0, 1)$;
  for *timestep* $k = 0$ **to** $K - 1$ **do**
    Generate action
    $$a_{t_k} \sim \pi^\psi(\cdot \mid t_k, X_{t_k}, B_{0,t_k}, B_{1,t_k}) \propto \exp\{\tfrac{1}{\tau B_{1,t_k}} q^\psi(t_k, X_{t_k}, B_{0,t_k}, B_{1,t_k}, \cdot)\};$$
    Simulate (5) from t_k to t_{k+1} and observe new state $(X_{t_{k+1}}, B_{0,t_{k+1}}, B_{1,t_{k+1}})$;
    Store test functions:
    $$\xi_{t_k} = \frac{\partial J^\theta}{\partial \theta}(t_k, X_{t_k}, B_{0,t_k}, B_{1,t_k}), \quad \zeta_{t_k} = \frac{\partial q^\psi}{\partial \psi}(t_k, X_{t_k}, B_{0,t_k}, B_{1,t_k}, a_{t_k});$$
    Store value function and q-function:
    $$J_{t_k}^\theta = J^\theta(t_k, X_{t_k}, B_{0,t_k}, B_{1,t_k}), \quad q_{t_k}^\psi = q^\psi(t_k, X_{t_k}, B_{0,t_k}, B_{1,t_k}, a_{t_k});$$
  end
  Compute incremental updates:
  $$\Delta\theta = \sum_{k=0}^{K-1} \xi_{t_k}\big[J_{t_{k+1}}^\theta - J_{t_k}^\theta - q_{t_k}^\psi \Delta t_k\big], \quad \Delta\psi = \sum_{k=0}^{K-1} \zeta_{t_k}\big[J_{t_{k+1}}^\theta - J_{t_k}^\theta - q_{t_k}^\psi \Delta t_k\big];$$
  Update θ and ψ:
  $$\theta \leftarrow \theta + l_j^\theta \Delta\theta, \quad \psi \leftarrow \psi + l_j^\psi \Delta\psi;$$
end
return J^θ, q^ψ

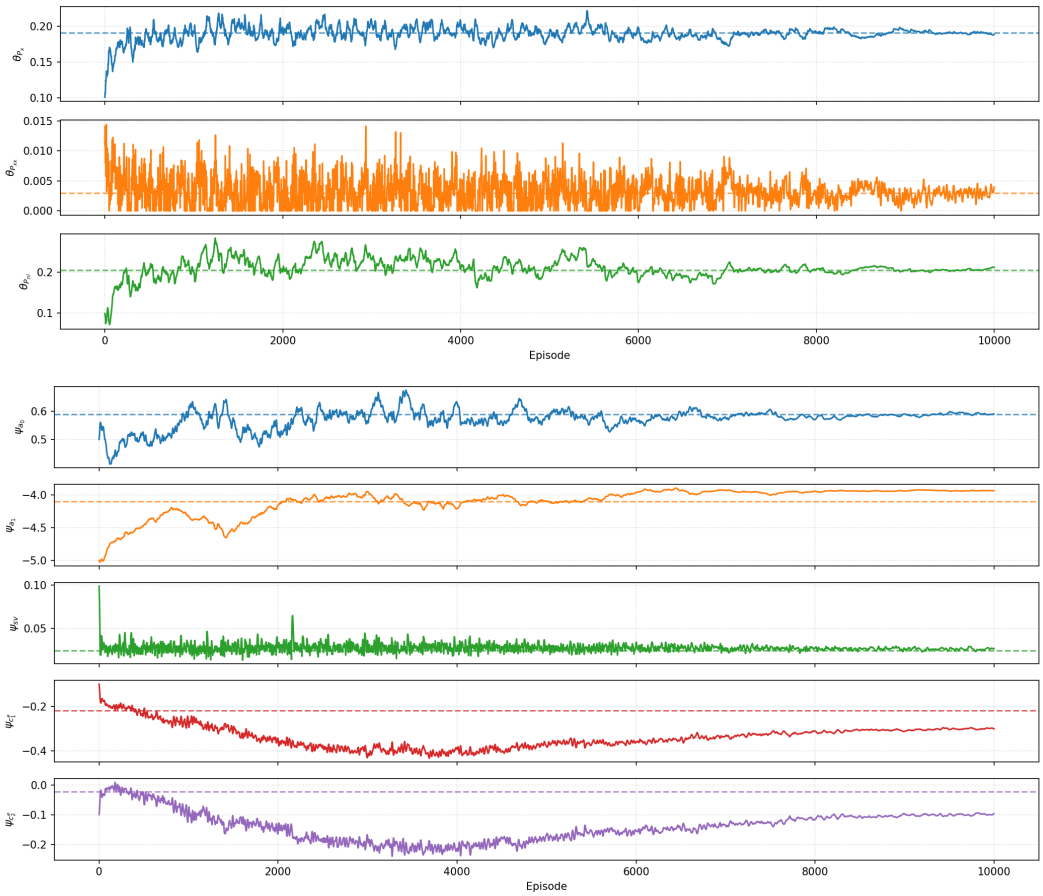

Figure 1: Convergence of model parameters. The first three are parameters of the value function J^θ, and the last five are parameters of the q-function q^ψ.

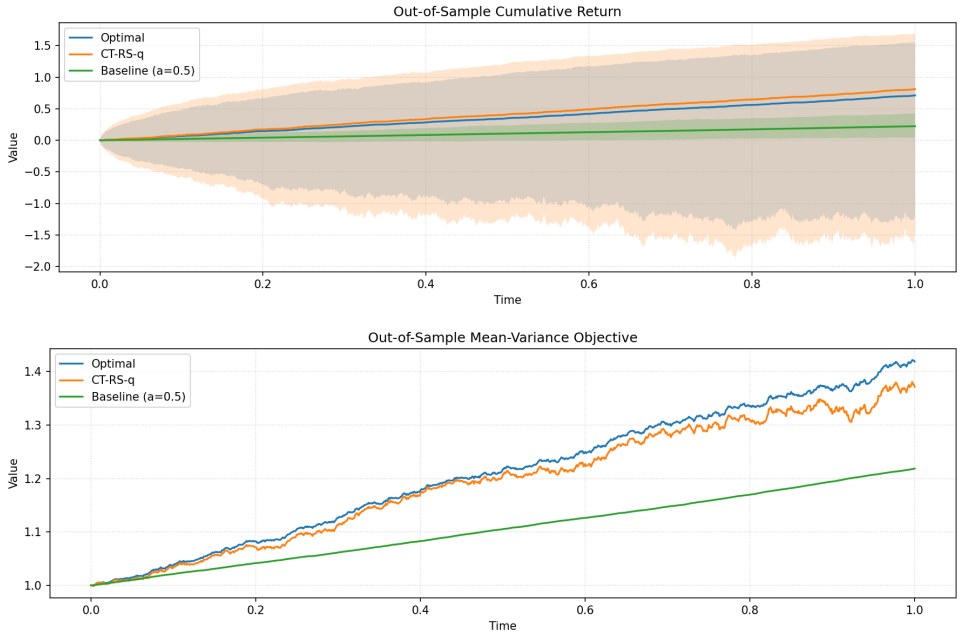

Figure 2: Curves of the cumulative return and the mean-variance objective for three policies.

A Theory of Risk-Sensitive Q-Learning

In this section, we provide a theoretical foundation of risk-sensitive q-learning and present more comprehensive results in addition to Section 4.

Notation. For the augmented SDE, we denote Π_{HD} and Π_{Mkv} as the set of all history-dependent policies and all Markov policies, respectively. For a policy $\pi\colon \mathcal{X}\times\mathcal{A}\to[0,1]$ and a function $f\colon \mathcal{X}\times\mathcal{A}\to\mathbb{R}$, we use $f(\cdot,\pi)=\int_{\mathcal{A}} f(\cdot,a)\pi(a\mid\cdot)\mathrm{d}a$ to denote the function value averaged over the policy.

A.1 Feynman-Kac formula and HJB equation

Let $\mathcal{L}^a_{\text{aug}}$ be the infinitesimal generator associated with the diffusion process governed by (5):

$$\begin{aligned}\mathcal{L}^a_{\text{aug}} f(t,x,b_0,b_1) &= \partial_t f(t,x,b_0,b_1) + \mu_{\text{aug}}(t,x,b_0,b_1,a)\partial_x f(t,x,b_0,b_1)\\ &\quad + \frac{1}{2}\sigma^2_{\text{aug}}(t,x,b_0,b_1,a)\partial^2_{xx} f(t,x,b_0,b_1).\end{aligned} \tag{13}$$

Recall that the augmented value function can be written as follows:

$$\begin{aligned}J^\pi(t,x,b_0,b_1) = \mathbb{E}\Big[&\varphi\big(B^\pi_{0,T}+B^\pi_{1,T}h(X^\pi_T)\big)\\ &-\tau\int_t^T B^\pi_{1,s}\log\pi(a^\pi_s\mid\mathcal{F}^X_s)\mathrm{d}s \mid X^\pi_t=x, B^\pi_{0,t}=b_0, B^\pi_{1,t}=b_1\Big].\end{aligned} \tag{14}$$

Given a Markov policy $\pi\in\Pi_{\text{Mkv}}$, its augmented value function J^π satisfies the following PDE:

$$\begin{aligned}\int_{\mathcal{A}}\big[\mathcal{L}^a_{\text{aug}}J^\pi(t,x,b_0,b_1) - \tau b_1\log\pi(a\mid t,x,b_0,b_1)\big]\pi(a\mid t,x,b_0,b_1)\mathrm{d}a = 0,\\ J^\pi(T,x,b_0,b_1)=\varphi(b_0+b_1h(x)),\end{aligned} \tag{15}$$

which is the well-known Feynman-Kac formula in the exploratory RL setting.

On the other hand, the optimal value function J^* satisfies the following HJB equation:

$$\begin{aligned}\sup_{\pi\in\Pi_{\text{Mkv}}}\int_{\mathcal{A}}\big[\mathcal{L}^a_{\text{aug}}J^*(t,x,b_0,b_1) - \tau b_1\log\pi(a\mid t,x,b_0,b_1)\big]\pi(a\mid t,x,b_0,b_1)\mathrm{d}a = 0,\\ J^*(T,x,b_0,b_1)=\varphi(b_0+b_1h(x)),\end{aligned} \tag{16}$$

Solving the above HJB equation yields the relationship between the optimal value function J^* and the optimal policy π^*, which we state in the following proposition.

Proposition A.1. *The optimal value function J^* and the optimal policy π^* satisfy the following relationship:*

$$\pi^*(a\mid t,x,b_0,b_1)=\exp\left\{\frac{\mathcal{L}^a_{aug}J^*(t,x,b_0,b_1)}{\tau b_1}\right\},\quad \int_{\mathcal{A}}\exp\left\{\frac{\mathcal{L}^a_{aug}J^*(t,x,b_0,b_1)}{\tau b_1}\right\}\mathrm{d}a=1. \tag{17}$$

A.2 *Lowercase* q-function

We next focus on deriving the q-function with respect to an augmented value function $J^\pi(t,x,b_0,b_1)$, analogous to Q-functions in the conventional RL setting. The concept of the *lowercase* q-function was proposed in Jia and Zhou (2023), where the authors provided a rigorous justification for a recent conjecture (Gao et al., 2022; Zhou, 2021) that the counterpart of Q-functions in continuous-time RL is the Hamiltonian of the dynamics. Below we extend the theory to the risk-sensitive scenario.

Given a Markov policy $\pi\in\Pi_{\text{Mkv}}$, a fixed action $a\in\mathcal{A}$ and a small constant $\Delta t>0$, consider a perturbed policy as follows: it takes the action a on $[t,t+\Delta t)$, and then follows π on $[t+\Delta t,T]$. The cumulative reward under such a perturbed policy then becomes

$$Z^\pi_{\Delta t}(t,x,a)=\int_t^{t+\Delta t} e^{-\delta(s-t)}r(s,X^a_s,a)\mathrm{d}s + e^{-\delta\Delta t}Z^\pi(t+\Delta t,X^a_{t+\Delta t}),\quad X^a_t=x, \tag{18}$$

and we introduce the corresponding Δt-parameterized Q-function as

$$Q^{\pi}_{\Delta t}(t,x,b_0,b_1,a) = \mathbb{E}[\varphi(b_0 + b_1 Z^{\pi}_{\Delta t}(t,x,a))] + \tau b_1 e^{-\delta \Delta t}\mathbb{E}[\mathrm{Ent}(\pi; t+\Delta t, X^{\pi}_{t+\Delta t})]. \quad (19)$$

Recall that in (3) and (6) an entropy term is included to incentivize exploration using stochastic policies. However, in defining $Q^{\pi}_{\Delta t}(t,x,b_0,b_1,a)$ we exclude the policy's entropy on the interval $[t, t+\Delta t)$, because a deterministic constant action a is applied whose entropy is always zero.

It is obvious that when $\Delta t \to 0$, the Δt-parameterized Q-function converges to the value function $J^{\pi}(t,x,b_0,b_1)$. However, the first-order term of Δt crucially reflects the advantage of the action a over the current policy π, which shares the same meaning as Q-functions in discrete-time RL. Following previous works (Jia and Zhou, 2023), we define such a first-order term as the q-function. To proceed, we define the infinitesimal generator of the SDE (1) as $\mathcal{L}^a$:

$$\mathcal{L}^a f(t,x) = \partial_t f(t,x) + \mu(t,x,a)\partial_x f(t,x) + \frac{1}{2}\sigma^2(t,x,a)\partial^2_{xx} f(t,x). \quad (20)$$

Definition A.1. The *q-function* of the augmented SDE (5)-(6) associated with a policy $\pi \in \Pi_{\text{Mkv}}$ is defined as follows:

$$q^{\pi}(t,x,b_0,b_1,a) = \{\mathcal{L}^a J^{\pi} + b_1 r(t,x,a)\partial_{b_0} J^{\pi} - b_1 \delta \partial_{b_1} J^{\pi}\}(t,x,b_0,b_1). \quad (21)$$

Note that any q-function q^{π} is related with the value function J^{π} of the same policy. In addition, we define the optimal q-function as $q^*(t,x,b_0,b_1,a) = q^{\pi^*}(t,x,b_0,b_1,a)$. Below we present some important properties of the q-function.

Proposition A.2. *The Δt-parameterized Q-function $Q^{\pi}_{\Delta t}$ satisfies that*

$$Q^{\pi}_{\Delta t}(t,x,b_0,b_1,a) = J^{\pi}(t,x,b_0,b_1) + q^{\pi}(t,x,b_0,b_1,a)\Delta t + o(\Delta t). \quad (22)$$

Proposition A.3. *The q-function q^{π} satisfies that*

$$\int_{\mathcal{A}} \{q^{\pi}(t,x,b_0,b_1,a) - \tau b_1 \log \pi(a \mid t,x,b_0,b_1)\}\pi(a \mid t,x,b_0,b_1)\mathrm{d}a = 0. \quad (23)$$

Proposition A.4. *The q-function q^{π} and the value function J^{π} satisfy that*

$$q^{\pi}(t,x,b_0,b_1,a) = \mathcal{L}^a_{aug} J^{\pi}(t,x,b_0,b_1). \quad (24)$$

A.3 Martingale characterization

The definition of the q-function enables us to design various kinds of martingale characterization of the value function, based on different algorithmic requirements. Specifically, we present below the martingale characterization theorems for: (i) on-policy policy evaluation in Theorem A.5, (ii) off-policy policy evaluation in Theorems A.6 and A.7, and (iii) off-policy policy optimization in Theorem A.8. Their proofs are similar to that of Theorem 4.1, so we omit them for brevity. As before, we define the discounted cumulative reward up to time s as Y^{π}_s:

$$Y^{\pi}_s = \int_t^s e^{-\delta(u-t)} r(u, X^{\pi}_u, a^{\pi}_u)\mathrm{d}u, \quad s \in [t,T]. \quad (25)$$

Theorem A.5. *Let a policy $\pi \in \Pi_{Mkv}$, its corresponding value function J^{π} and q-function q^{π}, a function $\hat{J} \in C^{1,2}([0,T)\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}_+) \cap C([0,T]\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}_+)$ with polynomial growth, and a continuous function $\hat{q}\colon [0,T]\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}_+\times\mathcal{A}\to\mathbb{R}$ be given such that for any $(t,x,b_0,b_1) \in [0,T]\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}_+$,*

$$\hat{J}(T,x,b_0,b_1) = \varphi(b_0 + b_1 h(x)), \quad \hat{q}(t,x,b_0,b_1,\pi) + \tau b_1 \,\mathrm{Ent}(\pi(\cdot \mid t,x,b_0,b_1)) = 0. \quad (26)$$

Then:

(i) $\hat{q} = q^{\pi}$ if and only if for any $(t,x) \in [0,T]\times\mathbb{R}^d$, the following process is an $(\mathcal{F},\mathbb{P})$-martingale:

$$J^{\pi}(s, X^{\pi}_s, Y^{\pi}_s, e^{-\delta(s-t)}) - \int_t^s \hat{q}(u, X^{\pi}_u, Y^{\pi}_u, e^{-\delta(u-t)}, a^{\pi}_u)\mathrm{d}u, \quad s \in [t,T]. \quad (27)$$

(ii) $\hat{J} = J^\pi$ *and* $\hat{q} = q^\pi$ *respectively, if and only if for any* $(t,x) \in [0,T] \times \mathbb{R}^d$*, the following process is an* $(\mathcal{F}, \mathbb{P})$*-martingale:*

$$\hat{J}(s, X_s^\pi, Y_s^\pi, e^{-\delta(s-t)}) - \int_t^s \hat{q}(u, X_u^\pi, Y_u^\pi, e^{-\delta(u-t)}, a_u^\pi)\mathrm{d}u, \quad s \in [t,T]. \tag{28}$$

Theorem A.6. *Let a policy* $\pi \in \Pi_{Mkv}$*, its corresponding value function* J^π *and q-function* q^π*, and a continuous function* $\hat{q}\colon [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ \times \mathcal{A} \to \mathbb{R}$ *be given. Then:*

(i) If $\hat{q} = q^\pi$*, then for any* $\pi' \in \Pi_{Mkv}$ *and any* $(t,x) \in [0,T] \times \mathbb{R}^d$*, the following process is an* $(\mathcal{F}, \mathbb{P})$*-martingale:*

$$J^\pi(s, X_s^{\pi'}, Y_s^{\pi'}, e^{-\delta(s-t)}) - \int_t^s \hat{q}(u, X_u^{\pi'}, Y_u^{\pi'}, e^{-\delta(u-t)}, a_u^{\pi'})\mathrm{d}u, \quad s \in [t,T]. \tag{29}$$

(ii) If there exists $\pi' \in \Pi_{Mkv}$ *such that* (29) *is an* $(\mathcal{F}, \mathbb{P})$*-martingale for any* $(t,x) \in [0,T] \times \mathbb{R}^d$*, then* $\hat{q} = q^\pi$*.*

Theorem A.7. *Let a policy* $\pi \in \Pi_{Mkv}$*, its corresponding value function* J^π *and q-function* q^π*, a function* $\hat{J} \in C^{1,2}([0,T) \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+) \cap C([0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+)$ *with polynomial growth, and a continuous function* $\hat{q}\colon [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ \times \mathcal{A} \to \mathbb{R}$ *be given such that for any* $(t,x,b_0,b_1) \in [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$*,*

$$\hat{J}(T,x,b_0,b_1) = \varphi(b_0 + b_1 h(x)), \quad \hat{q}(t,x,b_0,b_1,\pi) + \tau b_1 \operatorname{Ent}(\pi(\cdot \mid t,x,b_0,b_1)) = 0. \tag{30}$$

Then:

(i) If $\hat{J} = J^\pi$ *and* $\hat{q} = q^\pi$ *respectively, then for any* $\pi' \in \Pi_{Mkv}$ *and any* $(t,x) \in [0,T] \times \mathbb{R}^d$*, the following process is an* $(\mathcal{F}, \mathbb{P})$*-martingale:*

$$\hat{J}(s, X_s^{\pi'}, Y_s^{\pi'}, e^{-\delta(s-t)}) - \int_t^s \hat{q}(u, X_u^{\pi'}, Y_u^{\pi'}, e^{-\delta(u-t)}, a_u^{\pi'})\mathrm{d}u, \quad s \in [t,T]. \tag{31}$$

(ii) If there exists $\pi' \in \Pi_{Mkv}$ *such that* (31) *is an* $(\mathcal{F}, \mathbb{P})$*-martingale for any* $(t,x) \in [0,T] \times \mathbb{R}^d$*, then* $\hat{J} = J^\pi$ *and* $\hat{q} = q^\pi$ *respectively.*

Theorem A.8. *Let a function* $\hat{J}^* \in C^{1,2}([0,T) \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+) \cap C([0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+)$ *with polynomial growth and a continuous function* $\hat{q}^*\colon [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ \times \mathcal{A} \to \mathbb{R}$ *be given such that for any* $(t,x,b_0,b_1) \in [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$*,*

$$\hat{J}^*(T,x,b_0,b_1) = \varphi(b_0 + b_1 h(x)), \quad \int_{\mathcal{A}} \exp\left\{\frac{\hat{q}^*(t,x,b_0,b_1,a)}{\tau b_1}\right\}\mathrm{d}a = 1. \tag{32}$$

Then:

(i) If $\hat{J}^*$ *and* $\hat{q}^*$ *are respectively the optimal value function and the optimal q-function, then for any* $\pi \in \Pi_{Mkv}$ *and any* $(t,x) \in [0,T] \times \mathbb{R}^d$*, the following process is an* $(\mathcal{F}, \mathbb{P})$*-martingale:*

$$\hat{J}^*(s, X_s^\pi, Y_s^\pi, e^{-\delta(s-t)}) - \int_t^s \hat{q}^*(u, X_u^\pi, Y_u^\pi, e^{-\delta(u-t)}, a_u^\pi)\mathrm{d}u, \quad s \in [t,T]. \tag{33}$$

Moreover, $\hat{\pi}^*(a \mid t,x,b_0,b_1) = \exp\left\{\frac{\hat{q}^*(t,x,b_0,b_1,a)}{\tau b_1}\right\}$ *is the optimal policy in this case.*

(ii) If there exists $\pi \in \Pi_{Mkv}$ *such that for any* $(t,x) \in [0,T] \times \mathbb{R}^d$*,* (33) *is an* $(\mathcal{F}, \mathbb{P})$*-martingale, then* $\hat{J}^*$ *and* $\hat{q}^*$ *are respectively the optimal value function and the optimal q-function.*

B Details on Dynamic Portfolio Selection

In this section, we provide a more comprehensive introduction to dynamic portfolio selection. Consider a market with two risky assets, with their prices following the log-normal dynamics:

$$\mathrm{d}S_{1,t} = S_{1,t}(r_1\mathrm{d}t + \sigma_1\mathrm{d}W_{1,t}), \quad \mathrm{d}S_{2,t} = S_{2,t}(r_2\mathrm{d}t + \sigma_2\mathrm{d}W_{2,t}), \tag{34}$$

where $r_1, r_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 \in \mathbb{R}_+$, and $(W_{1,t})_{t\geq 0}, (W_{2,t})_{t\geq 0}$ are two independent Brownian motions. Suppose we have a \$1 budget at $t = 0$ and would like to invest all the money in these two assets; at each time t, we are allowed to reallocate the money between assets. Denoting the proportion of money invested on the first asset as a_t, our budget X_t follows another log-normal process:

$$\mathrm{d}X_t = X_t\{a_t r_1 + (1-a_t)r_2\}\mathrm{d}t + X_t\{a_t\sigma_1\mathrm{d}W_{1,t} + (1-a_t)\sigma_2\mathrm{d}W_{2,t}\}. \tag{35}$$

Here, a_t's are possible to take values outside the unit interval $[0, 1]$, as we allow shorting of an asset.

We are concerned with our budget X_T at the end of the whole period without any discount in time; i.e., $\delta = 0$. We choose the mean-variance risk measure as our objective:

$$\mathrm{MV}(X_T) = \mathbb{E}[X_T] - \frac{\alpha}{2}\mathrm{Var}(X_T). \tag{36}$$

Then the corresponding utility function and reward functions are as follows:

$$\varphi(x) = x - \frac{\alpha}{2}x^2, \quad r(t, x, a) = 0, \quad h(x) = x. \tag{37}$$

The remainder of this section is organized as follows. Appendix B.1 gives analytical formulae for the optimal value function and its corresponding q-function with zero exploration factor; i.e., $\tau = 0$. According to the forms of functions, we design their parameterization with well specification at optimum in Appendix B.2. In Appendix B.3, we numerically justify that the optimal parameter in our parameterization is a stable point, which indicates local convergence from near the optimum.

B.1 Analytical solution to optimal control

The following proposition gives the analytical formulae of the optimal control, the optimal value function, and its corresponding q-function. We assume zero exploration factor, $\tau = 0$, for simplicity. Note that both the value function and the q-function are quadratic in x and a.

Proposition B.1. *The optimal control to* (35)-(37)*, the optimal value function, and its corresponding q-function are*

$$\begin{aligned}
&a^*(t, x, b_0, b_1) = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} - \frac{r_1 - r_2}{\sigma_1^2+\sigma_2^2}\left(1 + \frac{c_1}{2c_2 x}\right),\\
&J^*(t, x, b_0, b_1) = c_0 + c_1 x + c_2 x^2,\\
&q^*(t, x, b_0, b_1, a) = (\sigma_1^2+\sigma_2^2)c_2 x^2\{a - a^*(t, x, b_0, b_1)\}^2,
\end{aligned}$$

where c_0, c_1, c_2 are defined as follows:

$$\begin{aligned}
&c_0 = c_0(t, b_0, b_1) = b_0\left(1 - \frac{\alpha}{2}b_0\right) + \frac{(1-\alpha b_0)^2 P_{nl}}{2\alpha(P_{xx}+P_{nl})}\left[1 - e^{-2(P_{xx}+P_{nl})(T-t)}\right],\\
&c_1 = c_1(t, b_0, b_1) = (1-\alpha b_0)b_1 e^{(P_x - 2P_{nl})(T-t)},\\
&c_2 = c_2(t, b_0, b_1) = -\frac{\alpha}{2}b_1^2 e^{2(P_x + P_{xx} - P_{nl})(T-t)},\\
&P_x = \frac{r_1\sigma_2^2 + r_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \quad P_{xx} = \frac{\sigma_1^2\sigma_2^2}{2(\sigma_1^2+\sigma_2^2)}, \quad P_{nl} = \frac{(r_1-r_2)^2}{2(\sigma_1^2+\sigma_2^2)}.
\end{aligned}$$

Proof of Proposition B.1. Recall the optimal value function is

$$J^*(t, x, b_0, b_1) = \sup_a \mathbb{E}\left[(b_0 + b_1 X_T) - \frac{\alpha}{2}(b_0 + b_1 X_T)^2 \mid X_t = x\right].$$

Its HJB equation is as follows:

$$\sup_a \left\{ \partial_t J^* + \{ar_1 + (1-a)r_2\} x \partial_x J^* + \frac{1}{2}\{a^2\sigma_1^2 + (1-a)^2\sigma_2^2\} x^2 \partial_{xx}^2 J^* \right\} = 0.$$

For now we assume $\partial_{xx}^2 J^* < 0$, so the left-hand side is a quadratic function with respect to a and achieves optimality at

$$a^* = \frac{\sigma_2^2 x^2 \partial_{xx}^2 J^* - (r_1 - r_2) x \partial_x J^*}{(\sigma_1^2 + \sigma_2^2) x^2 \partial_{xx}^2 J^*} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} - \frac{(r_1 - r_2)\partial_x J^*}{(\sigma_1^2 + \sigma_2^2) x \partial_{xx}^2 J^*},$$

and by plugging back into the HJB equation we obtain

$$\partial_t J^* + \underbrace{\frac{r_1\sigma_2^2 + r_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}}_{P_x} x \partial_x J^* + \underbrace{\frac{\sigma_1^2\sigma_2^2}{2(\sigma_1^2 + \sigma_2^2)}}_{P_{xx}} x^2 \partial_{xx} J^* - \underbrace{\frac{(r_1 - r_2)^2}{2(\sigma_1^2 + \sigma_2^2)}}_{P_{nl}} \frac{(\partial_x J^*)^2}{\partial_{xx} J^*} = 0. \tag{38}$$

Assume the value function has a quadratic form in x:

$$J^*(t, x, b_0, b_1) = c_0(t, b_0, b_1) + c_1(t, b_0, b_1)x + c_2(t, b_0, b_1)x^2,$$

which satisfies the terminal condition:

$$c_0(T, b_0, b_1) = b_0 - \frac{\alpha}{2}b_0^2, \quad c_1(T, b_0, b_1) = b_1 - \alpha b_0 b_1, \quad c_2(T, b_0, b_1) = -\frac{\alpha}{2}b_1^2.$$

The partial derivatives with respect to t and x are thus given by

$$\partial_t J^* = \dot{c}_0 + \dot{c}_1 x + \dot{c}_2 x^2, \quad \partial_x J^* = c_1 + 2c_2 x, \quad \partial_{xx}^2 J^* = 2c_2.$$

Plug these into (38) and note that it holds for any $x \in \mathbb{R}^d$, so the following differential equations should be satisfied:

$$\begin{aligned}
\dot{c}_0 - P_{nl}\frac{c_1^2}{2c_2} = 0, &\quad c_0(T, b_0, b_1) = b_0 - \frac{\alpha}{2}b_0^2, \\
\dot{c}_1 + (P_x - 2P_{nl})c_1 = 0, &\quad c_1(T, b_0, b_1) = b_1 - \alpha b_0 b_1, \\
\dot{c}_2 + 2(P_x + P_{xx} - P_{nl})c_2 = 0, &\quad c_2(T, b_0, b_1) = -\frac{\alpha}{2}b_1^2.
\end{aligned}$$

Solving these equations yields

$$\begin{aligned}
c_0(t, b_0, b_1) &= b_0\left(1 - \frac{\alpha}{2}b_0\right) + \frac{(1 - \alpha b_0)^2 P_{nl}}{2\alpha(P_{xx} + P_{nl})}\left[1 - e^{-2(P_{xx} + P_{nl})(T-t)}\right], \\
c_1(t, b_0, b_1) &= (1 - \alpha b_0) b_1 e^{(P_x - 2P_{nl})(T-t)}, \\
c_2(t, b_0, b_1) &= -\frac{\alpha}{2}b_1^2 e^{2(P_x + P_{xx} - P_{nl})(T-t)}.
\end{aligned}$$

Note that $c_2 < 0$, so the previous assumption $\partial_{xx}^2 J^* < 0$ is verified. Hence, the corresponding optimal control is

$$a^*(t, x, b_0, b_1) = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} - \frac{r_1 - r_2}{\sigma_1^2 + \sigma_2^2}\left(1 + \frac{c_1}{2c_2 x}\right).$$

Finally, we turn to the q-function. Applying (21) yields

$$\begin{aligned}
&q^*(t, x, b_0, b_1, a) \\
&= \mathcal{L}^a J^*(t, x, b_0, b_1) + b_1 r(t, x, a)\partial_{b_0} J^*(t, x, b_0, b_1) - b_1 \delta \partial_{b_1} J^*(t, x, b_0, b_1) \\
&= (\dot{c}_0 + \dot{c}_1 x + \dot{c}_2 x^2) + \{ar_1 + (1-a)r_2\} x(c_1 + 2c_2 x) + \{a^2\sigma_1^2 + (1-a)^2\sigma_2^2\} c_2 x^2 \\
&= (\sigma_1^2 + \sigma_2^2) c_2 x^2 a^2 + \{(r_1 - r_2)(c_1 x + 2c_2 x^2) - 2\sigma_2^2 c_2 x^2\} a + \text{remainder} \\
&= (\sigma_1^2 + \sigma_2^2) c_2 x^2 \left(a + \frac{(r_1 - r_2)(c_1 x + 2c_2 x^2) - 2\sigma_2^2 c_2 x^2}{2(\sigma_1^2 + \sigma_2^2) c_2 x^2}\right)^2 \\
&= (\sigma_1^2 + \sigma_2^2) c_2 x^2 \{a - a^*(t, x, b_0, b_1)\}^2,
\end{aligned}$$

where the second to last equality is because the remainder term is independent with a and $\max_a q^*(t, x, b_0, b_1, a) = 0$. □

B.2 Parameterization of trainable functions

This section discusses how we parameterize the value function and the q-function.

The value function is parameterized as $J^\theta = c_0^\theta + c_1^\theta x + c_2^\theta x^2$ with the parameter $\theta = (\theta_{P_x}, \theta_{P_{xx}}, \theta_{P_{nl}})$, where c_0^θ, c_1^θ, c_2^θ are defined as follows:

$$
\begin{aligned}
c_0^\theta &= c_0^\theta(t, b_0, b_1) = b_0\left(1 - \frac{\alpha}{2}b_0\right) + \frac{(1-\alpha b_0)^2\theta_{P_{nl}}}{2\alpha(\theta_{P_{xx}} + \theta_{P_{nl}})}\left[1 - e^{-2(\theta_{P_{xx}}+\theta_{P_{nl}})(T-t)}\right], \\
c_1^\theta &= c_1^\theta(t, b_0, b_1) = (1-\alpha b_0)b_1 e^{(\theta_{P_x} - 2\theta_{P_{nl}})(T-t)}, \\
c_2^\theta &= c_2^\theta(t, b_0, b_1) = -\frac{\alpha}{2}b_1^2 e^{2(\theta_{P_x}+\theta_{P_{xx}}-\theta_{P_{nl}})(T-t)}.
\end{aligned}
$$

The q-function is parameterized as $q^\psi(t, x, b_0, b_1, a) = \psi_{sv} c_2^\psi x^2 (a - a^\psi)^2$ with the parameter $\psi = (\psi_{a_0}, \psi_{a_1}, \psi_{sv}, \psi_{c_1^e}, \psi_{c_2^e})$, where

$$
\begin{aligned}
c_1^\psi &= c_1^\psi(t, b_0, b_1) = (1-\alpha b_0)b_1 e^{\psi_{c_1^e}(T-t)}, \\
c_2^\psi &= c_2^\psi(t, b_0, b_1) = -\frac{\alpha}{2}b_1^2 e^{\psi_{c_2^e}(T-t)}, \\
a^\psi &= \psi_{a_0} - \psi_{a_1}\left(1 + \frac{c_1^\psi}{2c_2^\psi x}\right).
\end{aligned}
$$

Note that our parameterization well specifies the ground truth model when the following optimal parameters are achieved:

$$
\begin{aligned}
&\theta^*_{P_x} = P_x, \quad \theta^*_{P_{xx}} = P_{xx}, \quad \theta^*_{P_{nl}} = P_{nl}, \\
&\psi^*_{a_0} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad \psi^*_{a_1} = \frac{r_1 - r_2}{\sigma_1^2 + \sigma_2^2}, \quad \psi^*_{sv} = \sigma_1^2 + \sigma_2^2, \\
&\psi^*_{c_1^e} = P_x - 2P_{nl}, \quad \psi^*_{c_2^e} = 2(P_x + P_{xx} - P_{nl}).
\end{aligned}
$$

B.3 Numerical analysis of temporal difference around optimum

Figure 3 shows temporal difference of parameters around the optimum. The environment parameters are set as $r_1 = 0.15$, $r_2 = 0.25$, $\sigma_1 = 0.1$, $\sigma_2 = 0.12$, and the hyper-parameters are set as $T = 1.0$, $\Delta t = 0.001$, $\alpha = 1.0$. We simulate $N = 10,000$ trajectories starting from the initial budget $X_0 = 1.0$. In this setting, the value of optimal parameters $\theta^* = (\theta^*_{P_x}, \theta^*_{P_{xx}}, \theta^*_{P_{nl}})$ and $\psi^* = (\psi^*_{a_0}, \psi^*_{a_1}, \psi^*_{sv}, \psi^*_{c_1^e}, \psi^*_{c_2^e})$ are as follows:

$$
\begin{gathered}
\theta^*_{P_x} = 0.1910, \quad \theta^*_{P_{xx}} = 0.0030, \quad \theta^*_{P_{nl}} = 0.2049, \\
\psi^*_{a_0} = 0.5902, \quad \psi^*_{a_1} = -4.0984, \quad \psi^*_{sv} = 0.0244, \quad \psi^*_{c_1^e} = -0.2189, \quad \psi^*_{c_2^e} = -0.0220.
\end{gathered}
$$

For each parameter, it is perturbed around its optimal value by a small amount, while other parameters remain fixed. The dot at zero in each plot corresponds to the optimum, and the dashed black lines are the zero horizontal and vertical lines on which the update for the parameter is zero.

From the plots in Figure 3 we find that, for any of the eight parameters in the value function or the q-function, its temporal difference crosses zero from above to below, which indicates the optimum is a stable point. This phenomenon numerically provides guarantees for convergence to optimum of the given optimization algorithm.

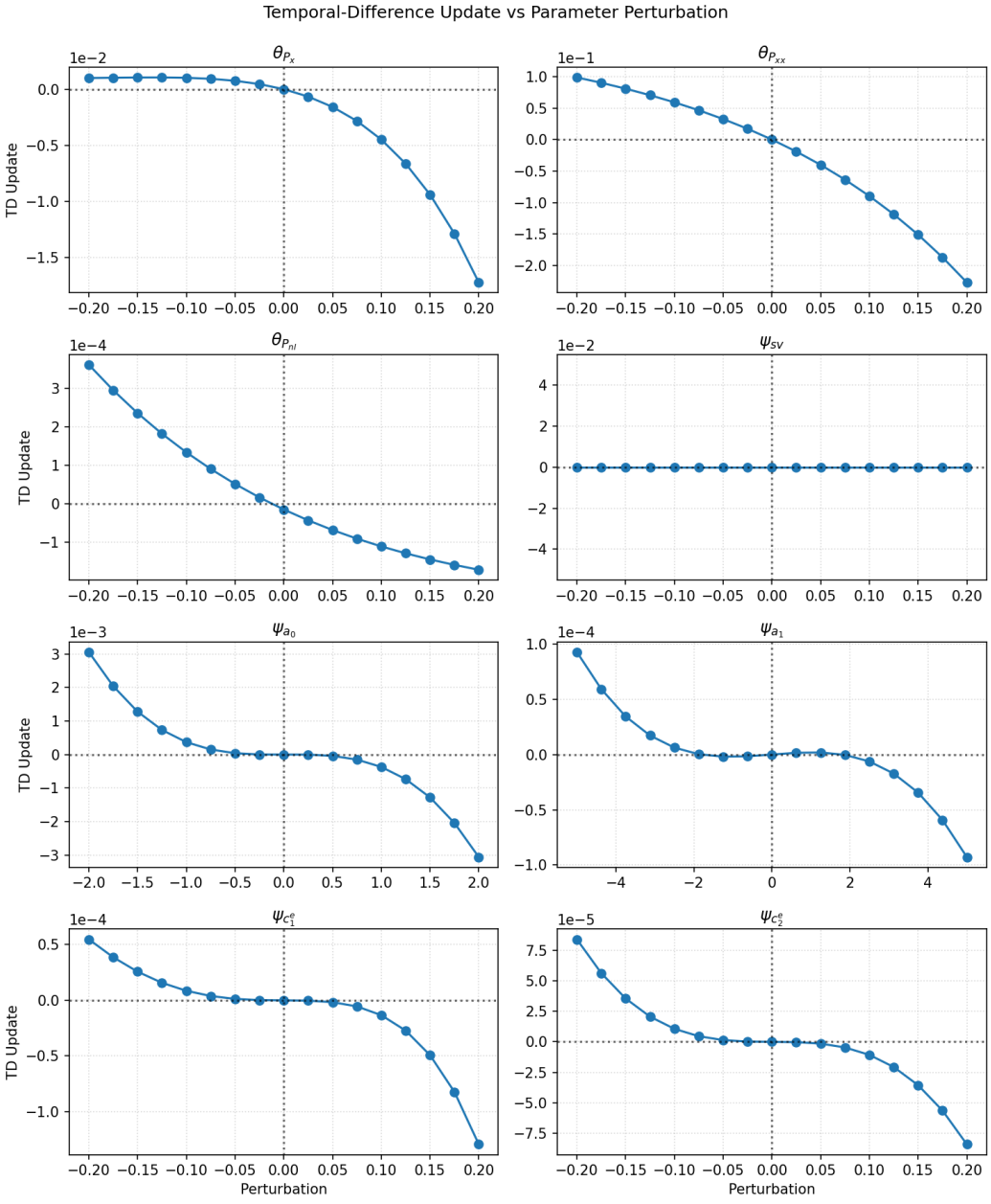

Figure 3: Temporal difference of parameters around optimum (zeros in the plots above).

C Proofs

C.1 Proof of Proposition 3.1

We adopt three steps to prove the proposition: (i) derivation of the augmented value function (6), (ii) Markovian optimality in the augmented SDE (5)-(6), and (iii) equivalence between optimal policies in the original SDE (1)-(3) and the augmented SDE (5)-(6).

Derivation of the augmented value function. Suppose the risk measure U is expectation and the discount factor for rewards is zero as in the conventional finite-horizon RL setting. The entropy-regularized value function should be

$$\begin{aligned}
J^\pi(t,x,b_0,b_1) &= \mathbb{E}\bigg[\int_t^T r_{\text{aug}}(s,X_s^\pi,B_{0,s}^\pi,B_{1,s}^\pi,a_s^\pi)\mathrm{d}s + h_{\text{aug}}(X_T^\pi,B_{0,T}^\pi,B_{1,T}^\pi) \\
&\qquad - \tau\int_t^T B_{1,s}^\pi \log\pi(a_s^\pi \mid \mathcal{F}_s^X)\mathrm{d}s \mid X_t^\pi = x, B_{0,t}^\pi = b_0, B_{1,t}^\pi = b_1\bigg] \\
&= \mathbb{E}\big[\varphi(B_{0,T}^\pi + B_{1,T}^\pi h(X_T^\pi)) \mid X_t^\pi = x, B_{0,t}^\pi = b_0, B_{1,t}^\pi = b_1\big] + \tau b_1 \operatorname{Ent}(\pi;t,x) \\
&= \mathbb{E}\bigg[\varphi\Big(\underbrace{b_0 + b_1\int_t^T e^{-\delta(s-t)}r(s,X_s^\pi,a_s^\pi)\mathrm{d}s}_{B_{0,T}^\pi} \\
&\qquad + \underbrace{b_1 e^{-\delta(T-t)}}_{B_{1,T}^\pi} h(X_T^\pi)\Big) \mid X_t^\pi = x, B_{0,t}^\pi = b_0, B_{1,t}^\pi = b_1\bigg] + \tau b_1 \operatorname{Ent}(\pi;t,x) \\
&= \mathbb{E}[\varphi(b_0 + b_1 Z^\pi(t,x))] + \tau b_1 \operatorname{Ent}(\pi;t,x),
\end{aligned}$$

which concludes the derivation of (6). Note that the entropy regularization $\operatorname{Ent}(\pi;t,x)$ is still δ-discounted across time to align with its original definition, despite absence of discount in rewards.

Markovian optimality in the augmented SDE. The Markovian optimality follows naturally from the conventional RL literature (Puterman, 2014). More detailed examples can be found in Jia and Zhou (2022a,b, 2023).

Equivalence between optimal policies. By the definition of OCE, we have

$$\begin{aligned}
J_0^*(t,x) &= \max_{\pi\in\Pi_{\text{HD}}}\left\{\max_{b\in\mathbb{R}}\{b + \mathbb{E}[\varphi(Z^\pi(t,x)-b)]\} + \tau\operatorname{Ent}(\pi;t,x)\right\} \\
&= \max_{b\in\mathbb{R}}\left\{b + \max_{\pi\in\Pi_{\text{HD}}}\{\mathbb{E}[\varphi(Z^\pi(t,x)-b)] + \tau\operatorname{Ent}(\pi;t,x)\}\right\} \\
&= \max_{b\in\mathbb{R}}\left\{b + \max_{\pi\in\Pi_{\text{HD}}} J^\pi(t,x,-b,1)\right\} \\
&= \max_{b\in\mathbb{R}}\left\{b + \max_{\pi\in\Pi_{\text{Mkv}}} J^\pi(t,x,-b,1)\right\} \\
&= \max_{b\in\mathbb{R}}\{b + J^*(t,x,-b,1)\}.
\end{aligned}$$

Denoting π^* as the Markovian optimal policy that achieves J^* in the last equality above, and letting $b^* = \operatorname{argmax}_{b\in\mathbb{R}}\{b + J^*(t,x,-b,1)\}$, we can construct a history-dependent policy of the original SDE (1)-(3) as follows:

$$\pi_0^*(a \mid \mathcal{F}_s^X) = \pi^*(a \mid s, X_s, B_{0,s}, B_{1,s}), \quad s\in[t,T], \tag{39}$$

where $(X_s, B_{0,s}, B_{1,s})$ follows the augmented SDE (5) starting from $X_t = x, B_{0,t} = -b^*, B_{1,t} = 1$. Hence, we have

$$\begin{aligned} J_0^{\pi_0^*}(t,x) &= \max_{b\in\mathbb{R}} \left\{ b + \mathbb{E}\big[\varphi\big(Z^{\pi_0^*}(t,x) - b\big)\big] \right\} + \tau \operatorname{Ent}(\pi; t, x) \\ &\geq b^* + \mathbb{E}\big[\varphi\big(Z^{\pi_0^*}(t,x) - b^*\big)\big] + \tau \operatorname{Ent}(\pi; t, x) \\ &= b^* + J^{\pi^*}(t, x, -b^*, 1) \\ &= b^* + J^*(t, x, -b^*, 1) \\ &= J_0^*(t,x), \end{aligned}$$

where we have used the definitions of J_0^π as (3) and J^π as (6), the relationship between π_0^* and π^* as (39), and the property of J_0^* derived above. This means π_0^* is the optimal policy for the original SDE (1)-(3), which concludes the proof.

Remark C.1. The optimal policy π_0^* depends on the starting time t of the SDE (1). In other words, for $0 < t_1 < t_2 < s < T$, the optimal policy at time s that attains $J_0^*(t_1, \cdot)$ is different from the one that attains $J_0^*(t_2, \cdot)$. This is because the augmented MDP includes the state b_1 which implicitly tracks the time duration from start. For notational simplicity, we omit the policy's dependence on the starting time t; we will always refer to an optimal policy along with an optimal value at a specific time t.

C.2 Proof of Theorem 4.1

If $\hat{J}^* = J^*$ and $\hat{q}^* = q^*$ are respectively the optimal value function and the optimal q-function, applying Itô's lemma to the process $J^*(s, X_s^\pi, Y_s^\pi, e^{-\delta(s-t)})$ yields

$$\begin{aligned} &J^*(s, X_s^\pi, Y_s^\pi, e^{-\delta(s-t)}) - J^*(t, x, 0, 1) - \int_t^s q^*(u, X_u^\pi, Y_u^\pi, e^{-\delta(u-t)}, a_u^\pi)\mathrm{d}u \\ &= \int_t^s \partial_x J^*(u, X_u^\pi, Y_u^\pi, e^{-\delta(u-t)}) \cdot \sigma(u, X_u^\pi, a_u^\pi)\mathrm{d}W_u, \end{aligned}$$

which is an $(\mathcal{F}, \mathbb{P})$-martingale.

Conversely, we consider the case where (9) is an $(\mathcal{F}, \mathbb{P})$-martingale, The second constraint in (8) implies that $\hat{\pi}^*(a \mid t, x, b_0, b_1) := \exp\{\hat{q}^*(t, x, b_0, b_1, a)/(\tau b_1)\}$ is a probability density function, and $\hat{q}^*(t, x, b_0, b_1, a) = \tau b_1 \log \hat{\pi}^*(a \mid t, x, b_0, b_1)$. Hence $\hat{q}^*(t, x, b_0, b_1, a)$ satisfies the second constraint in (30) with respect to the policy $\hat{\pi}^*$. It then follows from Theorem A.7 that $\hat{J}^*$ and $\hat{q}^*$ are respectively the value function and the q-function associated with $\hat{\pi}^*$. In addition, since $\hat{\pi}^*$ is the fixed point of the policy improvement iteration:

$$\hat{\pi}^*(a \mid t, x, b_0, b_1) = \exp\left\{ \frac{\hat{q}^*(t, x, b_0, b_1, a)}{\tau b_1} \right\} = \frac{\exp\{\hat{q}^*(t, x, b_0, b_1, a)/(\tau b_1)\}}{\int_{\mathcal{A}} \exp\{\hat{q}^*(t, x, b_0, b_1, a)/(\tau b_1)\}\mathrm{d}a},$$

we conclude that $\hat{\pi}^*$ is the optimal policy and thus $\hat{J}^*$ and $\hat{q}^*$ are the optimal value function and the optimal q-function, respectively.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There is no such research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.