To appear in *Quantitative Finance*, Vol. 00, No. 00, Month 20XX, 1–35

The Reinforcement Learning Kelly Strategy

R. JIANG†, D. SAUNDERS†, and C. WENG†*

†University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada

(*v3.0 released January 2022*)

The full Kelly portfolio strategy's deficiency in the face of estimation errors in practice can be mitigated by fractional or shrinkage Kelly strategies. This paper provides an alternative, the RL Kelly strategy, based on a reinforcement learning (RL) framework. RL algorithms are developed for the practical implementation of the RL Kelly strategy. Extensive simulation studies are conducted, and the results confirm the superior performance of the RL Kelly strategies.

Keywords: Kelly criterion; fractional Kelly strategy; portfolio selection; reinforcement learning.

JEL Classification: G11, C44, C61

1. Introduction

In the classical Merton portfolio optimization model (Merton 1971), an investor aims to maximize her utility by trading stocks and bonds. A common choice of the utility function is the log-utility. Maximizing the expected log-utility of the terminal portfolio value is the same as maximizing the expected log-return of the portfolio, and such optimality target is known ast the Kelly criterion (MacLean *et al.* 2010). For this criterion, the portfolio selection problem has been well studied, and closed-form solutions have been found in many models. It is well known that the full Kelly strategy, the optimal allocation strategy under the Kelly criterion, outperforms other strategies in terms of capital growth (MacLean *et al.* 2010). However, one important risk with the full Kelly strategy is that an investor may have to invest a large amount of money in stocks. This fact could lead to a substantial loss given a sequence of poor market returns.

Furthermore, the full Kelly strategy's optimality is sensitive to estimation errors. When estimation of model parameters is involved, the full Kelly strategy's empirical performance typically has a considerable deviation from the theoretical optimality results. As a remedy to mitigate the adverse effect of the estimation errors and improve the full Kelly strategy's performance, the fractional Kelly strategy was introduced. Under the fractional Kelly strategy, the portfolio weight is a fraction of that in the full Kelly strategy. Simulations and empirical evidence have shown the advantage of the fractional Kelly strategy over the full Kelly strategy (MacLean *et al.* 2010, 2011, Davis and Lleo 2013). Although there is no universal choice for the weight, a straightforward and common example of the fractional Kelly strategy is the half Kelly strategy (Nekrasov 2014, Han

We are grateful to two anonymous reviewers for several helpful comments and suggestions.
*Corresponding author. Email: chengguo.weng@uwaterloo.ca

et al. 2019), where the portfolio weight is half of that in the full Kelly strategy. This simple strategy can reduce the portfolio risk in a bad scenario significantly (Ziemba 2016). As a member of the fractional Kelly strategy family, the shrinkage Kelly strategy is also an alternative to adjust for parameter estimation error. One can either shrink the estimated expected stock return towards the risk-free return (Rising and Wyner 2012) or directly shrink the portfolio weight (Han *et al.* 2019). Another potential modification to the Kelly strategy is applying machine learning methods in portfolio selection problems. Shen *et al.* (2019) improve the Kelly strategy by ensemble learning. They combine the bootstrap aggregating algorithm and random subspace method to reduce the estimation risk at a single step for multivariate portfolios. The algorithm is sequentially applied to empirical data and shown to outperform several competing strategies.

In this paper, we apply reinforcement learning (RL) to tackle the practical challenge of the Kelly strategy when faced with unknown model parameters. In RL, agents take actions and receive rewards from the environment. They start with knowing very little about the environment and dynamically learn from interactions with the environment. Then they use the knowledge to maximize their rewards or objectives, e.g., expected log-return in the Kelly problem. This framework is more realistic than traditional portfolio selection models, where market parameters are assumed known a priori to investors. An important consideration in RL is to balance exploitation and exploration in the action process. At each decision point, the agent can either fully use the experience to execute the optimal action, i.e., exploiting the experience, or take a random action, i.e., exploring the environment. The benefit for the agent to explore is that more information about the environment is collected through exploration to find a better path towards higher long-term rewards. Wang *et al.* (2019) formulate the exploration-exploitation trade-off in a control problem. In particular, they adopt an entropy-regularization method to regularize the efforts in exploration and apply it in linear-quadratic control problems.

We apply the entropy-regularization RL framework to the Kelly portfolio problem. In this problem, we assume the underlying model dynamics are known to the investor (agent) to be a geometric Brownian motion, while the model parameters are unknown. The reward is the investment return from a given trading strategy, and the investor needs to learn how to find the optimal strategy to achieve the highest expected terminal log-return. We include a general time-varying temperature parameter in the regularization term to balance the degree of exploration and exploitation in the resulting RL algorithm. We consider both the Kelly portfolio problem for controlling the amount of investment in the stock, and the portion of wealth invested in the stock. The equivalence between the two formulations is not as apparent as the problem under the classical formulation. Indeed, given the same temperature parameter, they lead to different investment strategies for the two exploratory versions. In our study, we derive the optimal exploratory solution as a Gaussian distribution with parameters depending on time and portfolio wealth. By virtue of the derived closed-form solutions, we identify a relationship in the temperature parameter for the two exploratory versions to yield the same exploratory investment strategy and the same exploratory wealth process. It is worth noting that the resulting value functions are not identical even when we set the temperature parameters to yield the same exploratory investment strategy from both.

In our study, we consider three specific functional forms for the temperature parameter in the exploratory Kelly portfolio problems and develop implementable RL algorithms with the aid of the obtained closed-form solutions and value functions. The variance term in the Gaussian distribution of optimal control under the three functional forms of temperature parameter shows different time-varying patterns: increasing, constant, and decreasing over time. We call the resulting portfolio strategies the RL Kelly strategies. We apply the three RL algorithms in extensive simulation studies. In particular, we conduct a simulation study to confirm the convergence of our RL algorithms, and we then compare their performance with the fully Kelly, the fractional Kelly, and the shrinkage Kelly (by Han *et al.* (2019)) strategies. The simulation results show that the RL Kelly strategies

yield significantly better and more robust performance than these benchmark Kelly strategies even under model misspecification (when the stock actually follows a Heston's model). Thus, the RL Kelly strategy provides a practical improvement to those existing Kelly strategies.

The entropy-regularization RL framework has been applied to several investment problems in recent literature including Wang and Zhou (2019, 2020) and Dai *et al.* (2020). Wang and Zhou (2019, 2020) apply the RL method to a mean-variance problem. This method benefits the investor in the mean-variance space by achieving the target expected return faster than several other estimation methods. Dai *et al.* (2020) also adopt a mean-variance framework but base their analysis on the log-return of the portfolio and study the equilibrium solution, instead of the pre-commitment solution that Wang and Zhou (2019, 2020) investigate. The exploratory mean-variance problem in Dai *et al.* (2020) reduces to an exploratory Kelly problem as a special case, but our study in this paper differs from Dai *et al.* (2020) and makes contributions in several different aspects. First, while the control in Dai *et al.* (2020) is the investment portion of wealth, we study the exploratory Kelly problem by considering both the amount of investment in stock and the portion of wealth in stock as the control. We derive explicit solutions for both formulations. Second, while the discussion in Dai *et al.* (2020) mainly focuses on a constant temperature parameter and covers the case with exponentially decaying temperature parameter, our study for both formulations is for a general time-varying temperature parameter. Third, we clarify the condition for the two formulations to have the same exploratory investment strategy, which otherwise is not as apparent as their equivalence under the classical formulation. When the temperature parameters from the two formulations satisfy a certain equation (see equation (35)), both formulations will yield the same exploratory investment strategy.

The RL algorithm of Wang and Zhou (2020) addresses the minimum variance of the terminal wealth problem when the terminal mean is targeted. The RL algorithm of Dai *et al.* (2020) finds the equilibrium strategy under the log-mean-variance criterion. Our RL algorithms borrow the same idea of using temporal difference error to update parameters as in Wang and Zhou (2020) and Dai *et al.* (2020), but have a different design. First, their algorithms follow an episodic framework, where the updated values of model parameters from one episode are used as the initial values of parameters for the next episode. The investment strategy rule is updated only at the start of each episode and remains unchanged throughout each episode. In our paper, we interpret one episode as one investment time horizon considered for the Kelly problem. Our RL algorithms update the investment strategy over each trading period to make the strategy more practical. So, our algorithm is a one-step online algorithm. Second, in our simulation studies, we treat each episode independently and start from the same initial guess of model parameters in simulating for each episode. The independence of all episodes run in the simulation allows us to assess the performance of an RL algorithm applied over one episode.

The remainder of the paper is structured as follows. Section 2 introduces the classical Kelly criterion problem. Section 3 presents the exploratory Kelly problem that takes the dollar amount invested in stock as the control. Section 4 introduces the exploratory Kelly problem that uses the portion of wealth as the control. Section 5 creates RL algorithms, and section 6 contains the simulation studies. Section 7 concludes the paper. Appendix A collects the proofs for some theoretical results in the main body of the paper. Appendix B includes the RL algorithms and some technical details under two special time-decaying temperature parameters.

2. Kelly Criterion Problem

We consider a frictionless market and a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{0\leq t\leq T}, \mathbb{P})$ in a finite time horizon $[0, T]$. The market allows short-selling and leverage without extra cost. We assume

that there are only two assets in the market: one riskless asset (bond) B_t and one risky asset (stock) S_t. The risk-free interest rate is r, so that $dB_t = rB_t dt$. The stock price follows a geometric Brownian motion (GBM) with constant parameters μ and σ:

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

where $\{W_t, 0 \leq t \leq T\}$ is a standard Brownian motion.

An investor with an initial wealth of x_0 trades in the market to maximize her discounted terminal wealth. Denote the discounted amount invested in the stock at time t by $u_t \in \mathbb{R}$ and the corresponding discounted wealth process by $x_t \equiv x_t^u \in \mathbb{R}_+$. Under the self-financing condition, the discounted wealth process x_t satisfies:

$$dx_t^u = \rho\sigma u_t dt + \sigma u_t dW_t \tag{1}$$

where $\rho = \frac{\mu - r}{\sigma}$ is the Sharpe ratio of the stock. We assume the investor aims to establish a trading strategy according to the Kelly criterion. In other words, the investor's optimal trading strategy is solved from the following optimization problem:

$$\max_{u \in \mathcal{A}(0,x_0)} \mathbb{E}[U(x_T^u)] = \max_{u \in \mathcal{A}(0,x_0)} \mathbb{E}[\log x_T^u], \tag{2}$$

where U is the logarithmic function, i.e., $U(x) = \log x$ for $x \geq 0$, and $\mathcal{A}(0, x_0)$ is the set of admissible controls (i.e., $\mathbb{R}$-valued measurable $\mathcal{F}_t$-adapted and square integrable processes).

Through either the Hamilton-Jacobi-Bellman (HJB) equation (Merton (1971)), or the martingale method (e.g., Goll and Kallsen (2000)), we have the optimal strategy:

$$u_t^*(x) = \frac{\rho x}{\sigma}, \tag{3}$$

and the optimal wealth process:

$$dx_t^* = \rho^2 x_t^* dt + \rho x_t^* dW_t, \text{ or equivalently } x_t^* = x_0 e^{\frac{\rho^2}{2}t + \rho W_t}. \tag{4}$$

Thus, the optimal expected terminal log-return is given by:

$$\mathbb{E}[\log x_T^*] = \mathbb{E}\left[\log x_0 + \frac{\rho^2}{2}T + \rho W_T\right] = \log x_0 + \frac{\rho^2}{2}T. \tag{5}$$

3. Exploratory Kelly Amount Problem

In this section, we extend the classical Kelly criterion problem into an RL framework. We call it the exploratory version of the Kelly criterion problem, or simply, the exploratory (Kelly) problem. In the section, we use the amount of investment in stock as the control process and refer to the resulting exploratory problem as the "(exploratory Kelly) amount problem" when it becomes necessary or helpful to distinguish the exploratory formulation from the one using portion of wealth as the control process.

3.1. *Motivation*

The exploratory formulation in Wang *et al.* (2019), and Wang and Zhou (2019, 2020) is motivated by the trade-off between exploitation and exploration. In RL, when an agent is going to take an action, they will either exploit the current knowledge or explore the environment. If exploitation is selected, they will choose an action that maximizes the short-term rewards based on the experience so far. This optimal short-term action is also called the *greedy* action. However, a short-term greedy action may not always lead to the greatest long-term rewards due to the inadequacies of the current level of knowledge. Consequently, it is preferable to occasionally explore the environment randomly to improve the level of knowledge. The agent always faces a trade-off between exploitation and exploration when trying to accumulate the largest long-term reward. One approach to solving the RL problem is to employ an ε-*greedy* policy, i.e., choosing the greedy action with a probability of $1-\varepsilon$ and random actions with a probability of ε, for some $\varepsilon \in (0,1)$. A larger ε leads to more exploration, while a smaller ε favors employing the greedy strategy.

For our Kelly criterion problem, it seems that there is no need to explore the environment since the full Kelly strategy dominates other strategies. However, the domination has been found to fail in a practical scenario. Indeed, fractional Kelly strategies perform better than the full Kelly strategy in practice. This is partially due to the large bets of the full Kelly strategy, which would be very risky in a short period. Another reason is the estimation error of model parameters, i.e., μ and σ in our setting. Estimation errors of mean returns affect portfolio selection problems more than those of the covariances (Kallberg and Ziemba (1984)). A full Kelly strategy with biased estimates would be dominated by a fractional Kelly strategy with unbiased estimates (Han *et al.* (2019)). The deficient practical performance of the full Kelly strategy motivates us to apply exploration methods for higher rewards.

3.2. *Exploratory Wealth Process*

In the Kelly criterion problem, the greedy action is given by (3). In the exploratory problem, we consider random actions like in the ε-*greedy* scheme. Random actions could be formulated by a control distribution $\pi(u)$, i.e., every action is randomly drawn from the control distribution, and the control distribution should include the greedy action in a way that the greedy action has the highest chance to be executed. Therefore, we are now interested in finding the optimal control distribution instead of greedy actions.

In RL, a policy is how an agent behaves in different states (Sutton and Barto (2018)). In our scenario, the state corresponds to the current (discounted) wealth, and the policy corresponds to the control distribution $\pi(u)$. Given a control distribution π, every draw from it is a classical control like one in equation (3). Every classical control will receive a reward from the environment. Then, we could estimate the rewarding mechanism of the environment by drawing N classical controls. As N goes to infinity, we would be very close to the true rewarding mechanism.

Suppose at time t, we have a control distribution π_t and N independent sample classical controls u^i, $i = 1, 2, ..., N$, drawn from π_t. $\{x_t^i, t \in [0,T]\}$ is the wealth process under the control $\{u_t^i, t \in [0,T]\}$ for $i = 1, 2, \ldots, N$. The key idea is to view x_t^i as independent samples drawn from a new wealth process X_t^π. We denote X_t^π as the exploratory version of the controlled wealth process. This new wealth process, by the idea of RL, can be approached by sample paths x_t^i, $i = 1, 2, ..., N$ as N goes to infinity. Following the procedure in section 2.1 of Wang *et al.* (2019), we get the dynamics

of the exploratory wealth process X_t^π

$$\begin{aligned} dX_t^\pi &= \int_{\mathbb{R}} \rho\sigma u \pi_t(u)\, du dt + \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \pi_t(u) du}\, dW_t \\ &= \rho\sigma\mu_t\, dt + \sigma\sqrt{\mu_t^2 + \sigma_t^2}\, dW_t \\ &=: \alpha(\pi_t)\, dt + \beta(\pi_t)\, dW_t \end{aligned} \tag{6}$$

where

$$\alpha(\pi_t) = \rho\sigma\mu_t, \qquad \beta(\pi_t) = \sigma\sqrt{\mu_t^2 + \sigma_t^2} \tag{7}$$

and

$$\mu_t = \int_{\mathbb{R}} u\pi(u)\, du, \qquad \sigma_t^2 = \int_{\mathbb{R}} u^2 \pi_t(u)\, du - \mu_t^2. \tag{8}$$

We also assume $\mathbb{E}\left[\int_0^T \sqrt{\mu_t^2 + \sigma_t^2}\, dt\right] < \infty$ to ensure that X_t^π is well defined in (6).

If an agent invests following an RL policy $\{\pi_t\}$, then, with the highest chance, the agent would execute μ_t, i.e., the mean of the control distribution, and meanwhile the agent would have a chance to explore the environment by taking other possible actions. It is worth noting, however, that the agent's wealth is not fully explained by the exploratory wealth process. By empirically executing an RL policy, the wealth process is a realization from the draws of the control distribution and the stock price process. Both are random and independent of each other. The exploratory wealth process only incorporates the random effect from the stock price process, or equivalently from the Brownian motion W_t. The exploratory wealth process describes the average of wealth paths from a given exploratory investment strategy π_t.

3.3. *Trade-Off Between Exploitation and Exploration*

If the control distribution gives a larger probability mass to a single control rule, e.g., classical control, the agent will explore less and execute the single control rule more frequently. An extreme case of the control distribution is that it gives probability one to a single classical control. In this case, the agent would not explore anymore and the single control would be the optimal classical control. So, the control distribution needs to be regulated to maintain a certain degree of exploration in a learning procedure.

The need of regulating the level of exploration leads to the application of differential entropy, which has been widely used in information theory to measure a random variable's average level of uncertainty or information (Cover and Thomas (1991)). More uncertainty of the control distribution corresponds to a larger entropy. The differential entropy has indeed been used by Wang *et al.* (2019), Wang and Zhou (2020, 2019) and Dai *et al.* (2020) to regularize exploration for linear-quadratic control problems with uncertainty. In particular, it has been used by Wang and Zhou (2020) and Dai *et al.* (2020) for a continuous-time mean-variance portfolio allocation problem.

The entropy for control π_t is defined as:

$$\mathcal{H}(\pi_t) = -\int_{\mathbb{R}} \pi_t(u) \log\, \pi_t(u) du. \tag{9}$$

Because we study the problem in the time horizon $[0,T]$, the aggregated entropy of the control distribution process $\{\pi_t, t\in[0,T]\}$ is the integral of the differential entropy over the whole investment time horizon $[0,T]$.

Our exploratory Kelly amount problem modifies the classical optimization problem (2) by using the exploratory wealth process and an entropy regularization term. It is defined as follows:

$$\begin{aligned}&\max_{\pi\in\mathcal{A}(0,x_0)}\mathbb{E}\left[\log X_T^\pi+\int_0^T\lambda_a(t)\mathcal{H}(\pi_t)\,dt\right]\\ =&\max_{\pi\in\mathcal{A}(0,x_0)}\mathbb{E}\left[\log X_T^\pi-\int_0^T\lambda_a(t)\int_{\mathbb{R}}\pi_t(u)\log\pi_t(u)\,du\,dt\right]\end{aligned}\tag{10}$$

where the exogenous parameter $\lambda_a(t)>0$ regularizes the level of exploration and is called the temperature parameter in the RL literature. The temperature parameter balances exploitation and exploration in an RL framework. A larger $\lambda_a(t)$ encourages more exploration as the resulting control rule solved from (10) would lead to a larger value for the entropy $\mathcal{H}(\pi_t)$. We attach the subscript "a" in the notation $\lambda_a(t)$ to indicate the temperature parameter is for the exploratory amount problem, as in the subsequent sections we need to distinguish it from the temperature parameter for the exploratory portion problem, which is formulated with the portion of wealth as the control.

$\mathcal{A}(0,x_0)$ in (10) is the set of admissible control distribution processes on $[0,T]$. For fixed $(s,x)\in[0,T]\times\mathbb{R}_+$, a control distribution process $\pi=\{\pi_t, s\le t\le T\}$ belongs to $\mathcal{A}(s,x)$ if (Wang and Zhou (2020))

(1) for each $s\le t\le T$, $\pi_t\in\mathcal{P}(\mathbb{R})$ almost surely, where $\mathcal{P}(\mathbb{R})$ denotes the set of $\mathbb{R}$-valued probability density functions;
(2) for each $\mathbb{A}\in\mathcal{B}(\mathbb{R})$, $\{\int_{\mathbb{A}}\pi_t(u)\,du,\ s\le t\le T\}$ is $\mathcal{F}_t$-progressively measurable;
(3) $\mathbb{E}\left[\int_s^T\sqrt{\mu_t^2+\sigma_t^2}\,dt\right]<\infty$, $0\le s\le T$;
(4) $\mathbb{E}\left[|\log X_T^\pi-\int_s^T\lambda_a(t)\int_{\mathbb{R}}\pi_t(u)\log\pi_t(u)\,du\,dt|\ \middle|\ X_s^\pi=x\right]<\infty$, $0\le s\le T$.

Since π_t is a probability density for all $t\in[0,T]$, it must satisfy

$$\pi_t(u)\ge 0,\ \text{for all } u\in\mathbb{R}\ \text{ and }\ \int_{\mathbb{R}}\pi_t(u)du=1.$$

Below we discuss the solution of the exploratory amount problem (10). The value function of this optimization problem is defined as:

$$V^a(t,x;\lambda_a(t))=\max_{\pi\in\mathcal{A}(t,x)}\mathbb{E}\left[\log X_T^\pi-\int_t^T\lambda_a(v)\int_{\mathbb{R}}\pi_v(u)\log\pi_v(u)\,du\,dv\ \middle|\ X_t^\pi=x\right].\tag{11}$$

A standard application of the Dynamic Programming Principle yields the following HJB equation for the value function:

$$v_t(t,x)+\max_{\pi_t\in\mathcal{P}(\mathbb{R})}\left\{\alpha(\pi_t)v_x(t,x)+\frac{1}{2}\beta^2(\pi_t)v_{xx}(t,x)-\lambda_a(t)\int_{\mathbb{R}}\pi_t(u)\log\pi_t(u)du\right\}=0\tag{12}$$

with terminal condition $v(T,x)=\log x$. In the above, v_t, v_x and v_{xx} denote the corresponding partial derivatives of the function $v(t,x)$.

THEOREM 3.1 *The maximization problem in equation* (12) *possesses the following density function as a solution:*

$$\pi_t^*(u;x,\lambda_a(t)) = \frac{\exp\left\{\frac{1}{\lambda_a(t)}\left[\frac{1}{2}\sigma^2 v_{xx}(t,x)u^2 + \rho\sigma v_x(t,x)u\right]\right\}}{\int_{\mathbb{R}}\exp\left\{\frac{1}{\lambda_a(t)}\left[\frac{1}{2}\sigma^2 v_{xx}(t,x)u^2 + \rho\sigma v_x(t,x)u\right]\right\}du}. \tag{13}$$

Proof. See Appendix A.1. □

Equation (13) indicates that $\pi_t^*(u;x,\lambda_a(t))$ is a Gaussian density if $v_{xx}(t,x) < 0$, which actually holds as we can tell shortly from Theorem 3.2.

Now that

$$\mu_t^* := \int_{\mathbb{R}} u\pi_t^*(u)\,du = -\frac{\rho v_x(t,x)}{\sigma v_{xx}(t,x)}$$

and

$$\sigma_t^* := \sqrt{\int_{\mathbb{R}} u^2\pi_t^*(u)du - (\mu_t^*)^2} = \sqrt{-\frac{\lambda_a(t)}{\sigma^2 v_{xx}(t,x)}},$$

substituting these into the expressions for $\alpha(\pi_t)$ and $\beta(\pi_t)$ in (7) and using the control $\pi_t^*(u;x,\lambda_a(t))$ in equation (13), we can simplify the HJB equation (12) into:

$$v_t(t,x) - \frac{\rho^2 v_x^2(t,x)}{2v_{xx}(t,x)} - \frac{\lambda_a(t)}{2}\log\left(-\frac{\sigma^2 v_{xx}(t,x)}{2\pi\lambda_a(t)}\right) = 0 \tag{14}$$

with $v(T,x) = \log x$. The above partial differential equation (PDE) is actually the Merton-type PDE in the classical optimization problem plus a term resulting from the entropy penalization. A similar PDE arises in the mean-variance portfolio allocation problem in Wang and Zhou (2020) but with a different terminal condition.

Define the real-valued function

$$f(t) = 1 + \int_t^T \lambda_a(s)ds, \;\; t \in [0,T], \tag{15}$$

and let g_a be a real-valued function with $g_a(T) = 0$ and derivative:

$$g_a'(t) = -\frac{\rho^2}{2}f(t) + \frac{\lambda_a(t)}{2}\log\frac{\sigma^2 f(t)}{2\pi\lambda_a(t)}. \tag{16}$$

THEOREM 3.2 *For the exploratory optimization problem* (10)*,*

(a) the value function is

$$V^a(t,x;\lambda_a(t)) = f(t)\log x + g_a(t), \tag{17}$$

(b) the optimal control follows a Gaussian distribution

$$\pi_t^*(u; x, \lambda_a(t)) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho x}{\sigma},\ \frac{x^2\lambda_a(t)}{\sigma^2 f(t)}\right), \tag{18}$$

(c) the exploratory wealth process $X_t^{\pi^*}$ under the optimal control π^* satisfies

$$dX_t^{\pi^*} = \rho^2 X_t^{\pi^*} dt + \sqrt{\rho^2 + \frac{\lambda_a(t)}{f(t)}} \cdot X_t^{\pi^*} dW_t, \tag{19}$$

or equivalently,

$$X_t^{\pi^*} = x_0 \exp\left\{\frac{\rho^2}{2}t + \frac{1}{2}\log\frac{1+\int_t^T \lambda_a(s)ds}{1+\int_0^T \lambda_a(s)ds} + \int_0^t \sqrt{\rho^2 + \frac{\lambda_a(s)}{f(s)}} dW_s\right\}, \tag{20}$$

(d) the expected terminal log-return is

$$\mathbb{E}[\log X_T^{\pi^*}] = \log x_0 + \frac{\rho^2}{2}T - \frac{1}{2}\log\left(1 + \int_0^T \lambda_a(s)ds\right), \tag{21}$$

(e) the relative loss of expected terminal log-return is

$$\frac{\mathbb{E}[\log x_T^*] - \mathbb{E}[\log X_T^{\pi^*}]}{\mathbb{E}[\log x_T^*]} = \frac{\log(1+\int_0^T \lambda_a(s)ds)}{2\log x_0 + \rho^2 T}, \tag{22}$$

where $\mathbb{E}[\log x_T^*]$ is the expected log-return of terminal portfolio under the classical Kelly strategy and given in equation (5).

Proof. See Appendix A.2 □

According to part (b) of Theorem 3.2, the exploratory optimal control π^* for the amount problem (10) is centered at the classical optimal control u^* given in equation (3). The variance term in the optimal Gaussian distribution control depends on

$$\frac{\lambda_a(t)}{f(t)} = \frac{\lambda_a(t)}{1+\int_t^t \lambda_a(s)ds}.$$

The variance term determines the level of exploration. So, Theorem 3.2 sheds important light on how different time-varying properties of the control distribution can be designed by using different time-decaying $\lambda_a(t)$, or equivalently $f(t)$. However, not all time-decaying temperature processes lead to a time-decaying variance, for example, linearly decaying $\lambda_a(t)$ (see Appendix A.3) and exponentially decaying $\lambda_a(t)$ (see next subsection). The following proposition suggests the conditions of $f(t)$ for an appropriate temperature process.

PROPOSITION 3.3 *A temperature process $\lambda_a(t)$ can be characterized by $f(t)$ as $\lambda_a(t) = -f'(t)$. A necessary and sufficient condition for $\lambda_a(t)$ to be time-decaying and lead to a time-decaying variance is that $f(t)$ is strictly log-convex.*

Proof. The proof is straightforward from the definitions of $\lambda_a(t)$ and the variance of the control distribution, $\frac{x^2\lambda_a(t)}{\sigma^2 f(t)}$. □

3.4. *Exploratory Solutions under Several Specific Temperature Parameters*

In this section, we present results for three specific forms of temperature parameter. The variance term shows different time-varying patterns over time under these three forms.

3.4.1. **Constant $\lambda_a(t)$.** When the temperature parameter for problem (10), $\lambda_a(t)$, is a constant $\lambda > 0$, equations (15) and (16), together with the terminal condition $g_a(T) = 0$, imply

$$f(t) = 1 + (T-t)\lambda$$

and

$$g_a(t) = -\int_t^T \left[-\frac{\rho^2}{2}[1+\lambda(T-s)] + \frac{\lambda}{2}\log\frac{\sigma^2}{2\pi\lambda} + \frac{\lambda}{2}\log\left[1+(T-s)\lambda\right]\right] ds.$$

Computing the integral for g_a and applying Theorem 3.2, we get the value function given by

$$\begin{aligned} V^a(t,x;\lambda) =& [1+\lambda(T-t)]\log x - \frac{1+\lambda(T-t)}{2}\log\left[1+\lambda(T-t)\right] \\ & - \frac{\lambda\rho^2}{4}(T^2-t^2) + \left[\frac{\rho^2}{2} + \frac{\lambda}{2}\left(\rho^2 T - \log\frac{\sigma^2}{2\pi e\lambda}\right)\right](T-t), \end{aligned} \tag{23}$$

and the optimal exploratory amount control given by

$$\pi_t^*(u;x,\lambda) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho x}{\sigma},\ \frac{\lambda x^2}{\sigma^2(1+\lambda(T-t))}\right). \tag{24}$$

3.4.2. **Power-Decaying $\lambda_a(t)$.** In practice, as the agent collects more information from the environment, their attitude towards exploration may change over time. In light of this, state-dependent or time-dependent temperature parameters have been adopted in the literature (Ishii *et al.* (2002); Wang and Zhou (2020); Dai *et al.* (2020)). For our exploratory amount problem, one feasible temperature process is the power-decaying $\lambda_a(t)$ defined as follows:

$$\lambda_a(t) = \lambda_0 \frac{(T+\lambda_1)^{\lambda_0}}{(t+\lambda_1)^{\lambda_0+1}} \tag{25}$$

with constants $\lambda_0 > 0$ and $\lambda_1 > 0$. Its corresponding $f(t)$ is $f(t) = (\frac{T+\lambda_1}{t+\lambda_1})^{\lambda_0}$, which is log-convex (and therefore convex). Under this particular power-decaying temperature process, the

value function for the amount problem (10) is given by

$$V^a(t,x;\lambda_a(t)) = \begin{cases} (\frac{T+\lambda_1}{t+\lambda_1})^{\lambda_0}\log x + \frac{\rho^2(T+\lambda_1)}{2(\lambda_0-1)}\left[(\frac{T+\lambda_1}{t+\lambda_1})^{\lambda_0-1}-1\right] \\ \quad -\frac{1}{2\lambda_0}\left[(\frac{T+\lambda_1}{t+\lambda_1})^{\lambda_0}-1\right] - \frac{1}{2}(\frac{T+\lambda_1}{t+\lambda_1})^{\lambda_0}\log\frac{\sigma^2(t+\lambda_1)}{2\pi\lambda_0} + \frac{1}{2}\log\frac{\sigma^2(T+\lambda_1)}{2\pi\lambda_0}, & \lambda_0 \neq 1 \\ \\ \frac{T+\lambda_1}{t+\lambda_1}\log x + \frac{\rho^2(T+\lambda_1)}{2}\log\frac{T+\lambda_1}{t+\lambda_1} - \frac{T+\lambda_1}{2(t+\lambda_1)} + \frac{1}{2} \\ \quad -\frac{T+\lambda_1}{2(t+\lambda_1)}\log\frac{\sigma^2(t+\lambda_1)}{2\pi} + \frac{1}{2}\log\frac{\sigma^2(T+\lambda_1)}{2\pi}, & \lambda_0 = 1 \end{cases} \tag{26}$$

and the optimal control is given by

$$\pi_t^*(u;x,\lambda_a(t)) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho x}{\sigma},\ \frac{\lambda_0 x^2}{\sigma^2(t+\lambda_1)}\right) \tag{27}$$

which has a time-decreasing variance $\frac{\lambda_0 x^2}{\sigma^2(t+\lambda_1)}$.

3.4.3. **Exponentially Decaying $\lambda_a(t)$.** Consider the optimization problem (10) with an exponentially decaying temperature parameter

$$\lambda_a(t) = \lambda_0 e^{\lambda_0(T-t)}, \text{ where } \lambda_0 > 0.$$

Applying Theorem 3.2, we get the value function

$$V^a(t,x;\lambda_a(t)) = e^{\lambda_0(T-t)}\log x + \left(\frac{\rho^2}{2\lambda_0} - \frac{1}{2}\log\frac{\sigma^2}{2\pi\lambda_0}\right)\left(e^{\lambda_0(T-t)}-1\right),$$

and the optimal amount control

$$\pi_t^*(u;x,\lambda_a(t)) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho x}{\sigma},\ \frac{\lambda_0 x^2}{\sigma^2}\right)$$

which has a time-constant variance $\frac{\lambda_0 x^2}{\sigma^2}$.

4. Exploratory Kelly Portion Problem

In portfolio selection problems, the investor can either control the amount of wealth or the proportion of wealth invested in the stock. In the classical problem, these two choices are equivalent, producing the same optimal strategy, value function, and expected terminal log-return. In section 3, the exploratory problem is formulated in terms of controlling the amount invested in the stock and is regularized using the differential entropy (9). Now, we revisit the problem with a regularization based on the portion of wealth invested in the stock. We call the resulting control problem the "exploratory (Kelly) portion problem" to it from distinguish the formulation based on the amount of investment.

Recall that $u_t \in \mathbb{R}$ represents the amount of wealth invested in the stock. Given that total wealth is x_t, the portion of wealth invested in the stock is therefore $z_t = u_t/x_t \in \mathbb{R}$. We denote the

associated control distribution for the portion of wealth by $\phi_t(\cdot)$ and still use $\pi_t(\cdot)$ for the control distribution of the amount of wealth. Then, they relate to each other by:

$$\phi_t(z) = \pi_t(zx_t)x_t, \quad z \in \mathbb{R}.$$

Therefore the entropy regularizing the new control distribution is:

$$\mathcal{H}(\phi_t) = -\int_{\mathbb{R}} \phi_t(z) \log \phi_t(z)\, dz = \mathcal{H}(\pi_t) - \log x_t, \tag{28}$$

which includes the extra term "$\log x_t$". For a solution to the exploratory problem with the entropy applied to ϕ_t, we can apply the above relationship and solve the problem via π_t:

$$\begin{aligned} V^p(t,x;\lambda_p(t)) &:= \max_{\phi\in\mathcal{A}(t,x)} \mathbb{E}\left[\log X_T^\phi + \int_t^T \lambda_p(v)\mathcal{H}(\phi_v)\, dv \;\middle|\; X_t^\phi = x\right] \\ &= \max_{\pi\in\mathcal{A}(t,x)} \mathbb{E}\left[\log X_T^\pi + \int_t^T \lambda_p(v)(\mathcal{H}(\pi_v) - \log X_v^\pi)\, dv \;\middle|\; X_t^\pi = x\right] \\ &= \max_{\pi\in\mathcal{A}(t,x)} \mathbb{E}\left[\log X_T^\pi - \int_t^T \lambda_p(v)\int_{\mathbb{R}} \pi_v(z)\log\pi_v(z)\, dz\, dv - \int_t^T \lambda_p(v)\log X_v^\pi\, dv \;\middle|\; X_t^\pi = x\right]. \end{aligned} \tag{29}$$

Thus, V^p satisfies the HJB equation

$$v_t(t,x) + \max_{\phi_t\in\mathcal{P}(\mathbb{R})}\left\{\alpha(\phi_t)xv_x(t,x) + \frac{1}{2}\beta^2(\phi_t)x^2v_{xx}(t,x) - \lambda_p(t)\int_{\mathbb{R}}\phi_t(z)\log\phi_t(z)\, dz\right\} = 0$$

or equivalently,

$$v_t(t,x) + \max_{\pi_t\in\mathcal{P}(\mathbb{R})}\left\{\alpha(\pi_t)v_x(t,x) + \frac{1}{2}\beta^2(\pi_t)v_{xx}(t,x) - \lambda_p(t)\int_{\mathbb{R}}\pi_t(z)\log\pi_t(z)\, dz - \lambda_p(t)\log x\right\} = 0$$

with terminal condition $v(T,x) = \log x$.

Equations (28) and (29) demonstrate why the solutions turn out not to be equivalent when using the amount of investment and using the portion of wealth as the control in the exploratory Kelly problem while they are under the classical formulation. The distribution for the portion variable yields a smaller entropy compared with that for the amount variable.

Let g_p be a real function with derivative

$$g_p'(t) = -\frac{\rho^2}{2} + \frac{\lambda_p(t)}{2}\log\frac{\sigma^2}{2\pi\lambda_p(t)},$$

and terminal condition $g_p(T) = 0$.

THEOREM 4.1 *For the exploratory optimization problem* (29)*,*

(a) the value function is

$$V^p(t,x;\lambda_p(t)) = \log x + g_p(t), \tag{30}$$

(b) the optimal control follows a Gaussian distribution

$$\phi_t^*(u; x, \lambda_p(t)) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho}{\sigma},\ \frac{\lambda_p(t)}{\sigma^2}\right), \tag{31}$$

(c) the exploratory wealth process $X_t^{\phi^*}$ under the optimal control ϕ^* satisfies

$$dX_t^{\phi^*} = \rho^2 X_t^{\phi^*} dt + \sqrt{\rho^2 + \lambda_p(t)} \cdot X_t^{\phi^*} dW_t, \ \ \textit{with}\ \ X_0^{\phi^*} = x_0,$$

or equivalently,

$$X_t^{\phi^*} = x_0 \exp\left\{\frac{\rho^2}{2}t - \frac{1}{2}\int_0^t \lambda_p(s)ds + \int_0^t \sqrt{\rho^2 + \lambda_p(s)}dW_s\right\}, \tag{32}$$

(d) the expected terminal log-return is

$$\mathbb{E}[\log X_T^{\phi^*}] = \log x_0 + \frac{\rho^2}{2}T - \frac{1}{2}\int_0^T \lambda_p(s)ds, \tag{33}$$

(e) the relative loss of expected terminal log-return is

$$\frac{\mathbb{E}[\log x_T^*] - \mathbb{E}[\log X_T^{\phi^*}]}{\mathbb{E}[\log x_T^*]} = \frac{\int_0^T \lambda_p(s)ds}{2\log x_0 + \rho^2 T}. \tag{34}$$

Proof. The proof is parallel to that of Theorem 3.2 and hence, omitted. □

Remark 1 A comparison between Theorems 3.2 and 4.1 leads to the following interesting observations:

(a) The variance term in the optimal control distribution is more directly related to the temperature parameter for the portion problem than the amount problem. The variance term is proportional to the temperature parameter for the portion problem.

When $\lambda_p(t)$ is set to be identical to $\lambda_a(t)$, the variance of the investment amount is larger in the portion problem than in the amount problem (see equations (18) and (31)) since $f(t) = 1 + \int_t^T \lambda_a(s)ds \geq 1$. Accordingly, the expected terminal log-return is smaller, and the relative loss is larger in the portion problem. A smaller exploratory variance for the amount solution is attributed to its relatively smaller magnitude in the entropy term. As indicated in equation (29), its entropy is smaller than that of the corresponding amount control by $\log x_t$, and therefore, its inclusion in the optimality objective discourages exploration compared with the formulation based on the amount variable.

(b) Equations (22) and (34) indicate that, as long as the temperature parameter is set to decrease to zero over time, the relative loss in expected terminal log-return diminishes to zero when the investment time horizon becomes infinitely long.

(c) If we set temperature parameters in the two exploratory Kelly problems to satisfy

$$\lambda_p(t) = \frac{\lambda_a(t)}{1 + \int_t^T \lambda_a(s)ds}, \ \ t \in [0, t], \tag{35}$$

then, both problems yield the same exploratory wealth process (see equations (20) and (32)). Furthermore, under the condition (35), the optimal amount control is equivalent to the optimal portion control in the sense that both are Gaussian distributions and the parameters for one are scaled by the portfolio wealth x from the other. However, it is worth noting that the value functions differ between the two exploratory Kelly problems even if the temperature parameters satisfy condition (35).

For the three temperature processes $\lambda_a(t)$ given in section 3.4 for the amount problem, the equivalent temperature process $\lambda_p(t)$ for the portion problem is respectively as follows:

1. **Constant $\lambda_a(t)$.** When the temperature parameter for problem (10), $\lambda_a(t)$, is a constant $\lambda > 0$, the temperature parameter for the equivalent portion control problem (29) is given by

$$\lambda_p(t) = \frac{\lambda}{1 + \lambda(T-t)}.$$

In this case, the value function for the portion problem is given by

$$V^p(t, x; \lambda_p(t)) = \log x + \frac{\rho^2}{2}(T-t) + \frac{1}{4}\left[\left(\log\frac{\sigma^2}{2\pi\lambda}\right)^2 - \left(\log\frac{\sigma^2(1+\lambda(T-t))}{2\pi\lambda}\right)^2\right]$$

and the optimal exploratory portion control is

$$\phi_t^*(u; x, \lambda_p(t)) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho}{\sigma},\ \frac{\lambda}{\sigma^2(1+\lambda(T-t))}\right).$$

2. **Power-Decaying $\lambda_a(t)$.** When the temperature for the amount problem is $\lambda_a(t) = \lambda_0 \frac{(T+\lambda_1)^{\lambda_0}}{(t+\lambda_1)^{\lambda_0+1}}$ with constants $\lambda_0 > 0$ and $\lambda_1 > 0$, the equivalent temperature parameter for the portion problem is given by

$$\lambda_p(t) = \frac{\lambda_0}{t + \lambda_1}.$$

In this case, the value function for the portion problem is given by

$$V^p(t, x; \lambda_p(t)) = \log x + \frac{\rho^2}{2}(T-t) + \frac{\lambda_0}{4}\left[\left(\log\frac{\sigma^2(t+\lambda_1)}{2\pi\lambda_0}\right)^2 - \left(\log\frac{\sigma^2(T+\lambda_1)}{2\pi\lambda_0}\right)^2\right]$$

and the optimal exploratory portion control is

$$\phi_t^*(u; x, \lambda_p(t)) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho}{\sigma},\ \frac{\lambda_0}{\sigma^2(t+\lambda_1)}\right).$$

3. **Exponentially Decaying $\lambda_a(t)$.** When the temperature for the amount problem is $\lambda_a(t) = \lambda_0 e^{\lambda_0(T-t)}$ with constant $\lambda_0 > 0$, the equivalent temperature parameter for the portion problem is given by $\lambda_p(t) = \lambda_0$. In this case, the value function for the portion problem is given by

$$V^p(t, x; \lambda_p(t)) = \log x + \left(\frac{\rho^2}{2} - \frac{\lambda_0}{2}\log\frac{\sigma^2}{2\pi\lambda_0}\right)(T-t),$$

and the optimal exploratory portion control is

$$\phi_t^*(u; x, \lambda_p(t)) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho}{\sigma},\ \frac{\lambda_0}{\sigma^2}\right).$$

5. Exploratory RL Algorithms

5.1. *Policy Improvement*

In a common RL problem, an agent learns from the environment through iterations between policy evaluation and policy improvement (Sutton and Barto (2018)). We previously formulated the procedure to find the optimal value function using the exploratory version of the wealth process and entropy regularization. In this subsection, we study the policy improvement to complete the essential iterations in an RL framework. We will focus on the amount problem (10) and consider a constant temperature parameter because the results can be obtained in parallel for the portion problem (29) and for both problems with a general time-varying temperature parameter.

For a policy of a particular type, the following theorem guarantees that it could be improved to a Gaussian policy. The theorem is modified from Wang and Zhou (2020) to our scenario of the Kelly criterion problem.

THEOREM 5.1 *Suppose π_t is an admissible control policy and $V^\pi(t,x)$, $(t,x) \in [0,T] \times \mathbb{R}_+$ is its corresponding value function satisfying $V^\pi_{xx}(t,x) < 0$. Suppose a new control policy defined as*

$$\tilde{\pi}_t(u; x) \sim \mathcal{N}\left(u \;\middle|\; -\frac{\rho V^\pi_x(t,x)}{\sigma V^\pi_{xx}(t,x)},\ -\frac{\lambda}{\sigma^2 V^\pi_{xx}(t,x)}\right) \tag{36}$$

is also admissible under the same choice of λ. Then we have $V^{\tilde{\pi}}(t,x) \geq V^\pi(t,x)$. That is, we can improve policy π_t by an admissible Gaussian policy (36).

Proof. See Appendix A.4. □

Since Theorem 5.1 suggests that the improved policy is Gaussian, below we illustrate how exactly we improve Gaussian policies to the form of (24). The improvement could be achieved by updating only the parameters of the Gaussian control, i.e., the mean and the variance. Assume we start with a simple Gaussian control:

$$\pi_t^0(u; x) \sim \mathcal{N}\left(u \;\middle|\; \beta_1 x,\ \frac{cx^2}{1 + b(T-t)}\right) \tag{37}$$

with $c > 0$ and $b > 0$ guaranteeing a positive variance. To apply the update in (36), we shall calculate the value function $V^{\pi^0}(t,x)$ and its derivatives. We start from the PDE:

$$V_t^{\pi^0}(t,x) + \int_{\mathbb{R}} \left[\rho\sigma u V_x^{\pi^0}(t,x) + \frac{1}{2}\sigma^2 u^2 V_{xx}^{\pi^0}(t,x) - \lambda \log \pi_t^0(u)\right] \pi_t^0(u)\, du = 0, \text{ with } V^{\pi^0}(T,x) = \log x.$$

Substituting the form of the control distribution π^0 into the above PDE yields:

$$V_t^{\pi^0} + \rho\sigma\beta_1 x V_x^{\pi^0} + \frac{1}{2}\sigma^2\left(\beta_1^2 + \frac{c}{1 + b(T-t)}\right) x^2 V_{xx}^{\pi^0} + \frac{\lambda}{2}\log\frac{2\pi e c x^2}{1 + b(T-t)} = 0 \tag{38}$$

with the terminal condition $V^{\pi^0}(T,x) = \log x$. Following the same procedure as for solving PDE (14), we can obtain a solution to the above PDE:

$$\begin{aligned} V^{\pi^0}(t,x) = {} & [1+\lambda(T-t)]\log x - \left[\frac{\lambda+\sigma^2 c(1-\lambda/b)}{2b} + \frac{\lambda}{2}(T-t)\right]\log\left[1+b(T-t)\right] \\ & + \left[\left(\rho\sigma\beta_1 - \frac{1}{2}\sigma^2\beta_1^2\right)(1+\lambda T) + \frac{\lambda}{2}\log 2\pi c e^2 - \frac{\lambda\sigma^2 c}{2b}\right](T-t) \\ & - \frac{\lambda}{2}\left(\rho\sigma\beta_1 - \frac{1}{2}\sigma^2\beta_1^2\right)(T^2-t^2). \end{aligned} \tag{39}$$

Calculating and substituting the corresponding derivatives of the above $V^{\pi^0}(t,x)$ into equation (36), we obtain the update:

$$\pi_t^1(u;x) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho x}{\sigma},\ \frac{\lambda x^2}{\sigma^2(1+\lambda(T-t))}\right) \tag{40}$$

which is exactly the optimal control defined in equation (24).

However, it is worth noting that the above two-step procedure is not a directly implementable scheme for policy improvement since it requires the true model parameter values. The value function V^{π^0} depends on the true values of the parameters σ and ρ. This motivates us to develop iterative algorithms to update our belief on the model parameters over time. The iterative algorithms will be discussed in the following section.

5.2. *Temporal Difference Error Minimization Algorithm*

The previous discussion about the policy evaluation and policy improvement completes the requirements of RL procedures. In this subsection we build algorithms for the exploratory optimization problems. Our discussion will be focused on the exploratory amount problem with a constant temperature parameter λ. Algorithms for extensions to a time-varying temperature parameter and to the portion problem follow in the same fashion. The design of the algorithms is adapted from Wang and Zhou (2020). However, our algorithm is an one-step online algorithm, different from the offline algorithm used in Wang and Zhou (2020).

Theorem 5.1 suggests that the main task is to update model parameters since the optimal controls are from the Gaussian distribution family. We parametrize the value function V as $V^\pi(t,x;\boldsymbol{\alpha})$ and control $\pi_t(u)$ as $\pi_t(u;\boldsymbol{\beta})$ to facilitate the discussion of parameter updating in the algorithm, where $\boldsymbol{\alpha} = (\alpha_1,\alpha_2)$ and $\boldsymbol{\beta} = (\beta_1,\beta_2)$ with each element of the two vectors specified later. The temperature parameter λ represents the weight an agent puts in exploration against exploitation. So λ is exogenous and pre-specified by the agent.

In view of the discussion following Theorem 5.1 in section 5.1, we start with a simple Gaussian distribution $\pi_t(u;\boldsymbol{\beta})$ parametrized by $\boldsymbol{\beta}$, with mean $\beta_1 x$ and variance

$$\frac{x^2 e^{-2\beta_2-1}}{2\pi[1+\lambda(T-t)]} \tag{41}$$

for some constants $\beta_1 < 0$ and $\beta_2 > 0$. The parametrization for the above variance is to get a neat

expression for its entropy:

$$-\int_{\mathbb{R}} \pi_t(u;\boldsymbol{\beta}) \log \pi_t(u;\boldsymbol{\beta})\, du = \log x - \frac{1}{2}\log\left[1+\lambda(T-t)\right] - \beta_2. \tag{42}$$

Recall that the value function under the Gaussian control in (37) is given by (39). So, for the control $\pi_t(u;\boldsymbol{\beta})$, we set $c = e^{-2\beta_2-1}$ and $b = \lambda$ in (39) to get the value function as follows:

$$\begin{aligned}
&[1+\lambda(T-t)]\log x - \frac{1+\lambda(T-t)}{2}\log\left[1+\lambda(T-t)\right] \\
&+\left[\left(\rho\sigma\beta_1 - \frac{1}{2}\sigma^2\beta_1^2\right)(1+\lambda T) + \frac{\lambda}{2}\log 2\pi c e^2 - \frac{\sigma^2\beta_2}{2}\right](T-t) \\
&-\frac{\lambda}{2}\left(\rho\sigma\beta_1 - \frac{1}{2}\sigma^2\beta_1^2\right)(T^2-t^2).
\end{aligned} \tag{43}$$

On the other hand, equation (23) suggests the following form for the value function:

$$V^{\pi}(t,x;\boldsymbol{\alpha}) = [1+\lambda(T-t)]\log x - \frac{1+\lambda(T-t)}{2}\log\left[1+\lambda(T-t)\right] + \alpha_1(t^2-T^2) + \alpha_2(t-T), \tag{44}$$

for some constants α_1 and α_2.

In the iterative algorithm, we start from some initialized values for the model parameters ρ and σ. The specification of these two parameter values would give us initial values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The initial values for $\boldsymbol{\beta}$ can be obtained through comparing the mean $\beta_1 x$ and the variance term in (41) with the counterparts in equation (24). The initial values of $\boldsymbol{\alpha}$ can be derived by comparing the above parametric form of $V^{\pi}(t,x;\boldsymbol{\alpha})$ with that of $V(t,x;\lambda)$ in equation (23).

Given initial values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we now discuss how to update the parameters iteratively. We update the parameter β_1 using a heuristic relationship with the other parameters α_1, α_2 and β_2. We update α_1, α_2 and β_2 through a minimization procedure using the gradient descent algorithm that we will describe in detail later.

For the update of β_1, we note that $\beta_1 x$ is the mean in the proposed Gaussian policy and $\frac{\rho}{\sigma}x$ is the mean in the optimal Gaussian policy. So, we work to find an expression for $\frac{\rho}{\sigma}$ in terms of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We first calculate the entropy under the optimal control distribution π_t^* in (24):

$$-\int_{\mathbb{R}} \pi_t^*(u;\boldsymbol{\beta}) \log \pi_{t_i}^*(u;\boldsymbol{\beta})\, du = \log x - \frac{1}{2}\log\left[1+\lambda(T-t)\right] + \frac{1}{2}\log\frac{2\pi e\lambda}{\sigma^2} \tag{45}$$

and then compare its expression with (42) to get:

$$\beta_2 = -\frac{1}{2}\log\frac{2\pi e\lambda}{\sigma^2}, \text{ or equivalently } \sigma^2 = 2\pi\lambda e^{2\beta_2+1}. \tag{46}$$

We consequently equate the coefficients for t^2 in equations (43) and (44) to obtain

$$\alpha_1 = \frac{\lambda}{2}\left(\rho\sigma\beta_1 - \frac{1}{2}\sigma^2\beta_1^2\right)$$

which, along with (46), gives

$$\frac{\rho}{\sigma} = \frac{2\alpha_1}{\lambda\sigma^2\beta_1} + \frac{1}{2}\beta_1 = \frac{\alpha_1 e^{-2\beta_2-1}}{\pi\lambda^2\beta_1} + \frac{1}{2}\beta_1. \tag{47}$$

Since $\beta_1 x$ is the mean in the proposed Gaussian policy and $\frac{\rho}{\sigma}x$ is the mean in the optimal Gaussian policy, we apply equation (47) to do the update:

$$\beta_1 \leftarrow \left(\frac{\alpha_1 e^{-2\beta_2-1}}{\pi\lambda^2\beta_1} + \frac{1}{2}\beta_1\right). \tag{48}$$

For updating other parameters β_2, α_1 and α_2, we follow the idea of Doya (2000) and Wang and Zhou (2020) by minimizing the cumulative continuous-time temporal difference (TD) error. Suppose π_t is the optimal control and V^π is the corresponding value function. Then V^π satisfies the following equation according to Bellman's principle:

$$V^\pi(t,x) = \mathbb{E}\left[V^\pi(s, X_s^\pi) - \lambda\int_t^s \int_{\mathbb{R}} \pi_\tau(u)\log\pi_\tau(u)\,du\,d\tau \;\middle|\; X_t^\pi = x\right], \quad s \in (t,T]. \tag{49}$$

The continuous-time TD error measures the difference between the two sides of the equation as s approaches t (Doya (2000)):

$$\varepsilon_t = \dot{V}^\pi(t,x;\boldsymbol{\alpha}) - \lambda\int_{\mathbb{R}} \pi_t(u;\boldsymbol{\beta})\log\pi_t(u;\boldsymbol{\beta})\,du \tag{50}$$

where $\dot{V}^\pi$ is the partial derivative of V^π with respect to time t. Then, the cumulative continuous-time TD error to time t is defined as

$$C_t(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\mathbb{E}\left[\int_0^t |\varepsilon_s|^2\,ds\right] = \frac{1}{2}\mathbb{E}\left[\int_0^t |\dot{V}^\pi(s,x;\boldsymbol{\alpha}) - \lambda\int_{\mathbb{R}} \pi_s(u;\boldsymbol{\beta})\log\pi_s(u;\boldsymbol{\beta})du|^2\,ds\right]. \tag{51}$$

which is a function of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

To implement the RL algorithm numerically, we need to get an approximation to the TD error. We partition the time interval $[0,T]$ into $\{t_i,\ i = 0,\dots,n\}$ with $t_0 = 0$, $t_{i+1} = t_i + \Delta t$ and $t_n = T$ for a constant Δt. Let x_i denote the state value at time t_i for $i = 0,\dots,n$, and write $\mathcal{S}_i = \{(t_j,x_j); j = 0,\dots,i\}$ for the information up to time t_i. We approximate the TD error up to time t_i by (with a slight abuse of notation in the subscript of C):

$$C_i(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^\pi(t_j,x_j;\boldsymbol{\alpha}) - \lambda\int_{\mathbb{R}} \pi_{t_j}(u;\boldsymbol{\beta})\log\pi_{t_j}(u;\boldsymbol{\beta})\,du\right)^2\Delta t. \tag{52}$$

The derivative of the value function V^π at time t_i is approximated by

$$\begin{aligned}\dot{V}^\pi(t_i,x_i;\boldsymbol{\alpha}) &= \frac{V^\pi(t_{i+1},x_{i+1};\boldsymbol{\alpha})-V^\pi(t_i,x_i;\boldsymbol{\alpha})}{\Delta t}\\ &= \frac{[1+\lambda(T-t_{i+1})]\log x_{i+1}-[1+\lambda(T-t_i)]\log x_i}{\Delta t}\\ &- \frac{[1+\lambda(T-t_{i+1})]\log[1+\lambda(T-t_{i+1})]-[1+\lambda(T-t_i)]\log[1+\lambda(T-t_i)]}{2\Delta t}\\ &+ \frac{\alpha_1(t_{i+1}^2-t_i^2)+\alpha_2\Delta t}{\Delta t}.\end{aligned} \tag{53}$$

Using the parametrized control and (42), we take the TD error approximation as follows:

$$C_i(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^\pi(t_j,x_j;\boldsymbol{\alpha})+\lambda\log x_j-\frac{\lambda}{2}\log[1+\lambda(T-t_j)]-\lambda\beta_2\right)^2\Delta t. \tag{54}$$

We then use the Stochastic Gradient Descent Algorithm in Goodfellow *et al.* (2016) to update α_1, α_2 and β_2 at time t_i. The gradients with respect to the parameters are calculated as follows:

$$\frac{\partial C_i}{\partial\alpha_1} = \sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^\pi(t_j,x_j;\boldsymbol{\alpha})+\lambda\log x_j-\frac{\lambda}{2}\log[1+\lambda(T-t_j)]-\lambda\beta_2\right)(t_{j+1}^2-t_j^2) \tag{55}$$

$$\frac{\partial C_i}{\partial\alpha_2} = \sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^\pi(t_j,x_j;\boldsymbol{\alpha})+\lambda\log x_j-\frac{\lambda}{2}\log[1+\lambda(T-t_j)]-\lambda\beta_2\right)\Delta t \tag{56}$$

$$\frac{\partial C_i}{\partial\beta_2} = \sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^\pi(t_j,x_j;\boldsymbol{\alpha})+\lambda\log x_j-\frac{\lambda}{2}\log[1+\lambda(T-t_j)]-\lambda\beta_2\right)(-\lambda\Delta t) \tag{57}$$

Supposing θ_α and θ_β are learning rates for updating (α_1,α_2) and β_2, we update them by $(\alpha_1,\alpha_2)' - \theta_\alpha\nabla_\alpha C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ and $\beta_2-\theta_\beta\nabla_\beta C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$. Once we obtain an update for $(\alpha_1,\alpha_2,\beta_2)$, we update β_1 using (48) and keep the iterative procedure until a termination criterion is satisfied. The pseudocode for the online updating procedure is summarized in Algorithm 1.

Algorithm 1: RL Algorithm with Amount Control

Input: Market parameters (μ,σ,r,ρ), learning rates θ_α, θ_β, initial wealth x_0, investment horizon T, discretization Δt, exploration rate λ.
Initialization: $i=1$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$
while $i\le\frac{T}{\Delta t}$ **do**
  Sample (t_i,x_i) under $\pi(u;\boldsymbol{\beta})$
  Update set of samples $\mathcal{S}_i=\{(t_j,x_j);j=0,\dots,i\}$
  Update $(\alpha_1,\alpha_2)'$ as $(\alpha_1,\alpha_2)' - \theta_\alpha\nabla_\alpha C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ using (55) and (56)
  Update β_2 as $\beta_2-\theta_\beta\nabla_\beta C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ using (57)
  Update β_1 using (48)
  Update $\pi_t(u;x,\boldsymbol{\alpha},\boldsymbol{\beta})$ as $\mathcal{N}\left(u \;\middle|\; \beta_1 x,\ \frac{e^{-2\beta_2-1}x^2}{2\pi(1+\lambda(T-t))}\right)$
  $i=i+1$
end

Algorithms for other exploratory problems are given in Appendix B.

6. Simulation Studies

This section implements the RL algorithms with simulated data and compares them with several benchmark Kelly portfolio strategies.

6.1. *Portfolio Strategies and Simulation Setting*

We consider the following seven portfolio strategies:

(1) Oracle: Kelly strategy with actual parameter values,
(2) Plug-in: Kelly strategy with maximum likelihood estimation (MLE),
(3) Shrinkage: Shrinkage Kelly strategy proposed by Han *et al.* (2019),
(4) Fractional: Fractional Kelly strategy with MLE,
(5) RL-Amount: RL algorithm 1 with amount control and constant λ,
(6) RL-Portion: RL algorithm 2 with portion control and constant λ, and
(7) RL-Decay: RL algorithm 3 with amount control and power-decaying $\lambda_a(t)$.

The Oracle strategy is defined in equation (3) with the true values used for the parameters ρ and σ. The Oracle strategy is not a legitimate investment strategy because it uses true parameter values which are unknown to us in practice. This strategy is expected to perform the best since it is subject to no estimation risk and no cost of exploration, although its performance is not attainable for practical use. The Plug-in strategy follows a growing window framework to form the MLE for the parameters μ and σ (or equivalently ρ and σ), and then substitute the resulting MLE into equation (3) for the portfolio weight. In the Shrinkage and Fractional Kelly strategies, portfolio weights are also based on MLE but multiplied with fractional weights. The fraction in the Shrinkage Kelly strategy is defined in equation (11) in Han *et al.* (2019). It is proposed to mitigate the estimation error from MLE. The Shrinkage Kelly strategy is validated to outperform several fractional Kelly strategies with empirical studies (Han *et al.* (2019)). When there are more sample data for estimation, the fraction becomes closer to 1. For the Fractional Kelly strategy, we test nine fractional candidates from 0.1 to 0.9 with a step size of 0.1, by $M = 2{,}000$ independent simulations of stock returns. We consider model parameters $(\mu, \sigma, r, \rho) = (0.2, 0.1, 0.02, 1.8)$, investment time horizon $T = 1$ year and the initial portfolio wealth $x_0 = 1$ as the benchmark setting in our simulation study. We discretize the investment time horizon into 252 sub-intervals (i.e., the discretization length $\Delta t = 1/252$) with each subinterval representing one trading day in the stock market. For all the fractional candidates along with a fraction of one (i.e., the full Kelly strategy), we plot the corresponding average terminal log-return in Figure 1. For the benchmark market setting, we choose the fraction of 0.7, under which the portfolio performance is the best, in terms of the average terminal log-return. For other market settings, we repeat the same selection procedure to choose the best fraction.

Other strategies are also evaluated using 2,000 independent paths of stock returns. Over each simulated path, all the strategies (2)-(7) start with an initial estimation of parameters and updates parameters through time based on observed data. We set the initial estimation of the model parameters to be the MLE from 100 simulated data points.

For the three RL based strategies, we set the default learning rates $\theta_\alpha = \theta_\beta = 0.0005$, following Wang and Zhou (2020). For strategies RL-Amount and RL-Portion with constant temperature parameter, we set $\lambda = 0.5$ as the default. For the RL-Decay strategy with power-decaying temperature parameter, we set $\lambda_0 = 0.1$ and $\lambda_1 = 0.236$ in equation (25) which gives $\lambda_a(0) = 0.5$. Similar to the fraction selection for the Fractional strategy, we choose the default λ from several candidates (see Table 1). From 2,000 simulations, the RL-Amount strategy, with λ varying from 0.05 to 0.5,

yields similar performance in terms of the average terminal log-return and its standard error. The default λ of 0.5 is not the best temperature parameter for the benchmark market setting, but a fair choice in comparing with MLE based strategies (i.e., strategies Plug-in, Shrinkage, and Fractional).

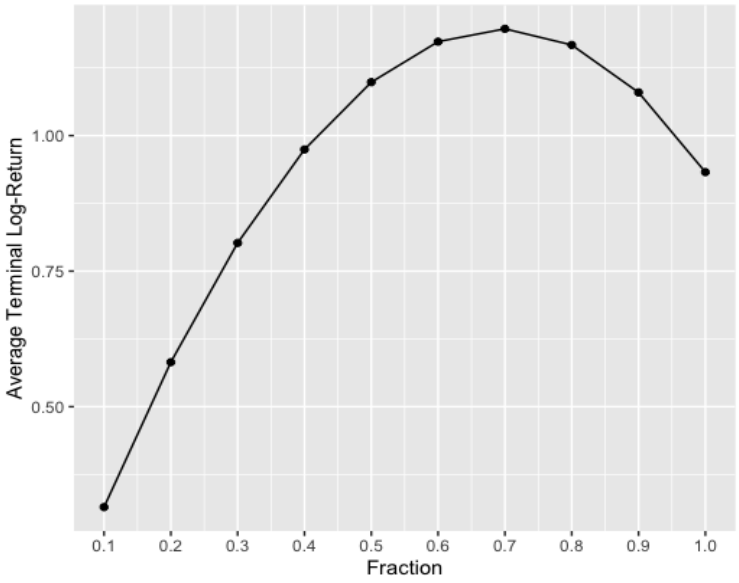

Figure 1. Fraction Selection for the Fractional Strategy

Table 1. Selection of Default λ: RL-Amount Strategy

λ	0.025	0.05	0.1	0.25	0.5	1
Average Terminal Log-return	1.01	1.39	1.49	1.51	1.43	1.24
Standard Error	0.12	0.07	0.05	0.04	0.04	0.05

6.2. *Model Convergence*

Before comparing our RL strategies with the MLE based strategies, we first investigate the convergence of the exploratory algorithm with simulated data. We focus on the RL-Decay strategy, since it has the desirable time-decaying control variance. We run the Oracle strategy and RL-Decay strategy for different time horizons $T \in \{10/252, 1/12, 1/4, 1/2, 3/4, 1, 5/4, 3/2, 7/4, 2, 3, 4\}$, with other parameters set as default, and then calculate the loss of the RL-Decay strategy relative to the Oracle strategy based on 8,000 independent replications. Figure 2 illustrates the convergence of the relative loss to zero. The relative loss diminishes quickly as T increases. Particularly, it is close to 0 when the investment time horizon is longer than one year. It decreases to around 2% when the time horizon is two years. These simulation results mean that the relative performance of the RL strategy improves quickly over time and it performs almost as well as if we know the true parameter values in the Oracle strategy when the investment time horizon is as long as one year.

We also study the convergence of the RL algorithm under an episodic framework. In contrast to the proposed online algorithms, an episodic algorithm only updates the model parameter values after one episode. The learned parameter values are then used throughout the next episode. For

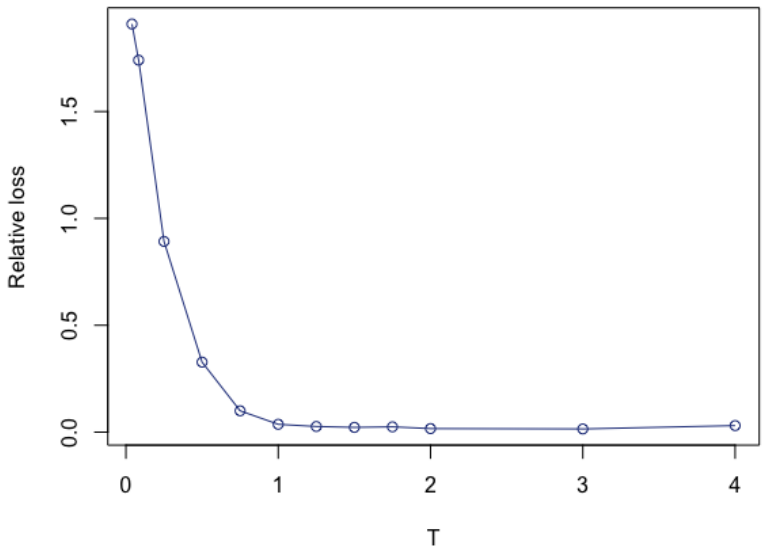

Figure 2. RL-Decay Model Convergence

one simulation, we start from a random set of initial parameter values[1] and run for 150 episodes of length one year to get the terminal (year-end) utility at the end of each episode. Then in another independent simulation, we repeat this procedure to get another 150 terminal (year-end) utility values. In total, we repeat for 4,000 independent simulations of 150 episodes. Hence, we have 4,000 values of the terminal (year-end) utility at the end of each episode. Their average is taken to estimate the expected terminal utility for each episode. If the RL algorithm works, then as k increases, the average terminal utility of the kth episode is anticipated to be close to the theoretically optimal value in (21). The result of the RL-Decay model is shown in Figure 3. The solid line is the average terminal utility at the end of each episode. Under the benchmark parameters, the theoretically optimal terminal utility is 1.54 (dashed line). As indicated by the graph, after six episodes, the average (year-end) terminal utility keeps fluctuating around the theoretically optimal one. This validates the rapid convergence of the algorithm under the episodic framework.

6.3. *Simulation Results under the Benchmark Parameters*

We now implement our RL Kelly strategies as well as some MLE based Kelly strategies with the online algorithms. We carry out these experiments because they are closer to the situation in practice where decisions are made frequently. Figure 4 shows the distributions of the terminal log-return under each of the strategies (2)-(7) compared to the Oracle strategy (1). Table 2 summarizes the average terminal log-return from each strategy. The first row is the theoretically optimal value of the expected terminal log-return in a classical Kelly criterion problem which assumes the true model parameter values are known. We also report the standard errors of the estimates. The last two columns are the theoretical and estimated values of the cost of the strategy, which is the relative difference between the average terminal log-return of a specific trading strategy and the theoretical

[1]The initial model parameter values are chosen as the MLEs from 100 simulated data points.

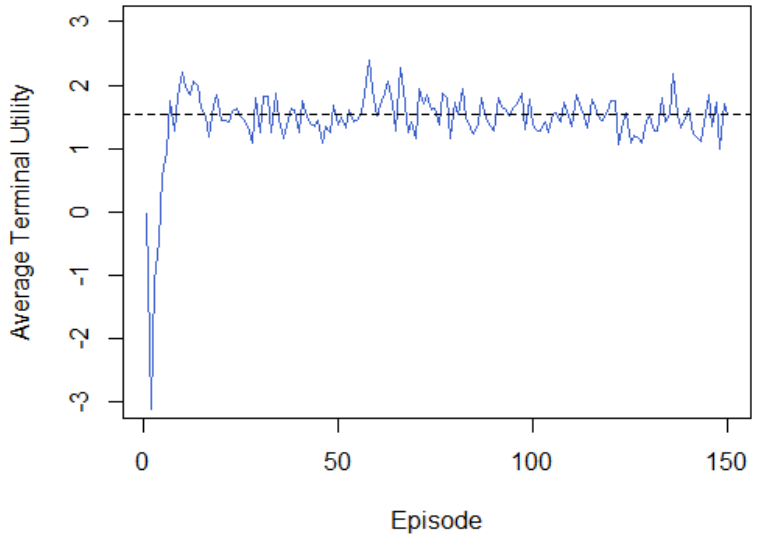

Figure 3. RL-Decay Model Convergence Under Episodic Framework

optimal terminal log-return. Note that an inevitable issue in simulations is that the wealth may go negative, due to the uncertainty in either the stochastic stock process or the unconstrained short-selling or leverage of the RL Kelly strategy. In this case, the log-return is meaningless. We adopt the reflection approach to replace the negative wealth by its absolute value. The impact of the reflection method on our results is limited. Among 2,000 simulations, each with 252 values of wealth, only four values are reflected under the MLE based strategies. Under the RL-Amount strategy, only two values are reflected, both greater than -0.002. There are no cases for the other two RL strategies.

Table 2. Model Performance: Terminal Log-Return

Model	*Mean*	*Std. Error*	*Cost*	$\widehat{Cost}$
Theoretical	1.62			
Oracle	1.68	0.04	0.00	0.04
Plug-in	0.93	0.05	0.00	0.42
Shrinkage	0.98	0.04	0.00	0.39
Fractional	1.20	0.04	0.00	0.26
RL-Amount	1.43	0.04	0.13	0.12
RL-Portion	1.40	0.04	0.15	0.14
RL-Decay	1.59	0.04	0.05	0.02

The results in Figure 4 and Table 2 validate the practical merit of the fractional and shrinkage Kelly strategy over the full Kelly strategy. Furthermore, they also confirm the outperformance of the RL strategies over the three MLE based strategies. Figure 4 indicates that the distributions of terminal log-return from the three MLE based strategies are shifted to the left compared to those from RL strategies. The average terminal log-return reported in Table 2 also shows that the RL strategies yield a higher terminal log-return than the MLE based strategies on average. Moreover, the simulated results reported in Table 2 also confirm the benefit of using time-decaying λ. With

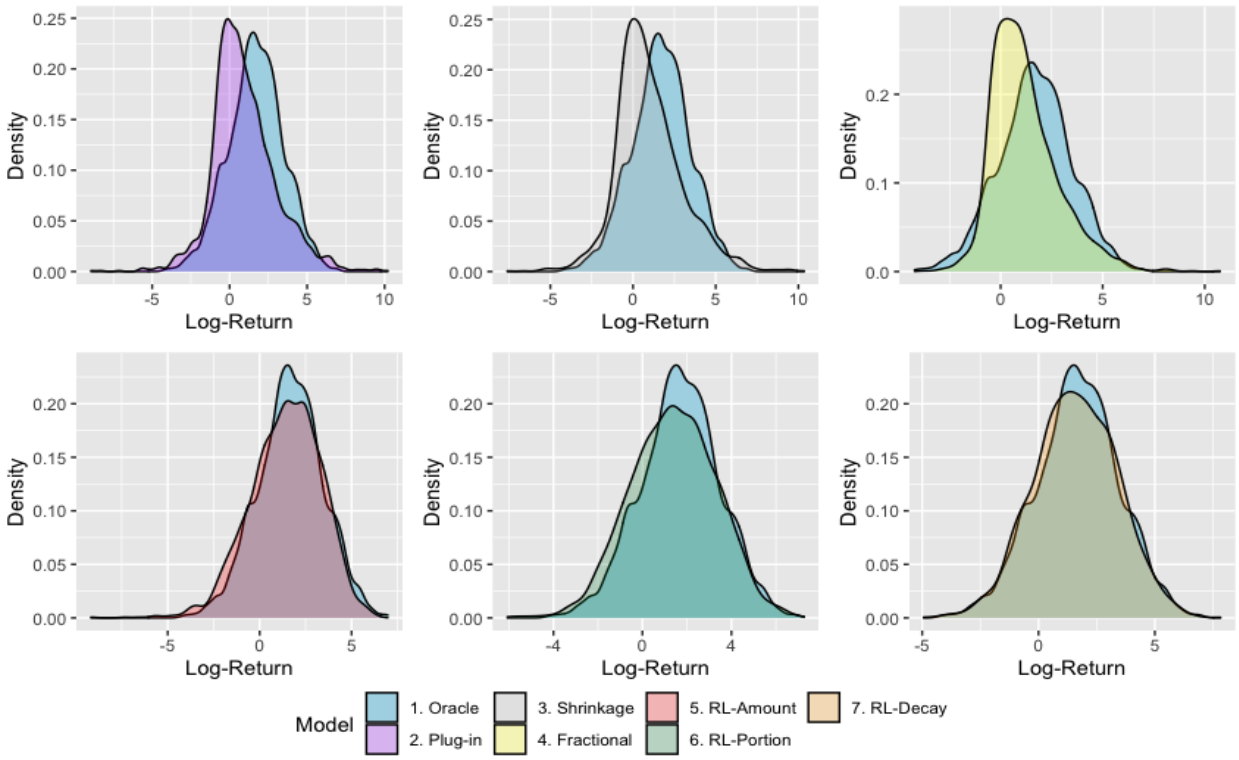

Figure 4. Model performance

a power-decaying λ, the agent is subject to less cost due to exploration and achieves a higher expected terminal log-return.

Table 3 reports the statistical summary of the terminal wealth from the 2,000 simulations: average value, standard deviation, skewness, and quantiles at levels of 0.1%, 1%, 5%, 95%, 99% and 99.9%, respectively. To obtain wealth, we modify the reflection approach. For example, when the wealth becomes -0.1, we build an additional account to borrow 0.2 from the bank. The total wealth is still -0.1 but the investment wealth becomes 0.1. Then we invest based on the new wealth of 0.1. At the end of the period, we report the total wealth. It is still possible to have negative terminal wealth. From the result, MLE based strategies have higher average terminal wealth, as well as significantly large standard deviations. They also have thicker tails of the wealth distribution, especially the Plug-in and Shrinkage strategies. These are due to the nature of the (fractional) Kelly strategy, in particular that it usually bets a large amount of money. Hence, in a few extreme cases where a sequence of the simulated stock returns are relatively high, the MLE based strategies benefit greatly from the aggressive investment. However, on the other hand, the aggressive investment could lead to negative wealth when the stock returns are relatively low. The Fractional strategy mitigates these effects by smaller investments. As a result, the distribution is centered at a smaller mean and becomes more leptokurtic and skewed. Compared with the MLE based strategies, the three RL strategies learn the entire wealth distribution better. Their wealth distributions are closer to that under the Oracle strategy. Particularly, the RL strategy with power-decaying temperature parameter has close quantiles to those under the Oracle model.

The MLE based strategies and RL strategies are essentially updating market parameters over time. For each strategy, one simulation yields one pair of estimates $(\hat{\mu}, \hat{\sigma}^2)$ by the end of the investment time horizon. Since we have run 2,000 independent simulations for each strategy, we obtained 2,000 pairs of estimates from each strategy. Table 4 shows the mean and its standard error

Table 3. Model Performance: Terminal Wealth

Model	*Mean*	*Std. Deviation*	*Skewness*	$q_{0.001}$	$q_{0.01}$	$q_{0.05}$	$q_{0.95}$	$q_{0.99}$	$q_{0.999}$
Theoretical	25.53	126.47	136.38	0.02	0.08	0.26	97.59	332.76	1316.09
Oracle	23.14	63.54	8.26	0.02	0.07	0.27	96.12	265.81	936.50
Plug-in	68.29	830.54	23.61	-2.48	0.02	0.15	98.51	694.85	12862.84
Shrinkage	70.14	897.16	25.70	-0.73	0.03	0.17	99.68	691.45	13208.59
Fractional	54.72	1088.69	39.64	0.05	0.16	0.46	72.02	415.23	4526.27
RL-Amount	17.80	41.21	6.33	0.00	0.03	0.16	74.36	212.80	403.23
RL-Portion	21.69	69.22	9.96	0.00	0.04	0.17	85.79	333.25	692.60
RL-Decay	22.24	77.68	18.51	0.02	0.07	0.28	89.90	302.69	765.29

of those 2,000 estimates (for both μ and σ^2) from each strategy.[1] Not surprisingly, all are close to the true parameter values after taking the average over the 2,000 estimates. But the estimation of parameter μ from the RL strategies is consistently more robust than that from the MLE based strategies, noting that the MLE based strategies have a higher standard error.

Table 4. Model Performance: Parameter Estimation

Model	μ		σ^2	
	Mean	*Std. Error*	*Mean*	*Std. Error*
Oracle	0.2		0.1	
Plug-in	0.2030	0.0018	0.0998	0.0001
RL-Amount	0.2025	0.0004	0.1025	0.0002
RL-Portion	0.2056	0.0005	0.1024	0.0002
RL-Decay	0.2054	0.0003	0.1023	0.0001

6.4. *Sensitivity Tests*

To assess the robustness of the outperformance of the RL strategies over the three MLE based strategies, we repeat simulations for all strategies under different market settings, i.e., different values of μ and σ. We report the results in Table 5.

We choose four different values for σ and seven different values for μ, which yield 28 market scenarios in total, including the benchmark setting where $(\mu, \sigma) = (0.2, 0.1)$. The fractions used in the Fractional strategy are again chosen from 9 candidates, by repeating the selection procedure under the benchmark setting. The temperature parameters for the RL strategies are still the same. For each scenario, we compare the average terminal log-return between the RL strategies and the MLE based strategies. The best performance under each scenario is labeled with a superscript asterisk. Among all the 28 scenarios, the RL strategies outperform MLE based strategies under 24 settings. Particularly, the RL-Decay strategy outperforms all the three MLE based strategies in all 24 cases. The other two RL strategies beat all MLE based strategies in 19 cases, even though the fractions for strategy (4) are chosen based on ex-post information.

Four exceptions where RL strategies do not outperform are under settings of $(\mu, \sigma) = (-0.1, 0.01)$, $(0.2, 0.01)$, $(0, 0.1)$ and $(0, 0.15)$ which yield extreme values of ρ^2. However, the performance of the RL-Decay strategy is still comparable with the best one in these cases. Under the setting of $(\mu, \sigma) = (0, 0.1)$ and $(0, 0.15)$, the differences between the RL-Decay strategy and the best strategy are less than 0.004. Under the other two cases, the relative differences are less than 5%.

To sum up, RL strategies have robust performance, in terms of the relatively high average terminal log-return, under different market scenarios. In cases where MLE based strategies fail to

[1]MLE based strategies (2), (3) and (4) share the same estimates. Hence, only one set of results is reported, named Plug-in.

Table 5. Average Terminal Log-Return under Different Market Settings

Model	$\sigma = 0.01$						
	$\mu = -0.2$	$\mu = -0.1$	$\mu = -0.05$	$\mu = 0$	$\mu = 0.05$	$\mu = 0.1$	$\mu = 0.2$
Theoretical	242.00	72.00	24.50	2.00	4.50	32.00	162.00
Oracle	229.97	72.72	21.41	1.91	4.53	29.05	165.87
Plug-in	231.18	73.55*	20.86	1.12	3.61	29.12	166.63*
Shrinkage	227.59	72.38	21.03	1.18	3.68	29.04	164.50
Fractional	216.63	68.92	21.28	1.40	3.89	28.77	155.94
RL-Amount	232.07*	71.05	21.75	1.80	4.17	29.39*	156.78
RL-Portion	230.88	70.97	21.70	1.78	4.17	28.99	157.73
RL-Decay	229.70	71.17	21.89*	1.97*	4.40*	28.95	158.74
Model	$\sigma = 0.05$						
	$\mu = -0.2$	$\mu = -0.1$	$\mu = -0.05$	$\mu = 0$	$\mu = 0.05$	$\mu = 0.1$	$\mu = 0.2$
Theoretical	9.68	2.88	0.98	0.08	0.18	1.28	6.48
Oracle	9.02	2.76	0.92	0.06	0.20	1.34	6.43
Plug-in	7.80	1.92	0.19	-0.62	-0.48	0.61	5.34
Shrinkage	7.96	1.99	0.24	-0.58	-0.44	0.65	5.46
Fractional	8.34	2.18	0.53	0.00	0.04	0.88	5.73
RL-Amount	8.78	2.65	0.80	-0.13	0.00	1.09	5.94
RL-Portion	8.91	2.63	0.77	-0.14	-0.02	1.06	5.99
RL-Decay	9.23*	2.83*	0.96*	0.06*	0.16*	1.26*	6.27*
Model	$\sigma = 0.1$						
	$\mu = -0.2$	$\mu = -0.1$	$\mu = -0.05$	$\mu = 0$	$\mu = 0.05$	$\mu = 0.1$	$\mu = 0.2$
Theoretical	2.42	0.72	0.25	0.02	0.04	0.32	1.62
Oracle	2.31	0.67	0.22	0.01	0.06	0.35	1.68
Plug-in	1.51	-0.04	-0.47	-0.67	-0.63	-0.34	0.93
Shrinkage	1.56	0.00	-0.43	-0.63	-0.58	-0.30	0.98
Fractional	1.77	0.34	0.05	-0.01*	0.00	0.12	1.20
RL-Amount	2.21	0.54	0.06	-0.18	-0.14	0.15	1.43
RL-Portion	2.19	0.51	0.04	-0.28	-0.16	0.12	1.40
RL-Decay	2.38*	0.70*	0.23*	-0.01	0.03*	0.30*	1.59*
Model	$\sigma = 0.15$						
	$\mu = -0.2$	$\mu = -0.1$	$\mu = -0.05$	$\mu = 0$	$\mu = 0.05$	$\mu = 0.1$	$\mu = 0.2$
Theoretical	1.08	0.32	0.11	0.01	0.02	0.14	0.72
Oracle	1.02	0.29	0.09	0.00	0.03	0.16	0.77
Plug-in	0.28	-0.41	-0.60	-0.68	-0.65	-0.52	0.06
Shrinkage	0.33	-0.36	-0.55	-0.63	-0.61	-0.48	0.10
Fractional	0.61	0.09	0.01	-0.01*	0.00	0.03	0.40
RL-Amount	0.89	0.16	-0.11	-0.14	-0.16	-0.02	0.55
RL-Portion	0.87	0.12	-0.12	-0.22	-0.20	-0.06	0.51
RL-Decay	1.05*	0.30*	0.09*	-0.01	0.00*	0.12*	0.70*

Note: $*$: the best model among MLE based and RL models.

achieve relatively high terminal log-return, RL strategies outperform them. On the other hand, in extreme cases of ρ^2 where MLE based strategies obtain relatively high average log-return, RL strategies also have comparable performance.

6.5. *Performance under Heston's Model*

The simulation studies in the preceding subsections confirm the outperformance of our RL strategies over the MLE based strategies under the correctly specified stock price model (i.e., the geometric Brownian motion). To test the practical feasibility of the RL strategies, we consider Heston's model

for the stock price:

$$dS_t = \mu S_t dt + \sqrt{L_t} S_t dW_t$$
$$dL_t = \kappa(\nu - L_t)dt + \xi\sqrt{L_t} d\tilde{W}_t$$

where $\tilde{W}_t$ is a Brownian motion correlated with W_t: $\text{Cov}(W_t, \tilde{W}_t) = \tilde{\rho}t$.

The experimental procedure is the same as before except that the stock price paths are simulated from Heston's model with parameters $\mu = 0.2$, $\nu = 0.01$, $\tilde{\rho} = -0.3$, $\kappa = 2$ and $\xi \in \{0.001, 0.01, 0.05, 0.1, 0.15, 0.2\}$. We exclude the Oracle strategy from the analysis because we implement investment strategies derived based on the geometric Brownian motion but test them on data from Heston's model. We use the Plug-in strategy as the benchmark and report the relative performance of the average terminal log-return for the other five strategies in Table 6. The average terminal log-returns are computed based on 2,000 independent replications. Zero for the relative performance measure means an equivalent performance with the Plug-in strategy, and the larger the relative performance, the higher the average terminal log-return for a strategy. The issue that simulated wealth goes to negative is not very significant in the results, and we reflect the negative values as we did in the previous simulation studies. For each different ξ value, at most five wealth values out of the 2,000 simulations are reflected, and they are all greater than -0.06. The issue is even less significant for a smaller ξ in the Helston's model. From the results, the RL-Decay strategy still has the best performance compared to the other strategies. RL-Amount and RL-Portion strategies have better or comparable performance to the MLE based strategies in all scenarios.

Table 6. Performance Relative to the Plug-in Strategy: Heston's Model

Strategy	$\xi = 0.001$	$\xi = 0.01$	$\xi = 0.05$	$\xi = 0.1$	$\xi = 0.15$	$\xi = 0.2$
Shrinkage	0.06	0.07	0.08	0.09	0.13	0.22
Fractional	0.33	0.35	0.43	0.59	0.85	1.05
RL-Amount	0.56	0.57	0.63	0.69	0.76	1.01
RL-Portion	0.52	0.54	0.61	0.70	0.84	1.10
RL-Decay	0.66	0.69	0.78	0.92	1.12	1.53

7. Conclusion

The performance of the full Kelly strategy in practice is not as superior as claimed in theory due to estimation errors in market parameters. Two alternatives to the full Kelly strategies are fractional and shrinkage Kelly strategies. Motivated by the practical deficiency, we extend the classical Kelly criterion problem to an RL framework. Based on the novel exploratory formulation (Wang *et al.* (2019), Wang and Zhou (2020)), we build two exploratory Kelly criterion problems respectively, taking the amount of investment and the portion of wealth as the control. The resulting optimal strategies, the RL Kelly strategies, are sequences of normal distributions that center at the classical optimal allocation.

We establish learning algorithms to implement the RL Kelly strategies. We use simulated data to compare the performance of our strategies with three MLE based strategies. Our results validate the practical advantage of the fractional and shrinkage Kelly strategies against the full Kelly strategy with plug-in MLE. The RL Kelly strategies perform even significantly better than the fractional and shrinkage strategies. They achieve higher average terminal log-return and obtain parameter

estimates close to the true values of parameters. Particularly, the RL strategy with a time-decaying $\lambda_a(t)$ is the best in that it not only achieves the highest average terminal log-return but also learns the entire terminal wealth distribution more precisely. Furthermore, the performance of the RL strategies is robust under different market settings and Heston's model. When the MLE based strategies perform well, the RL strategies also have comparable performance. When the MLE based strategies perform poorly, the RL strategies outperform them significantly.

Possible directions for future work include an extension of the portfolio to include multiple stocks for log-return maximization. Another interesting direction for future research is applying the framework to other optimality objectives (e.g., expected utility maximization) and/or other stock price processes. For these models, the resulting HJB equations may not have a closed-form solution, and efficient numerical methods would be needed to tackle the HJB equations in the context of RL algorithms.

References

Cover, T.M. and Thomas, J.A., *Elements of Information Theory*, 1991, Wiley.

Dai, M., Yuchao, D. and Jia, Y., Learning equilibrium mean-variance strategy. *Available at SSRN 3770818*, 2020.

Davis, M. and Lleo, S., Fractional Kelly strategies in continuous time: Recent developments. In *Handbook of the Fundamentals of Financial Decision Making: Part II*, pp. 753–787, 2013, World Scientific.

Doya, K., Reinforcement learning in continuous time and space. *Neural Computation*, 2000, **12**, 219–245.

Goll, T. and Kallsen, J., Optimal portfolios for logarithmic utility. *Stochastic Processes and their Applications*, 2000, **89**, 31–48.

Goodfellow, I., Bengio, Y. and Courville, A., *Deep Learning*, 2016, MIT press.

Han, Y., Yu, P.L.H. and Mathew, T., Shrinkage estimation of Kelly portfolios. *Quantitative Finance*, 2019, **19**, 277–287.

Ishii, S., Yoshida, W. and Yoshimoto, J., Control of exploitation–exploration meta-parameter in reinforcement learning. *Neural Networks*, 2002, **15**, 665–687.

Kallberg, J.G. and Ziemba, W.T., Mis-specifications in portfolio selection problems. In *Risk and Capital*, pp. 74–87, 1984, Springer.

MacLean, L.C., Thorp, E.O., Zhao, Y. and Ziemba, W.T., Medium term simulations of the full Kelly and fractional Kelly investment strategies. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pp. 543–561, 2011, World Scientific.

MacLean, L.C., Thorp, E.O. and Ziemba, W.T., Long-term capital growth: the good and bad properties of the Kelly and fractional Kelly capital growth criteria. *Quantitative Finance*, 2010, **10**, 681–687.

Merton, R., Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 1971, **3**, 373–413.

Nekrasov, V., Kelly criterion for multivariate portfolios: A model-free approach. *Available at SSRN 2259133*, 2014.

Rising, J.K. and Wyner, A.J., Partial Kelly portfolios and shrinkage estimators. In *Proceedings of the 2012 IEEE International Symposium on Information Theory Proceedings*, pp. 1618–1622, 2012.

Shen, W., Wang, B., Pu, J. and Wang, J., The Kelly growth optimal portfolio with ensemble learning. In *Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 1134–1141, 2019.

Sutton, R.S. and Barto, A.G., *Reinforcement Learning: An Introduction*, 2018, MIT press.

Wang, H., Zariphopoulou, T. and Zhou, X.Y., Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 2019.

Wang, H. and Zhou, X.Y., Large scale continuous-time mean-variance portfolio allocation via reinforcement learning. *Available at SSRN 3428125*, 2019.

Wang, H. and Zhou, X.Y., Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 2020, **30**, 1273–1308.

Weng, C. and Zhuang, S.C., CDF formulation for solving an optimal reinsurance problem. *Scandinavian Actuarial Journal*, 2017, **2017**, 395–418.

Ziemba, W.T., Understanding the Kelly capital growth investment strategy. *Investment Strategies*, 2016, **3**, 49–55.

Appendix A: Proofs of Results

A.1. *Proof of Theorem 3.1*

The Lagrangian function for the maximization problem is given by

$$\begin{aligned}\mathcal{L}(\pi_t,\eta) &:= \rho\sigma v_x(t,x)\int_{\mathbb{R}} u\pi_t(u)du + \frac{1}{2}\sigma^2 v_{xx}(t,x)\int_{\mathbb{R}} u^2\pi_t(u)du \\ &\quad -\lambda_a(t)\int_{\mathbb{R}}\pi_t(u)\log\pi_t(u)du + \eta\int_{\mathbb{R}}\pi_t(u)du \\ &=: \int_{\mathbb{R}} L(u,\pi_t(u))du, \end{aligned} \tag{A1}$$

where η denotes the Lagrangian multiplier, and $L(u,\pi_t(u))$ is given by

$$L(u,\pi_t(u)) = \rho\sigma v_x(t,x)u\pi_t(u) + \frac{1}{2}\sigma^2 v_{xx}(t,x)u^2\pi_t(u) - \lambda_a(t)\pi_t(u)\log\pi_t(u) + \eta\pi_t(u).$$

By a standard Lagrangian duality argument (e.g., see Lemma 4.3 of Weng and Zhuang (2017)), if $\pi^* := \pi^*_{\eta^*} \in \mathcal{P}(\mathbb{R})$ maximizes $\mathcal{L}(\pi,\eta^*)$ for some $\eta^* \in \mathbb{R}$ satisfying $\int_{\mathbb{R}}\pi^*(u)du = 1$, then π^* is a solution to the maximization problem in equation (12). Consequently, we focus on analyzing the optimizer(s) of $\mathcal{L}(\cdot,\eta)$ before we show the optimality of π^*_t in equation (13).

To derive a maximizer of $\mathcal{L}(\cdot,\eta)$, we apply a pointwise maximization procedure and analyze the integrand in (A1), $L(u,\pi_t(u))$. Since $\pi\log\pi$ is convex in π while the other items in the expression of $L(u,\pi)$ are linear in π, $L(u,\pi)$ is concave as a function of π. Accordingly, the first order optimality condition is sufficient to determine its maximizers, whereby we take the partial derivative of $L(u,\pi_t(u))$ with respect to $\pi_t(u)$ and equate it to zero to get:

$$\rho\sigma v_x(t,x)u + \frac{1}{2}\sigma^2 v_{xx}(t,x)u^2 - \lambda_a(t)\log\pi_t(u) - \lambda_a(t) + \eta = 0,$$

which gives

$$\pi_t(u) = \exp\left(\frac{1}{\lambda_a(t)}\left[\frac{1}{2}\sigma^2 v_{xx}(t,x)u^2 + \rho\sigma v_x(t,x)u\right] - \lambda_a(t) + \eta\right).$$

Taking η to scale $\pi_t(u)$ to satisfy the constraint $\int_{\mathbb{R}}\pi_t(u)du = 1$ yields the desired optimality of π^*_t in (13).

A.2. *Proof of Theorem 3.2*

We start from conjecturing the solution to the PDE (14) in the form $v(t,x) = f(t)\log x + g_a(t)$ for some functions f and g defined on $[0,T]$ with conditions $f(T) = 1$ and $g_a(T) = 0$. This yields

$v_x = x^{-1}f(t)$, $v_{xx} = -x^{-2}f(t)$ and $v_t = f'(t)\log(x) + g_a'(t)$. It is straightforward to use equation (13) to verify equation (18) for the optimal control $\pi_t^*(u; x, \lambda_a(t))$. Furthermore, substituting the expressions of v_t, v_x and v_{xx} (in terms of f and g_a) into the PDE (14) yields:

$$\begin{aligned}
v_t(t,x) &- \frac{\rho^2(v_x(t,x))^2}{2v_{xx}(t,x)} - \frac{\lambda_a(t)}{2}\log\left(-\frac{\sigma^2 v_{xx}(t,x)}{2\pi\lambda_a(t)}\right) \\
&= f^{'}(t)\log x + g_a'(t) + \frac{\rho^2}{2}f(t) - \frac{\lambda_a(t)}{2}\log\frac{\sigma^2 x^{-2}f(t)}{2\pi\lambda_a(t)} \\
&= f^{'}(t)\log x + g_a'(t) + \frac{\rho^2}{2}f(t) - \frac{\lambda_a(t)}{2}\log x^{-2} - \frac{\lambda_a(t)}{2}\log\frac{\sigma^2 f(t)}{2\pi\lambda_a(t)} \\
&= (f^{'}(t) + \lambda_a(t))\log x + g_a'(t) + \frac{\rho^2}{2}f(t) - \frac{\lambda_a(t)}{2}\log\frac{\sigma^2 f(t)}{2\pi\lambda_a(t)} \\
&= 0.
\end{aligned}$$

The above equation implies the following ordinary differential equations (ODEs):

$$\begin{cases} f^{'}(t) + \lambda_a(t) = 0, \\ g_a'(t) + \frac{\rho^2}{2}f(t) - \frac{\lambda_a(t)}{2}\log\frac{\sigma^2 f(t)}{2\pi\lambda_a(t)} = 0, \end{cases}$$

with terminal conditions $f(T) = 1$ and $g(T) = 0$, which are the same as equations (15) and (16).

Using equations (6)-(8), it is easy to get the SDE in equation (19) for the exploratory wealth process under the optimal control, and the verification for results in equations (21) and (22) also follows trivially.

A.3. *Linearly Decaying $\lambda_a(t)$*

THEOREM A.1 *Consider the optimization problem* (10) *with a linearly time-decaying* $\lambda_a(t)$*:*

$$\lambda_a(t) = -2\lambda_0 t + \lambda_1$$

with $\lambda_0, \lambda_1 > 0$ *and* $2\lambda_0 T < \lambda_1$ *to ensure* $\lambda_a(t) > 0$, $\forall t \in [0, T]$*. Then, the value function is given by*

$$\begin{aligned}
V^a(t,x) =& f(t)\log x + \frac{\lambda_0\rho^2}{6}(T^3 - t^3) - \frac{\lambda_1\rho^2}{4}(T^2 - t^2) + \frac{\rho^2}{2}(-\lambda_0 T^2 + \lambda_1 T + 1)(T - t) \\
&+ \left(\frac{1}{2} - \log\frac{\sigma^2}{2\pi}\right)\frac{f(t) - 1}{2} - \frac{f(t)}{2}\log f(t) + \frac{\lambda_a^2(t)}{8\lambda_0}\log\lambda_a(t) - \frac{\lambda_a^2(T)}{8\lambda_0}\log\lambda_a(T)
\end{aligned}$$

where $f(t) = -\lambda_0(T^2 - t^2) + \lambda_1(T - t) + 1$*, and the optimal control is given by*

$$\pi_t^{\lambda*}(u; x) \sim \mathcal{N}\left(u \;\middle|\; \frac{\rho x}{\sigma},\ \frac{x^2}{\sigma^2}\frac{-2\lambda_0 t + \lambda_1}{-\lambda_0(T^2 - t^2) + \lambda_1(T - t) + 1}\right)$$

for which the variance

(1) increases in $[0, T]$ *if* $(2\lambda_0 T - \lambda_1)^2 - 2\lambda_0 \geq 0$*;*

(2) decreases in $[0,T]$ *if* $(\lambda_1-\lambda_0 T)^2+\lambda_0^2T^2-2\lambda_0\le 0$*;*
(3) first increases then decreases in $[0,T]$*, otherwise.*

Proof. We apply Theorem 3.2 and only show how we derive the expression for f and g_a in the conjectured value function $V^a(t,x;\lambda_a(t)) = f(t)\log x + g_a(t)$ and the variance in the optimal Gaussian control distribution.

By virtue of Theorem 3.2, we apply $f^{'}(t)=-\lambda_a(t)$ to get

$$f(t) = -\lambda_0(T^2-t^2)+\lambda_1(T-t)+1.$$

Therefore, the variance of the optimal control is given by

$$\frac{x^2\lambda_a(t)}{\sigma^2 f(t)} = \frac{x^2}{\sigma^2}\frac{-2\lambda_0 t+\lambda_1}{-\lambda_0(T^2-t^2)+\lambda_1(T-t)+1} =: \frac{x^2}{\sigma^2}\zeta(t)$$

for which the derivative is

$$\zeta^{'}(t) = \frac{\kappa(t)}{(-\lambda_0(T^2-t^2)+\lambda_1(T-t)+1)^2}$$

with

$$\kappa(t) := 2\lambda_0^2t^2-2\lambda_0\lambda_1 t+\lambda_1^2+2\lambda_0^2T^2-2\lambda_0\lambda_1 T-2\lambda_0.$$

Therefore, we can focus on the function $\kappa(t)$ and investigate its sign for the changing pattern of the variance term. Clearly κ is decreasing in $t\in[0,T]$ since $\lambda_1>2\lambda_0 T$. Furthermore, we observe that $\kappa(0)=(\lambda_1-\lambda_0T)^2+\lambda_0^2T^2-2\lambda_0$ and $\kappa(T)=(2\lambda_0T-\lambda_1)^2-2\lambda_0$. So, we have

$$\begin{cases}\zeta^{'}(t)\ge 0,\ \forall t\in[0,T], & \text{if } (2\lambda_0T-\lambda_1)^2-2\lambda_0\ge 0,\\ \zeta^{'}(t)\le 0,\ \forall t\in[0,T], & \text{if } (\lambda_1-\lambda_0T)^2+\lambda_0^2T^2-2\lambda_0\le 0,\\ \zeta^{'}(t)\ge 0 \text{ in } [0,\tilde{t}] \text{ and } \le 0 \text{ in } [\tilde{t},T] \text{ for some } \tilde{t}\in(0,T), & \text{otherwise.}\end{cases}$$

The above properties of $\zeta^{'}(t)$ immediately imply the desired monotonicity of the variance of the optimal control as stated in the theorem.

For $g_a(t)$, we also apply Theorem 3.2 to get

$$\begin{aligned} g_a^{'}(t) &= -\frac{\rho^2}{2}f(t)+\frac{\lambda_a(t)}{2}\log\frac{\sigma^2 f(t)}{2\pi\lambda_a(t)}\\ &= -\frac{\rho^2}{2}f(t)-\frac{f^{'}(t)}{2}\log f(t)-\frac{f^{'}(t)}{2}\log\frac{\sigma^2}{2\pi}-\frac{\lambda_a(t)}{2}\log\lambda_a(t).\end{aligned}$$

Then the expression for $g_a(t)$ follows from the facts that

$$\int f(t)\log f(t)\,dt = f(t)\log f(t)-f(t)+C$$

and

$$\int \lambda_a(t)\log\lambda_a(t)\,dt = -\frac{\lambda_a^2(t)}{4\lambda_0}\log\lambda_a(t) + \frac{\lambda_a^2(t)}{8\lambda_0} + C.$$

□

A.4. *Proof of Theorem 5.1*

Since π is admissible, we have, for $\forall\ (t,x)\in[0,T]\times\mathbb{R}_+$,

$$V_t^\pi(t,x) + \int_{\mathbb{R}}\left(\rho\sigma u V_x^\pi(t,x) + \frac{1}{2}\sigma^2u^2V_{xx}^\pi(t,x) - \lambda\log\pi_t(u)\right)\pi_t(u)\,du = 0.$$

From the results in and after Theorem 3.1, the control $\tilde{\pi}$ satisfies

$$\begin{aligned}
V_t^\pi(t,x)+&\int_{\mathbb{R}}\left(\rho\sigma u V_x^\pi(t,x) + \frac{1}{2}\sigma^2u^2V_{xx}^\pi(t,x) - \lambda\log\tilde{\pi}_t(u)\right)\tilde{\pi}_t(u)\,du\\
=&V_t^\pi(t,x) + \max_{\hat{\pi}\in\mathcal{P}(\mathbb{R})}\left\{\int_{\mathbb{R}}\left(\rho\sigma u V_x^\pi(t,x) + \frac{1}{2}\sigma^2u^2V_{xx}^\pi(t,x) - \lambda\log\hat{\pi}_t(u)\right)\hat{\pi}_t(u)\,du\right\} \qquad \text{(A2)}\\
\geq&V_t^\pi(t,x) + \int_{\mathbb{R}}\left(\rho\sigma u V_x^\pi(t,x) + \frac{1}{2}\sigma^2u^2V_{xx}^\pi(t,x) - \lambda\log\pi_t(u)\right)\pi_t(u)\,du = 0.
\end{aligned}$$

Let $\{X_t^{\tilde{\pi}},\ 0\leq t\leq T\}$ denote the exploratory wealth process under the control $\tilde{\pi}$. For a fixed pair $(t,x)\in[0,T]\times\mathbb{R}_+$ and $n\geq 1$, define stopping times

$$\tau_n := \inf\left\{s\geq t : \int_t^s \sigma^2\int_{\mathbb{R}} u^2\tilde{\pi}_v du (V_x^\pi(v,X_v^{\tilde{\pi}}))^2\,dv \geq n\right\},\ \ n=1,2,\ldots$$

and apply Itô's lemma to obtain, for $s\in[t,T]$,

$$\begin{aligned}
V^\pi(s\wedge\tau_n, X_{s\wedge\tau_n}^{\tilde{\pi}}) =& V^\pi(t,x) + \int_t^{s\wedge\tau_n} V_t^\pi(v,X_v^{\tilde{\pi}})\,dv\\
&+\int_t^{s\wedge\tau_n}\int_{\mathbb{R}}\left(\rho\sigma u V_x^\pi(v,X_v^{\tilde{\pi}}) + \frac{1}{2}\sigma^2u^2V_{xx}^\pi(v,X_v^{\tilde{\pi}})\right)\tilde{\pi}_v(u)\,du\,dv\\
&+\int_t^{s\wedge\tau_n}\sigma\sqrt{\int_{\mathbb{R}}u^2\tilde{\pi}_v du}\cdot V_x^\pi(v,X_v^{\tilde{\pi}})\,dW_v.
\end{aligned}$$

Rearranging the above equation and applying the inequality in (A2), we get

$$\begin{aligned}
V^\pi(t,x) =& \mathbb{E}\Bigg[V^\pi(s\wedge\tau_n, X^{\tilde\pi}_{s\wedge\tau_n}) - \int_t^{s\wedge\tau_n} V_t^\pi(v, X_v^{\tilde\pi})\,dv \\
&- \int_t^{s\wedge\tau_n}\int_{\mathbb{R}} \left(\rho\sigma u V_x^\pi(v, X_v^{\tilde\pi}) + \frac{1}{2}\sigma^2u^2V_{xx}^\pi(v, X_v^{\tilde\pi})\right)\tilde\pi_v(u)\,du\,dv \;\Bigg|\; X_t^{\tilde\pi} = x\Bigg] \\
\le& \mathbb{E}\left[V^\pi(s\wedge\tau_n, X^{\tilde\pi}_{s\wedge\tau_n}) - \int_t^{s\wedge\tau_n}\int_{\mathbb{R}} \lambda\tilde\pi_v(u)\log\tilde\pi_v(u)\,du\,dv \;\middle|\; X_t^{\tilde\pi} = x\right].
\end{aligned}$$

At time $s = T$, the above inequality holds since $s \in [t, T]$. As $n \to \infty$, $T \wedge \tau_n = T$. By the Dominated Convergence Theorem, we have, as $n \to \infty$,

$$\begin{aligned}
V^\pi(t,x) \le& \mathbb{E}\left[V^\pi(T\wedge\tau_n, X^{\tilde\pi}_{T\wedge\tau_n}) - \int_t^{T\wedge\tau_n}\int_{\mathbb{R}} \lambda\tilde\pi_v(u)\log\tilde\pi_v(u)\,du\,dv \;\middle|\; X_t^{\tilde\pi} = x\right] \\
=& \mathbb{E}\left[V^\pi(T, X_T^{\tilde\pi}) - \int_t^{T}\int_{\mathbb{R}} \lambda\tilde\pi_v(u)\log\tilde\pi_v(u)\,du\,dv \;\middle|\; X_t^{\tilde\pi} = x\right] \\
=& \mathbb{E}\left[\log X_T^{\tilde\pi} - \int_t^{T}\int_{\mathbb{R}} \lambda\tilde\pi_v(u)\log\tilde\pi_v(u)\,du\,dv \;\middle|\; X_t^{\tilde\pi} = x\right] \\
=& \mathbb{E}\left[V^{\tilde\pi}(T, X_T^{\tilde\pi}) - \int_t^{T}\int_{\mathbb{R}} \lambda\tilde\pi_v(u)\log\tilde\pi_v(u)\,du\,dv \;\middle|\; X_t^{\tilde\pi} = x\right] \\
=& V^{\tilde\pi}(t,x).
\end{aligned}$$

Appendix B: RL Algorithms

B.1. *RL Algorithm with Portion Control*

For the exploratory problem controlling the investment portion, we parametrize the value function as

$$V^\pi(t,x;\boldsymbol{\alpha}) = \log x + \alpha(T-t).$$

We also have the mean of the Gaussian control parametrized as $\beta_1 = \frac{\rho}{\sigma}$ and

$$-\int_{\mathbb{R}} \pi_t(u;\boldsymbol{\beta})\log\pi_t(u;\boldsymbol{\beta})du = -\beta_2 = \frac{1}{2}\log\frac{2\pi e\lambda}{\sigma^2}.$$

The updating scheme for β_1 goes as follows:

$$\beta_1 \leftarrow \left(\frac{2\alpha + 2\lambda\beta_2 + \lambda}{4\pi\lambda\beta_1}e^{-2\beta_2-1} + \frac{\beta_1}{2}\right). \tag{B1}$$

The gradients of the TD error in α and β_2 at time t_i are

$$\frac{\partial C_i}{\partial \alpha} = \sum_{(t_j,x_j)\in\mathcal{S}_i} \left(\dot{V}^\pi(t_j,x_j;\boldsymbol{\alpha}) - \lambda\beta_2\right)(-\Delta t). \tag{B2}$$

$$\frac{\partial C_i}{\partial \beta_2} = \sum_{(t_j,x_j)\in\mathcal{S}_i} \left(\dot{V}^\pi(t_j,x_j;\boldsymbol{\alpha}) - \lambda\beta_2\right)(-\lambda\Delta t) \tag{B3}$$

The algorithm is summarized by pseudocode in Algorithm 2.

Algorithm 2: RL Algorithm with Portion Control

Input: Market parameters (μ,σ,r,ρ), learning rate θ_α, θ_β, initial wealth x_0, investment horizon T, discretization Δt, exploration rate λ.
Initialization: $i=1$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$
while $i \le \frac{T}{\Delta t}$ **do**
  Sample (t_i,x_i) under $\pi(u;\boldsymbol{\beta})$
  Update set of samples $\mathcal{S}_i = \{(t_j,x_j); j=0,\dots,i\}$
  Update α as $\alpha - \theta_\alpha\nabla_\alpha C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ using (B2)
  Update β_2 as $\beta_2 - \theta_\beta\nabla_\beta C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ using (B3)
  Update β_1 using (B1)
  Update $\pi_t(u;x,\boldsymbol{\alpha},\boldsymbol{\beta})$ as $\mathcal{N}\left(u \,\middle|\, \beta_1,\ \frac{e^{-2\beta_2-1}}{2\pi}\right)$
  $i = i+1$
end

B.2. *RL Algorithm with Power-Decaying λ*

For the exploratory problem with a power-decaying λ and $\lambda_0 \neq 1$, we parametrize the value function as

$$V^\pi(t,x;\boldsymbol{\alpha}) = \left(\frac{T+\lambda_1}{t+\lambda_1}\right)^{\lambda_0}\log x + \alpha_1\frac{\left(\frac{T+\lambda_1}{t+\lambda_1}\right)^{\lambda_0-1}-1}{\lambda_0-1} + \alpha_2\left(\frac{T+\lambda_1}{t+\lambda_1}\right)^{\lambda_0} - \frac{1}{2}\left(\frac{T+\lambda_1}{t+\lambda_1}\right)^{\lambda_0}\log(t+\lambda_1) + \alpha_3.$$

We also have the mean of the Gaussian control parametrized as $\beta_1 = \frac{\rho}{\sigma}$ and

$$-\int_{\mathbb{R}} \pi_t(u;\boldsymbol{\beta})\log\pi_t(u;\boldsymbol{\beta})du = \log x - \frac{1}{2}\log(t+\lambda_1) - \beta_2$$

where $\beta_2 = -\frac{1}{2}\log\frac{2\pi e\lambda_0}{\sigma^2}$. The updating scheme for β_1 goes as follows:

$$\beta_1 \leftarrow \left(\frac{\alpha_1 e^{-2\beta_2-1}}{2\pi\lambda_0(T+\lambda_1)\beta_1} + \frac{\beta_1}{2}\right). \tag{B4}$$

The updating scheme for α_3 also applies the terminal condition and goes as follows:

$$\alpha_3 \leftarrow \left(-\alpha_2 + \frac{1}{2}\log\left(T+\lambda_1\right)\right). \tag{B5}$$

The gradients of the TD error in α_1, α_2 and β_2 at time t_i are

$$\frac{\partial C_i}{\partial \alpha_1} = \sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^{\pi}(t_j,x_j;\boldsymbol{\alpha}) - \lambda(t_{j+1})\beta_2\right)\frac{1}{\lambda_0-1}\left(\left(\frac{T+\lambda_1}{t_{j+1}+\lambda_1}\right)^{\lambda_0-1} - \left(\frac{T+\lambda_1}{t_j+\lambda_1}\right)^{\lambda_0-1}\right) \tag{B6}$$

$$\frac{\partial C_i}{\partial \alpha_2} = \sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^{\pi}(t_j,x_j;\boldsymbol{\alpha}) - \lambda(t_{j+1})\beta_2\right)\left(\left(\frac{T+\lambda_1}{t_{j+1}+\lambda_1}\right)^{\lambda_0} - \left(\frac{T+\lambda_1}{t_j+\lambda_1}\right)^{\lambda_0}\right). \tag{B7}$$

$$\frac{\partial C_i}{\partial \beta_2} = \sum_{(t_j,x_j)\in\mathcal{S}_i}\left(\dot{V}^{\pi}(t_j,x_j;\boldsymbol{\alpha}) - \lambda(t_{j+1})\beta_2\right)\left(-\lambda(t_{j+1})\Delta t\right) \tag{B8}$$

The algorithm is summarized by pseudocode in Algorithm 3.

Algorithm 3: RL Algorithm with Power-Decaying λ

Input: Market parameters (μ,σ,r,ρ), learning rate θ_α, θ_β, initial wealth x_0, investment horizon T, discretization Δt, exploration rates λ_0, λ_1.
Initialization: $i=1$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$
while $i \le \frac{T}{\Delta t}$ **do**
  Sample (t_i, x_i) under $\pi(u;\boldsymbol{\beta})$
  Update set of samples $\mathcal{S}_i = \{(t_j,x_j); j=0,\dots,i\}$
  Update $(\alpha_1,\alpha_2)'$ as $(\alpha_1,\alpha_2)' - \theta_\alpha \nabla_\alpha C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ using (B6) and (B7)
  Update α_3 using (B5)
  Update β_2 as $\beta_2 - \theta_\beta \nabla_\beta C_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ using (B8)
  Update β_1 using (B4)
  Update $\pi_t(u;x,\boldsymbol{\alpha},\boldsymbol{\beta})$ as $\mathcal{N}\left(u \;\middle|\; \beta_1 x,\ \frac{e^{-2\beta_2-1}x^2}{2\pi(t+\lambda_1)}\right)$
  $i = i+1$
end