Article

Not peer-reviewed version

Exploratory Dividend Optimization with Entropy Regularization

Sang Hu * and Zihan Zhou

Posted Date: 28 November 2023

doi: 10.20944/preprints202311.1796.v1

Keywords: Dividend optimization; entropy regularization; distributional control; exploratory HJB

Article

Exploratory Dividend Optimization with Entropy Regularization

Sang Hu * and Zihan Zhou

School of Data Science, The Chinese University of Hong Kong, Shenzhen, China 518172; zihanzhou@link.cuhk.edu.cn

* Correspondence: husang@cuhk.edu.cn

Abstract: This paper studies the dividend optimization problem in the entropy regularization framework by following the same continuous-time reinforcement learning setting as in Wang *et al.* (2020). The exploratory HJB is established and the optimal exploratory dividend policy is a truncated exponential distribution. We show that, for suitable choices of the maximal dividend paying rate and the temperature parameter, the value function of the exploratory dividend optimization problem could be significantly different from the value function in the classical dividend optimization problem. In particular, the value function of the exploratory dividend optimization problem could be classified into three cases based on its monotonicity. Numerical examples are also presented to show the impact of temperature parameter on the solution.

Keywords: dividend optimization; entropy regularization; distributional control; exploratory HJB

1. Introduction

The risk management problem of an insurance company has been studied extensively in the literature. It dates back to the Cramér-Lundberg (C-L) model of Lundberg (1903) which described the surplus process of the insurance company in terms of two cash flows: premiums received and claims paid. Consider an insurance company with claims arriving at Poisson rate ν, i.e., the total number of claims N_t up to time t is Poisson distributed with parameter νt. Denote by ξ_i the size of the i-th claim, where $\{\xi_i\}$'s are independently and identically distributed with $\mathbb{E}[\xi_i] = \mu_1$ and $\mathbb{E}[\xi_i^2] = \mu_2$ for some constants $\mu_1, \mu_2 > 0$. Let $\tilde{X}_t$ denote the surplus process of the insurance company. Then

$$\tilde{X}_t = x_0 + \zeta t - \sum_{i=1}^{N_t} \xi_i,$$

where x_0 is the initial surplus level, and ζ is the premium rate which is the amount of premium received by the insurance company per unit of time.

De Finetti (1957) first proposed the dividend optimization problem: An insurance company maximizes the expectation of cumulative discounted dividends until the ruin time by choosing dividend strategies, that is, when and how much of the surplus should be distributed as dividends to the shareholders. De Finetti (1957) derived that the optimal dividend policy under a simple discrete random walk model should be a barrier strategy. Gerber (1969) then generalized the dividend problem from a discrete-time model to the classical C-L model and showed that the optimal dividend strategy should be band strategy, which degenerates to a barrier strategy for exponentially distributed claim size.

With the development of technical tools such as dynamic programming, the dividend optimization problem has been analyzed under the stochastic control framework. In particular, $\tilde{X}_t$ in C-L model can be approximated by a diffusion process X_t that evolves according to

$$dX_t = \mu dt + \sigma dW_t,$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

where $\mu := \zeta - \nu\mu_1$, $\sigma := \sqrt{\nu\mu_2}$, and $\{W_t\}$ is a standard Brownian motion; see, e.g., Schmidli (2007). It is worth noting that the diffusion approximation for the surplus process works well for large insurance portfolios, where an individual claim is relatively small compared to the size of the surplus. Under the drifted Brownian motion model the optimal dividend strategy is a barrier strategy, and if the dividend rate is further upper bounded, the optimal dividend strategy is threshold-type; see, e.g., Jeanblanc-Picqué and Shiryaev (1995) and Asmussen and Taksar (1997). Other extensions on the dividend optimization problem include Jgaard and Taksar (1999), Asmussen *et al.* (2000), Azcue and Muler (2005), Azcue and Muler (2010), Gaier *et al.* (2003), Kulenko and Schmidli (2008), Yang and Zhang (2005), Choulli *et al.* (2003), Gerber and Shiu (2006), Avram *et al.* (2007) and Yin and Wen (2013), etc.

Previous literature studied the dividend optimization problem based on the complete information of the environment, i.e., all the model parameter values are known. This assumption is no longer valid if the environment is a black-box or the model parameter values are unknown. One way to handle this issue is to use the past information to estimate the model parameters and then solve the problem with the estimated parameters. However, the optimal strategy in the classical dividend optimization problem is a barrier-type or threshold-type, which is extremely sensitive to the model parameter values; a slight change in the model parameters would lead to a totally different strategy.[1]

In contrast to the traditional approach that separates the estimation and optimization, reinforcement learning aims to learn the optimal strategy through trial-and-error interactions with the unknown environment without estimating the model parameters. In particular, one takes different actions in the unknown territory and receives feedbacks to learn the optimal action and use it to further interact with the environment. In recent years, reinforcement learning had successful applications in many fields such as health-care, autonomous control, natural language processing, and video games; see, e.g., Zhao *et al.* (2009), Komorowski *et al.* (2018), Mirowski *et al.* (2016), Zhu *et al.* (2017), Radford *et al.* (2017), Paulus *et al.* (2017), Mnih *et al.* (2015), Jaderberg *et al.* (2019), Silver *et al.* (2016), Silver *et al.* (2017). There is no doubt that reinforcement learning has become one of the most popular and fastest-growing fields today.

Exploration and exploitation are the key concepts in reinforcement learning, and they proceed simultaneously. On one hand, exploitation is to utilize the so-far-known information to derive the current optimal strategy which might not be optimal from the long-term view. On the other hand, exploration emphasizes learning from trial-and-error interactions with the environment to improve its knowledge for the sake of long-term benefit. While the optimal strategy of the classical dividend optimization problem is deterministic when the model parameter values are fully known, randomized strategies are considered to encourage exploration of other actions in the unknown environment. Although the exploration causes a cost in the short term, it helps to learn the optimal (or near-optimal) strategy and bring benefit from the long-term point of view.

Obviously, how to balance the trade-off between exploitation and exploration is an important issue. The ε-greedy strategy is a frequently used randomized strategy in reinforcement learning. It balances the exploration and exploitation by illustrating that the agent should stick to the current optimal policy most of the time, while the agent could sometimes randomly take other non-optimal actions to explore the environment; see, e.g., Auer *et al.* (2002). Boltzmann exploration is another randomized strategy extensively studied in RL literature. Instead of assigning constant probabilities to different actions based on current information, Boltzmann exploration uses the Boltzmann distribution to allocate the probability to different actions, where the probability of each action is positively related to its reward. In other words, agent should choose action with higher expected rewards with higher probability; see, e.g., Cesa-Bianchi *et al.* (2017).

---

[1] For example, the dividend paying rate under the threshold strategy is the maximal rate if the surplus exceeds the threshold; otherwise, it pays nothing. Since the threshold is determined by the model parameters, the change in the estimated parameters may dramatically change the dividend paying rate from zero to the maximal rate, or conversely.

 doi:10.20944/preprints202311.1796.v1

Another way to introduce a randomized strategy is to intentionally include a regularization term to encourage exploration. Entropy is a frequently used criterion in the RL family that measures the level of exploration. The entropy regularization framework directly incorporates entropy as a regularization term into the original objective function to encourage exploration; see, e.g., Todorov (2006), Ziebart *et al.* (2008), Nachum *et al.* (2017). In the entropy regularization framework, the weight of exploration is determined by the coefficient imposed on the entropy, which is called the temperature parameter. The larger the temperature parameter, the greater the weight of exploration. A temperature parameter that is too large may result in too much focus on exploring the environment and little effort in exploiting the current information; otherwise, if the temperature parameter is too small, one may stick to the current optimal strategy without the opportunity to explore better solutions. Therefore, careful selections of the temperature parameter is important for designing reinforcement learning algorithms.

While most existing literature in reinforcement learning focus on the Markov decision process, recently Wang *et al.* (2020) extended the entropy regularization framework to the continuous-time setting. The authors showed that the optimal distributional control is Gaussian distribution in the linear-quadratic stochastic control problem. In the series work, Wang and Zhou (2020) studied continuous-time mean-variance portfolio selection problem under the entropy-regularized RL framework and showed that the precommitted strategies are Gaussian distributions with time-decaying variance. Dai *et al.* (2023) considered the equilibrium mean-variance problem with log return target and showed that the optimal control is Gaussian distribution with the variance term not necessarily decaying in time.

This paper studies the dividend optimization problem in the entropy regularization framework to encourage the exploration in the unknown environment. We follow the same setting as in Wang *et al.* (2020) which use Shannon's differential entropy. The key idea is to use distribution as the control to solve the entropy-regularized dividend optimization problem. Consequently, the optimal dividend policy is a randomization over the possible dividend paying rates. We derive the so-called exploratory HJB and establish the theoretical results to guarantee the existence of the solution. We obtain that the optimal exploratory dividend policy is a truncated exponential distribution whose parameter depends on the surplus level and the temperature parameter. We show that, for suitable choices of the maximal dividend paying rate and the temperature parameter, the value function of the exploratory dividend optimization problem could be significantly different from the value function in the traditional problem. In particular, we classify the value function of the exploratory dividend optimization problem into three cases based on its monotonicity.

Recently, Bai *et al.* (2023) also study the optimal dividend problem under the continuous time diffusion model. The authors then use a policy improvement argument along with policy evaluation devices to construct approximating sequences of the optimal strategy. The difference is that in their paper the feasible controls are open-loop, while we consider feedback controls only. We show that the value function is decreasing when the maximal dividend paying rate is relatively small compared to the temperature parameter, where in their paper the maximal dividend paying rate is assumed to be larger than one and thus the value function is always increasing.

The rest of the paper is organized as follows. In Section 2, we introduce the formulation of the entropy-regularized dividend optimization problem. In Section 3, we present the exploratory HJB and the theoretical results to solve the exploratory dividend problem. We then discuss the three cases of the value function for the exploratory dividend problem in Section 4. Some numerical examples to show the impact of parameters on the optimal dividend policy and the value function are presented in Section 5. Section 6 concludes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

2. Problem

2.1. The Model

Suppose an insurance company has surplus X_t at time t, with

$$dX_t = \mu dt + \sigma dW_t, \quad X_0 = x, \tag{1}$$

where $\mu > 0$, $\sigma > 0$, and $\{W_t\}_{t\geq 0}$ is a standard Brownian motion defined on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$. As remarked by Asmussen and Taksar (1997), such a surplus process (1) can be viewed either as a direct modelling with drifted Brownian motion or as an approximation to the classical compound Poisson model.

A dividend strategy or policy is defined as $\mathbf{a} = \{a_t\}_{t\geq 0}$, where a_t is the dividend paying rate at time t, i.e., the cumulative amount of dividends paid from time t_1 to time t_2 is given by $\int_{t_1}^{t_2} a_t dt$. We consider herein the Markov feedback controls, i.e., $a_t = a(X_t)$, where $a(\cdot)$ is a function of the surplus level X_t. Note that a_t is nonnegative for any t. Further, we assume that a_t is upper bounded by a positive constant M, which is consistent with the assumption made in the literature. We give the formal definition of the admissible dividend policy below.

Definition 1. *A dividend policy* $\mathbf{a}$ *is said to be admissible if* $\{a_t\}_{t\geq 0}$ *is* $\mathcal{F}_t$*-adapted and* $a_t \in [0, M]$ *for all* $t \geq 0$.

Denote by $\mathcal{A}$ the set of admissible dividend policies. For an insurance company whose surplus process evolves according to (1) and pays the dividend according to policy $\mathbf{a} = \{a_t\}_{t\geq 0} \in \mathcal{A}$, the controlled surplus process for this insurance company is

$$dX_t^{\mathbf{a}} = (\mu - a_t)dt + \sigma dW_t, \quad X_0^{\mathbf{a}} = x. \tag{2}$$

Define the ruin time to be the first time that the surplus level hits zero, i.e.,

$$\tau_x^{\mathbf{a}} := \inf\{t \geq 0 : X_t^{\mathbf{a}} \leq 0 \mid X_0^{\mathbf{a}} = x\}.$$

For an insurance company starting with initial surplus $x \in [0, \infty)$, the problem is to find the optimal dividend policy that maximizes the expected value of exponentially discounted dividends to be accumulated until the ruin time, that is,

$$J_{cl}(x, \mathbf{a}) := \mathbb{E}\left[\int_0^{\tau_x^{\mathbf{a}}} e^{-\rho t} a(X_t^{\mathbf{a}}) dt\right],$$

where $\rho > 0$ is the discounting rate. Then the optimal dividend problem is

$$\sup_{\mathbf{a}\in\mathcal{A}} J_{cl}(x, \mathbf{a}). \tag{3}$$

2.2. Classical Optimal Dividend Problem

First, we briefly review the results of solving the dividend optimization problem (3) classically. Let $V_{cl}(x)$ be the value function of the dividend optimization problem:

$$V_{cl}(x) := \sup_{\mathbf{a}\in\mathcal{A}} J_{cl}(x, \mathbf{a}).$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

Assume that the value function $V_{cl}(x)$ is twice-continuously differentiable. The standard dynamic programming approach leads to the following Hamilton-Jacobi-Bellman equation,

$$\rho V_{cl}(x) = \sup_{a\in[0,M]} \left\{ a + (\mu - a)V'_{cl}(x) + \frac{1}{2}\sigma^2 V''_{cl}(x) \right\}, \tag{4}$$

with boundary condition $V_{cl}(0) = 0$. It can be easily seen that the optimal dividend paying rate at surplus level x is

$$a^*(x) = \begin{cases} 0, & \text{if } V'_{cl}(x) > 1, \\ M, & \text{if } V'_{cl}(x) \leq 1. \end{cases} \tag{5}$$

Assume that $V_{cl}(x)$ is a concave function. Then there exists a nonnegative constant x_b such that $V'_{cl}(x) \leq 1$ when $x \geq x_b$ and $V'_{cl}(x) > 1$ when $0 \leq x < x_b$. Substitute (5) into (4), then it turns into the following ODEs:

$$\begin{cases} \frac{1}{2}\sigma^2 V''_{cl}(x) + \mu V'_{cl}(x) - \rho V_{cl}(x) = 0, & \text{if } 0 \leq x < x_b, \\ \frac{1}{2}\sigma^2 V''_{cl}(x) + (\mu - M)V'_{cl}(x) - \rho V_{cl}(x) + M = 0, & \text{if } x \geq x_b. \end{cases} \tag{6}$$

Combining with the boundary condition, one can derive that

$$V_{cl}(x) = \begin{cases} C_1 \left(e^{\theta_1 x} - e^{-\theta_2 x} \right), & \text{if } 0 \leq x < x_b, \\ \frac{M}{\rho} - C_2 e^{-\theta_3 x}, & \text{if } x \geq x_b, \end{cases} \tag{7}$$

where

$$\theta_1 = \frac{-\mu + \sqrt{\mu^2 + 2\rho\sigma^2}}{\sigma^2},\ \theta_2 = \frac{\mu + \sqrt{\mu^2 + 2\rho\sigma^2}}{\sigma^2},\ \theta_3 = \frac{(\mu - M) + \sqrt{(\mu - M)^2 + 2\rho\sigma^2}}{\sigma^2}.$$

C_1, C_2 and x_b are determined by the smooth pasting conditions, i.e.,

$$\begin{cases} C_1 \left(e^{\theta_1 x_b} - e^{-\theta_2 x_b} \right) = \frac{M}{\rho} - C_2 e^{-\theta_3 x_b}, \\ C_1 \left(\theta_1 e^{\theta_1 x_b} + \theta_2 e^{-\theta_2 x_b} \right) = 1, \\ C_2 \theta_3 e^{-\theta_3 x_b} = 1. \end{cases} \tag{8}$$

If $\frac{M}{\rho} - \frac{1}{\theta_3} > 0$, there exists unique solution to (8). In this case, $V_{cl}(x)$ is given by (7), where C_1, C_2, x_b are determined uniquely through (8). Consequently, the optimal dividend policy is to pay the maximal rate M when surplus level x exceeds threshold x_b and to pay nothing if else. If $\frac{M}{\rho} - \frac{1}{\theta_3} \leq 0$, then $V_{cl}(x) = \frac{M}{\rho}(1 - e^{-\theta_3 x})$. In this case, the optimal dividend policy is always to pay the maximal rate M. Detailed proofs can be found in Asmussen and Taksar (1997). It is also straightforward to check that the optimal value function $V_{cl}(x)$ is concave on x and always smaller than M/ρ which is the limit of $V_{cl}(x)$ as x going to infinity. Figure 1 below illustrates the value function and the corresponding optimal dividend policy under the following parameter values: $\mu = 1$, $\sigma = 1$, $\rho = 0.3$, $M = 0.6$ (left panels), $M = 1.2$ (middle panels), and $M = 1.8$ (right panels), respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

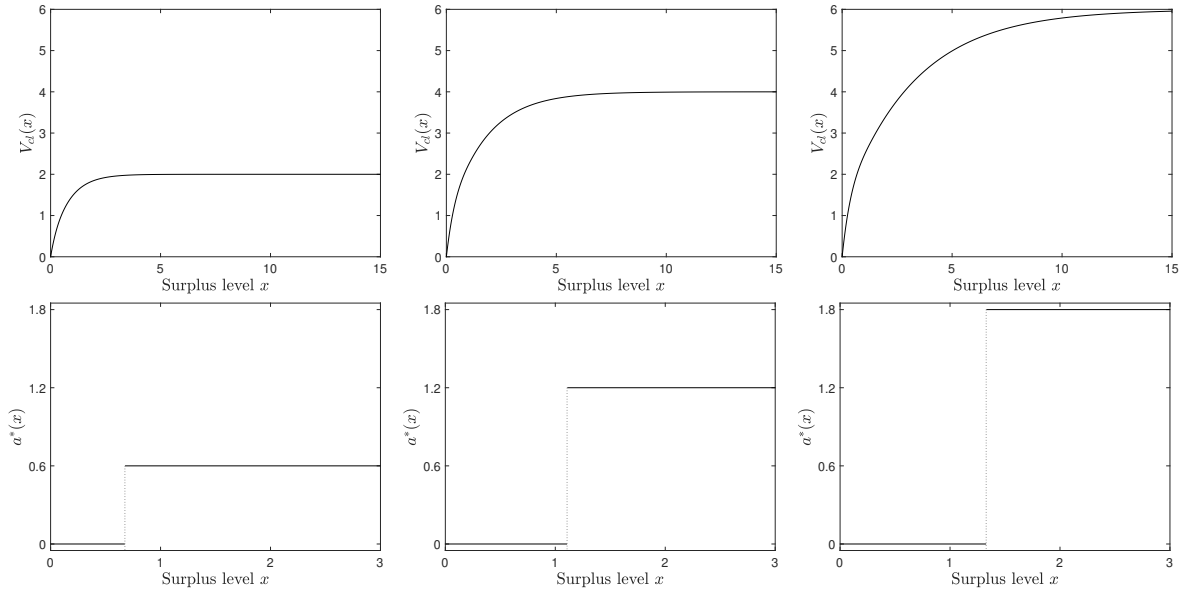

Figure 1. The classical value functions (top) and the optimal dividend-paying rate (bottom) for $\mu = 1$, $\sigma = 1$, $\rho = 0.3$, $M = 0.6$ (left panels), $M = 1.2$ (middle panels), and $M = 1.8$ (right panels), respectively.

2.3. Exploratory Formulation

The above optimal dividend policy (5) is implemented based on the complete information, i.e., the model parameters μ and σ are known. In reality, however, it is difficult to know exactly the values of μ and σ, due to the uncertainty in premium rate, claim arrival process and claim size. Therefore, we consider a technique named reinforcement learning to learn the optimal (or near-optimal) dividend paying strategy through trial-and-error interactions with the unknown territory.

Whereas the majority of the work in reinforcement learning consider Markov decision process in discrete time, we follow the pioneering work of Wang *et al.* (2020) who model the reinforcement learning in continuous time as a relaxed stochastic control problem. At time t with surplus level X_t the dividend paying rate a is randomly sampled according to a distribution $\pi_t := \pi(a; X_t)$, where $\pi(\cdot;\cdot) : [0, M] \times [0, \infty) \mapsto [0, \infty)$, satisfying $\int_0^M \pi(a; x)da = 1$ for any $x \in [0, \infty)$. We call $\boldsymbol{\pi} := \{\pi_t\}_{t\geq 0}$ the distributional dividend policy. Following the same procedure as in Wang *et al.* (2020), we derive the exploratory dynamic of the surplus process under $\boldsymbol{\pi}$ to be

$$dX_t^{\boldsymbol{\pi}} = \left(\mu - \int_0^M a\pi(a; X_t^{\boldsymbol{\pi}})da\right) dt + \sigma dW_t, \quad X_0^{\boldsymbol{\pi}} = x, \tag{9}$$

and the expected value of total discounted dividends under exploration to be

$$\mathbb{E}\left[\int_0^{\tau_x^{\boldsymbol{\pi}}} e^{-\rho t} \int_0^M a\pi(a; X_t^{\boldsymbol{\pi}})dadt\right],$$

where the ruin time is

$$\tau_x^{\boldsymbol{\pi}} := \inf\{t \geq 0 : X_t^{\boldsymbol{\pi}} \leq 0 \mid X_0^{\boldsymbol{\pi}} = x\}.$$

In addition to the expected value of total discounted dividends under exploration, Shannon's differential entropy is introduced into the objective to encourage exploration. For a given distribution π, the entropy is defined as

$$\mathcal{H}(\pi) := -\int_0^M \pi(a) \ln \pi(a)da. \tag{10}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

Thus, the objective of entropy-regularized exploratory dividend problem is

$$\begin{aligned} J(x,\boldsymbol{\pi}) :&= \mathbb{E}\left[\int_0^{\tau_x^{\pi}} e^{-\rho t}\left(\int_0^M a\pi(a;X_t^{\pi})da + \lambda\mathcal{H}(\pi_t)\right)dt\right] \\ &= \mathbb{E}\left[\int_0^{\tau_x^{\pi}} e^{-\rho t}\left(\int_0^M \left(a - \lambda\ln\pi(a;X_t^{\pi})\right)\pi(a;X_t^{\pi})da\right)dt\right], \end{aligned} \tag{11}$$

where $\lambda > 0$ is the so-called temperature parameter. Note that λ controls the weight to be put on the exploration and is exogenously given. If $\lambda = 0$, the distribution degenerates to the Dirac measure, which is the solution to classical optimal dividend problem without exploration. The entropy-regularized exploratory dividend problem is

$$\sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} J(x,\boldsymbol{\pi}), \tag{12}$$

where $\boldsymbol{\Pi}$ is the set of admissible exploratory dividend policies. We give the formal definition of admissible exploratory dividend policy $\boldsymbol{\pi}$ below.

Definition 2. *An exploratory dividend policy $\boldsymbol{\pi}$ is admissible if the following conditions are satisfied:*

(i) $\pi(\cdot;x) \in \Pi_{[0,M]}$ *for any* $x \in [0,\infty)$, *where* $\Pi_{[0,M]}$ *is a set of probability density functions with support* $[0,M]$;
(ii) the stochastic differential equation (9) has a unique solution $\{X_t^{\pi}\}_{t\geq 0}$ *under* $\boldsymbol{\pi}$;
(iii) $\mathbb{E}\left[\int_0^{\tau_x^{\pi}} e^{-\rho t}\left|\int_0^M \left(a - \lambda\ln\pi(a;X_t^{\pi})\right)\pi(a;X_t^{\pi})da\right|dt\right] < \infty$.

The following proposition will be used later.

Proposition 1. *For any distribution π on support $[0,M]$, the entropy $\mathcal{H}(\pi) \leq \ln M$.*

3. Exploratory HJB Equation

To solve exploratory optimal dividend problem (12), we first derive the corresponding HJB equation, the so-called exploratory HJB; see Wang *et al.* (2020), Tang *et al.* (2022), etc.

Let $V(x)$ be the value function of entropy-regularized exploratory dividend problem, that is,

$$V(x) := \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} J(x,\boldsymbol{\pi}).$$

Assume that the value function $V(x)$ is twice-continuously differentiable. Following the standard arguments in dynamic programming, we derive the exploratory HJB equation below:

$$\rho V(x) = \sup_{\pi\in\Pi_{[0,M]}}\left\{\int_0^M \left[a(1-V'(x)) - \lambda\ln\pi(a;x)\right]\pi(a;x)da\right\} + \mu V'(x) + \frac{1}{2}\sigma^2 V''(x), \tag{13}$$

with boundary condition

$$V(0) = 0. \tag{14}$$

3.1. Exploratory Dividend Policy

To solve the supremum in (13) together with the constraint that $\int_0^M \pi(a;x)da = 1$, we introduce the Lagrange multiplier η:

$$\sup_{\pi\in\Pi_{[0,M]}}\left\{\int_0^M \left[a(1-V'(x)) - \lambda\ln\pi(a;x) - \eta\right]\pi(a;x)da + \eta\right\}.$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

Maximizing the integrand above pointwisely and using the first-order condition leads to the solution

$$\pi^*(a;x) = \exp\left(a\left(\frac{1-V'(x)}{\lambda}\right) - 1 - \frac{\eta}{\lambda}\right),\ a \in [0,M].$$

Because $\int_0^M \pi^*(a;x)da = 1$, we solve that

$$\pi^*(a;x) = \frac{1}{Z_M((1-V'(x))/\lambda)} \exp\left(a\left(\frac{1-V'(x)}{\lambda}\right)\right),\ a \in [0,M], \tag{15}$$

where

$$Z_M(y) := \begin{cases} \frac{e^{My}-1}{y}, & y \neq 0 \\ M, & y = 0. \end{cases} \tag{16}$$

Recall that the classical optimal dividend policy given in (5) is two-threshold strategy, i.e., it pays nothing, $a^*(x) = 0$, if $V_{cl}'(x) > 1$ or pays the maximal rate, $a^*(x) = M$, if $V_{cl}'(x) \leq 1$. In contrast, the exploratory dividend policy is not restricted to two extreme actions only but gives the probability to take certain actions. This result is very similar to Gao *et al.* (2022) in which the authors study the temperature control problem for Langevin diffusions by incorporating randomization of the temperature control and regularizing its entropy. The classical optimal control of such a problem is of the bang-bang type, whereas the exploratory control is a state-dependent, truncated exponential distribution. Likewise, the optimal distribution $\pi^*(a;x)$ given in (15) is also a continuous version of Boltzmann distribution or Gibbs measure which is widely used in discrete reinforcement learning.

When $V'(x) > 1$, $\pi^*(a;x)$ is decreasing in a so it has large probability to take small dividend pay-out rate close to 0; when $V'(x) < 1$, $\pi^*(a;x)$ is increasing in a so it has large probability to take large dividend pay-out rate close to M ; when $V'(x) = 1$, it degenerates to a uniform distribution on $[0,M]$. In other words, the optimal exploratory dividend policy is an "exploration" of the classical dividend pay-out policy: it searches around the current optimal dividend rate given by the classical solution, 0 or M, with the probability to take a certain rate decreasing as it moving away from the classical solution.

The exploratory surplus process under the optimal policy is well-posed. Note that the optimal distributional policy is $\boldsymbol{\pi^*} = \{\pi_t^*\}_{t\geq 0}$, where

$$\pi_t^* := \pi^*(a; X_t^{\pi^*}) = \frac{1}{Z_M((1-V'(X_t^{\pi^*}))/\lambda)} \exp\left(a\left(\frac{1-V'(X_t^{\pi^*})}{\lambda}\right)\right). \tag{17}$$

Applying the optimal distributional policy (17) into the exploratory surplus process (9), we obtain that

$$\begin{aligned} dX_t^{\pi^*} &= \left(\mu - \int_0^M a\pi^*(a;X_t^{\pi^*})da\right)dt + \sigma dW_t \\ &= \left[\mu - \left(M - \frac{\lambda}{1-V'(X_t^{\pi^*})} + \frac{M}{e^{M(1-V'(X_t^{\pi^*}))/\lambda} - 1}\right)\mathbf{1}_{V'(X_t^{\pi^*})\neq 1} - \left(\frac{M}{2}\right)\mathbf{1}_{V'(X_t^{\pi^*})=1}\right]dt \\ &\quad + \sigma dW_t. \end{aligned} \tag{18}$$

Since $\int_0^M a\pi^*(a;X_t^{\pi^*})da \in [0,M]$, the SDE (18) has bounded drift and constant volatility. As a result, there exists unique solution $\{X_t^{\pi^*}\}$ to (18).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

3.2. Verification Theorem

Substituting the optimal distribution $\pi^*(a;x)$ as shown in (15) into the HJB equation (13), we have the following equation for $V(x)$:

$$\rho V(x) = \mu V'(x) + \frac{1}{2}\sigma^2 V''(x) + \lambda \ln Z_M((1 - V'(x))/\lambda),$$

or equivalently,

$$\begin{aligned}\rho V(x) =& \mu V'(x) + \frac{1}{2}\sigma^2 V''(x) \\ &+ \lambda \ln\left(\frac{\lambda}{1 - V'(x)}\left(e^{M(1-V'(x))/\lambda} - 1\right)\mathbf{1}_{V'(x)\neq 1} + M\mathbf{1}_{V'(x)=1}\right).\end{aligned} \tag{19}$$

The following verification theorem shows that $V(x)$ that solves (19) is indeed the value function of the exploratory dividend problem (12).

Theorem 1. *Assume there exists twice-continuously differentiable function V that solves (19) with boundary condition (14), and $|V|, |V'|$ are bounded. Then V is the value function of entropy-regularized exploratory dividend problem (12) under exponential discounting.*

Theorem 1 shows that solution to the exploratory HJB equation (19) could be the value function of exploratory dividend problem (12). On the other hand, a similar argument could show that the value function shall also satisfy (19), while the optimal exploratory dividend strategy is given by (17). To establish a rigorous statement, we need the following result. The next proposition shows that the value function $V(x)$ converges as x going to infinity.

Proposition 2. *Let V be the value function of (12) and suppose the optimal exploratory dividend strategy is (17). Then as x going to infinity, $V(x)$ converges to a constant, i.e.,*

$$\lim_{x\to\infty} V(x) = \frac{\lambda \ln \lambda + \lambda \ln(e^{M/\lambda} - 1)}{\rho}. \tag{20}$$

3.3. Solution to Exploratory HJB

Compared with the differential equation (6) which solves the classical value function, the exploratory HJB equation (19) has a nonlinear term $\ln Z_M((1 - V'(x))/\lambda)$, which makes it difficult to be solved explicitly. The theorem below guarantees the existence and uniqueness of solution $V(x)$.

Theorem 2. *There exists a unique twice-continuously differentiable function $V(x)$ that solves (19) with boundary condition (14) and (20). Moreover, $\lim_{\lambda\to 0} |V(x) - V_{cl}(x)| = 0$ for all $x \in [0, \infty)$, where $V_{cl}(x)$ is the value function of classical dividend problem.*

Theorem 2 follows from the results in Tang *et al.* (2022, Theorem 3.9, 3.10) and in Strulovici and Szydlowski (2015, Proposition 1). It is straightforward to check that the conditions to guarantee the existence and uniqueness of the solution to (19) and its twice-continuous differentiability are satisfied.

Theorem 2 also states that when λ becomes smaller, the exploratory value function converges to the classical value function. Indeed, a stronger convergence is established by Tang *et al.* (2022) that V converges to V_{cl} locally uniformly as λ going to 0. Note that the parameter λ is the weight to be put on the exploration in contrast to the exploitation. If it is more close to 0, the entropy term has smaller effect on the total objective value and the optimal exploratory distribution $\pi^*(a;x)$ in (15) are more concentrated and close to the Dirac distribution – the optimal solution to the classical dividend

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

optimization problem. Then not surprisingly, the exploratory value function $V(x)$ also converges to the classical value function $V_{cl}(x)$ as λ going to 0.

Now, thanks to Theorem 2, we have $V(x)$ that solves the exploratory HJB equation (19). On the other hand, it is straightforward to show that according to (20), if $M < \lambda \ln(1/\lambda + 1)$, the limit of $V(x)$ is negative; if $M > \lambda \ln(1/\lambda + 1)$, the limit of $V(x)$ is positive; if $M = \lambda \ln(1/\lambda + 1)$, the limit of $V(x)$ is zero. The next theorem shows that indeed, we classify $V(x)$ into three cases based on its monotonicity.

Theorem 3. *Let $V(x)$ be the solution to (19) with boundary condition (14) and (20). Then $V(x)$ is monotone. To be more specific,*

(i) if $M < \lambda \ln(1/\lambda + 1)$, $V(x)$ is non-increasing.
(ii) if $M > \lambda \ln(1/\lambda + 1)$, $V(x)$ is non-decreasing.
(iii) if $M = \lambda \ln(1/\lambda + 1)$, $V(x) \equiv 0$.

The following corollary is a direct result from above theorem.

Corollary 1. *Let $V(x)$ be the solution to (19) with boundary condition (14) and (20). Then $|V(x)|$ and $|V'(x)|$ are bounded.*

Note that in Theorem 1 we need $|V|$ and $|V'|$ to be bounded so that V – solution to (19) – is indeed the value function of exploratory optimal dividend problem. Corollary 1 verifies the boundedness conditions are satisfied. In other words, the solution to the exploratory HJB equation (19) is the value function of the exploratory dividend problem.

4. Discussion

In view of Theorem 3, value functions can be classified into three cases according to the monotonicity: (1) $M < \lambda \ln(1/\lambda + 1)$; (2) $M > \lambda \ln(1/\lambda + 1)$; (3) $M = \lambda \ln(1/\lambda + 1)$. The following proposition will be useful in analyzing the properties of value functions.

Proposition 3. *(a) Define*

$$d_1(\lambda) := \lambda \ln(1/\lambda + 1)\, \mathbf{1}_{\lambda > 0} + 0 \cdot \mathbf{1}_{\lambda = 0}, \quad \lambda \in [0, \infty).$$

Then $d_1(\lambda)$ is increasing. Therefore, $\lim_{\lambda \to 0} d_1(\lambda) = d_1(0) = 0$ and $\lim_{\lambda \to \infty} d_1(\lambda) = 1$;
(b) Define

$$d_2(\lambda) := \lambda \ln(\lambda / (\lambda - 1))\, \mathbf{1}_{\lambda > 1} + \infty \cdot \mathbf{1}_{\lambda \in [0,1]}, \quad \lambda \in [0, \infty).$$

Then $d_2(\lambda) > d_1(\lambda)$, and $d_2(\lambda)$ is decreasing on $\lambda > 1$. Therefore, $\lim_{\lambda \to 1} d_2(\lambda) = \infty$ and $\lim_{\lambda \to \infty} d_2(\lambda) = 1$.

Case 1: $M < d_1(\lambda)$.

The value function in this case is non-increasing and thus non-positive, as a sharp contrast to the results of classical dividend problem. To see the reason, on one hand, note that for $\lambda > 0$,

$$\ln M < \ln d_1(\lambda) = \ln\left(\lambda \ln\left(\frac{1}{\lambda} + 1\right)\right) \leq \ln\left(\lambda \cdot \frac{1}{\lambda}\right) = 0.$$

Then due to Proposition 1, the entropy term is negative, that is, $\mathcal{H}(\pi) \leq \ln M < 0$. On the other hand, when $d_1(\lambda)$ is large, it implies that the exploration parameter λ is relatively large compared with the maximal dividend paying rate M. Then the negative entropy has a large weight in the total objective value, dominating the total expected dividends and leading to a negative value function.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

Case 2: $M > d_1(\lambda)$.

When $M > d_1(\lambda)$, the value function is non-decreasing, which is closer to the increasing value function in classical dividend optimization problem than it does in Case 1. This is because a relatively small λ compared with M decreases the weight of entropy term in the total objective value. Note that in classical dividend optimization the limit of value function is M/ρ, while in the current exploratory dividend optimization the limit of value function is given in (20). Therefore, if (i) $d_1(\lambda) < M \leq d_2(\lambda)$, the limit of $V(x)$ is no larger than that of $V_{cl}(x)$; if (ii) $d_2(\lambda) < M$, then $V(x)$ asymptotically achieves a higher value than that of the classical dividend optimization. Then if $\lambda > 1$ and $M > \lambda \ln \lambda - \lambda \ln(\lambda - 1) = d_2(\lambda)$, the limit of $V(x)$ is larger than that of $V_{cl}(x)$.

As shown in Proposition 3, for any $\lambda \geq 0$, $d_1(\lambda) < \lim_{\lambda\to\infty} d_1(\lambda) = 1$. Therefore, when $M \geq 1$, it always belongs to Case 2 for any $\lambda \geq 0$. On the other hand, for any $\lambda \geq 0$, $d_2(\lambda) > \lim_{\lambda\to\infty} d_2(\lambda) = 1$. Therefore, when $M \leq 1$ or $\lambda \leq 1$, since $d_2(\lambda) > M$, it cannot be Case 2 (ii); when $\lambda \leq 1$ and $M \geq 1$, it is always Case 2 (i). Note that $\lambda = 0$ corresponds to the classical dividend optimization and $d_1(0) = 0$, $d_2(0) = \infty$ by definition. Since $d_1(0) < M < d_2(0)$ for any positive constant M, classical dividend optimization can be viewed as a special Case 2 (i). It implies that exploratory dividend optimization is a generalization of the classical dividend optimization.

Case 3: $M = d_1(\lambda)$.

As shown in Theorem 3, the value function in this case should be constantly zero. This is because λ compared with M happens to strike a balance between exploitation and exploration such that the total expected dividends is offset by the entropy.

Figure 2 depicts the different cases of value functions given different combinations of M and λ. When $M < d_1(\lambda)$, the value function falls into Case 1 area. When $M > d_1(\lambda)$, the value function corresponds to Case 2, which can be further classified into two cases based on the comparison of M and $d_2(\lambda)$, i.e., whether the value function asymptotically achieves a higher value than that of the classical problem. When $M = d_1(\lambda)$, the value function should be Case 3 type.

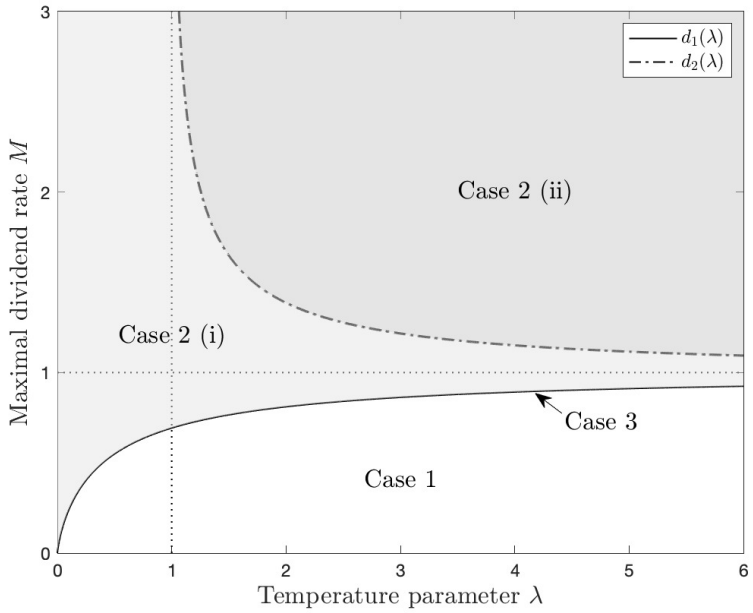

Figure 2. Cases of value functions given M and λ.

5. Numerical Examples

In this section, we present numerical examples of optimal exploratory policy and corresponding value function which solves exploratory HJB equation (19) based on the theoretical results obtained

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

in the previous sections.[2] To have a clear vision on the weight of cumulative dividends and that of entropy in the total objective value, we further decompose $V(x)$ into two parts: the expected total discounted dividends under the optimal exploratory dividend policy

$$Dv(x) := \mathbb{E}\left[\int_0^{\tau_x^{\pi^*}} e^{-\rho t}\left(\int_0^M a\pi^*(a; X_t^{\pi^*})da\right)dt\right];$$

and the expected total weighted discounted entropy under the optimal exploratory dividend policy

$$Entr(x) := \lambda\mathbb{E}\left[\int_0^{\tau_x^{\pi^*}} e^{-\rho t}\mathcal{H}(\pi_t^*)dt\right],$$

where the entropy of π^* is derived via substituting the optimal distribution (15) into the definition of entropy (10), i.e.,

$$\mathcal{H}(\pi^*) = \ln(Z_M((1-V'(x))/\lambda)) - \frac{Me^{M(1-V'(x))/\lambda}}{Z_M((1-V'(x))/\lambda)} + 1. \tag{21}$$

Hence, $V(x) = Dv(x) + Entr(x)$. We show examples of three cases, respectively, with commonly used parameters: $\mu = 1, \sigma = 1, \rho = 0.3$.

First, let $\lambda = 1.5$, $M = 0.6$. Then $M < d_1(\lambda)$ and it belongs to Case 1. Note that $V(x)$ in this case is decreasing and non-positive, as a sharp contrast to the results of classical dividend problem. The figure on the top row, left column of Figure 3 plots the corresponding value function and its two components $Dv(x)$ and $Entr(x)$.[3] The figure on the middle row, left column of Figure 3 plots the mean of the optimal distribution $\pi^*(\cdot; x)$, which is decreasing on x. The figure on the bottom row, left column of Figure 3 shows the density function of the optimal distribution with respect to different surplus level x. Because $V'(x) \leq 0$, the optimal distribution is a truncated exponential distribution with rate $-\lambda/(1-V'(x)) < 0$ for any $x \geq 0$. Therefore, it is more likely to pay high dividend rate. Furthermore, as surplus x increases, the density function becomes more flat because $V'(x)$ is increasing to 0 and the rate $-\lambda/(1-V'(x))$ is decreasing on x.

Second, let $\lambda = 1.5$, $M = 1.2$. Then $d_1(\lambda) < M < d_2(\lambda)$ and it belongs to Case 2 (i). The figure on the top row, middle column of Figure 3 shows the corresponding value function, $Dv(x)$ and $Entr(x)$. In contrast to Case 1, $Entr(x)$ in this case becomes positive since M is sufficiently large, making the value function $V(x)$ positive. The figure on the middle row, middle column of Figure 3 plots the mean of the optimal distribution $\pi^*(\cdot; x)$, which is increasing on x. The figure on the bottom row, middle column of Figure 3 shows the density function of optimal distribution with respect to different surplus level x. When x is small, it is more likely to choose a low dividend paying rate, because paying too high dividend rate would probably cause the insurance company to go bankruptcy and harms the shareholder's benefit in the long run. When x becomes larger, it is more likely to pay high dividend rate.

Third, let $\lambda = 1.5$, $M = 1.8$. Then $M > d_2(\lambda)$ and it belongs to Case 2 (ii). The figure on the top row, right column of Figure 3 shows the corresponding value function, $Dv(x)$ and $Entr(x)$. In this case, the limit of $V(x)$ is higher than that of the classical value function $V_{cl}(x)$, which is $M/\rho = 6$. Note that the expected total discounted dividends under exploratory policy $Dv(x)$ does not exceed that of the classical policy $V_{cl}(x)$, because the classical optimal dividend policy fully exploits the known environment. For sufficiently large M and λ, $Entr(x)$ is large enough to make $V(x)$ larger than $V_{cl}(x)$.

---

[2] We apply the shooting method which adjusts the initial value of first-order derivative such that the boundary conditions (14) and (20) are satisfied and use "ode45" function in Matlab to find the numerical solution to (19).

[3] For each initial surplus x, we discretize the continuous time into small pieces ($\Delta t = 0.0005$) and sample 2000 independent surplus processes $X_t^{\pi^*}$ to simulate $Dv(x)$ and $Entr(x)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

The figure on the middle row, right column of Figure 3 plots the mean of the optimal distribution $\pi^*(\cdot;x)$ and the figure on the bottom row, right column of Figure 3 plots the density function of the optimal distribution, which are similar to that of Case 2 (i).

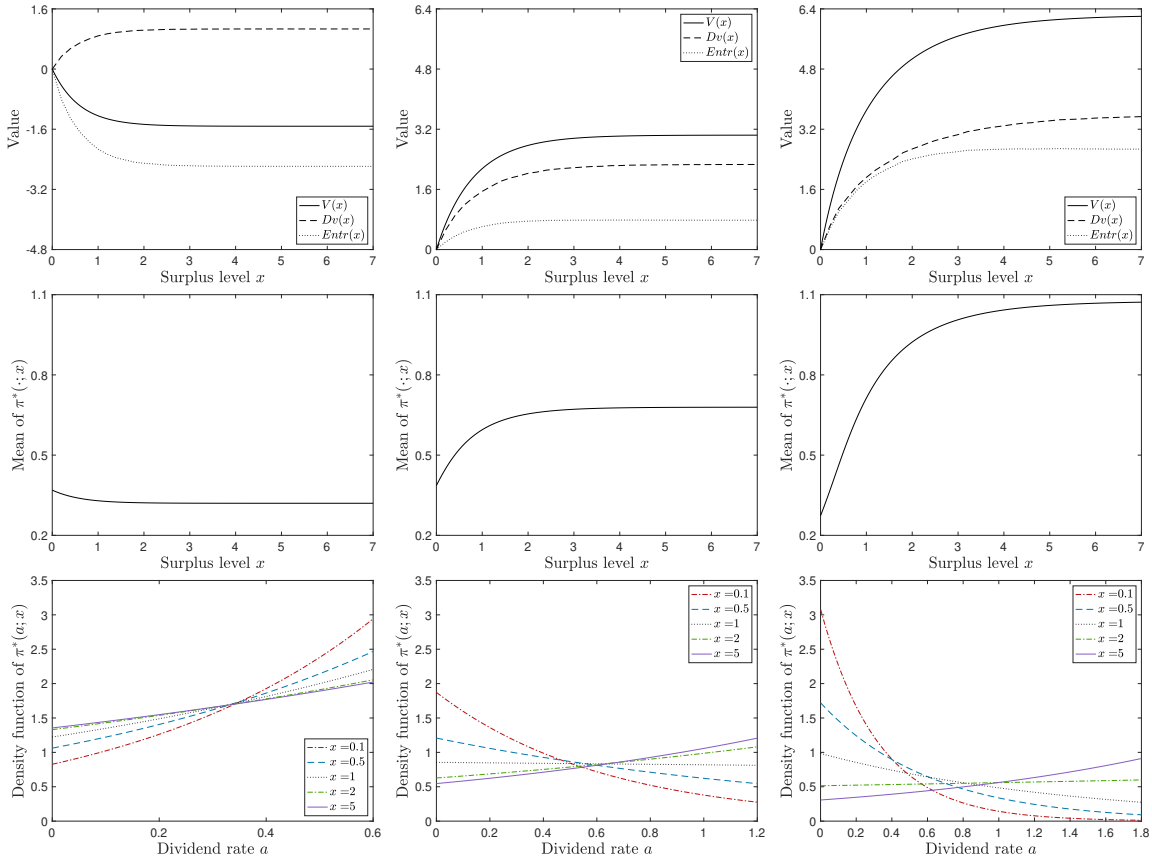

Figure 3. Let $\mu = 1$, $\sigma = 1$, $\rho = 0.3$, $\lambda = 1.5$. Let $M = 0.6$ (left column); $M = 1.2$ (middle column); $M = 1.8$ (right column), respectively. The figures on the top row show the value function $V(x)$, the expected total discounted dividends $Dv(x)$ and the expected total weighted discounted entropy $Entr(x)$. The figures on the middle row show the mean of the optimal distribution $\pi^*(\cdot;x)$. The figures on the bottom row show the density function of the optimal distribution with respect to different surplus level x.

When $\lambda = 1.5$, $M = 0.7662$, it belongs to Case 3 and the value function in this case should be constantly zero.

We also vary the value of λ while keeping the other parameter values unchanged. Figure 4 shows the value function under different values of λ with $M = 0.6$ and $M = 1.2$ respectively. Note that when $\lambda = 0$, $V(x)$ degenerates to the classical value function $V_{cl}(x)$. For $M = 0.6$, it is Case 2 (ii) when λ is small and then becomes Case 3 and Case 1 as λ getting larger. As aforementioned, it cannot be Case 2 (ii) since $M < 1$. Indeed, the left panel of Figure 4 shows the value function could not exceed the classical one as λ getting smaller. On the other hand, for $M = 1.2$, it can only be Case 2 and even Case 2 (ii) if λ is large enough. The right panel of Figure 4 shows the value function is always increasing on x for different values of λ and it can exceed the classical value function for a sufficiently large λ.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

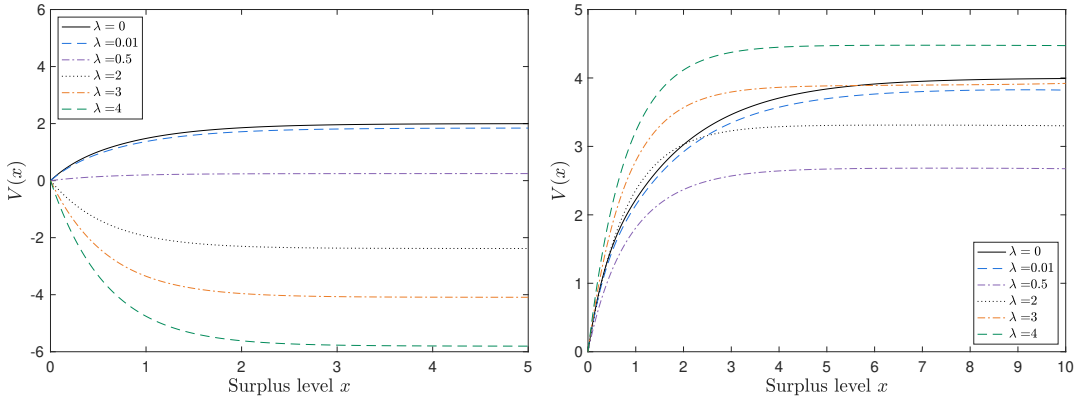

Figure 4. The value function $V(x)$ given different values of λ with $M = 0.6$ (left) and $M = 1.2$ (right)

6. Conclusion

This paper studies the dividend optimization problem in the entropy regularization framework. In an unknown environment, the entropy is incoprated into the objective function to encourage the exploration and an exploratory dividend policy is introduced. We establish the exploratory HJB equation, we find that the optimal distributional control is a truncated exponential distribution. Compared to the classical value function, the value function in the exploratory dividend problem is classified into three cases. The monotonicity of the value function is determined by the maximal dividend paying rate and the temperature parameter which controls the weight of exploration.

One future research direction is to consider the exploratory dividend policy under the non-exponential discounting which makes the problem time-inconsistent. Furthermore, reinsurance could also be considered as part of the insurance company's strategy in addition to the dividend policy, which is more technically challenging under the entropy regularization framework. Finally, one could take other definitions of the entropy, instead of the Shannon's differential, as a measure of the level of exploration in reinforcement learning.

Appendix A. Proof

Proof of Proposition 1 By definition (10),

$$\begin{aligned}\mathcal{H}(\pi) &= -\int_0^M \pi(a)\ln\pi(a)da = \int_0^M \pi(a)\ln\left(\frac{1}{\pi(a)}\right)da \\ &\leq \ln\left(\int_0^M \pi(a)\frac{1}{\pi(a)}da\right) = \ln M,\end{aligned}$$

where the inequality is due to Jensen's inequality. Q.E.D.

Proof of Theorem 1 Let $\tilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}$ be an exploratory dividend policy. Because V solves (13), for any $x \in [0,\infty)$,

$$\begin{aligned}0 &= \sup_{\pi\in\Pi_{[0,M]}}\left\{\int_0^M \left(a - \lambda\ln\pi(a;x) - aV'(x)\right)\pi(a;x)da\right\} + \mu V'(x) + \frac{1}{2}\sigma^2 V''(x) - \rho V(x) \\ &\geq \int_0^M \left(a - \lambda\ln\tilde{\pi}(a;x) - aV'(x)\right)\tilde{\pi}(a;x)da + \mu V'(x) + \frac{1}{2}\sigma^2 V''(x) - \rho V(x).\end{aligned}$$

which shows that

$$-\rho V(x) + \left(\mu - \int_0^M a\tilde{\pi}(a;x)da\right)V'(x) + \frac{1}{2}\sigma^2 V''(x) \leq -\int_0^M \left(a - \lambda\ln\tilde{\pi}(a;x)\right)\tilde{\pi}(a;x)da. \quad \text{(A1)}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

Applying Itô's Lemma on $e^{-\rho t}V(X_t^{\tilde{\pi}})$,

$$\begin{aligned}
V(x) &= e^{-\rho(T\wedge\tau_x^{\tilde{\pi}})}V(X^{\tilde{\pi}}_{T\wedge\tau_x^{\tilde{\pi}}}) \\
&\quad - \int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\left(-\rho V(X_t^{\tilde{\pi}}) + \left(\mu - \int_0^M a\tilde{\pi}(a;X_t^{\tilde{\pi}})da\right)V'(X_t^{\tilde{\pi}}) + \frac{1}{2}\sigma^2 V''(X_t^{\tilde{\pi}})\right)dt \\
&\quad - \int_0^{T\wedge\tau_x^{\tilde{\pi}}} \sigma e^{-\rho t}V'(X_t^{\tilde{\pi}})dW_t \\
&\geq e^{-\rho(T\wedge\tau_x^{\tilde{\pi}})}V(X^{\tilde{\pi}}_{T\wedge\tau_x^{\tilde{\pi}}}) \\
&\quad + \int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M \left(a - \lambda\ln\tilde{\pi}(a;X_t^{\tilde{\pi}})\right)\tilde{\pi}(a;X_t^{\tilde{\pi}})dadt \\
&\quad - \int_0^{T\wedge\tau_x^{\tilde{\pi}}} \sigma e^{-\rho t}V'(X_t^{\tilde{\pi}})dW_t,
\end{aligned}$$

where the inequality is due to (A1). Then taking expectation on both sides,

$$\begin{aligned}
V(x) \geq\ & \mathbb{E}\left[e^{-\rho(T\wedge\tau_x^{\tilde{\pi}})}V(X^{\tilde{\pi}}_{T\wedge\tau_x^{\tilde{\pi}}})\right] + \mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M \left(a - \lambda\ln\tilde{\pi}(a;X_t^{\tilde{\pi}})\right)\tilde{\pi}(a;X_t^{\tilde{\pi}})dadt\right] \\
& - \mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} \sigma e^{-\rho t}V'(X_t^{\tilde{\pi}})dW_t\right]. \qquad \text{(A2)}
\end{aligned}$$

For the first term on the right hand side of (A2), noting that $|V|$ is bounded, then by bounded convergence theorem,

$$\lim_{T\to\infty}\mathbb{E}\left[e^{-\rho(T\wedge\tau_x^{\tilde{\pi}})}V(X^{\tilde{\pi}}_{T\wedge\tau_x^{\tilde{\pi}}})\right] = \mathbb{E}\left[e^{-\rho(\tau_x^{\tilde{\pi}})}V(X^{\tilde{\pi}}_{\tau_x^{\tilde{\pi}}})\right] = 0.$$

For the second term on the right hand side of (A2), since $\tilde{\pi}$ is admissible and satisfies Definition 2 *(iii)*,

$$\begin{aligned}
&\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M \left(a - \lambda\ln\tilde{\pi}(a;X_t^{\tilde{\pi}})\right)\tilde{\pi}(a;X_t^{\tilde{\pi}})dadt\right] \\
=&\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M a\tilde{\pi}(a;X_t^{\tilde{\pi}})dadt\right] - \lambda\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M \left(\ln\tilde{\pi}(a;X_t^{\tilde{\pi}})\tilde{\pi}(a;X_t^{\tilde{\pi}}) + 1\right)dadt\right] \\
&+ \lambda\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}Mdt\right].
\end{aligned}$$

Because $\int_0^M a\tilde{\pi}(a;X_t^{\tilde{\pi}})da$ is non-negative, by monotone convergence theorem,

$$\lim_{T\to\infty}\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M a\tilde{\pi}(a;X_t^{\tilde{\pi}})dadt\right] = \mathbb{E}\left[\int_0^{\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M a\tilde{\pi}(a;X_t^{\tilde{\pi}})dadt\right].$$

Noting that $y\ln y + 1 \geq y > 0$ for any $y \in (0,\infty)$, by monotone convergence theorem,

$$\begin{aligned}
&\lim_{T\to\infty}\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M \left(\ln\tilde{\pi}(a;X_t^{\tilde{\pi}})\tilde{\pi}(a;X_t^{\tilde{\pi}}) + 1\right)dadt\right] \\
&\qquad = \mathbb{E}\left[\int_0^{\tau_x^{\tilde{\pi}}} e^{-\rho t}\int_0^M \left(\ln\tilde{\pi}(a;X_t^{\tilde{\pi}})\tilde{\pi}(a;X_t^{\tilde{\pi}}) + 1\right)dadt\right].
\end{aligned}$$

and $\lim_{T\to\infty}\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} e^{-\rho t}Mdt\right] = \mathbb{E}\left[\int_0^{\tau_x^{\tilde{\pi}}} e^{-\rho t}Mdt\right]$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

For the third term on the right hand side of (A2), noting that $|V'|$ is bounded, the stochastic integral $\left\{\int_0^s \sigma e^{-\rho t} V'(X_t^{\tilde{\pi}}) dW_t\right\}_{s\geq 0}$ is a martingale, then by optional sampling theorem,

$$\mathbb{E}\left[\int_0^{T\wedge\tau_x^{\tilde{\pi}}} \sigma e^{-\rho t} V'(X_t^{\tilde{\pi}}) dW_t\right] = 0.$$

Thus, letting $T \to \infty$ on both sides of (A2),

$$V(x) \geq \mathbb{E}\left[\int_0^{\tau_x^{\tilde{\pi}}} e^{-\rho t} \int_0^M \left(a - \lambda \ln \tilde{\pi}(a; X_t^{\tilde{\pi}})\right) \tilde{\pi}(a; X_t^{\tilde{\pi}}) da dt\right] = J(x, \tilde{\pi}).$$

Since $\tilde{\pi}$ is arbitrarily chosen, $V(x)$ becomes an upper bound of the optimal value of $J(x;\cdot)$.

On the other hand, the above inequality becomes an equality if the supremum in (13) is achieved, that is, $\tilde{\pi} = \pi^*$, where π^* is given by (15). Thus, $V(x)$ is the value function. Q.E.D.

Define function $G_{\lambda,M}$ to be

$$G_{\lambda,M}(y) := \left[\left(M - \frac{1}{y} + \frac{M}{e^{My}-1}\right)\mathbf{1}_{y\neq 0} + \left(\frac{M}{2}\right)\mathbf{1}_{y=0}\right](1-\lambda y) + \lambda \ln Z_M(y). \tag{A3}$$

where function Z_M is given in (16).

Lemma A1. *The function $G_{\lambda,M}(y)$ defined in (A3) is maximized when $y = 1/\lambda$, and*

$$G_{\lambda,M}(1/\lambda) = \lambda \ln \lambda + \lambda \ln(e^{M/\lambda} - 1).$$

Moreover, $G_{\lambda,M}(1/\lambda) < 0$ when $M < \lambda \ln(1/\lambda + 1)$, $G_{\lambda,M}(1/\lambda) > 0$ when $M > \lambda \ln(1/\lambda + 1)$, and $G_{\lambda,M}(1/\lambda) = 0$ when $M = \lambda \ln(1/\lambda + 1)$.

Proof. Take the first-order derivative of function $G_{\lambda,M}$:

$$\begin{aligned} G'_{\lambda,M}(y) &= \frac{1}{y^2} - \frac{M^2 e^{My}}{(e^{My}-1)^2} - \lambda M - \lambda \frac{M(e^{My}-1) - M^2 y e^{My}}{(e^{My}-1)^2} - \lambda \frac{(e^{My}-1) - My e^{My}}{y(e^{My}-1)} \\ &= \frac{(1-\lambda y)\left(e^{2My} - (2+M^2y^2)e^{My} + 1\right)}{y^2(e^{My}-1)^2} = \frac{(1-\lambda y) f_1(y)}{y^2(e^{My}-1)^2},\ y \neq 0, \end{aligned}$$

where $f_1(y) := e^{2My} - (2+M^2y^2)e^{My} + 1$, $y \neq 0$. Take the first-order derivative of f_1:

$$f_1'(y) = 2Me^{2My} - 2Me^{My} - M^3y^2e^{My} - 2M^2ye^{My} = Me^{My} f_2(y),\ y \neq 0,$$

where $f_2(y) := 2e^{My} - M^2y^2 - 2My - 2$, $y \neq 0$. Take the first-order derivative of f_2:

$$f_2'(y) = 2Me^{My} - 2M^2y - 2M = 2M f_3(y),\ y \neq 0,$$

where $f_3(y) := e^{My} - My - 1$, $y \neq 0$. Take the first-order derivative of f_3:

$$f_3'(y) = Me^{My} - M,\ y \neq 0.$$

Note that $f_3'(y) > 0$ for $y > 0$ and $f_3'(y) < 0$ for $y < 0$. Hence, $f_3(y)$ is increasing on $y > 0$ and decreasing on $y < 0$, and $f_3(y) > 0$. Then $f_2'(y) > 0$, which means that $f_2(y)$ is increasing. As a result, $f_2(y) > 0$ for $y > 0$ and $f_2(y) < 0$ for $y < 0$. Hence, $f_1'(y) > 0$ for $y > 0$ and $f_1'(y) < 0$ for $y < 0$, which means that $f_1(y)$ is increasing on $y > 0$ and decreasing on $y < 0$. As a result, $f_1(y) > 0$ for $y \neq 0$.

 doi:10.20944/preprints202311.1796.v1

The above analysis shows that $G'_{\lambda,M}(y)$ is positive when $1-\lambda y>0$, i.e., $y<1/\lambda$, and negative when $1-\lambda y<0$, i.e., $y>1/\lambda$. Thus the maximum is obtained at $y=1/\lambda$:

$$\max G_{\lambda,M}(y)=G_{\lambda,M}(1/\lambda)=\lambda\ln Z_M(1/\lambda)=\lambda\ln\lambda+\lambda\ln(e^{M/\lambda}-1).$$

Moreover, when $M<\lambda\ln(1/\lambda+1)$, $G_{\lambda,M}(1/\lambda)<\lambda\ln\lambda+\lambda\ln(e^{\ln(1/\lambda+1)}-1)=0$; when $M>\lambda\ln(1\lambda+1)$, $G_{\lambda,M}(1/\lambda)>0$; when $M=\lambda\ln\left(\frac{1}{\lambda}+1\right)$, $G_{\lambda,M}(1/\lambda)=0$. Q.E.D.

Proof of Proposition 2 With the optimal distributional policy given in (17), substituting (17) into the objective (11) leads to

$$\begin{aligned}
V(x)=J(x,\boldsymbol{\pi}^*)&=\mathbb{E}\left[\int_0^{\tau_x^{\pi^*}}e^{-\rho t}\int_0^M\left(a-\lambda\ln\pi^*(a;X_t^{\pi^*})\right)\pi^*(a;X_t^{\pi^*})dadt\right]\\
&=\mathbb{E}\left[\int_0^{\tau_x^{\pi^*}}e^{-\rho t}\int_0^M\left(a-\lambda a\left(\frac{1-V'(X_t^{\pi^*})}{\lambda}\right)+\lambda\ln Z_M\left(\frac{1-V'(X_t^{\pi^*})}{\lambda}\right)\right)\pi^*(a;X_t^{\pi^*})dadt\right]\\
&=\mathbb{E}\Bigg[\int_0^{\tau_x^{\pi^*}}e^{-\rho t}\Bigg\{\left[\left(M-\frac{\lambda}{1-V'(X_t^{\pi^*})}+\frac{M}{e^{M(1-V'(X_t^{\pi^*}))/\lambda}-1}\right)\mathbf{1}_{V'(X_t^{\pi^*})\neq1}+\left(\frac{M}{2}\right)\mathbf{1}_{V'(X_t^{\pi^*})=1}\right]\\
&\qquad\left(1-\lambda\left(\frac{1-V'(X_t^{\pi^*})}{\lambda}\right)\right)+\lambda\ln Z_M\left(\frac{1-V'(X_t^{\pi^*})}{\lambda}\right)\Bigg\}dt\Bigg]\\
&=\mathbb{E}\left[\int_0^{\tau_x^{\pi^*}}e^{-\rho t}G_{\lambda,M}\left(\frac{1-V'(X_t^{\pi^*})}{\lambda}\right)dt\right],
\end{aligned}$$

where $G_{\lambda,M}$ is defined in (A3).

On one hand,

$$V(x)=\mathbb{E}\left[\int_0^{\tau_x^{\pi^*}}e^{-\rho t}G_{\lambda,M}\left(\frac{1-V'(X_t^{\pi^*})}{\lambda}\right)dt\right]\le\mathbb{E}\left[\int_0^{\tau_x^{\pi^*}}e^{-\rho t}\left(\lambda\ln\lambda+\lambda\ln(e^{M/\lambda}-1)\right)dt\right],$$

where the inequality follows from Lemma A1. Letting $x\to\infty$ and by dominated convergence theorem,

$$\begin{aligned}
\lim_{x\to\infty}V(x)&\le\lim_{x\to\infty}\mathbb{E}\left[\int_0^{\tau_x^{\pi^*}}e^{-\rho t}\left(\lambda\ln\lambda+\lambda\ln(e^{M/\lambda}-1)\right)dt\right]\\
&=\mathbb{E}\left[\int_0^{\infty}e^{-\rho t}\left(\lambda\ln\lambda+\lambda\ln(e^{M/\lambda}-1)\right)dt\right]=\frac{\lambda\ln\lambda+\lambda\ln(e^{M/\lambda}-1)}{\rho}.
\end{aligned}$$

On the other hand, consider an exploratory policy $\hat{\boldsymbol{\pi}}=\{\hat{\pi}_t\}_{t\ge0}$, where

$$\hat{\pi}_t=\hat{\pi}(a;X_t^{\hat{\boldsymbol{\pi}}})=\frac{e^{a/\lambda}}{\lambda(e^{M/\lambda}-1)},\ a\in[0,M].$$

Then

$$\begin{aligned}
V(x)\ge J(x,\hat{\boldsymbol{\pi}})&=\mathbb{E}\left[\int_0^{\tau_x^{\hat{\pi}}}e^{-\rho t}\left(\int_0^M\left(a-\lambda\ln\hat{\pi}(a;X_t^{\hat{\boldsymbol{\pi}}})\right)\hat{\pi}(a;X_t^{\hat{\boldsymbol{\pi}}})da\right)dt\right]\\
&=\mathbb{E}\left[\int_0^{\tau_x^{\hat{\pi}}}e^{-\rho t}\left(\lambda\ln\lambda+\lambda\ln(e^{M/\lambda}-1)\right)dt\right].
\end{aligned}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

Letting $x \to \infty$ and by dominated convergence theorem,

$$\lim_{x\to\infty} V(x) \geq \frac{\lambda \ln \lambda + \lambda \ln(e^{M/\lambda} - 1)}{\rho},$$

which then together with the previous inequality leads to (20). Q.E.D.

Define a function h as

$$h(x) = \begin{cases} \ln \frac{e^{k(1-x)}-1}{1-x}, & x \neq 1, \\ \ln k, & x = 1, \end{cases} \tag{A4}$$

where $k > 0$ is given.

Lemma A2. *The function h defined in (A4) satisfies following properties:*

(i) $h(x)$ is continuous and decreasing in x;
(ii) there exists a unique $x_0 \in \mathbb{R}$ such that $h(x_0) = 0$;
(iii) $|h(x)| < k|x| + c$, for some constant $c \in \mathbb{R}$ which depends on k only;
(iv) $|h(x_1) - h(x_2)| < k|x_1 - x_2|, \forall x_1, x_2 \in \mathbb{R}$.

Proof. We first show that function $h(x)$ is continuous at $x = 1$. By L'Hôpital rule, $\lim_{x\to 1} \frac{e^{k(1-x)}-1}{1-x} = k$. Hence, $\lim_{x\to 1} h(x) = \ln k = h(1)$.

Taking the first-order derivative of h, for $x \neq 1$,

$$h'(x) = \frac{1-x}{e^{k(1-x)}-1} \frac{-k(1-x)e^{k(1-x)} + e^{k(1-x)} - 1}{(1-x)^2} = \frac{h_1(x)}{(1-x)(e^{k(1-x)}-1)},$$

where $h_1(x) = e^{k(1-x)} - 1 - k(1-x)e^{k(1-x)}$. Then

$$h_1'(x) = -ke^{k(1-x)} + ke^{k(1-x)} + k^2(1-x)e^{k(1-x)} = k^2(1-x)e^{k(1-x)},$$

which is positive when $x < 1$ and negative when $x > 1$. Therefore, $h_1(x)$ is increasing on $x < 1$ then decreasing on $x > 1$ and $h_1(x) < \lim_{x\to 1} h_1(x) = 0$. Combining with the fact that $(1-x)(e^{k(1-x)} - 1) > 0$ for $x \neq 1$, we show that $h'(x) < 0$ for $x \neq 1$. It then completes the proof of (i) that $h(x)$ is decreasing in x.

To show (ii), note that $\lim_{x\to-\infty} h(x) > 0$ and $\lim_{x\to\infty} h(x) < 0$. By the continuity and monotonicity of $h(x)$, there must exist a unique $x_0 \in \mathbb{R}$ such that $h(x_0) = 0$. In particular, when $k = 1$, $x_0 = 1$.

Note that for $x \neq 1$, $e^{k(1-x)} - 1 > k(1-x)$. which implies

$$e^{k(1-x)} - 1 - k(1-x)e^{k(1-x)} > -k(1-x)(e^{k(1-x)} - 1),$$

Combining with the fact that $(1-x)(e^{k(1-x)} - 1) > 0$ for $x \neq 1$,

$$h'(x) = \frac{e^{k(1-x)} - 1 - k(1-x)e^{k(1-x)}}{(1-x)(e^{k(1-x)} - 1)} > -k.$$

Based on the previous results, for $x < x_0$,

$$|h(x)| = h(x) = h(x_0) - \int_x^{x_0} h'(y)\mathrm{d}y = -\int_x^{x_0} h'(y)\mathrm{d}y < \int_x^{x_0} k\mathrm{d}y = k(x_0 - x);$$

 doi:10.20944/preprints202311.1796.v1

similarly, for $x \geq x_0$,

$$|h(x)| = -h(x) = -h(x_0) - \int_{x_0}^{x} h'(y)\mathrm{d}y = -\int_{x_0}^{x} h'(y)\mathrm{d}y < \int_{x_0}^{x} k\mathrm{d}y = k(x - x_0).$$

To show (iii),

$$|h(x)| < k|x - x_0| \leq k|x| + k|x_0|, \qquad \forall x \in \mathbb{R}.$$

It remains to prove (iv). Without loss of generality, we assume $x_1 \geq x_2$. Then

$$|h(x_1) - h(x_2)| = h(x_2) - h(x_1) = -\int_{x_1}^{x_2} h'(y)\mathrm{d}y < \int_{x_2}^{x_1} k\mathrm{d}y = k|x_1 - x_2|.$$

Q.E.D.

Proof of Theorem 2 It is straightforward to show that Assumption 3.8 in Tang *et al.* (2022) hold for our exploratory dividend problem. The well-posedness of SDE (18) for the optimal exploratory surplus process is also established. Then, by applying the results of Tang *et al.* (2022, Theorem 3.9, 3.10), the existence and uniqueness of solution to (19) and convergence of V to V_{cl} are established.

To show the twice-continuously differentiability of $V(x)$, we apply the results in Strulovici and Szydlowski (2015, Proposition 1) (with the infinite domain). We rewrite the HJB equation (19) into the following form:

$$V''(x) + H(V(x), V'(x)) = 0,$$

where

$$\begin{aligned} H(p,q) :&= \frac{2}{\sigma^2}\left[-\rho p + \mu q + \lambda \ln\left(\frac{\lambda}{1-q}\left(e^{M(1-q)/\lambda} - 1\right)\mathbf{1}_{q\neq 1} + M\mathbf{1}_{q=1}\right)\right] \\ &= \frac{2}{\sigma^2}\left[-\rho p + \mu q + \lambda \ln\lambda + \lambda h(q)\right], \end{aligned}$$

and h is defined in (A4) with $k = M/\lambda$. According to Proposition 1 in Strulovici and Szydlowski (2015), if H satisfies Condition 1-3, then there exists a twice-continuously differentiable solution to the HJB equation.

To check Condition 1 in Strulovici and Szydlowski (2015, Proposition 1), note that for $p, q \in \mathbb{R}$,

$$\begin{aligned} |H(p,q)| &\leq \frac{2}{\sigma^2}\left[\rho|p| + \mu|q| + \lambda|\ln\lambda| + \lambda|h(q)|\right] \\ &< \frac{2}{\sigma^2}\left[\rho|p| + \mu|q| + \lambda|\ln\lambda| + M|q| + c\right], \end{aligned}$$

where the second inequality comes from Lemma A2 (iii), and $c \in \mathbb{R}$ is a constant. Taking $L_1 := \frac{2}{\sigma^2}\max(\lambda|\ln\lambda| + c, \rho, \mu + M)$, we have

$$|H(p,q)| \leq L_1(1 + |p| + |q|).$$

Secondly, for $p, \tilde{p}, q, \tilde{q} \in \mathbb{R}$,

$$\begin{aligned} |H(p,q) - H(\tilde{p},\tilde{q})| &\leq \frac{2}{\sigma^2}\left[\rho|p - \tilde{p}| + \mu|q - \tilde{q}| + \lambda|h(q) - h(\tilde{q})|\right] \\ &< \frac{2}{\sigma^2}\left[\rho|p - \tilde{p}| + \mu|q - \tilde{q}| + M|q - \tilde{q}|\right], \end{aligned}$$

 doi:10.20944/preprints202311.1796.v1

where the second inequality comes from Lemma A2 (iv). Taking $L_2 := \frac{2}{\sigma^2}\max(\rho, \mu + M)$, we have

$$|H(p,q) - H(\tilde{p},\tilde{q})| \leq L_2(|p-\tilde{p}| + |q-\tilde{q}|).$$

To check Condition 2, note that for all $q \in \mathbb{R}$, $H(\cdot, q)$ is nonincreasing in p.
It remains to check Condition 3. For each $\bar{K} > 0$, choose $K_1, K_2 > \bar{K}$ such that

$$K_1 \geq \max\left(\frac{(M+\mu)K_2 + \lambda\ln\lambda + \lambda c}{\rho}, \frac{(M+\mu)K_2 - \lambda\ln\lambda + \lambda c}{\rho}\right), \tag{A5}$$

where c is a constant satisfying Lemma A2 (iii). Then for all $p \in \mathbb{R}, \epsilon \in \{-1, 1\}$,

$$\begin{aligned} H(K_1 + K_2|p|, \epsilon K_2) &= \frac{2}{\sigma^2}\left[-\rho K_1 - \rho K_2|p| + \mu\epsilon K_2 + \lambda\ln\lambda + \lambda h(\epsilon K_2)\right] \\ &< \frac{2}{\sigma^2}\left[-\rho K_1 + \mu K_2 + \lambda\ln\lambda + \lambda h(\epsilon K_2)\right] \\ &< \frac{2}{\sigma^2}\left[-\rho K_1 + \mu K_2 + \lambda\ln\lambda + MK_2 + \lambda c\right] < 0, \end{aligned}$$

where the third inequality is due to Lemma A2 (iii) and last inequality due to (A5). Secondly,

$$\begin{aligned} H(-K_1 - K_2|p|, \epsilon K_2) &= \frac{2}{\sigma^2}\left[\rho K_1 + \rho K_2|p| + \mu\epsilon K_2 + \lambda\ln\lambda + \lambda h(\epsilon K_2)\right] \\ &> \frac{2}{\sigma^2}\left[\rho K_1 - \mu K_2 + \lambda\ln\lambda + \lambda h(\epsilon K_2)\right] \\ &> \frac{2}{\sigma^2}\left[\rho K_1 - \mu K_2 + \lambda\ln\lambda - MK_2 - \lambda c\right] > 0, \end{aligned}$$

where the third inequality is due to Lemma A2 *(iii)* and last inequality due to (A5). Q.E.D.

Proof of Theorem 3 Note that (19) can be rewritten as

$$\rho V(x) = \frac{\sigma^2}{2}V''(x) + \mu V'(x) + \lambda h(V'(x)) + \lambda\ln\lambda, \tag{A6}$$

where h is defined in (A4) with $k = M/\lambda$.

First, suppose $M < \lambda\ln(1/\lambda + 1)$. Then $\lambda\ln\lambda + \lambda\ln(e^{M/\lambda} - 1)) < 0$. According to (20), $\lim_{x\to\infty} V(x) < 0$. Define $x_0 := \inf\{x \geq 0 : V'(x+) \neq 0\}$. Note that $V(x)$ is not a constant in this case and hence, $V'(x)$ does not always equal to 0, which implies that $x_0 < \infty$.

Assume that $V'(x_0+) > 0$. Since $V(x_0) = V(0) = 0$, there must exist some interval such that $V(x)$ is decreasing in order to reach its negative limit, which means that there exists some point such that $V'(x)$ changes its sign from positive to negative. Define this point as

$$x_1 := \inf\{x > x_0 : V'(x) = 0, V'(x+) < 0\}.$$

Hence, $V''(x_1) \leq 0$. Then according to (A6),

$$\begin{aligned} \rho V(x_1) &= \frac{\sigma^2}{2}V''(x_1) + \mu V'(x_1) + \lambda h(V'(x_1)) + \lambda\ln\lambda \\ &= \frac{\sigma^2}{2}V''(x_1) + \lambda\ln\left(e^{\frac{M}{\lambda}} - 1\right) + \lambda\ln\lambda < 0, \end{aligned}$$

which implies that $V(x_1) < 0$. But a contradiction happens because $V'(x)$ is non-negative on $[0, x_1]$, which leads to $V(x_1) > 0$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1796.v1

Then, assume that $V'(x_0+) < 0$ and there exists some point such that $V'(x) > 0$. Define x_2 as

$$x_2 := \inf\{x > x_0 : V'(x) = 0, V'(x+) > 0\}.$$

Hence, $V''(x_2) \geq 0$. According to (A6),

$$\begin{aligned}\rho V(x_2) &= \frac{\sigma^2}{2}V''(x_2) + \mu V'(x_2) + \lambda h(V'(x_2)) + \lambda \ln \lambda \\ &= \frac{\sigma^2}{2}V''(x_2) + \lambda \ln\left(e^{\frac{M}{\lambda}} - 1\right) + \lambda \ln \lambda \\ &\geq \lambda \ln\left(e^{\frac{M}{\lambda}} - 1\right) + \lambda \ln \lambda.\end{aligned}$$

Therefore,

$$V(x_2) \geq \frac{\lambda \ln \lambda + \lambda \ln(e^{M/\lambda} - 1)}{\rho} = \lim_{x\to\infty} V(x).$$

Since $V'(x_2+) > 0$, $V(x)$ is strictly increasing in a local neighborhood after x_2. Then, after point x_2 there should exists some interval such that $V(x)$ is strictly decreasing in order to achieve the limit. Define x_3 as

$$x_3 := \inf\{x > x_2 : V'(x) = 0, V'(x+) < 0\}.$$

Hence, $V''(x_3) \leq 0$. Note that $V'(x)$ is strictly positive in a local neighborhood after x_2 and non-negative on $[x_2, x_3]$, thus $V(x_3) > V(x_2)$. Then according to (A6),

$$\begin{aligned}V''(x_2) &= \frac{2}{\sigma^2}\left(\rho V(x_2) - \mu V'(x_2) - \lambda h(V'(x_2)) - \lambda \ln \lambda\right) \\ < \frac{2}{\sigma^2}&\left(\rho V(x_3) - \mu V'(x_3) - \lambda h(V'(x_3)) - \lambda \ln \lambda\right) = V''(x_3),\end{aligned}$$

which is a contradiction. Therefore, $V'(x) \leq 0$ and $V(x)$ is decreasing.

For the other two cases, the proof is similar. Q.E.D.

Proof of Corollary 1 Because according to Theorem 3 $V(x)$ is monotone and its limit as shown in (20) is finite, it is straightforward that $|V(x)|$ and $|V'(x)|$ are bounded. Q.E.D.

Proof of Proposition 3 *(a)* Taking the first-order derivative of d_1, for $\lambda > 0$,

$$d_1'(\lambda) = \ln\left(\frac{1}{\lambda} + 1\right) + \lambda \cdot \frac{-\lambda^{-2}}{1/\lambda + 1} = -\ln\left(\frac{\lambda}{\lambda+1}\right) + \left(\frac{\lambda}{\lambda+1} - 1\right) = \omega\left(\frac{\lambda}{\lambda+1}\right),$$

where $\omega(x) := -\ln x + (x-1)$, $x > 0$. Since $\omega'(x) = -1/x + 1 < 0$ for $x \in (0,1)$, $\omega(x)$ is decreasing on $x \in (0,1)$. Therefore, $\omega(x) > \omega(1) = 0$ for $x \in (0,1)$, which shows that $d_1(\lambda)$ is increasing. By L'Hôpital rule,

$$\lim_{\lambda\to 0} d_1(\lambda) = \lim_{\lambda\to 0} \frac{-\lambda^{-2}/(1/\lambda+1)}{-\lambda^{-2}} = \lim_{\lambda\to 0} \frac{1}{1/\lambda+1} = 0, \quad \lim_{\lambda\to\infty} d_1(\lambda) = \lim_{\lambda\to\infty} \frac{1}{1/\lambda+1} = 1.$$

(b) Note that for $\lambda > 1$, $(\lambda+1)/\lambda < \lambda/(\lambda-1)$. Therefore, $d_1(\lambda) < d_2(\lambda)$.
Taking the first-order derivative of d_2, for $\lambda > 1$,

$$d_2'(\lambda) = \ln\left(\frac{\lambda}{\lambda-1}\right) + \lambda \cdot \frac{\lambda-1}{\lambda} \cdot \frac{\lambda-1-\lambda}{(\lambda-1)^2} = \ln\left(\frac{\lambda}{\lambda-1}\right) - \left(\frac{\lambda}{\lambda-1} - 1\right) = -\omega\left(\frac{\lambda}{\lambda-1}\right).$$

Since $\omega'(x) = -1/x + 1 > 0$ for $x > 1$, $\omega(x)$ is increasing on $x > 1$. Therefore, $\omega(x) > \omega(1) = 0$ for $x > 1$, which shows that $d_2(\lambda)$ is decreasing. By L'Hôpital rule,

$$\lim_{\lambda\to 1} d_2(\lambda) = \lim_{\lambda\to 1} \frac{\lambda-1}{\lambda} \cdot \frac{\lambda-1-\lambda}{(\lambda-1)^2} \cdot \frac{1}{-\lambda^{-2}} = \lim_{\lambda\to 1} \frac{\lambda}{\lambda-1} = \infty, \lim_{\lambda\to\infty} d_2(\lambda) = \lim_{\lambda\to\infty} \frac{\lambda}{\lambda-1} = 1.$$

Q.E.D.

References

Wang, H.; Zariphopoulou, T.; Zhou, X.Y. Reinforcement Learning in Continuous Time and Space: A Stochastic Control Approach. *Journal of Machine Learning Research* **2020**, *21*, 1–34.

Lundberg, F. *Approximerad framställning af sannolikhetsfunktionen. Återförsäkring af kollektivrisker. Akademisk afhandling*; Almqvist &Wiksells, 1903.

De Finetti, B. Su un'impostazione alternativa della teoria collettiva del rischio. Transactions of the XVth international congress of Actuaries. New York, 1957, Vol. 2, pp. 433–443.

Gerber, H.U. Entscheidungskriterien für den zusammengesetzten Poisson-Prozess. PhD thesis, ETH Zurich, 1969.

Schmidli, H. *Stochastic control in insurance*; Springer Science & Business Media, 2007.

Jeanblanc-Picqué, M.; Shiryaev, A.N. Optimization of the flow of dividends. *Uspekhi Matematicheskikh Nauk* **1995**, *50*, 25–46.

Asmussen, S.; Taksar, M. Controlled diffusion models for optimal dividend pay-out. *Insurance: Mathematics and Economics* **1997**, *20*, 1–15.

Jgaard, B.H.; Taksar, M. Controlling risk exposure and dividends payout schemes: insurance company example. *Mathematical Finance* **1999**, *9*, 153–182.

Asmussen, S.; Højgaard, B.; Taksar, M. Optimal risk control and dividend distribution policies. Example of excess-of loss reinsurance for an insurance corporation. *Finance and Stochastics* **2000**, *4*, 299–324.

Azcue, P.; Muler, N. Optimal reinsurance and dividend distribution policies in the Cramér-Lundberg model. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* **2005**, *15*, 261–308.

Azcue, P.; Muler, N. Optimal investment policy and dividend payment strategy in an insurance company. *The Annals of Applied Probability* **2010**, *20*, 1253–1302.

Gaier, J.; Grandits, P.; Schachermayer, W. Asymptotic ruin probabilities and optimal investment. *The Annals of Applied Probability* **2003**, *13*, 1054–1076.

Kulenko, N.; Schmidli, H. Optimal dividend strategies in a Cramér–Lundberg model with capital injections. *Insurance: Mathematics and Economics* **2008**, *43*, 270–278.

Yang, H.; Zhang, L. Optimal investment for insurer with jump-diffusion risk process. *Insurance: Mathematics and Economics* **2005**, *37*, 615–634.

Choulli, T.; Taksar, M.; Zhou, X.Y. A diffusion model for optimal dividend distribution for a company with constraints on risk control. *SIAM Journal on Control and Optimization* **2003**, *41*, 1946–1979.

Gerber, H.U.; Shiu, E.S. On optimal dividend strategies in the compound Poisson model. *North American Actuarial Journal* **2006**, *10*, 76–93.

Avram, F.; Palmowski, Z.; Pistorius, M.R. On the optimal dividend problem for a spectrally negative Lévy process. *The Annals of Applied Probability* **2007**, *17*, 156–180.

Yin, C.; Wen, Y. Optimal dividend problem with a terminal value for spectrally positive Levy processes. *Insurance: Mathematics and Economics* **2013**, *53*, 769–773.

Zhao, Y.; Kosorok, M.R.; Zeng, D. Reinforcement learning design for cancer clinical trials. *Statistics in medicine* **2009**, *28*, 3294–3315.

Komorowski, M.; Celi, L.A.; Badawi, O.; Gordon, A.C.; Faisal, A.A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine* **2018**, *24*, 1716–1720.

Mirowski, P.; Pascanu, R.; Viola, F.; Soyer, H.; Ballard, A.J.; Banino, A.; Denil, M.; Goroshin, R.; Sifre, L.; Kavukcuoglu, K.; others. Learning to navigate in complex environments. arXiv:1611.03673.

Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J.J.; Gupta, A.; Fei-Fei, L.; Farhadi, A. Target-driven visual navigation in indoor scenes using deep reinforcement learning. 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017, pp. 3357–3364.

Radford, A.; Jozefowicz, R.; Sutskever, I. Learning to generate reviews and discovering sentiment. arXiv:1704.01444.

Paulus, R.; Xiong, C.; Socher, R. A deep reinforced model for abstractive summarization. arXiv:1705.04304.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; others. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.

Jaderberg, M.; Czarnecki, W.M.; Dunning, I.; Marris, L.; Lever, G.; Castaneda, A.G.; Beattie, C.; Rabinowitz, N.C.; Morcos, A.S.; Ruderman, A.; others. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* **2019**, *364*, 859–865.

Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; others. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; others. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359.

Auer, P.; Cesa-Bianchi, N.; Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning* **2002**, *47*, 235–256.

Cesa-Bianchi, N.; Gentile, C.; Lugosi, G.; Neu, G. Boltzmann exploration done right. *Advances in neural information processing systems* **2017**, *30*.

Todorov, E. Linearly-solvable Markov decision problems. *Advances in neural information processing systems* **2006**, *19*.

Ziebart, B.D.; Maas, A.L.; Bagnell, J.A.; Dey, A.K.; others. Maximum entropy inverse reinforcement learning. Aaai. Chicago, IL, USA, 2008, Vol. 8, pp. 1433–1438.

Nachum, O.; Norouzi, M.; Xu, K.; Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems* **2017**, *30*.

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

Dai, M.; Dong, Y.; Jia, Y. Learning equilibrium mean-variance strategy. *Mathematical Finance* **2023**, *33*, 1166–1212.

Bai, L.; Gamage, T.; Ma, J.; Xie, P. Reinforcement Learning for optimal dividend problem under diffusion model. arXiv:math/2309.10242.

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

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

Strulovici, B.; Szydlowski, M. On the smoothness of value functions and the existence of optimal strategies in diffusion models. *Journal of Economic Theory* **2015**, *159*, 1016–1055.