EXPLORATORY OPTIMAL STOPPING: A SINGULAR CONTROL FORMULATION

JODI DIANETTI, GIORGIO FERRARI, AND RENYUAN XU

Abstract. This paper explores continuous-time and state-space optimal stopping problems from a reinforcement learning perspective. We begin by formulating the stopping problem using randomized stopping times, where the decision maker's control is represented by the probability of stopping within a given time—specifically, a bounded, non-decreasing, càdlàg control process. To encourage exploration and facilitate learning, we introduce a regularized version of the problem by penalizing it with the cumulative residual entropy of the randomized stopping time. The regularized problem takes the form of an $(n+1)$-dimensional degenerate singular stochastic control with finite-fuel. We address this through the dynamic programming principle, which enables us to identify the unique optimal exploratory strategy. For the specific case of a real option problem, we derive a semi-explicit solution to the regularized problem, allowing us to assess the impact of entropy regularization and analyze the vanishing entropy limit. Finally, we propose a reinforcement learning algorithm based on policy iteration. We show both policy improvement and policy convergence results for our proposed algorithm.

Keywords: Optimal stopping, singular stochastic control, free boundary problem, entropy regularization, reinforcement learning

AMS subject classification: 60G40, 93E20, 35R35, 68T01

Contents

1. Introduction 2
1.1. Our work and contributions 3
1.2. Related literature 5
1.3. Outline of the paper 6
1.4. General notation 6
2. Exploratory formulation and entropy regularization of OS problem 7
2.1. Exploratory formulation via singular controls 8
2.2. Entropy regularization 9
2.3. Vanishing entropy limit 10
3. Solving the entropy regularized OS problem via dynamic programming 12
3.1. Preliminary estimates 14
3.2. Proof of Theorem 3.2: solution of the HJB 15
3.3. Proof of Theorem 3.2: uniqueness for the HJB 23
3.4. Proof of Theorem 3.3 24
4. A real option problem: analytical solutions and reinforcement learning algorithms 25
4.1. Analytical solution 25
4.2. Reinforcement learning algorithm 30
Appendix A. Proof of Lemma 2.3 40
Appendix B. Proof of the auxiliary estimates 41
References 46

Date: August 20, 2024.

1. Introduction

In optimal stopping (OS) problems, a decision maker chooses a time to take a given action based on an adaptively observed stochastic process in order to optimize an expected performance criterion. Problems of this type are found in the area of Statistics, where the action taken may be to test a hypothesis or to estimate a parameter; in the area of Operations Research, where the action may be to replace a machine, hire a secretary, or reorder stock; and in Mathematical Finance, in the context of pricing American contingent claims or in irreversible investment problems (we refer to the textbook [46] and to the references therein).

Despite the wide-ranging applications of optimal stopping across various fields, most of the existing literature relies on the *full knowledge* of the system, including both the underlying process and the reward function. Little is known about addressing these problems in a model-free context, where decision-makers interact with an unknown system and learn to improve their decisions over time. In this regard, machine learning techniques, particularly reinforcement learning (RL), offer a promising framework. However, the existing literature has largely focused on “frequently controlling the system”, leaving the RL framework for optimal stopping largely unexplored.

Generally speaking, RL aims to learn optimal strategies to control an unknown dynamical system (or random environment) by interacting with the system through exploration and exploitation [53]. In recent years, the availability of vast amounts of data and advances in computational resources have spurred extensive research in both theoretical developments and practical applications of RL. While much of the existing literature on RL has focused on discrete-time problems, continuous-time RL is rapidly growing and presents unique challenges and technical difficulties in algorithmic design, stability analysis, and efficiency guarantees. A key question is how to balance the trade-off between exploring unknown information in the continuous-time system and exploiting current information to take near-optimal actions. In their seminal work, [61] propose adding Shannon’s entropy to the objective function of a continuous-time linear-quadratic control problem to encourage exploration, leading to an optimal policy of Gaussian-type with the mean capturing the exploited policy and the variance quantifying the degree of exploration. Since then, there have been studies on continuous-time RL through the lens of partial differential equations and control theory under entropy regularization, primarily focusing on *regular controls*, such as controlling the drift and volatility coefficients of the underlying stochastic processes [3, 30, 31, 32, 40, 57, 60, 61].

One key challenge that distinguishes OS from regular control is that OS involves a non-smooth decision—whether to stop or continue—while regular controls gradually change the dynamics through drift and/or volatility terms. As a result, gradient-based RL algorithms, although popular for regular controls, cannot be directly applied to the “stop-or-continue” decisions in our setting [50, 52]. To overcome this difficulty, we replace these sharp rules based on the hitting times by stopping probabilities, leading to a *randomized stopping time* that follows the stopping probability. Unlike the heuristic of a fuzzy boundary considered in [50], where the agent stops with a probability proportional to the distance to the boundary, our approach to randomized stopping time is based on principled guidance—regularized objective function with cumulative residual entropy. Another advantage of randomized stopping time is its exploratory nature; it stops at different scenarios according to certain probability, thereby collecting more information from the unknown environment in the RL regime. Compared to classic RL settings, exploration is particularly essential for optimal stopping problems because the terminal reward can only be collected upon making the stopping decision, contributing to the *reward sparsity* challenge in RL [14, 29]. Therefore, randomized stopping time helps the decision-maker learn more about the unknown terminal reward.

Following the above idea, this paper proposes a framework to address continuous time and state-space OS problems via continuous-time RL. We first embed the perspective of exploration-exploitation into OS, and then use this construction to design a new type of RL algorithm for which we provide a convergence analysis.

1.1. **Our work and contributions.** In this section, we provide an informal discussion on the key ideas and contributions of this work. The precise assumptions and statements of the results can be found in Sections 2-4.

The classical OS problem (without regularization) is given, for any initial state $x \in \mathbb{R}^n$, as

$$V(x) := \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-\rho s}\pi(X_s^x)\mathrm{d}s + e^{-\rho\tau}G(X_\tau^x)\right],$$
$$\text{subject to} \quad \mathrm{d}X_t^x = b(X_t^x)\mathrm{d}t + \sigma(X_t^x)dW_t, \quad X_0^x = x,$$

where the maximization is performed over stopping times τ of the (augmented) filtration generated by the Brownian motion W, $\rho > 0$, and the functions b, σ, π, G are the data of the problem (satisfying the assumptions in Section 2).

To model exploration, we follow the ideas in [59] and let the agent randomize the choice of the stopping time. The resulting strategy are chacterized by nondecreasing processes $\xi = (\xi_t)_t$ with $0 \leq \xi_t \leq 1$, where ξ_t is the probability that the agent stops before time t according to $\xi_t = \mathbb{P}(\tau \leq t|\mathcal{F}_t^W)$. We refer to such processes ξ as *singular controls*. However, allowing for randomized stopping times *does not change* the optimal value function and, more importantly, it does not imply that optimal actions are necessarily randomized/exploratory (see Proposition 2.4). From the RL point of view, this non-exploratory behavior of the optimal control suggests that, without proper modification of the objective function, *optimizing* and *collecting information* do not naturally come together.

To overcome the above-mentioned non-exploratory phenomenon and to incentivize exploration, we reward intermediate values of ξ_t by perturbing the original OS problem via the cumulative residual entropy (CRE) [48] of the probability measures ξ

$$\mathrm{CRE}(\xi) := -\int_0^{\infty} e^{-\rho t}\left(1-\xi_t\right)\log\left(1-\xi_t\right)\mathrm{d}t.$$

The adoption of CRE in RL is *novel* in the literature, as Shannon's entropy and Kullback–Leibler (KL) divergence are more often used. It is worth noting that this CRE criterion is more suitable for encouraging randomized optimal stopping time by postponing the exercise time via a large cumulative probability.

The resulting *entropy regularized OS problem* takes the form of the singular control problem:

$$V^{\lambda}(x) := \sup_{\xi} \mathbb{E}\Bigg[\int_0^{\infty} e^{-\rho t}\underbrace{\left(\pi(X_t^x)(1-\xi_t)\mathrm{d}t + G(X_t^x)\mathrm{d}\xi_t\right)}_{\text{exploitation}} - \lambda\int_0^{\infty} e^{-\rho t}\underbrace{\left(1-\xi_t\right)\log\left(1-\xi_t\right)}_{\text{exploration}}\mathrm{d}t\Bigg],$$

where $\lambda > 0$ is a temperature parameter that balances the exploitation and exploration. For any $\lambda > 0$, such a problem admits a unique optimal control ξ^{λ}. The proposed regularization naturally approximates the original OS problem. Indeed, we show that (see formal statements in Propositions 2.9 and 2.10):

- $\sup_x |V^{\lambda}(x) - V(x)| \leq \lambda(\rho e)^{-1} \to 0$ as $\lambda \to 0$.
- For any sequence $(\lambda_k)_k \to 0$, the sequence $(\xi^{\lambda_k})_k$ converges (up to subsequences) to an optimal singular control ξ^* of the original OS with randomized stopping times.

These convergence results have also been discovered in [61] for the linear-quadratic control with Shannon's entropy, using explicit formulas for the value functions and optimal controls.

A solution to the entropy regularized OS problem is obtained by applying the dynamic programming principle (DPP). To this end, we introduce the additional controlled state variable

$Y_t^{y,\xi} = y - \xi_t$, $y \in [0,1]$, and define, for $x \in \mathbb{R}^n$ and $\lambda > 0$, the extended problem

$$V^\lambda(x,y) := \sup_{\xi_t \le y} \mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\big(\pi(X_t^x)Y_t^{y,\xi} - \lambda Y_t^{y,\xi}\log(Y_t^{y,\xi})\big)\mathrm{d}t + \int_0^\infty e^{-\rho t} G(X_t^x)\mathrm{d}\xi_t\bigg].$$

This is an $(n+1)$-dimensional *degenerate* singular stochastic control problem with finite-fuel. Notice that $V^\lambda(x,1) = V^\lambda(x)$. In particular, we show that (see Theorems 3.2 and 3.3 for the formal statements):

- The value function V^λ is $W^{2,2}_{loc}(\mathbb{R}^n \times (0,1))$ and it is the unique solution to the Hamilton-Jacobi-Bellman (HJB) variational inequality
$$\max\big\{ (\mathcal{L}_x - \rho)\, V^\lambda(x,y) + \pi(x)y - \lambda y \log y, -V^\lambda_y(x,y) + G(x)\big\} = 0,$$
with $V^\lambda(x,0) = 0$, where $\mathcal{L}_x f(x,y) := b(x) D_x f(x,y) + \frac{1}{2}\operatorname{tr}\big(\sigma\sigma^*(x) D_x^2 f(x,y)\big)$.
- The optimal control ξ^λ is of reflecting type and is given by $\xi_t^\lambda := \sup_{s\le t}\big(y - g_\lambda(X_s^x)\big)^+$, with the reflecting free boundary defined as $g_\lambda(x) := \sup\big\{y \in [0,1] \,|\, -V^\lambda_y(x,y) + G(x) < 0\big\}$.

These results illustrate that the optimal control of the entropy regularized OS problem is not related to any strict stopping time; that is, by introducing the entropy regularization, we obtain optimal strategies of exploratory type. From the RL point of view, the exploratory behavior of the optimal control is desirable, as it suggests that *optimizing* and *collecting information* from the environment can actually be achieved together.

From the technical point of view, showing the regularity of V^λ (required for the characterization of ξ^λ) involves several challenges. In particular, while the regularity in x is studied via semi-convexity estimates and PDE arguments, the regularity in y is obtained via a probabilistic connection with a different but related OS problem.

In a benchmark real option example, we can solve the entropy regularized OS problem semi-explicitly (see Theorem 4.2 for the details). For this particular example, we can characterize the free boundary g_λ and show its convergence, as $\lambda \to 0$, to the free boundary of the original OS problem.

Our theoretical analysis suggests a new approach to design RL algorithms for OS problems, which aims at learning the boundary of the λ-entropy regularized problem rather than the boundary of the original OS problem. Notice that the error introduced by the temperature parameter λ is of order $\mathcal{O}(\lambda)$. The proposed RL framework consists of two steps: a model-based step and a model-free step. In the first step, where all model parameters are known, we design a Policy Iteration algorithm to learn g_λ. In the second step, where the model parameters are unknown, we combine the Policy Iteration with a Policy Evaluation procedure. Notably, we do not directly estimate model parameters in the design, which enhances robustness against model misspecification and environmental shifts [1, 19, 28]. Our new approach also overcomes the instability issue often encountered when directly learning the boundary of the original OS problem [50].

We establish the theoretical foundation of the algorithm for the real option example. Specifically, in the kth iteration of the first step, we use Hessian information of $V^\lambda_{g_k}$ (i.e., the value function associated with the policy using reflecting boundary g_k) to update g_{k+1}, which lies in a region where $V^\lambda_{g_k}$ has a *better regularity*. Mathematically, define the new boundary $g_{k+1}(x)$ as:

$$\begin{cases} \max\Big\{y^* < g_k(x) \,\Big|\, \partial_{xy} V^\lambda_{g_k}(x,y) = 0\Big\} \text{ if } \partial^-_{xy} V^\lambda_{g_k}(x, g_k(x)) < 0, \\ g_{k+1}(x) = g_k(x) \qquad \text{o.w.,} \end{cases}$$

where we define the notation $\partial^-_{xy} f(x,y) := \lim_{h\to 0-} \frac{\partial_x f(x,y+h)-\partial_x f(x,y)}{h}$ as the left y-derivative of $\partial_x f$ (if exists). With such an updating scheme, we have the following results (see Proposition 4.5 and Theorem 4.9 for formal statements):

- **Policy improvement:** $V^\lambda_{g_{k+1}}(x,y) \geq V^\lambda_{g_k}(x,y)$ for all $k \in \mathbb{N}$.
- **Policy convergence:** $\lim_{k\to\infty} V^\lambda_{g_k}(x,y) = V^\lambda(x,y)$.

Policy improvement is a result of Itô's formula in the weak version. A similar result is shown in [61] for linear-quadratic controls. On the other hand, it is well-recognized in the literature that policy convergence is much harder to establish, with only a few exemptions for regular controls in the literature [3, 30, 40, 60]. Our result on policy convergence is established based on an induction argument that iteratively verifies the regularity, convexity, and a few estimation bounds held along the entire updating scheme.

In the second step, we combine the Policy Iteration with Policy Evaluation, where the value function of a given policy with reflecting barrier g_k is estimated using sample trajectories. Finally, we demonstrate the performance of the proposed algorithm on a few numerical examples.

1.2. Related literature.

Our work is related to different streams of literature.

Randomized optimal stopping and singular control. First of all, we relate to the literature on randomized stopping times. These typically appear when proving the existence of equilibria in zero-sum or nonzero-sum optimal stopping games (see [12, 37, 51, 59], among others) and can be thought of as the "mixed strategies" in stopping times. Loosely speaking, according to a randomized stopping time, a decision maker chooses a probability of stopping and at any time decides or not to stop according to that distribution. As discussed in [51] and [59], different but equivalent definitions of randomized stopping times can be developed. In this paper, we follow the approach in [59], which identifies randomized stopping times with nondecreasing càdlàg processes starting from zero at the initial time and bounded above by one. This in turn leads to the fact that the exploratory formulation of the optimal stopping problem is mathematically equivalent to an $(n+1)$-dimensional degenerate singular stochastic control with finite-fuel (see, e.g., [11, 34, 36]).

Continuous-time RL for regular control. Initial studies on continuous-time RL primarily focused on algorithm design and the potential for time discretization [17, 45, 56].

Following the seminal work of [61], which provided a mathematical perspective within the framework of linear-quadratic control, there has been a recent surge in employing stochastic control theory and partial differential equations (PDEs) to establish the theoretical foundation of continuous-time RL [3, 30, 32, 31, 40, 57, 60]. In particular, [57] generalized the framework in [61] and studied the exploratory Hamilton–Jacobi–Bellman (HJB) equation arising from a general entropy-regularized stochastic control problem; [32] developed q-Learning algorithm in continuous time; [31] established policy gradient and actor-critic learning frameworks in continuous time and space. Despite these developments, the overall convergence results of control-inspired or PDE-inspired algorithms, such as the Policy Improvement Algorithm (PIA), remain largely unexplored. Very recently, [30] proved a qualitative convergence result for the PIA with bounded coefficients when the diffusion term is not controlled. [60] generalized the result to unbounded coefficients or controlled diffusion terms by a uniform estimate for the value sequence generated by PIA. [40] provides a much simpler proof of the convergence of the PIA using Feynman-Kac-type probabilistic representation instead of sophisticated PDE estimates. Moreover, [3] showed the convergence of the PIA for an optimal dividend problem, taking advantage that the state process is one dimensional and taking non-negative values. In addition to the PIA method, another popular continuous-time RL algorithm is the policy gradient method. [49] demonstrated a linear convergence of this method for certain finite-horizon non-linear control problems, while [24] showed its convergence for finite-horizon exploratory linear-quadratic control problems.

Another important aspect of quantifying the effectiveness of continuous-time RL algorithms is regret analysis. For theoretical developments in this area, see [6, 26, 54, 44, 55].

We note that all the works mentioned above focus on regular controls.

Machine learning for optimal stopping. RL for optimal stopping is closely related to the challenging RL scenario of sparse rewards [14, 29]. More specifically, the terminal reward $G(X_\tau)$ can only be collected upon making the stopping decision, contributing to the sparsity of the rewards. This sparsity significantly complicates the learning process compared to the more straightforward regular control problems in continuous-time settings or classical Markov decision processes in discrete time.

When model parameters are fully known to the decision maker, [50] and [52] developed deep-learning-based learning algorithms to learn the stopping boundaries. See also [2] for solving high-dimensional singular control problems using deep learning methods.

It is worth noting that [13] proposed a comprehensive framework for policy gradient methods tailored to continuous-time RL. This framework leverages the connection between stochastic control problems and randomized problems, allowing for applications beyond diffusion models, such as regular, impulse, and optimal stopping/switching problems. However, no theoretical convergence results have been established yet. The most relevant paper to our setting is [16], which considers a regularized American Put Option problem under Shannon's entropy in one dimension, which involves a single risky asset that follows a geometric Brownian motion. The author employs an intensity control formulation to encourage exploration and demonstrates the convergence of PIA for the regularized problem under a given temperature parameter. In comparison to the latter work, we highlight some novelties of our approach. Notably, our analysis of the regularized problem is based on regularity analysis that combines PDE arguments and a probabilistic connection with another OS problem. In contrast, the analysis in [16] relies primarily on explicit calculations, which are challenging to extend to higher dimensions. Moreover, within our framework, we can show the convergence of the optimal policies of the regularized problems to the original OS-optimal policy as the temperature parameter λ tends to zero. While our convergence holds regardless of the dimension of the problem, such a convergence is not clear in [16] even for the specific American Put Option problem.

1.3. **Outline of the paper.** Section 2 examines the properties of the original OS problem, introduces the entropy-regularized version, and explores its relationship with the original problem. The entropy-regularized problem is further explored in Section 3, including a regularity analysis of the value function and the construction of the optimal policy. Section 4 then shifts focus to a real option example, establishing its analytical properties and developing corresponding RL algorithms. Finally, Appendix A contains the proof of the uniqueness of the optimal stopping time, while the proofs of some auxiliary technical estimates are collected in Appendix B.

1.4. **General notation.** For $q \in [1, \infty]$, and a measure space $(E, \mathcal{E}, m)$, we define the set $\mathbb{L}^q = \mathbb{L}^q(E) = \mathbb{L}^q(E, m)$ of measurable functions $f : E \to \mathbb{R}$ s.t. $\int_E |f|^q dm < \infty$, if $q < \infty$, and $\operatorname{ess\,sup}_E |f| < \infty$ for $q = \infty$. For $d \in \mathbb{N} \setminus \{0\}$, an open set $B \subset \mathbb{R}^d$, $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}^d$, and a function $f : B \to \mathbb{R}$, we denote by $D^\alpha f := D_1^{\alpha_1} ... D_d^{\alpha_d} f$ the weak derivative of f, where $D_i f := f_{x_i} := \partial f / \partial x_i$, and we set $|\alpha| := \alpha_1 + ... + \alpha_d$. Sometimes, we will denote by $D_x f$ and $D_x^2 f$ the gradient and the Hessian matrix of f (in the weak sense), respectively. Define the set $C(B)$ of continuous functions $f : B \to \mathbb{R}$, the set $C^\ell(B)$ of functions $f : B \to \mathbb{R}$ with continuous ℓ-order derivatives, and the Sobolev spaces $W^{\ell,q}(B)$ of functions $f \in \mathbb{L}^q(B)$ with $\sum_{|\alpha| \leq \ell} \int_E |D^\alpha f|^q dm < \infty$, if $q < \infty$, and $\sum_{|\alpha| \leq \ell} \operatorname{ess\,sup}_E |D^\alpha f| < \infty$, for $q = \infty$. Finally, define the Sobolev spaces $W^{\ell,q}_{loc}(B)$ of functions f s.t. $f \in W^{\ell,q}(D)$ for each bounded open set $D \subset B$.

2. Exploratory formulation and entropy regularization of OS problem

Consider a discount factor $\rho > 0$ and, for $n \in \mathbb{N} \setminus \{0\}$, continuous functions $b : \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : \mathbb{R}^n \to \mathbb{R}^{n \times m}$, $\pi, G : \mathbb{R}^n \to \mathbb{R}$. Let $W = (W_t)_t$ be an m-dimensional Brownian motion on a complete probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ and let $\mathbb{F}^W = (\mathcal{F}^W_t)_t$ be the right-continuous extension of the filtration generated by W. Introduce the second-order differential operator

$$\mathcal{L}_x \phi(x) := b(x) D_x \phi(x) + \frac{1}{2} \operatorname{tr}(\sigma^* \sigma(x) D^2_x \phi(x)), \quad \phi \in W^{2,2}_{loc}(\mathbb{R}^n),$$

and denote by $\mathcal{T}$ the set of $\mathbb{F}^W$-stopping times.

For an initial condition $x \in \mathbb{R}^n$ and a stopping time $\tau \in \mathcal{T}$, define the profit functional

$$J(x;\tau) := \mathbb{E}\left[\int_0^\tau e^{-\rho s} \pi(X^x_s) \mathrm{d}s + e^{-\rho \tau} G(X^x_\tau)\right],$$

$$\text{subject to } \mathrm{d}X^x_t = b(X^x_t)\mathrm{d}t + \sigma(X^x_t)\mathrm{d}W_t, \quad X^x_0 = x,$$

and consider the optimal stopping problem

$$V(x) := \sup_{\tau \in \mathcal{T}} J(x;\tau). \tag{2.1}$$

A stopping time τ^* is said to be optimal if $J(x;\tau^*) = V(x)$ and the function V is referred to as the value function.

In order to have a well-defined problem, we enforce the some requirements on the data of the problem.

Assumption 2.1. *The following conditions hold true:*

1. *There exists a constant $L > 0$ such that, for $\phi = b, \sigma$, we have*
$$|\phi(x)| \le L(1 + |x|), \quad |\phi(\bar{x}) - \phi(x)| \le L|\bar{x} - x|, \quad \textit{for any } x, \bar{x} \in \mathbb{R}^n.$$
2. *There exist $p \ge 2$ and $K > 0$ such that, for $\phi = G, \pi$, we have*
$$|\phi(x)| \le K(1 + |x|^p), \quad \textit{for any } x \in \mathbb{R}^n.$$
3. *ρ is large enough: $\rho > p(\frac{3}{2}L + (2p-1)L^2)$.*

The following result is classical.

Proposition 2.2. *Under Assumption 2.1, there exists an optimal stopping time.*

Proof. We limit ourselves to directing the reader to a reference. More specifically, Assumption 2.1 implies (see estimate (3.5) below and the related (3.3)) that, for a suitable constant $C < \infty$, we have

$$\mathbb{E}\left[\int_0^\infty e^{-\rho t} |\pi(X^x_t)| \mathrm{d}t + \sup_{t \ge 0} e^{-\rho t} |G(X^x_t)|\right] < C(1 + |x|^p).$$

Hence, the result follows by Corollary 2.9 at p. 46 in [46] (notice that, differently from [46], we do not require stopping times to be finite). □

Some of the later results (see in particular Proposition 2.4 and Proposition 2.10 below) enjoy a sharper statement under uniqueness of the optimal stopping time, which is often satisfied in the one-dimensional case. For the sake of illustration, we discuss the following sufficient conditions.

Lemma 2.3. *Under Assumption 2.1, if*

$$\begin{aligned}&n=1,\ G\in C^2(\mathbb{R}),\ \sigma^2(x)\geq c>0 \text{ and } D_xG(x)\leq C(1+|x|) \text{ for some } c,C\in(0,\infty),\\ &\text{the function } \hat{\pi}:=\pi+(\mathcal{L}_x-\rho)G \text{ is strictly increasing and there exists } \bar{x} \text{ s.t. } \hat{\pi}(\bar{x})=0,\end{aligned} \tag{2.2}$$

then the optimal stopping time is unique.

The proof of this lemma relies on classical arguments in OS and it is provided in Appendix A.

2.1. **Exploratory formulation via singular controls.** Inspired by [59], we introduce a notion of randomized/exploratory stopping times. Define the set of *singular controls* as the set of processes

$$\mathcal{A}(1):=\Big\{\xi:\Omega\times[0,\infty)\to[0,1],\ \mathbb{F}^W\text{-adapted, nondecreasing, càdlàg, with } \xi_{0-}=0\Big\},$$

and assume the probability space $(\Omega,\mathcal{F},\mathbb{P})$ to be large enough to accommodate a W-independent uniformly distributed random variable $U:\Omega\to[0,1]$.

Given any $\xi\in\mathcal{A}(1)$, we consider a random variable τ^ξ by setting $\tau^\xi:=\inf\{t\geq 0\,|\,\xi_t>U\}$, with the convention $\inf\emptyset=+\infty$, and we will refer to it as *randomized/exploratory stopping time*. Notice that τ^ξ is not necessarily an $\mathbb{F}^W$-stopping time. However, if $\tau\in\mathcal{T}$, the process $\xi^\tau\in\mathcal{A}(1)$ defined by $\xi^\tau:=(\mathbb{1}_{\{t\geq\tau\}})_t$ is such that $\tau=\tau^{\xi^\tau}$, which gives a natural inclusion of $\mathcal{T}$ into $\mathcal{A}(1)$. Moreover, we have

$$\mathbb{P}\big(\tau^\xi\leq t\mid\mathcal{F}^W_t\big)=\mathbb{P}\left(U\leq\xi_t\mid\mathcal{F}^W_t\right)=\int_0^{\xi_t}\mathrm{d}u=\xi_t,$$

so that ξ_t can be interpreted as the probability of stopping before time t. Notice also that $\xi_\infty:=\lim_{t\to\infty}\xi_t$ is not necessarily equal to 1: indeed, the related randomized stopping time τ^ξ is not necessarily finite.

For $\xi\in\mathcal{A}(1)$, by evaluating the randomized stopping time τ^ξ, we obtain

$$\begin{aligned}J(x;\tau^\xi)=&\mathbb{E}\bigg[\int_0^{\tau^\xi}e^{-\rho t}\pi(X^x_t)\mathrm{d}t+e^{-\rho\tau^\xi}G\left(X^x_{\tau^\xi}\right)\bigg]\\ =&\mathbb{E}\left[\int_0^\infty e^{-\rho t}\pi(X^x_t)\mathbb{1}_{\{t\leq\tau^\xi\}}\mathrm{d}t+\int_0^\infty e^{-\rho t}G(X^x_t)\mathbb{P}\big(\tau^\xi\in\mathrm{d}t\mid\mathcal{F}^W\big)\right]\\ =&\mathbb{E}\left[\int_0^\infty e^{-\rho t}\pi(X^x_t)\mathbb{1}_{\{\xi_t<U\}}\mathrm{d}t+\int_0^\infty e^{-\rho t}G(X^x_t)\mathrm{d}\xi_t\right]\\ =&\mathbb{E}\left[\int_0^\infty e^{-\rho t}\pi(X^x_t)\left(1-\xi_t\right)\mathrm{d}t+\int_0^\infty e^{-\rho t}G(X^x_t)\mathrm{d}\xi_t\right].\end{aligned}$$

This suggests to define a profit functional in terms of the singular controls $\xi\in\mathcal{A}(1)$ by

$$J^0(x;\xi):=\mathbb{E}\left[\int_0^\infty e^{-\rho t}\pi(X^x_t)\left(1-\xi_t\right)\mathrm{d}t+\int_0^\infty e^{-\rho t}G(X^x_t)\mathrm{d}\xi_t\right],\quad \xi\in\mathcal{A}(1),$$

as a natural extension of J to randomized stopping times.

However, allowing for randomized stopping times does not change the optimal value and (crucial for our story) it does not imply that optimal actions are necessarily randomized, as discussed in the next proposition.

Proposition 2.4. *Under Assumption 2.1, for any $x\in\mathbb{R}^n$ we have*

$$V(x)=\sup_{\xi\in\mathcal{A}(1)}J^0(x;\xi)=J^0(x;\xi^*),$$

with $\xi^*:=(\mathbb{1}_{\{t\geq\tau^*\}})_t$ and τ^* optimal for $J(x;\cdot)$. Moreover, if the optimal stopping time τ^* is unique (e.g., if (2.2) holds), then $\xi^*:=(\mathbb{1}_{\{t\geq\tau^*\}})_t$ is the unique optimal control for $J^0(x;\cdot)$.

Proof. For generic $\xi \in \mathcal{A}(1)$ and $z \in [0,1]$, we can define the stopping time $\tau^{\xi}(z)$ as

$$\tau^{\xi}(z) := \inf\{t \geq 0 \,|\, \xi_t \geq z\},$$

with the convention $\inf \emptyset := +\infty$. Indeed, notice that $\tau^{\xi}(z) = +\infty$ if $z > \xi_{\infty}$. By using Fubini's theorem and then the change of variable formula, we have

$$\begin{aligned}
&J^0(x;\xi)\\
&=\mathbb{E}\left[\int_0^{\infty} e^{-\rho t}\pi(X_t^x)\left(1-\xi_{\infty}+\int_t^{\infty}\mathrm{d}\xi_s\right)\mathrm{d}t+\int_0^{\infty} e^{-\rho t}G(X_t^x)\mathrm{d}\xi_t\right]\\
&=\mathbb{E}\left[(1-\xi_{\infty})\int_0^{\infty} e^{-\rho t}\pi(X_t^x)\mathrm{d}t+\int_0^{\infty}\left(\int_0^t e^{-\rho s}\pi(X_s^x)\mathrm{d}s+e^{-\rho t}G(X_t^x)\right)\mathrm{d}\xi_t\right]\\
&=\mathbb{E}\left[(1-\xi_{\infty})\int_0^{\infty} e^{-\rho t}\pi(X_t^x)\mathrm{d}t+\int_0^{\xi_{\infty}}\left(\int_0^{\tau^{\xi}(z)} e^{-\rho s}\pi(X_s^x)\mathrm{d}s+e^{-\rho\tau^{\xi}(z)}G(X^x_{\tau^{\xi}(z)})\right)\mathrm{d}z\right]\\
&=\int_0^1\mathbb{E}\bigg[\int_0^{\tau^{\xi}(z)} e^{-\rho s}\pi(X_s^x)\mathrm{d}s+e^{-\rho\tau^{\xi}(z)}G(X^x_{\tau^{\xi}(z)})\bigg]\mathrm{d}z=\int_0^1 J(x;\tau^{\xi}(z))\mathrm{d}z.
\end{aligned}$$

Thus, since $\tau^{\xi}(z)$ is a stopping time, by the optimality of τ^* we have

$$J^0(x;\xi) \leq \int_0^1 \sup_{\tau\in\mathcal{T}} J(x;\tau)\mathrm{d}z = J(x;\tau^*) = V(x) = J^0(x;\xi^*),$$

where the last equality follows from the fact that ξ^* is such that $\tau^{\xi^*}(z) = \tau^*$ for any $z \in (0,1]$. This proves the first part of the proposition.

Assume now that the optimal stopping time τ^* is unique. By repeating the previous argument with ξ being optimal, we obtain that $J(x;\tau^{\xi}(z)) = J(x;\tau^*)$ dz-a.e. in $(0,1)$. By uniqueness of the optimal stopping time, we deduce that $\tau^{\xi}(z) = \tau^*$ dz-a.e. in $(0,1)$, which in turn imples $\xi = \xi^*$. □

Remark 2.5 (Non-exploratory behavior of the optimal controls)**.** *We elaborate on Proposition 2.4 from the RL and exploration point of view. Under the uniqueness of the optimal stopping time τ^*, the optimal singular control ξ^* in fact corresponds to the strict stopping time τ^*. Thus, no randomization is needed for the optimal strategy. In particular, once an action becomes necessary, the agent stops the process, which prevents the gradual collection of information on the performance of other actions. From the RL point of view, this non-exploratory behavior of the optimal control ξ^* suggests that optimizing and collecting information do not naturally come together, and a modification is necessary in order to overcome this phenomenon (see the related Remark 3.5 below for a discussion on the consequences and benefits of our entropy regularization).*

2.2. **Entropy regularization.** In light of Proposition 2.4 and of the related Remark 2.5, we introduce a regularization term to incentivize exploration/randomization. This is achieved by incorporating an entropy term to regularize the problem, drawing motivation from the RL literature.

Since the sample paths of the processes $\xi \in \mathcal{A}(1)$ are not necessarily absolutely continuous with respect to the Lebesgue measure, we choose (a discounted version of) the *cumulative residual entropy* (see [48]) weighted by a parameter $\lambda > 0$; namely, we consider, for $\xi \in \mathcal{A}(1)$ the entropy

$$\begin{aligned}
\Lambda^{\lambda}(\tau^{\xi}) :=& -\lambda\int_0^{\infty} e^{-\rho t}\mathbb{P}\left(\tau^{\xi}\geq t\mid\mathcal{F}_t^W\right)\log\left(\mathbb{P}\left(\tau^{\xi}\geq t\mid\mathcal{F}_t^W\right)\right)\mathrm{d}t\\
=& -\lambda\int_0^{\infty} e^{-\rho t}\left(1-\xi_t\right)\log\left(1-\xi_t\right)\mathrm{d}t =: \Lambda^{\lambda}(\xi).
\end{aligned} \tag{2.3}$$

The entropy Λ^λ is nonnegative and achieves its highest values when the probability ξ_t is near the level e^{-1}, thus incentivising the use of randomized stopping times.

Building on this intuition, we define an entropy regularized OS problem

$$V^\lambda(x) := \sup_{\xi \in \mathcal{A}(1)} J^\lambda(x;\xi), \tag{2.4}$$

where the exploration–exploitation trade off is captured by the profit functional

$$J^\lambda(x;\xi) := \mathbb{E}\bigg[\int_0^\infty e^{-\rho t} \underbrace{\big(\pi(X_t^x)(1-\xi_t)\mathrm{d}t + G(X_t^x)\mathrm{d}\xi_t\big)}_{\text{exploitation}} - \lambda \int_0^\infty e^{-\rho t} \underbrace{(1-\xi_t)\log{(1-\xi_t)}}_{\text{exploration}} \mathrm{d}t \bigg].$$

A control $\xi^\lambda \in \mathcal{A}(1)$ is said to be optimal if $J^\lambda(x;\xi^\lambda) = V^\lambda(x)$.

Before studying the entropy regularized problem and its connections with the original stopping problem, we list here few remarks.

Remark 2.6. *For any $x \in \mathbb{R}^n$ and $\lambda > 0$, the existence of a unique optimal control ξ^λ can be established by exploiting the strict concavity of the functional $J^\lambda(x;\cdot)$ in the control variable ξ (see in particular Theorem 8 in [43]). Moreover, Theorem 3.3 below will characterize such an optimal control and illustrate that the randomized strategy maximizing J^λ does not correspond to a (classical) stopping time (see also Remark 3.5).*

Remark 2.7 (Connection between optimal stopping and singular control)**.** *It is well-known in the stochastic control literature that optimal stopping problems and singular control problems enjoy intimate connections. On the one hand, the derivative (in the controlled state variable) of the value function of a singular control problem can be written as the value function of an optimal stopping problem [7, 8, 10, 27, 35]. On the other hand, randomized stopping strategies can be naturally seen as finite-fuel singular controls (see [59] or the discussion above), which has the advantage of compactifying the space of stopping strategies (thus explaining its heavy use in the Dynkin game literature).*

In this framework, our entropy regularization adds two crucial properties on the latter connection. First, while solely considering randomized strategies still leads to a degenerate singular control problem (where the optimal actions are pure jumps, as in Proposition 2.4), our cumulative residual entropy regularizes the problem by making it concave and allows for a more explicit characterization of the optimal control in terms of a reflecting strategy (see Theorem 3.3 below). Secondly, this entropy regularization very well approximates the original OS problem, both in terms of value and optimal strategy (see Propositions 2.9 and 2.10 in the next subsection). For these reasons, our entropy regularization seems to be a useful tool in order to approximate/characterize the equilibria also in the context of Dynkin games, which we leave for future research.

Remark 2.8. *Our choice to work with a Markovian formulation is motivated by RL. Since the typical optimization problem that RL aims to solve (in the discrete time case) is a Markov decision problem, starting our study with a Markovian optimal stopping problem seems to be a natural choice. However, it should be pointed out that our entropy regularization formulation (and its subsequent analysis) also allows us to treat optimal stopping problems in a non-Markovian framework. In this regard, we limit ourselves to mentioning that (while we will solve the Markovian entropy-regularized problem using the DDP approach in Section 3 below), in order to tackle the non-Markovian entropy-regularized OS problem, one can rely on representation theorems via backward stochastic differential equations (see [4, 5] among others).*

2.3. **Vanishing entropy limit.** With elementary arguments, one can show that the entropy regularized problem approximates the original OS problem as the entropy tends to zero.

Proposition 2.9. *Under Assumption 2.1, for any $\lambda, \bar{\lambda} \in [0,1]$ we have*

$$\sup_{x\in\mathbb{R}^n} |V^{\lambda}(x) - V^{\bar{\lambda}}(x)| \leq |\lambda - \bar{\lambda}|(\rho e)^{-1}.$$

In particular, $V^{\lambda} \to V$ uniformly on $\mathbb{R}^n$, as $\lambda \to 0$.

Proof. Take $x \in \mathbb{R}$ and $\lambda, \bar{\lambda} \in [0,1]$. If $\lambda \geq \bar{\lambda}$, for every $\xi \in \mathcal{A}(1)$ we have $J^{\lambda}(x;\xi) \geq J^{\bar{\lambda}}(x;\xi)$, so that

$$0 \leq V^{\lambda}(x) - V^{\bar{\lambda}}(x).$$

Moreover, upper bounding the right-hand side above, we find

$$\begin{aligned}
0 \leq V^{\lambda}(x) - V^{\bar{\lambda}}(x) &= \sup_{\xi\in\mathcal{A}(1)} J^{\lambda}(x;\xi) - \sup_{\bar{\xi}\in\mathcal{A}(1)} J^{\bar{\lambda}}(x;\bar{\xi}) \\
&\leq \sup_{\xi\in\mathcal{A}(1)} \Big(J^{\lambda}(x;\xi) - J^{\bar{\lambda}}(x;\xi) \Big) \\
&= (\lambda - \bar{\lambda}) \sup_{\xi\in\mathcal{A}(1)} \mathbb{E}\bigg[\int_0^{\infty} e^{-\rho t}\big(- (1-\xi_t)\log(1-\xi_t)\big) \mathrm{d}t \bigg] \\
&= (\lambda - \bar{\lambda}) \frac{1}{\rho} \sup_{z\in[0,1]} (-z\log z) = (\lambda - \bar{\lambda})(\rho e)^{-1},
\end{aligned}$$

which is the claim bound. We conclude the proof by choosing $\bar{\lambda} = 0$ and taking limits as $\lambda \to 0$. □

Proposition 2.10. *For any $x \in \mathbb{R}^n$ and $\lambda > 0$, let ξ^{λ} be the unique optimal control as in Remark 2.6. Under Assumption 2.1, the following statements hold true:*

1. *(1) There exist a subsequence $(\lambda_k)_k$ with $\lambda_k \to 0$ and an optimal control ξ^* for $J^0(x;\cdot)$ such that $\xi^{\lambda_k} \to \xi^*$ weakly in $\mathbb{L}^2(\Omega \times [0,\infty))$ as $k \to \infty$.*
2. *(2) If the optimal stopping time τ^* is unique (e.g., if (2.2) holds), then $\xi^{\lambda} \to (\mathbb{1}_{\{t\geq\tau^*\}})_t$ weakly in $\mathbb{L}^2(\Omega \times [0,\infty))$ as $\lambda \to 0$.*

Proof. Since the family $(\xi^{\lambda})_{\lambda}$ is bounded, we can find a subsequence $(\lambda_k)_k$ with $\lambda_k \to 0$ and a limit point ξ^* such that $\xi^{\lambda_k} \to \xi^*$ weakly in $\mathbb{L}^2(\Omega \times [0,\infty))$ as $k \to \infty$. We need to show that ξ^* is optimal for $J^0(x;\cdot)$.

First of all, using Proposition 2.9, find

$$\begin{aligned}
|V(x) - J(x;\xi^{\lambda_k})| &\leq |V(x) - J^{\lambda_k}(x;\xi^{\lambda_k})| + |J^{\lambda_k}(x;\xi^{\lambda_k}) - J(x;\xi^{\lambda_k})| \\
&\leq |V(x) - V^{\lambda_k}(x)| + \mathbb{E}[\Lambda^{\lambda_k}(\xi^{\lambda_k})] \\
&\leq C\lambda_k \to 0, \quad \text{as } k \to \infty.
\end{aligned} \tag{2.5}$$

Moreover, since $\sup_{t\geq0} \xi_t^{\lambda_k} \leq 1$, we can employ (after some minimal adjustment to take care of our infinite time horizon $[0,\infty)$) Lemma 3.5 in [33] in order to find a subsequence (not relabelled) $(\lambda_k)_k$ with $\lambda_k \to 0$ and a limit point $\xi \in \mathcal{A}(1)$ such that, setting $\zeta_t^m := \frac{1}{m}\sum_{k=1}^m \xi_t^{\lambda_k}$, we have

$$\int_0^{\infty} e^{-\rho t} f_t \mathrm{d}\zeta_t^m \to \int_0^{\infty} e^{-\rho t} f_t \mathrm{d}\xi_t, \ \mathbb{P}\text{-a.s., as } m \to \infty, \text{ for any } f \in C_b([0,\infty)).$$

It is easy to show that $\xi = \xi^*$. Since $\zeta^m \in \mathcal{A}(1)$, by Proposition 2.4 and the concavity of $J^0(x;\cdot)$ we find

$$\begin{aligned}
V(x) \geq J^0(x;\zeta^m) &\geq \frac{1}{m}\sum_{k=1}^m J^0(x;\xi^{\lambda_k}) \\
&= \frac{1}{m}\sum_{k=1}^m J^{\lambda_k}(x;\xi^{\lambda_k}) + \frac{1}{m}\sum_{k=1}^m \big(J^0(x;\xi^{\lambda_k}) - J^{\lambda_k}(x;\xi^{\lambda_k})\big) \\
&= \frac{1}{m}\sum_{k=1}^m J^{\lambda_k}(x;\xi^{\lambda_k}) - \frac{1}{m}\sum_{k=1}^m \mathbb{E}[\Lambda^{\lambda_k}(\xi^{\lambda_k})] \\
&\geq \frac{1}{m}\sum_{k=1}^m J^{\lambda_k}(x;\xi^{\lambda_k}) - (e\rho)^{-1}\frac{1}{m}\sum_{k=1}^m \lambda_k \to V(x), \quad \text{as } m \to \infty,
\end{aligned}$$

where the limits in the right-hand side follows from (2.5). Thus, we obtain $\lim_m J^0(x;\zeta^m) = V(x)$. By using the convergence of ζ^m to ξ^* and the dominated convergence theorem, the latter limits implies that

$$J^0(x;\xi^*) = \lim_m J^0(x;\zeta^m) = V(x),$$

proving the optimality of ξ^* (*cf.* Proposition 2.4).

If the optimal stopping time τ^* is unique, then, by Proposition 2.4, the optimal randomized stopping time ξ^* is unique as well. Therefore, ξ^* is the unique limit point of any subsequence of $(\xi^\lambda)_\lambda$, so that the whole sequence converges to ξ^*. □

Remark 2.11. *From the RL algorithmic design point of view, it is important to notice that Proposition 2.9 quantifies the error which is made when replacing the original OS problem* (2.1) *with the entropy regularized OS problem* (2.4)*.*

Remark 2.12. *Regarding Propositions 2.9 and 2.10, it is natural to question whether, and in what sense, the reflecting boundary of the entropy-regularized OS problem* (2.4) *converges to the optimal stopping boundary of the original OS problem* (2.1)*. Although establishing this result at a general level may be challenging, Section 4 (specifically the discussion in Subsection 4.1.3) offers a positive answer in a classical example of an OS problem arising from real option evaluation.*

3. Solving the entropy regularized OS problem via dynamic programming

To approach problem (2.4) via DPP we introduce, for $y \in [0,1]$, the additional controlled state process $Y_t^{y,\xi} = y - \xi_t$ and the set of admissible controls

$$\mathcal{A}(y) := \{\xi \in \mathcal{A}(1) \text{ with } \xi_t \in [0,y] \text{ a.s. for any } t \geq 0\}. \tag{3.1}$$

Thus, for $(x,y) \in \mathbb{R}^n \times [0,1]$ and $\xi \in \mathcal{A}(y)$, define the profit functional

$$J^\lambda(x,y;\xi) := \mathbb{E}\left[\int_0^\infty e^{-\rho t}\big((\pi(X_t^x)Y_t^{y,\xi} - \lambda Y_t^{y,\xi}\log(Y_t^{y,\xi})\big)\mathrm{d}t + \int_0^\infty e^{-\rho t}G(X_t^x)\mathrm{d}\xi_t\right],$$

$$\begin{aligned}
\text{subject to } \ &\mathrm{d}X_t^x = b(X_t^x)\mathrm{d}t + \sigma(X_t^x)\mathrm{d}W_t, \quad X_0^x = x, \\
&\mathrm{d}Y_t^{y,\xi} = -\mathrm{d}\xi_t, \quad Y_{0-}^{y,\xi} = y.
\end{aligned}$$

We then consider the optimization problem

$$V^\lambda(x,y) := \sup_{\xi \in \mathcal{A}(y)} J^\lambda(x,y;\xi), \tag{3.2}$$

and we notice that $V^\lambda(x) = V^\lambda(x,1)$.

In order to solve the entropy regularized problem, we introduce the following requirements, which imply Assumption 2.1.

Assumption 3.1. *The following conditions hold true:*

(1) There exists a constant $L > 0$ such that, for $\phi = b, \sigma$, we have

$$|\phi(x)| \leq L(1 + |x|),$$
$$|\phi(\bar{x}) - \phi(x)| \leq L|\bar{x} - x|,$$
$$|\phi(\delta\bar{x} + (1-\delta)x) - \delta\phi(\bar{x}) - (1-\delta)\phi(x)| \leq L\delta(1-\delta)|\bar{x} - x|^2,$$

for any $x, \bar{x} \in \mathbb{R}^n$, $\delta \in [0,1]$.

(2) There exist $p \geq 2$ and $K > 0$ such that, for $\phi = G, \pi$, we have

$$|\phi(x)| \leq K(1 + |x|^p),$$
$$|\phi(\bar{x}) - \phi(x)| \leq K(1 + |\bar{x}|^{p-1} + |x|^{p-1})|\bar{x} - x|,$$
$$|\phi(\delta\bar{x} + (1-\delta)x) - \delta\phi(\bar{x}) - (1-\delta)\phi(x)| \leq K\delta(1-\delta)(1 + |\bar{x}|^{p-2} + |x|^{p-2})|\bar{x} - x|^2,$$

for any $x, \bar{x} \in \mathbb{R}^n$, $\delta \in [0,1]$.

(3) The matrix $a(x) := \sigma^*\sigma(x)$ is uniformly elliptic; that is, there exists a constant $\kappa_\sigma > 0$ such that

$$\sum_{i,j=1}^{n} a_{i,j}(x) z_i z_j \geq \kappa_\sigma |z|^2, \quad \text{for any } z \in \mathbb{R}^n.$$

(4) ρ is large enough: namely,

$$\rho > \max\Big\{ \frac{\hat{c}_0(4(p-1))}{2}, \hat{c}_1(2), \hat{c}_1(4), \frac{\hat{c}_0(2(p-1)) + \hat{c}_1(2)}{2}, \frac{\hat{c}_0(2p)}{2}, \hat{c}_0(p), \frac{\hat{c}_0(2(p-1)) + \hat{c}_2(2)}{2}, \hat{c}_2(2), \frac{\hat{c}_2(4)}{2}, \frac{\hat{c}_1(8)}{2}, \frac{\hat{c}_0(4(p-2))}{2} \Big\},$$

where, for generic q we define the constants

$$\begin{aligned} \hat{c}_0(q) &:= q\Big(\frac{3}{2}L + (q-1)L^2\Big), \\ \hat{c}_1(q) &:= q\Big(L + \frac{q-1}{2}L^2\Big), \\ \hat{c}_2(q) &:= 2q\Big(L + \frac{2q-1}{2}L^2\Big) + (2q-1)L + 2(q-1)^2L^2, \end{aligned} \tag{3.3}$$

which are related to the growth behavior of the underlying diffusion (see the estimates (3.5)-(3.10) below).

The following theorem characterizes the value function of the entropy regularized OS problem. The proof of the existence of a solution to the HJB equation is given in Subsection 3.2, while the proof of its uniqueness is given in Subsection 3.3.

Theorem 3.2. *The value function V^λ is $C(\mathbb{R}^n \times [0,1]) \cap W^{2,2}_{loc}(\mathbb{R}^n \times (0,1))$, is concave in y, is such that $V(x,y) \leq C(1 + |x|^p)$ and it solves in the a.e. sense the HJB equation*

$$\max\big\{ (\mathcal{L}_x - \rho)V^\lambda(x,y) + \pi(x)y - \lambda y \log y, -V^\lambda_y(x,y) + G(x) \big\} = 0,$$

with boundary condition $V^\lambda(x,0) = 0$.

Moreover, the HJB equation admits a unique solution in the class of $C(\mathbb{R}^n \times [0,1]) \cap W^{2,2}_{loc}(\mathbb{R}^n \times (0,1))$ functions which are concave in y and such that $V(x,y) \leq C(1 + |x|^p)$.

We next move to the characterization of the optimal control. Define the function $g_\lambda : \mathbb{R}^n \to [0,1]$ by

$$g_\lambda(x) := \sup \big\{ y \in [0,1] \,|\, -V_y^\lambda(x,y) + G(x) < 0 \big\}, \tag{3.4}$$

where we set $g_\lambda(x) := 1$ if $\{y \in [0,1] \,|\, -V_y^\lambda(x,y) + G(x) < 0\} = \emptyset$. Notice that the function g_λ is well defined by the concavity of V^λ. Intuitively, the function g_λ separates the "exploration region" $\mathcal{E}_\lambda := \{(x,y) \,|\, -V_y^\lambda(x,y) + G(x) < 0\}$ in which no action is required, from the "stopping region" $\mathcal{S}_\lambda := \{(x,y) \,|\, -V_y^\lambda(x,y) + G(x) = 0\}$, in which the agent has to act.

The following theorem characterizes the optimal control in terms of the function g_λ. Its proof can be essentially found in the Step 2 in Subsection 3.3 and it is briefly summarized in Subsection 3.4.

Theorem 3.3. *For any $(x,y) \in \mathbb{R}^n \times [0,1]$, there exists a unique optimal control $\xi^\lambda \in \mathcal{A}(y)$, which is given by the reflection policy at the boundary g_λ; that is,*

$$\xi_t^\lambda := \sup_{s \le t} \big(y - g_\lambda(X_s^x) \big)^+, \quad t \ge 0.$$

A few comments are in order.

Remark 3.4 (Optimal control and the related Skorokhod problem)**.** *The optimal control ξ^λ is the control with minimal total variation such that*

$$(X_t^x, Y_t^{y,\xi^\lambda}) \in \overline{\mathcal{E}_\lambda}, \quad \text{for any } t \ge 0, \ \mathbb{P}\text{-a.s.},$$

where $\overline{\mathcal{E}_\lambda}$ is the closure of $\mathcal{E}_\lambda$. Such a process is also known as the solution to the Skorokhod reflection problem for the underlying state process in the waiting region $\mathcal{E}_\lambda$. Note that characterizing the optimal singular control in terms of the related Skorokhod problem represents a challenging open problem in singular control theory (we refer to the discussion in [15] for further details).

Remark 3.5 (Exploratory behavior of optimal controls)**.** *Unlike the OS problem with exploratory strategies but without modification on the objective function (see Proposition 2.4 and the related Remark 2.5), Theorem 3.3 clarifies that the optimal control of the entropy regularized OS problem is no longer a strict stopping time. In other words, introducing entropy regularization leads to optimal strategies of an exploratory nature—more precisely, strategies that are exclusively exploratory.*

From an RL perspective, the exploratory behavior of the optimal control is crutial, as it allows for simultaneous optimization and information gathering from the environment. Specifically, once the state process reaches the boundary of the exploration region, not only is action required, but, given the regularity of the free boundary g_λ (as demonstrated in the example in Section 4), an infinite number of infinitesimal actions are necessary to prevent the state process from entering the stopping region. This enables the agent to continuously gather information on the optimality of selecting a particular reflecting boundary in a non-episodic manner.

3.1. **Preliminary estimates.** Before diving into the proof of Theorem 3.2, we discuss some estimates that will be used several times in the sequel.

First, for a generic $q \ge 1$ and $\hat{c}_0(q), \hat{c}_1(q), \hat{c}_2(q)$ as in (3.3), we have the estimates

$$\mathbb{E}[|X_t^x|^q] \le C(1+|x|^q)e^{\hat{c}_0(q)t}, \quad q \ge 1, \tag{3.5}$$

$$\mathbb{E}[|X_t^{\bar{x}} - X_t^x|^q] \le |\bar{x} - x|^q e^{\hat{c}_1(q)t}, \quad q \ge 2, \tag{3.6}$$

$$\mathbb{E}\big[\big|\delta X_t^{\bar{x}} + (1-\delta)X_t^x - X_t^{\delta\bar{x}+(1-\delta)x}\big|^q\big] \le C\delta(1-\delta)|\bar{x}-x|^{2q} e^{\hat{c}_2(q)t}, \quad q \ge 2, \tag{3.7}$$

for any $x, \bar{x} \in \mathbb{R}^n$, $\delta \in [0,1]$. Even if these estimates are obtained through standard stochastic calculus techniques, it is difficult to provide a reference in which the exact exponential orders of growth $\hat{c}_0, \hat{c}_1, \hat{c}_2$ appear (importantly, these constants are those appearing in our Condition 4 in Assumption 3.1). For this reason, we provide a proof in Appendix B.

When ρ is large enough (as in Condition 4 of Assumption 3.1), the estimates above together with the Burkholder-Davis-Gundy inequality imply that

$$\mathbb{E}\bigg[\sup_{t\le\tau} e^{-\rho t}|X_t^x|^q\bigg] \le C(1+|x|^q), \quad q = 2(p-1), p, 2(p-2), \tag{3.8}$$

$$\mathbb{E}\bigg[\sup_{t\le\tau} e^{-\rho t}|X_t^{\bar{x}} - X_t^x|^q\bigg] \le C|\bar{x}-x|^q, \quad q = 2, 4 \tag{3.9}$$

(3.10)
$$\mathbb{E}\bigg[\sup_{t\le\tau} e^{-\rho t}\big|\delta X_t^{\bar{x}} + (1-\delta)X_t^x - X_t^{\delta\bar{x}+(1-\delta)x}\big|^2\bigg] \le C\delta(1-\delta)|\bar{x}-x|^4,$$

for any $x, \bar{x} \in \mathbb{R}^n$, $\delta \in [0,1]$ and $\tau \in \mathcal{T}$. Again, we provide a proof in Appendix B.

Finally, using the growth of π and G in Condition 2 in Assumption 3.1 as well as the estimates (3.5) and (3.8), we find

$$\begin{aligned} V^\lambda(x,y) &\le \sup_{\xi\in\mathcal{A}(y)} \mathbb{E}\bigg[\int_0^\infty e^{-\rho t}|\Pi(X_t^x, Y_{t-}^{y,\xi})|\mathrm{d}t + \int_0^\infty e^{-\rho t}|G(X_t^x)|\mathrm{d}\xi_t\bigg] \\ &\le \sup_{\xi\in\mathcal{A}(y)} \mathbb{E}\bigg[\int_0^\infty e^{-\rho t}(1+|X_t^x|^p)\mathrm{d}t + \lim_{T\to\infty}\sup_{t\le T} e^{-\rho t}(1+|X_t^x|^p)\xi_T\bigg] \\ &\le C(1+|x|^p), \end{aligned} \tag{3.11}$$

where we have used the fact that $\xi_t \le y \le 1$ for any $\xi \in \mathcal{A}(y)$.

3.2. **Proof of Theorem 3.2: solution of the HJB.** The concavity of V^λ follows by straightforward arguments (see Remark 2.6 for a reference), while its growth rate was shown in (3.11). Thus, we will show in the sequel that V^λ has the claimed regularity and it solves the HJB equation in the almost everywhere sense.

The proof is divided into five steps. In the proof, we will often use the notation $\Pi(x,y) := y\pi(x) - \lambda y \log y$.

Step 1 (Penalization). Let $\beta : \mathbb{R} \to [0,\infty)$ be a C^∞ nondecreasing convex function such that

$$\beta(z) = \begin{cases} 0, & \text{if } z \le 0, \\ 2z-1, & \text{if } z \geqslant 1. \end{cases}$$

For any $\varepsilon > 0$, consider the control set $U_\varepsilon := \big\{(a^1, a^2) \in \mathbb{R}^2 \,|\, sa^1 - a^2 \le \frac{\beta(s)}{\varepsilon} \ \forall s \in \mathbb{R}$ and $0 \le a^2 \le \frac{1}{\varepsilon}\big\}$ and the functional

$$\begin{aligned} J^{\lambda,\varepsilon}(x,y;(u,\eta)) &:= \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\Pi\left(X_t^x, Y_t^{y,u}\right) - G(X_t^x)u_t - \eta_t\right)\mathrm{d}t\right], \quad (u,\eta) \in \mathcal{U}_\varepsilon(y) \\ \text{subject to} \quad \mathrm{d}X_t^x &= b(X_t^x)\mathrm{d}t + \sigma(X_t^x)\mathrm{d}W_t, \quad X_0^x = x, \\ \mathrm{d}Y_t^{y,u} &= -u_t\mathrm{d}t, \quad Y_0^{y,u} = y, \end{aligned}$$

where $\mathcal{U}_\varepsilon(y) := \{(u,\eta) : \Omega \times [0,\infty) \to U_\varepsilon \,|\, Y_t^{y,u} \geqslant 0 \ \forall t \geqslant 0, \mathbb{P}\text{-a.s.}\}$. For any $\varepsilon > 0$, define the optimization problem

$$V^{\lambda,\varepsilon}(x,y) := \sup_{(u,\eta)\in\mathcal{U}_\varepsilon(y)} J^{\lambda,\varepsilon}(x,y;(u,\eta)). \tag{3.12}$$

Similarly to (3.11), we can find C such that, for y small enough, one has

$$
\begin{aligned}
&|V^{\lambda,\varepsilon}(x,y)| \\
&\quad\le \sup_{(u,\eta)\in\mathcal{U}_\varepsilon(y)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left|\Pi\left(X_t^x, Y_t^{y,u}\right) - G(X_t^x)u_t - \eta_t\right|\mathrm{d}t\right] \\
&\quad\le \sup_{(u,\eta)\in\mathcal{U}_\varepsilon(y)} \mathbb{E}\Big[\int_0^\infty e^{-\rho t} y(|\pi(X_t^x)| - \lambda\log y)\mathrm{d}t + \lim_{T\to\infty}\sup_{t\le T}\left(e^{-\rho t}|G(X_t^x)|\right)\int_0^T u_t\mathrm{d}t\Big] \\
&\quad\le Cy\bigg(\mathbb{E}\Big[\int_0^\infty e^{-\rho t}(1+|X_t^x|^p - \lambda\log y)\mathrm{d}t + \lim_{T\to\infty}\sup_{t\le T} e^{-\rho t}(1+|X_t^x|^p)\Big]\bigg) \\
&\quad\le Cy(1+|x|^p-\lambda\log y)\to 0, \quad \text{as } y\to 0,
\end{aligned}
\tag{3.13}
$$

which implies the continuity of $V^{\lambda,\varepsilon}$ for $y=0$.

Standard arguments (see e.g. Chapter IV.10 in [22]) give that $V^{\lambda,\varepsilon}$ is a generalized solution solution to the HJB equation:

$$
\rho V^{\lambda,\varepsilon}(x,y) - \mathcal{L}_x V^{\lambda,\varepsilon}(x,y) - \Pi(x,y) - \frac{1}{\varepsilon}\beta\left(G(x) - V_y^{\lambda,\varepsilon}(x,y)\right) = 0,
\tag{3.14}
$$

with boundary condition $V^{\lambda,\varepsilon}(x,0)=0$.

Moreover, following the rationale of the proof of Theorem 3.2 in [62], one can show that $V(x,y) = \lim_{\varepsilon\to 0} V^{\lambda,\varepsilon}(x,y)$.

Step 2 (Uniform Lipschitz estimates for $V^{\lambda,\varepsilon}$ in x). Take $x,\bar{x}\in\mathbb{R}^n$, $y\in(0,1]$, $\zeta>0$ and $(u,\eta)\in\mathcal{U}_\varepsilon(y)$ such that $V^{\lambda,\varepsilon}(\bar{x},y)-\zeta\le J^{\lambda,\varepsilon}(\bar{x},y;u,\eta)$. Since (u,η) is suboptimal for (x,y), we have

$$
\begin{aligned}
V^{\lambda,\varepsilon}(\bar{x},y) - V^{\lambda,\varepsilon}(x,y) - \zeta &\le J^{\lambda,\varepsilon}(\bar{x},y;u,\eta) - J^{\lambda,\varepsilon}(x,y;u,\eta) \\
&= \mathbb{E}\Big[\int_0^\infty e^{-\rho t}\big(Y_t^{y,u}(\pi(X_t^{\bar{x}}) - \pi(X_t^x)) + u_t(G(X_t^{\bar{x}}) - G(X_t^x))\big)\mathrm{d}t\Big].
\end{aligned}
\tag{3.15}
$$

In order to continue this estimate, we notice that by monotone convergence theorem, we have

$$
\mathbb{E}\Big[\int_0^\infty e^{-\rho t}u_t(G(X_t^{\bar{x}}) - G(X_t^x))\mathrm{d}t\Big] \le \lim_{T\to\infty}\mathbb{E}\Big[\int_0^T e^{-\rho t}u_t|G(X_t^{\bar{x}}) - G(X_t^x)|\mathrm{d}t\Big].
\tag{3.16}
$$

Furthermore, since $(u,\eta)\in\mathcal{U}_\varepsilon(y)$, we have $\int_0^t u_s\mathrm{d}s\le y\le 1$. Thus, using Assumption 3.1 and Hölder inequality, we can estimate the right-hand side of (3.16) to obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\int_0^T e^{-\rho t}u_t|G(X_t^{\bar{x}}) - G(X_t^x)|\mathrm{d}t\Big] \\
&\quad\le \mathbb{E}\Big[\sup_{t\le T}\left(e^{-\rho t}|G(X_t^{\bar{x}}) - G(X_t^x)|\right)\int_0^T u_t\mathrm{d}t\Big] \\
&\quad\le C\mathbb{E}\Big[\sup_{t\le T}\left(e^{-\rho t}\big(1+|X_t^{\bar{x}}|^{p-1}+|X_t^x|^{p-1}\big)\big|X_t^{\bar{x}}-X_t^x\big|\right)\Big] \\
&\quad\le C\bigg(\mathbb{E}\Big[\sup_{t\le T}\Big(e^{-\rho t}\big(1+|X_t^{\bar{x}}|^{2(p-1)}+|X_t^x|^{2(p-1)}\big)\Big)\Big]\bigg)^{\frac{1}{2}}\bigg(\mathbb{E}\Big[\sup_{t\le T}\left(e^{-\rho t}|X_t^{\bar{x}}-X_t^x|^2\right)\Big]\bigg)^{\frac{1}{2}}.
\end{aligned}
$$

Finally, using the estimates in (3.8) and (3.9), we get

$$
\mathbb{E}\Big[\int_0^T e^{-\rho t}u_t|G(X_t^{\bar{x}}) - G(X_t^x)|\mathrm{d}t\Big] \le C(1+|x|^{p-1}+|\bar{x}|^{p-1})|\bar{x}-x|,
$$

which allows to conclude thanks to (3.16):

$$\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}u_t(G(X_t^{\bar{x}})-G(X_t^x))\mathrm{d}t\bigg]\leq C(1+|x|^{p-1}+|\bar{x}|^{p-1})|\bar{x}-x|. \tag{3.17}$$

We next estimate the term in (3.15) which involves π. Since $Y_t^{y,u}\leq y\leq 1$, using Assumption 3.1 and Hölder inequality we find

$$\begin{aligned}
&\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}Y_t^{y,u}(\pi(X_t^{\bar{x}})-\pi(X_t^x))\mathrm{d}t\bigg]\\
&\quad\leq C\mathbb{E}\bigg[\int_0^T e^{-\rho t}\big(1+|X_t^{\bar{x}}|^{p-1}+|X_t^x|^{p-1}\big)\big|X_t^{\bar{x}}-X_t^x\big|\mathrm{d}t\bigg]\\
&\quad\leq C\int_0^T e^{-\rho t}\Big(1+\big(\mathbb{E}\big[|X_t^{\bar{x}}|^{2(p-1)}\big]\big)^{\frac12}+\big(\mathbb{E}\big[|X_t^{x}|^{2(p-1)}\big]\big)^{\frac12}\Big)\Big(\mathbb{E}\big[|X_t^{\bar{x}}-X_t^x|^2\big]\Big)^{\frac12}\mathrm{d}t.
\end{aligned}$$

Finally, using the estimates in (3.5) and (3.6), we get

$$\begin{aligned}
&\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}Y_t^{y,u}(\pi(X_t^{\bar{x}})-\pi(X_t^x))\mathrm{d}t\bigg]\\
&\quad\leq C(1+|x|^{p-1}+|\bar{x}|^{p-1})|\bar{x}-x|\int_0^T e^{\left(-\rho+\frac{\hat{c}_0(2(p-1))}{2}+\frac{\hat{c}_1(2)}{2}\right)t}\mathrm{d}t.
\end{aligned}$$

Hence, thanks to Condition 4 in Assumption 3.1, we conclude that

$$\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}Y_t^{y,u}(\pi(X_t^{\bar{x}})-\pi(X_t^x))\mathrm{d}t\bigg]\leq C(1+|x|^{p-1}+|\bar{x}|^{p-1})|\bar{x}-x|. \tag{3.18}$$

Plugging (3.17) and (3.18) into (3.15), we obtain

$$V^{\lambda,\varepsilon}(\bar{x},y)-V^{\lambda,\varepsilon}(x,y)-\zeta\leq C(1+|x|^{p-1}+|\bar{x}|^{p-1})|\bar{x}-x|,$$

which, by the arbitrariness of ζ, gives the local Lipschitz property of V^ε in x.

Step 3 (Uniform Lipschitz estimates for $V^{\lambda,\varepsilon}$ in y). Take $x\in\mathbb{R}^n$, $y,\bar{y}\in(0,1]$, $\zeta>0$ and $(\bar{u},\bar{\eta})\in\mathcal{U}_\varepsilon(\bar{y})$ such that $V^{\lambda,\varepsilon}(x,\bar{y})-\zeta\leq J^{\lambda,\varepsilon}(x,\bar{y};\bar{u},\bar{\eta})$. We will distinguish the two cases $\bar{y}\leq y$ and $\bar{y}>y$.

If $\bar{y}\leq y$, then we have $Y_t^{y,\bar{u}}\geq 0$, so that $(\bar{u},\bar{\eta})\in\mathcal{U}_\varepsilon(y)$. By suboptimality of $(\bar{u},\bar{\eta})$ for the initial condition (x,y), we have

$$\begin{aligned}
V^{\lambda,\varepsilon}&(x,\bar{y})-V^{\lambda,\varepsilon}(x,y)-\zeta\leq J^{\lambda,\varepsilon}(x,\bar{y};\bar{u},\bar{\eta})-J^{\lambda,\varepsilon}(x,y;\bar{u},\bar{\eta})\\
&=\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\big(\pi(X_t^x)(\bar{y}-y)-\lambda(Y_t^{\bar{y},\bar{u}}\log Y_t^{\bar{y},\bar{u}}-Y_t^{y,\bar{u}}\log Y_t^{y,\bar{u}})\big)\mathrm{d}t\bigg]\\
&=C(1+|x|^p)|\bar{y}-y|+\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\big(-\lambda(Y_t^{\bar{y},\bar{u}}\log Y_t^{\bar{y},\bar{u}}-Y_t^{y,\bar{u}}\log Y_t^{y,\bar{u}})\big)\mathrm{d}t\bigg],
\end{aligned} \tag{3.19}$$

where we have used Condition 4 in Assumption 3.1 together with (3.5). In order to continue the latter estimate, let $\hat{z}$ be the maximum point of the function $z\mapsto-\lambda z\log z$ (i.e., $\hat{z}=e^{-1}$), and notice that $z\mapsto-\lambda z\log z$ is increasing in $[0,\hat{z}]$ and Lipschitz in $[\frac{\hat{z}}{2},1]$, with Lipschitz constant smaller than $C_\lambda:=1+|\log(\hat{z}/2)|$. Suppose now that $y-\bar{y}\leq\frac{\hat{z}}{2}$. If $Y_t^{\bar{y},\bar{u}}\leq\frac{\hat{z}}{2}$, then $Y_t^{y,\bar{u}}\leq\hat{z}$, and since $Y_t^{\bar{y},\bar{u}}\leq Y_t^{y,\bar{u}}$, we obtain $-\lambda(Y_t^{\bar{y},\bar{u}}\log Y_t^{\bar{y},\bar{u}}-Y_t^{y,\bar{u}}\log Y_t^{y,\bar{u}})\leq 0$ by monotonicity. On the other hand, if $Y_t^{\bar{y},\bar{u}}\geq\frac{\hat{z}}{2}$, then $Y_t^{y,\bar{u}}\geq\frac{\hat{z}}{2}$, and by Lipschitzianity we find $-\lambda(Y_t^{\bar{y},\bar{u}}\log Y_t^{\bar{y},\bar{u}}-Y_t^{y,\bar{u}}\log Y_t^{y,\bar{u}})\leq C_\lambda(Y_t^{y,\bar{u}}-Y_t^{\bar{y},\bar{u}})=C_\lambda(y-\bar{y})$. Going back to (3.19), this argument implies that

$$0\leq y-\bar{y}\leq\frac{\hat{z}}{2}\implies V^{\lambda,\varepsilon}(x,\bar{y})-V^{\lambda,\varepsilon}(x,y)-\zeta\leq C(1+|x|^p+|\log(\hat{z}/2)|)|\bar{y}-y|,$$

which give the local Lipschitz property of V^ε in y in the case $\bar{y}\leq y$.

On the other hand, if $\bar{y} \geq y$, then the control $(\bar{u}, \bar{\eta})$ is not necessarily admissible for y, as $Y_t^{y,\bar{u}}$ could become smaller than 0. Define $u_t := \bar{u}_t \frac{y}{\bar{y}}$ and $\eta_t := \bar{\eta}_t \frac{y}{\bar{y}}$. Notice that, if $s\bar{u}_t - \bar{\eta}_t \leq 0$, then $su_t - \eta_t = \frac{y}{\bar{y}}(s\bar{u}_t - \bar{\eta}_t) \leq 0 \leq \frac{1}{\varepsilon}\beta(s)$, while $su_t - \eta_t = \frac{y}{\bar{y}}(s\bar{u}_t - \bar{\eta}_t) \leq s\bar{u}_t - \bar{\eta}_t \leq \frac{1}{\varepsilon}\beta(s)$ if $s\bar{u}_t - \bar{\eta}_t > 0$. Thus, the process (u, η) takes values in the set U_ε. Moreover, since

$$Y_t^{y,u} = y - \int_0^t u_s \mathrm{d}s = \frac{y}{\bar{y}}\Big(\bar{y} - \int_0^t \bar{u}_s \mathrm{d}s\Big) = \frac{y}{\bar{y}} Y^{\bar{y},\bar{u}} \geq 0, \tag{3.20}$$

it follows that $(u, \eta) \in \mathcal{U}_\varepsilon(y)$. Hence, by the suboptimality of (u, η) for the initial condition (x, y), using (3.20) we find

$$\begin{aligned}
&V^{\lambda,\varepsilon}(x, \bar{y}) - V^{\lambda,\varepsilon}(x, y) - \zeta \\
&\quad \leq J^{\lambda,\varepsilon}(x, \bar{y}; \bar{u}, \bar{\eta}) - J^{\lambda,\varepsilon}(x, y; u, \eta) \\
&\quad = \mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\Big(-\lambda \int_0^1 \Big(1 + \log\big(rY_t^{\bar{y},\bar{u}} + (1-r)Y_t^{y,u}\big)\Big)(Y_t^{\bar{y},\bar{u}} - Y_t^{y,\bar{u}})\mathrm{d}r \\
&\qquad\qquad + \pi(X_t^x)(\bar{y} - y)\frac{Y_t^{\bar{y},\bar{u}}}{\bar{y}} + G(X_t^x)\frac{\bar{y} - y}{\bar{y}}\bar{u}_t - \frac{\bar{y} - y}{\bar{y}}\bar{\eta}_t\Big)\mathrm{d}t\bigg] \\
&\quad \leq \mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\Big(-\lambda \int_0^1 \Big(1 + \log(r\bar{y} + (1-r)y) + \log\big(\frac{Y_t^{y,\bar{u}}}{\bar{y}}\big)\Big)\frac{Y_t^{y,\bar{u}}}{\bar{y}}(\bar{y} - y)\Big)\mathrm{d}r \\
&\qquad\qquad + |\pi(X_t^x)|(\bar{y} - y) + G(X_t^x)\frac{\bar{y} - y}{\bar{y}}\bar{u}_t\Big)\mathrm{d}t\bigg] \\
&\quad \leq (\bar{y} - y)\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\Big(\lambda\Big(1 + |\log y| - \log\big(\frac{Y_t^{y,\bar{u}}}{\bar{y}}\big)\frac{Y_t^{y,\bar{u}}}{\bar{y}}\Big) + |\pi(X_t^x)| + \frac{G(X_t^x)}{\bar{y}}\bar{u}_t\Big)\mathrm{d}t\bigg] \\
&\quad \leq C(\bar{y} - y)\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\Big(\lambda(1 + |\log y|) + |\pi(X_t^x)| + \frac{G(X_t^x)}{\bar{y}}\bar{u}_t\Big)\mathrm{d}t\bigg].
\end{aligned} \tag{3.21}$$

Similarly to the arguments that lead to (3.17), we have

$$\begin{aligned}
\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\bar{u}_t G(X_t^x)\mathrm{d}t\bigg] &\leq \lim_{T\to\infty} \mathbb{E}\bigg[\int_0^T e^{-\rho t}\bar{u}_t |G(X_t^x)|\mathrm{d}t\bigg] \\
&\leq \lim_{T\to\infty} \mathbb{E}\bigg[\sup_{t\leq T}\big(e^{-\rho t}|G(X_t^x)|\big)\int_0^T \bar{u}_t \mathrm{d}t\bigg] \\
&\leq \lim_{T\to\infty} C\mathbb{E}\bigg[\sup_{t\leq T}\big(e^{-\rho t}(1 + |X_t^x|^p)\big)\bigg] \\
&\leq C(1 + |x|^p),
\end{aligned}$$

where we have used the estimate in (3.8). Plugging the last inequality into (3.21), we obtain

$$\begin{aligned}
V^{\lambda,\varepsilon}(x, \bar{y}) - V^{\lambda,\varepsilon}(x, y) - \zeta &\leq C|\bar{y} - y|\bigg(\frac{1 + |x|^p}{\bar{y}} + |\log y| + \mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\big(1 + |X_t^x|^p\big)\mathrm{d}t\bigg]\bigg) \\
&\leq C|\bar{y} - y|\Big(\frac{1 + |x|^p}{\bar{y}} + |\log y|\Big)\int_0^\infty e^{\big(-\rho + \hat{c}_0(p)\big)t}\mathrm{d}t \\
&\leq C|\bar{y} - y|\Big(\frac{1 + |x|^p}{\bar{y}} + |\log y|\Big),
\end{aligned}$$

where we have used the condition $\rho > \hat{c}_0(p)$ given in Assumption 3.1. This shows the local Lipschitz property of V^ε in y for $\bar{y} \geq y$.

Step 4 (Uniform estimates for $D_x^2 V^{\lambda,\varepsilon}$). We first show the following semiconvexity estimate: there exists a constant C such that

$$V^{\lambda,\varepsilon}(x^\delta, y) - \delta V^{\lambda,\varepsilon}(\bar{x}, y) - (1-\delta) V^{\lambda,\varepsilon}(x, y) \leq C\delta(1-\delta)(1 + |x|^{p-1} + |\bar{x}|^{p-1})|\bar{x} - x|^2, \tag{3.22}$$

for each $\delta \in [0,1]$, $\bar{x}, x \in \mathbb{R}^n$, $y \in \mathbb{R}$ and $\varepsilon > 0$, and where $x^\delta := \delta\bar{x} + (1-\delta)x$.

Take an arbitrary $\zeta > 0$ and $(u,\eta) \in \mathcal{U}_\varepsilon(y)$ such that $V^{\lambda,\varepsilon}(x^\delta, y) - \zeta \leq J^{\lambda,\varepsilon}(x^\delta, y; u, \eta)$. Notice that such a control is admissible also for the initial conditions $(\bar{x}, y)$ and (x, y). Since (u,η) is not necessarily optimal for (x,y) or $(\bar{x}, \bar{y})$, we have

$$\begin{aligned} & V^{\lambda,\varepsilon}(x^\delta, y) - \delta V^{\lambda,\varepsilon}(\bar{x}, y) - (1-\delta) V^{\lambda,\varepsilon}(x, y) - \zeta \\ & \leq J^{\lambda,\varepsilon}(x^\delta, y; u, \eta) - \delta J^{\lambda,\varepsilon}(\bar{x}, y; u, \eta) - (1-\delta) J^{\lambda,\varepsilon}(x, y; u, \eta) \\ & \leq \mathbb{E}\bigg[\int_0^\infty e^{-\rho t}\Big(Y_t^{y,u}\big(\pi(X_t^\delta) - \delta\pi(X_t^{\bar{x}}) - (1-\delta)\pi(X_t^x)\big) \\ & \qquad + u_t\big(G(X_t^\delta) - \delta G(X_t^{\bar{x}}) - (1-\delta) G(X_t^x)\Big)\mathrm{d}t\bigg]. \end{aligned} \tag{3.23}$$

Next, set $\hat{X}_t := \delta X_t^{\bar{x}} + (1-\delta)X_t^x$, $X_t^\delta := X_t^{x^\delta}$ and define, for a generic function $\varphi : \mathbb{R}^n \to \mathbb{R}$, the transformations

$$\begin{aligned} \Delta[\varphi]_t &:= |\varphi(X_t^\delta) - \varphi(\hat{X}_t)|, \\ \Gamma[\varphi]_t &:= |\varphi(\hat{X}_t) - \delta\varphi(X_t^{\bar{x}}) - (1-\delta)\varphi(X_t^x)|. \end{aligned}$$

Then applying Fatou's lemma, we rewrite (3.23) as

$$\begin{aligned} & V^{\lambda,\varepsilon}(x^\delta, y) - \delta V^{\lambda,\varepsilon}(\bar{x}, y) - (1-\delta) V^{\lambda,\varepsilon}(x, y) - \zeta \\ & \leq \lim_{T\to\infty} \mathbb{E}\bigg[\int_0^T e^{-\rho t}\Big(Y_t^{y,u}\Delta[\pi]_t + u_t\Delta[G]_t + Y_t^{y,u}\Gamma[\pi]_t + u_t\Gamma[G]_t)\Big)\mathrm{d}t\bigg]. \end{aligned} \tag{3.24}$$

We next estimate each term on the right-hand side of (3.24) separately.

First, using Condition 2 in Assumption 3.1, Hölder inequality, and the SDE estimates (3.5) and (3.7), we find

$$\begin{aligned} & \mathbb{E}\bigg[\int_0^T e^{-\rho t} Y_t^{y,u}\Delta[\pi]_t \mathrm{d}t\bigg] \\ & \leq C\mathbb{E}\bigg[\int_0^T e^{-\rho t}\big(1 + |\hat{X}_t|^{p-1} + |X_t^\delta|^{p-1}\big)\big|\hat{X}_t - X_t^\delta\big|\mathrm{d}t\bigg] \\ & \leq C\bigg(\int_0^T e^{-\rho t}\Big(1 + \big(\mathbb{E}\big[|\hat{X}_t|^{2(p-1)}\big]\big)^{\frac{1}{2}} + \big(\mathbb{E}\big[|X_t^\delta|^{2(p-1)}\big]\big)^{\frac{1}{2}}\Big)\Big(\mathbb{E}\big[|\hat{X}_t - X_t^\delta|^2\big]\Big)^{\frac{1}{2}}\mathrm{d}t \\ & \leq C\delta(1-\delta)(1 + |x|^{p-1} + |\bar{x}|^{p-1})|\bar{x} - x|^2 \int_0^T e^{\left(-\rho + \frac{\hat{c}_0(2(p-1))}{2} + \frac{\hat{c}_1(2)}{2}\right)t}\mathrm{d}t \\ & \leq C\delta(1-\delta)(1 + |x|^{p-1} + |\bar{x}|^{p-1})|\bar{x} - x|^2, \end{aligned} \tag{3.25}$$

where the last inequality follows by Condition 4 in Assumption 3.1.

Secondly, using Condition 2 in Assumption 3.1, Hölder inequality, and the SDE estimates (3.5) and (3.7), we find

$$
\begin{aligned}
&\mathbb{E}\bigg[\int_0^T e^{-\rho t}u_t\Delta[G]_t\mathrm{d}t\bigg] \\
&\leq \mathbb{E}\bigg[\sup_{t\leq T}\big(e^{-\rho t}\Delta[G]_t\big)\int_0^T u_t\mathrm{d}t\bigg] \\
&\leq C\mathbb{E}\bigg[\sup_{t\leq T}\Big(e^{-\rho t}\big(1+|\hat{X}_t|^{p-1}+|X_t^\delta|^{p-1}\big)\big|\hat{X}_t-X_t^\delta\big|\Big)\bigg] \\
&\leq C\Big(\mathbb{E}\Big[\sup_{t\leq T}\Big(e^{-\rho t}\big(1+|\hat{X}_t|^{2(p-1)}+|X_t^\delta|^{2(p-1)}\big)\Big)\Big]\Big)^{\frac{1}{2}}\Big(\mathbb{E}\Big[\sup_{t\leq T}\big(e^{-\rho t}|\hat{X}_t-X_t^\delta|^2\big)\Big]\Big)^{\frac{1}{2}} \\
&\leq C\delta(1-\delta)(1+|x|^{p-1}+|\bar{x}|^{p-1})|\bar{x}-x|.
\end{aligned} \tag{3.26}
$$

where the last inequality follows from the estimates (3.8) and (3.10).

Thirdly, using Condition 2 in Assumption 3.1, Hölder inequality, and the SDE estimates (3.5) and (3.6), we find

$$
\begin{aligned}
&\mathbb{E}\bigg[\int_0^T e^{-\rho t}Y_t^{y,u}\Gamma[\pi]_t\mathrm{d}t\bigg] \\
&\leq C\delta(1-\delta)\mathbb{E}\bigg[\int_0^T e^{-\rho t}\Big(1+|X_t^{\bar{x}}|^{p-2}+|X_t^x|^{p-2}\Big)|X_t^{\bar{x}}-X_t^x|^2\mathrm{d}t\bigg] \\
&\leq C\delta(1-\delta)\int_0^T e^{-\rho t}\Big(1+\big(\mathbb{E}[|X_t^{\bar{x}}|^{2(p-2)}]\big)^{\frac{1}{2}}+\big(\mathbb{E}[|X_t^x|^{2(p-2)}]\big)^{\frac{1}{2}}\Big) \\
&\qquad\qquad\times\Big(\mathbb{E}[|X_t^{\bar{x}}-X_t^x|^4]\Big)^{\frac{1}{2}}dt \\
&\leq C\delta(1-\delta)(1+|x|^{p-2}+|\bar{x}|^{p-2})|\bar{x}-x|^2\int_0^T e^{\left(-\rho+\frac{\hat{c}_0(2(p-2))}{2}+\frac{\hat{c}_1(4)}{2}\right)t}\mathrm{d}t \\
&\leq C\delta(1-\delta)(1+|x|^{p-2}+|\bar{x}|^{p-2})|\bar{x}-x|^2,
\end{aligned} \tag{3.27}
$$

where the finiteness of the integrals follows again from Condition 4 in Assumption 3.1.

Fourthly, using Condition 2 in Assumption 3.1, Hölder inequality, and the SDE estimates (3.5) and (3.6), we find

$$
\begin{aligned}
&\mathbb{E}\bigg[\int_0^T e^{-\rho t}u_t\Gamma[G]_t\mathrm{d}t\bigg] \\
&\leq \mathbb{E}\bigg[\sup_{t\leq T}\big(e^{-\rho t}\Gamma[G]_t\big)\int_0^T u_t\mathrm{d}t\bigg] \\
&\leq C\delta(1-\delta)\mathbb{E}\bigg[\sup_{t\leq T}\Big(e^{-\rho t}\big(1+|X_t^{\bar{x}}|^{p-2}+|X_t^x|^{p-2}\big)|X_t^{\bar{x}}-X_t^x|^2\Big)\bigg] \\
&\leq C\delta(1-\delta)\Big(\mathbb{E}\Big[\sup_{t\leq T}\Big(e^{-\rho t}\big(1+|X_t^{\bar{x}}|^{2(p-2)}+|X_t^x|^{2(p-2)}\big)\Big)\Big]\Big)^{\frac{1}{2}} \\
&\qquad\times\Big(\mathbb{E}\Big[\sup_{t\leq T}\Big(e^{-\rho t}|X_t^{\bar{x}}-X_t^x|^4\Big)\Big]\Big)^{\frac{1}{2}} \\
&\leq C\delta(1-\delta)(1+|x|^{p-2}+|\bar{x}|^{p-2})|\bar{x}-x|^2\int_0^T e^{\left(-\rho+\frac{\hat{c}_0(2(p-2))}{2}+\frac{\hat{c}_1(4)}{2}\right)t}\mathrm{d}t \\
&\leq C\delta(1-\delta)(1+|x|^{p-2}+|\bar{x}|^{p-2})|\bar{x}-x|^2,
\end{aligned} \tag{3.28}
$$

where we have used Condition 4 in Assumption 3.1.

Plugging (3.25), (3.26), (3.27) and (3.28) back into (3.24), we obtain (3.22).

We will now use (3.22) in the PDE (3.14) in order to obtain an estimate for the derivative $D_x^2 V^{\lambda,\varepsilon}$. Notice that (3.22) implies that for any $r>0$ and $B_r(0) := \{x \in \mathbb{R}^n \mid |x| < r\}$ on can find a constant $C_r > 0$ such that

$$\partial_{zz}^2 V^{\lambda,\varepsilon}(x) \geq -C_r, \quad \text{for any } x \in B_r(0),\ z \in \mathbb{R}^n \text{ with } |z| = 1. \tag{3.29}$$

Choose an orthonormal basis $z^1, ..., z^n$ of $\mathbb{R}^n$ such that

$$\mathrm{tr}(\sigma\sigma^*(x) D_x^2 V^{\lambda,\varepsilon}(x,y)) = \sum_{i=1}^n c_\sigma^i(x) \partial_{z^i z^i}^2 V^{\lambda,\varepsilon}(x,y),$$

where $c_\sigma^1, ..., c_\sigma^n$ are the eigenvalues of the symmetric, nonnegative definite matrix $\sigma\sigma^*(x)$. Since V^ε solves the PDE in (3.14), using the latter identity together with (3.29) and the estimates in the previous steps, unless to take a larger constant C_r (independently from ε) we obtain

$$\begin{aligned}
-C_r \leq & \sum_{i=1}^n c_\sigma^i(x) \partial_{z^i z^i}^2 V^{\lambda,\varepsilon}(x,y) \\
=& \rho V^{\lambda,\varepsilon}(x,y) - b(x) D_x V^{\lambda,\varepsilon}(x,y) - \Pi(x,y) - \frac{1}{\varepsilon}\beta\left(G(x) - V_y^{\lambda,\varepsilon}(x,y)\right) \\
\leq & \rho V^{\lambda,\varepsilon}(x,y) - b(x) D_x V^{\lambda,\varepsilon}(x,y) - \Pi(x,y) \leq C_r, \quad \text{for any } (x,y) \in B_r(0) \times (0,1).
\end{aligned}$$

This in turn implies that

$$0 \leq \frac{1}{\varepsilon}\beta\left(G(x) - V_y^{\lambda,\varepsilon}(x,y)\right) \leq C_r, \quad \text{for any } (x,y) \in B_r(0) \times (0,1),$$

so that, by using the uniform ellipticity of σ (*cf.* Condition 3 in Assumption 3.1), by Theorem 9.11 p. 235 in [25] we conclude that, for any $q > 0$, unless to enlarge the constant C_r (independently from ε), for any $y \in (0,1)$ we have

$$\int_{B_{r/2}(0)} |D_x^2 V^{\lambda,\varepsilon}(x,y)|^q \mathrm{d}x \leq C_r\left(1 + \int_{B_{r/2}(0)} \left|\frac{1}{\varepsilon}\beta\left(G(x) - V_y^{\lambda,\varepsilon}(x,y)\right)\right|^q \mathrm{d}x\right) \leq C_r, \tag{3.30}$$

thus giving the desired estimate.

Step 5 (Lipschitzianity of V_y^λ via optimal stopping). Following the arguments in [35] (see in particular Theorem 3.4) we can show that, for any $(x,y) \in \mathbb{R}^n \times (0,1]$, one has the representation

$$V_y^\lambda(x,y) = \sup_{\tau \in \mathcal{T}} \mathbb{E}\left[\int_0^\tau e^{-\rho t}\Pi_y\left(X_t^x, y\right)\mathrm{d}t + e^{-\rho\tau}G\left(X_\tau^x\right)\right], \tag{3.31}$$

that we will now employ to prove the local Lipschitzianity of V_y^λ.

We first show the Lipschitzianity in y. Take $x \in \mathbb{R}^n$, $\bar{y}, y \in (0,1]$ with $\bar{y} \leq y$. By the monotonicity of Π_y in y we have $0 \leq V_y^\lambda(x,\bar{y}) - V_y^\lambda(x,y)$. Moreover, for $\zeta > 0$ and $\tau \in \mathcal{T}$ such that $V_y^\lambda(x,\bar{y}) - \zeta \leq \mathbb{E}\left[\int_0^\tau e^{-\rho t}\Pi_y(X_t^x, \bar{y})\mathrm{d}t + e^{-\rho\tau}G(X_\tau^x)\right]$, we find

$$\begin{aligned}
V_y^\lambda(x,\bar{y}) - V_y^\lambda(x,y) - \zeta &\leq \mathbb{E}\left[\int_0^\tau e^{-\rho t}\left(\Pi_y\left(X_t^x, \bar{y}\right) - \Pi_y\left(X_t^x, y\right)\right)\mathrm{d}t\right] \\
&= \lambda\mathbb{E}\left[\int_0^\tau e^{-\rho t}(\log y - \log \bar{y})\mathrm{d}t\right] \\
&\leqslant \lambda\mathbb{E}\left[\int_0^\tau e^{-\rho t}\mathrm{d}t\right]\frac{1}{\bar{y}}(y - \bar{y}) \leqslant \lambda C\frac{1}{\bar{y}}(y - \bar{y}).
\end{aligned}$$

From the arbitrariness of ζ, we conclude that $0 \leqslant V_y^\lambda(x,\bar{y}) - V_y^\lambda(x,y) \leqslant C\frac{1}{\bar{y}}(y-\bar{y})$, showing the Lipschitzianity in y.

We next show the Lipschitzianity in x. For $\bar{x}, x \in \mathbb{R}^n$ and $y \in (0,1]$, take $\zeta > 0$ and $\tau \in \mathcal{T}$ such that $V_y^\lambda(\bar{x},y) - \zeta \le \mathbb{E}\left[\int_0^\tau e^{-\rho t}\Pi_y(X_t^{\bar{x}},y)\mathrm{d}t + e^{-\rho\tau}G(X_\tau^{\bar{x}})\right]$. We observe that

$$
\begin{aligned}
&V_y^\lambda(\bar{x},y) - V_y^\lambda(x,y) - \zeta \\
&\quad\le \mathbb{E}\bigg[\int_0^\tau e^{-\rho t}\left(\Pi_y(X_t^{\bar{x}},y) - \Pi_y(X_t^x,y)\right)\mathrm{d}t + e^{-\rho\tau}\left(G(X_\tau^{\bar{x}}) - G(X_\tau^x)\right)\bigg] \\
&\quad= \mathbb{E}\bigg[\int_0^\tau e^{-\rho t}\left(\pi(X_t^{\bar{x}}) - \pi(X_t^x)\right)\mathrm{d}t + e^{-\rho\tau}\left(G(X_\tau^{\bar{x}}) - G(X_\tau^x)\right)\bigg].
\end{aligned}
\tag{3.32}
$$

Thus, by repeating the arguments that lead to (3.17) and (3.18), we obtain

$$
V_y^\lambda(\bar{x},y) - V_y^\lambda(x,y) - \zeta \le C(1+|x|^{p-1}+|\bar{x}|^{p-1})|\bar{x}-x|.
$$

Hence, the local Lipschitzianity in x follows by the arbitrariness of ζ and exchanging the role of x and $\bar{x}$.

Step 6 (Regularity of V^λ and variational inequality). We begin by noticing that the continuity of V^λ in $y=0$ is shown as in (3.13).

By the estimates in Steps 2 and 3, for any $r, \delta > 0$ and any ball $B_r(0) := \{x \in \mathbb{R}^n |\ |x| < r\}$, we can find a subsequence $(\varepsilon_n)_n$ with $\varepsilon_n \to 0$ as $n \to \infty$ such that

$$
\begin{aligned}
&V^{\lambda,\varepsilon_n} \to V^\lambda \text{ as } n \to \infty, \text{ uniformly in } B_r(0)\times(\delta,1), \\
&(D_x V^{\lambda,\varepsilon_n}, V_y^{\lambda,\varepsilon_n}) \to (D_x V^\lambda, V_y^\lambda) \text{ as } n \to \infty, \text{ weakly in } \mathbb{L}^2(B_r(0)\times(\delta,1)).
\end{aligned}
\tag{3.33}
$$

Moreover, thanks to the estimate in Step 4, for any fixed $y \in (0,1]$ we can select a further subsequence $(\varepsilon_n^y)_n$ such that

$$
\begin{aligned}
&D_x V^{\lambda,\varepsilon_n^y}(\cdot,y) \to D_x V^\lambda(\cdot,y) \quad \text{as } n \to \infty, \text{ uniformly in } B_r(0), \\
&D_x^2 V^{\lambda,\varepsilon_n^y}(\cdot,y) \to D_x^2 V^\lambda(\cdot,y) \quad \text{as } n \to \infty, \text{ weakly in } \mathbb{L}^2(B_r(0)).
\end{aligned}
\tag{3.34}
$$

Therefore, taking limits in (3.14) we obtain that

$$
\text{for any } y \in (0,1) \quad (\rho - \mathcal{L}_x)V^\lambda(\cdot,y) - \Pi(\cdot,y) \ge 0 \quad \text{and} \quad V_y^\lambda(\cdot,y) - G \geqslant 0, \quad \mathrm{d}x\text{-a.e.}
$$

which shows the HJB equation with inequality; that is,

$$
\max\left\{ (\mathcal{L}_x - \rho)\, V^\lambda(x,y) + \pi(x)y - \lambda y\log y, -V_y^\lambda(x,y) + G(x)\right\} \le 0.
$$

To complete the proof, if $(\bar{x},\bar{y})$ is such that $V_y^\lambda(\bar{x},\bar{y}) - G(\bar{x}) > 0$, by continuity of V_y^λ (*cf.* Step 5) we can find $r > 0$ such that $V_y^\lambda(x,\bar{y}) - G(x) > 0$ for any $x \in B_r(0)$. Thus, for a.a. $x \in B_r(0)$ there exists $\bar{n}_x$ such that $V_y^{\lambda,\varepsilon_n^{\bar{y}}}(x,\bar{y}) - G(x) > 0$ for any $n \ge \bar{n}_x$, which gives $(\rho - \mathcal{L}_x)V^{\lambda,\varepsilon_n^{\bar{y}}}(x,\bar{y}) - \Pi(x,\bar{y}) = 0$ for any $n \ge \bar{n}_x$ so that

$$
\lim_n (\rho - \mathcal{L}_x)V^{\lambda,\varepsilon_n^{\bar{y}}}(\cdot,\bar{y}) - \Pi(\cdot,\bar{y}) = 0, \quad \text{a.e. in } x \in B_r(0).
$$

Since $(\rho - \mathcal{L}_x)V^\lambda(\cdot,\bar{y}) - \Pi(\cdot,\bar{y})$ is the weak limit in $\mathbb{L}^2(B_r(\bar{x}))$ of $\left((\rho-\mathcal{L}_x)V^{\lambda,\varepsilon_n^{\bar{y}}}(\cdot,\bar{y}) - \Pi(\cdot,\bar{y})\right)_n$, the latter limits in turn implies that $(\rho - \mathcal{L}_x)V^\lambda(\cdot,\bar{y}) - \Pi(\cdot,\bar{y})$ a.e. in $B_r(\bar{x})$, thus showing the HJB with equality and completing the proof.

3.3. **Proof of Theorem 3.2: uniqueness for the HJB.** Uniqueness of the solution follows from a verification argument, that we prove in two steps.

Step 1. Let $\tilde{V}$ be a solution of the HJB equation as in the statement of the theorem. Fix $(x,y) \in \mathbb{R}^n \times [0,1]$ and consider a generic $\xi \in \mathcal{A}(y)$. For $R > 0$, set $\tau_R := \inf\{t \geq 0 \,|\, |X_t^x| \geq R\}$. Thanks to the regularity of $\tilde{V}$, we can employ (a generalized) Itô's formula to the process $\big(e^{-\rho t}\tilde{V}(X_t^x, Y_t^{y,\xi})\big)_t$ (see [23, Chapter 8, Section VIII.4, Theorem 4.1] and [11, Theorem 4.2]), up to the stopping time $\tau_R \wedge T$, for some $T > 0$, to obtain

(3.35)
$$\begin{aligned}\mathbb{E}\Big[e^{-\rho\tau_R\wedge T}\tilde{V}(X_{\tau_R\wedge T}, Y^{y,\xi}_{\tau_R\wedge T})\Big] =& \tilde{V}(x,y) + \mathbb{E}\bigg[\int_0^{\tau_R\wedge T} e^{-\rho t}(\mathcal{L}_x - \rho)\tilde{V}(X_t^x, Y_{t-}^{y,\xi})\mathrm{d}t\bigg] \\ &+ \mathbb{E}\bigg[\int_0^{\tau_R\wedge T} e^{-\rho t}\, \tilde{V}_y(X_t^x, Y_{t-}^{y,\xi})\mathrm{d}Y_t^{y,\lambda}\bigg] \\ &+ \mathbb{E}\bigg[\sum_{0\leq t\leq \tau_R\wedge T}\Big(\Delta\tilde{V}(X_t^x, Y_t^{y,\xi}) - \tilde{V}_y(X_t^x, Y_{t-}^{y,\xi})\Delta Y_t^{y,\xi}\Big)\bigg],\end{aligned}$$

with the notations $\Delta\tilde{V}(X_t, Y_t^{y,\xi}) = \tilde{V}(X_t, Y_t^{y,\xi}) - \tilde{V}(X_t, Y_{t-}^{y,\xi})$ and $\Delta Y_t^{y,\xi} = Y_t^{y,\xi} - Y_{t-}^{y,\xi}$. The concavity of $\tilde{V}$ in y now implies that

$$\sum_{0\leq t\leq \tau_R\wedge T}\Big(\Delta\tilde{V}(X_t^x, Y_t^{y,\xi}) - \tilde{V}_y(X_t^x, Y_{t-}^{y,\xi})\Delta Y_t^{y,\xi}\Big) \leq 0.$$

Moreover, since $\tilde{V}$ solves the HJB equation, we have

$$(\mathcal{L}_x - \rho)\tilde{V}(X_t^x, Y_{t-}^{y,\xi}) \leq -\Pi(X_t^x, Y_{t-}^{y,\xi}) \quad \text{and} \quad -\tilde{V}_y(X_t^x, Y_{t-}^{y,\xi}) \leq -G(X_t^x).$$

Plugging the latter two inequalities into (3.35) and then rearranging the terms, we obtain

$$\begin{aligned}\mathbb{E}\bigg[\int_0^{\tau_R\wedge T} e^{-\rho t}\Pi(X_t^x, Y_{t-}^{y,\xi})\mathrm{d}t + \int_0^{\tau_R\wedge T} e^{-\rho t}G(X_t^x)\mathrm{d}\xi_t\bigg] \\ \leq \tilde{V}(x,y) - \mathbb{E}\Big[e^{-\rho\tau_R\wedge T}\tilde{V}(X_{\tau_R\wedge T}, Y^{y,\xi}_{\tau_R\wedge T})\Big].\end{aligned}$$

Finally, thanks to the growth conditions of π and G, to the growth of $\tilde{V}$, to the estimates (3.5) and (3.8), we can use the dominated convergence theorem to take limits, first as $T \to \infty$ and then as $R \to \infty$, in order to obtain

$$J^\lambda(x,y;\xi) \leq \tilde{V}(x,y).$$

By arbitrariness of ξ, we deduce that $V^\lambda(x,y) \leq \tilde{V}(x,y)$.

Step 2. We adapt the arguments in [9] to construct a control ξ^λ which satisfies

$$J^\lambda(x,y;\xi^\lambda) = \tilde{V}(x,y).$$

This identity, together with the previous step, implies that $V^\lambda = \tilde{V}$ and that ξ^λ is optimal.

Define the function $\tilde{g}_\lambda : \mathbb{R}^n \to [0,1]$ by

$$\tilde{g}_\lambda(x) := \sup\big\{y \in [0,1] \,|\, -\tilde{V}_y(x,y) + G(x) < 0\big\},$$

where we set $\tilde{g}_\lambda(x) := 1$ if $\{y \in [0,1] \,|\, -\tilde{V}_y(x,y) + G(x) < 0\} = \emptyset$. The function $\tilde{g}_\lambda$ is well defined by concavity of $\tilde{V}$. Noticing that $\{(x,y) \in \mathbb{R}^n \times [0,1] \,|\, y < \tilde{g}_\lambda(x)\} = \{(x,y) \,|\, -\tilde{V}_y(x,y) + G(x) < 0\}$, by continuity of $\tilde{V}_y$ we deduce that the set $\{(x,y) \in \mathbb{R}^n \times [0,1] \,|\, y < \tilde{g}_\lambda(x)\}$ is open, so that $\tilde{g}_\lambda$ is lower semi-continuous.

Next, for fixed $(x,y) \in \mathbb{R}^n \times [0,1]$, define the process ξ^λ by setting

$$\xi_t^\lambda := \sup_{s\leq t}\big(y - \tilde{g}_\lambda(X_s^x)\big)^+, \quad t \geq 0.$$

Such a process is clearly monotone nondecreasing, so that it admits left limits. To show its right-continuity, use the upper semi-continuity of $x \mapsto (y - \tilde{g}_\lambda(x))^+$ to obtain

$$\begin{aligned}\lim_{s\downarrow t} \xi_s^\lambda &= \xi_t^\lambda \vee \lim_{s\downarrow t} \sup_{t\le r\le s} \big(y - \tilde{g}_\lambda(X_r^x)\big)^+ \\ &= \xi_t^\lambda \vee \limsup_{s\downarrow t} \big(y - \tilde{g}_\lambda(X_s^x)\big)^+ \le \xi_t^\lambda \vee \big(y - \tilde{g}_\lambda(X_t^x)\big)^+ = \xi_t^\lambda.\end{aligned}$$

By monotonicity of ξ^λ, this implies that $\lim_{s\downarrow t} \xi_s^\lambda = \xi_t^\lambda$. Thus, since this process is clearly adapted, it is progressively measurable, showing the admissibility of ξ^λ; i.e., $\xi^\lambda \in \mathcal{A}(y)$.

We will now prove the optimality of ξ^λ by repeating, with some modifications, the passages in Step 1. For $Y^{y,\lambda} := Y^{y,\xi^\lambda}$, $\tau_R := \inf\{t \ge 0 \,|\, |X_t^x| \ge R\}$ and for $T > 0$ we find

(3.36)
$$\begin{aligned}\mathbb{E}\Big[e^{-\rho\tau_R\wedge T}\tilde{V}(X_{\tau_R\wedge T}, Y^{y,\lambda}_{\tau_R\wedge T})\Big] =&\tilde{V}(x,y) + \mathbb{E}\bigg[\int_0^{\tau_R\wedge T} e^{-\rho t}(\mathcal{L}_x - \rho)\tilde{V}(X_t^x, Y_{t-}^{y,\lambda})\mathrm{d}t\bigg] \\ &+ \mathbb{E}\bigg[\int_0^{\tau_R\wedge T} e^{-\rho t}\,\tilde{V}_y(X_t^x, Y_{t-}^{y,\lambda})\mathrm{d}Y_t^{y,\lambda}\bigg] \\ &+ \mathbb{E}\bigg[\sum_{0\le t\le\tau_R\wedge T}\Big(\Delta\tilde{V}(X_t^x, Y_t^{y,\lambda}) - \tilde{V}_y(X_t^x, Y_{t-}^{y,\lambda})\Delta Y_t^{y,\lambda}\Big)\bigg],\end{aligned}$$

with the notations $\Delta\tilde{V}(X_t, Y_t^{y,\lambda}) = \tilde{V}(X_t, Y_t^{y,\lambda}) - \tilde{V}(X_t, Y_{t-}^{y,\lambda})$ and $\Delta Y_t^{y,\lambda} = Y_t^{y,\lambda} - Y_{t-}^{y,\lambda}$.

Notice that, by construction of ξ^λ, if the process $Y^{y,\lambda}$ jumps at time t, then we have $\tilde{g}_\lambda(X_t^x) < \tilde{g}_\lambda(X_{t-}^x)$ and

$$\tilde{V}_y(X_t^x, \zeta) = G(X_t^x), \quad \text{for any } \zeta \in [Y_t^{y,\lambda}, Y_{t-}^{y,\lambda}].$$

Therefore, we obtain

$$\Delta\tilde{V}(X_t^x, Y_t^{y,\lambda}) - \tilde{V}_y(X_t^x, Y_{t-}^{y,\lambda})\Delta Y_t^{y,\lambda} = G(X_t^x)(Y_t^{y,\lambda} - Y_{t-}^{y,\lambda}) - G(X_t^x)(Y_t^{y,\lambda} - Y_{t-}^{y,\lambda}) = 0.$$

Using the latter equality in (3.36), together with the fact that the process ξ^λ increases only at times t in which $\tilde{V}_y(X_t^x, Y_{t-}^{y,\lambda}) = G(X_t^x)$ and the fact that $(\mathcal{L}_x - \rho)\tilde{V}(X_t^x, Y_{t-}^{y,\lambda}) = \Pi(X_t^x, Y_{t-}^{y,\lambda})$ $\mathbb{P}\otimes \mathrm{d}t$-a.e., we find

$$\begin{aligned}\tilde{V}(x,y) =&\mathbb{E}\bigg[\int_0^{\tau_R\wedge T} e^{-\rho t}\Pi(X_t^x, Y_{t-}^{y,\lambda})\mathrm{d}t + \int_0^{\tau_R\wedge T} e^{-\rho t}G(X_t^x)\mathrm{d}\xi_t^\lambda\bigg] \\ &+ \mathbb{E}\Big[e^{-\rho\tau_R\wedge T}\tilde{V}(X_{\tau_R\wedge T}, Y^{y,\lambda}_{\tau_R\wedge T})\Big].\end{aligned}$$

Taking limits (first as $T \to \infty$ and then as $M \to \infty$) we conclude that

$$\tilde{V}(x,y) = J^\lambda(x,y;\xi^\lambda),$$

thus proving the optimality of ξ^λ. This completes the proof.

3.4. **Proof of Theorem 3.3.** In order to show Theorem 3.3, it is enough to repeat the arguments in Step 2 in Subsection 3.3 with the value function V^λ and the free boundary g_λ defined in (3.4), and using the fact that V^λ is a solution of the HJB equation (*cf.* Theorem 3.2).

4. A real option problem: analytical solutions and reinforcement learning algorithms

This section is devoted to a classical example of the optimal stopping problem arising from real option evaluation (see Chapter 5 in [47] and the references therein for further details). We provide analytical properties of the problem in Subsection 4.1 and develop the corresponding RL algorithms in Subsection 4.2.

We start with the introduction of the problem formulation. For parameters $\mu \in \mathbb{R}, \sigma, \kappa, \rho > 0$ and $\pi : \mathbb{R}_+ \to \mathbb{R}$, consider the optimal stopping problem

$$\sup_{\tau \in \mathcal{T}} \mathbb{E}\left[\int_0^\tau e^{-\rho s}\pi(X_s^x)\mathrm{d}s + \kappa e^{-\rho\tau}\right],$$
$$\text{subject to } \mathrm{d}X_t^x = \mu X_t^x \mathrm{d}t + \sigma X_t^x \mathrm{d}W_t, \quad X_0^x = x > 0.$$

In order to have well posedness of the problem, we assume $\rho > \mu$. Noticing that,

$$\mathbb{E}\left[\int_0^\tau e^{-\rho s}\pi(X_s^x)\mathrm{d}s + \kappa e^{-\rho\tau}\right] = \mathbb{E}\left[\int_0^\tau e^{-\rho s}\left(\pi(X_s^x) - \rho\kappa\right)\mathrm{d}s\right] + \kappa,$$

we define the OS problem

$$V(x) := \sup_{\tau \in \mathcal{T}} \mathbb{E}\left[\int_0^\tau e^{-\rho s}\left(\pi(X_s^x) - \rho\kappa\right)\mathrm{d}s\right]. \tag{4.1}$$

Despite the latter OS problem can be explicitly solved, we are interested here in illustrating the relations between the classical solution approach and the entropy regularization approach (see in particular the discussion in Subsection 4.1.3 below).

Following the strategy outlined in Section 3, introduce the extra state variable $y \in [0,1]$ and, for $\lambda > 0$, define the singular control problem

$$\begin{aligned} V^\lambda(x,y) &:= \sup_{\xi \in \mathcal{A}(y)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\big(\pi(X_t^x) - \rho\kappa)Y_t^{y,\xi} - \lambda Y_t^{y,\xi}\log(Y_t^{y,\xi})\big)\mathrm{d}t\right], \\ &\text{subject to } \mathrm{d}X_t^x = \mu X_t^x \mathrm{d}t + \sigma X_t^x \mathrm{d}W_t, \quad X_0^x = x, \\ &\qquad \mathrm{d}Y_t^{y,\xi} = -\,\mathrm{d}\xi_t, \quad Y_{0-}^{y,\xi} = y, \end{aligned} \tag{4.2}$$

with $\mathcal{A}(y)$ defined in (3.1). The HJB equation associated to V^λ reads as

$$\max\left\{(\mathcal{L}_x - \rho)V^\lambda + (\pi(x) - \rho\kappa)y - \lambda y \log y, -V_y^\lambda\right\} = 0, \quad \text{a.e. in } (0,\infty)\times(0,1),$$

with boundary condition $V(x,0) = 0$, $x \in (0,\infty)$, where $\mathcal{L}_x\phi(x) = \mu x\phi_x(x) + \frac{1}{2}\sigma^2 x^2\phi_{xx}(x)$.

The current problem differs from the general setting studied in Sections 2 and 3 due to the behavior of π at state 0. However, this model can be solved semi-explicitly with the –so called– *guess and verify* approach, which consists in guessing first a candidate value function and then proving that the candidate coincides with the true value function (see Theorem 4.2).

Assumption 4.1. *$\pi \in C^2(\mathbb{R}_+)$ with $\pi(0) = 0$ satisfies the following conditions:*

- *π is non-decreasing and concave, π has a sublinear growth, and $x\pi'(x)$ is strictly increasing. More specifically, there exists some $\theta \in (0,1)$ and $c > 0$ such that*

$$|\pi(x)| \le c(1 + |x|^\theta).$$

One example that satisfies Assumption 4.1 is $\pi(x) = c_0\, x^{\theta_0}$ with $\theta_0 \in (0,\theta]$ and $c_0 \in (0,c]$.

4.1. Analytical solution.

In this section, we analyze the solutions of the real option problem (4.2). Specifically, we construct a semi-explicit candidate solution in Subsection 4.1.1 and provide the verification theorem in Subsection 4.1.2.

4.1.1. *Construction of a candidate solution.* We will now search for a candidate solution $u : (0,\infty)\times[0,1]\to\mathbb{R}$ and a nondecreasing function $g_\lambda:(0,\infty)\to[0,1]$ such that

$$\begin{cases} (\mathcal{L}_x-\rho)\,u+(\pi(x)-\rho\kappa)y-\lambda y\log y=0, & 0<y\le g_\lambda(x),\\ -u_y=0, & y>g_\lambda(x), \end{cases} \tag{4.3}$$

and we will later verify that $u=V^\lambda$ (see Theorem 4.2 below).

If u and g_λ satisfy (4.3), for $0<y\le g_\lambda(x)$, we can integrate the equation to obtain

$$u(x,y)=A_2(y)x^{\alpha_-}+A_1(y)x^{\alpha_+}+\mathbb{E}\left[\int_0^\infty e^{-\rho t}\big((\pi(X_t^x)-\rho\kappa)y-\lambda y\log y\big)\mathrm{d}t\right], \tag{4.4}$$

with α_+,α_- solving $\frac{1}{2}\sigma^2\alpha(\alpha-1)+\mu\alpha-\rho=0$, and such that $\alpha_+>1$ (as $\rho>\mu$) and $\alpha_-<0$. Moreover, since $\rho>\mu$ and Assumption 4.1 hold, we obtain that

$$V^\lambda(x,y)\le C+C\mathbb{E}\bigg[\int_0^\infty e^{-\rho t}(X_t^x)^\theta dt\bigg]=C+Cx^\theta\mathbb{E}\bigg[\int_0^\infty e^{(-\rho+\theta(\mu+\frac{\sigma^2}{2}(\theta-1)))t}dt\bigg]\le C(1+x^\theta).$$

Thus, we take $A_1\equiv 0$, as otherwise u would explode as $x\uparrow\infty$, not matching the growth behavior of V^λ. Finally, when for $0<y\le g_\lambda(x)$, a direct integration leads to

$$u(x,y)=A_2(y)x^{\alpha_-}+H_\pi(x)y-\kappa y-\frac{\lambda}{\rho}y\log y,\quad\text{with}\quad H_\pi(x)=\mathbb{E}\left[\int_0^\infty e^{-\rho t}\pi(X_t^x)\mathrm{d}t\right].$$

To summerize, we have obtained the following candidate value function

$$u(x,y)=\begin{cases} A_2(y)x^{\alpha_-}+H_\pi(x)y-\kappa y-\frac{\lambda}{\rho}y\log y, & 0<y\le g_\lambda(x),\\ A_2\left(g_\lambda(x)\right)x^{\alpha_-}+\pi(x)\,y-\kappa g_\lambda(x)-\frac{\lambda}{\rho}g_\lambda(x)\log\left(g_\lambda(x)\right), & y>g_\lambda(x). \end{cases}$$

In order to determine A_2 and g_λ we impose

$$\lim_{z\uparrow g_\lambda(x)}u_y\left(x,z\right)=\lim_{z\uparrow g_\lambda(x)}u_{yx}\left(x,z\right)=0.$$

From these conditions, we derive the system

$$\begin{cases} A_2'(g_\lambda(x))\,x^{\alpha_-}+H_\pi(x)-\kappa-\frac{\lambda}{\rho}(1+\log g_\lambda(x))=0,\\ \alpha_-A_2'(g_\lambda(x))x^{\alpha_--1}+H_\pi'(x)=0, \end{cases}$$

which can be solved as

$$\begin{cases} g_\lambda(x)=\exp\left(\frac{-\frac{H_\pi'(x)}{\alpha_-}x+H_\pi(x)-\kappa-\frac{\lambda}{\rho}}{\frac{\lambda}{\rho}}\right)\\ A_2(y)=\int_{g_\lambda(0)}^y\frac{\kappa+\frac{\lambda}{\rho}\log(u)+\frac{\lambda}{\rho}-H_\pi(g_\lambda^{-1}(u))}{(g_\lambda^{-1}(u))^{\alpha_-}}\mathrm{d}u, \end{cases} \tag{4.5}$$

Notice that

$$g_\lambda'(x)=g_\lambda(x)\left(\frac{-1}{\alpha_-}(H_\pi'(x)x)'+H_\pi'(x)\right)>0, \tag{4.6}$$

as $(H_\pi'(x)x)'>0$ and $H_\pi'(x)\ge0$ according to Assumption 4.1. Hence g_λ is strictly increasing on $(0,\hat{x}_{g_\lambda}]$ with $\hat{x}_{g_\lambda}=\min\{x\in\mathbb{R}_+: g_\lambda(x)=1\}$.

Therefore, our construction of u and g_λ is valid only for $y\ge y^\lambda$ with

$$y^\lambda:=g_\lambda(0)=e^{-1-\frac{\kappa\rho}{\lambda}}\in(0,1). \tag{4.7}$$

For $y<y^\lambda$ we notice that the function

$$u_0(x,y):=H_\pi(x)y-\kappa y-\frac{\lambda}{\rho}y\log y,$$

is such that

$$\begin{cases} (\mathcal{L}_x-\rho)u_0+(x^\theta-\rho\kappa)y-\lambda y\log y=0, \\ -\partial_y u_0=-H_\pi(x)+\Big(\kappa+\frac{\lambda}{\rho}(\log y+1)\Big)<0, \quad \text{for } y<y^\lambda. \end{cases} \tag{4.8}$$

Thus, $u_0(x,y)$ solves the HJB and for $y<y^\lambda$ and we expect $\xi^\lambda\equiv 0$ to be optimal.

Summarizing the previous heuristic, set $F(x,y):=A_2(y)\,x^{\alpha_-}+H_\pi(x)y-\kappa y-\frac{\lambda}{\rho}y\log y$ and define the candidate value as

$$u(x,y):=\begin{cases} H_\pi(x)y-\kappa y-\frac{\lambda}{\rho}y\log y, & y<y^\lambda, \\ F(x,y), & y\le g_\lambda(x),\ y\ge y^\lambda, \\ F(x,b_\lambda^{-1}(x)), & y>g_\lambda(x),\ y\ge y^\lambda, \end{cases}$$

and we extend g_λ on $[0,\infty)$ as follows

$$g_\lambda(x)=\begin{cases} g_\lambda(x), & x\in[0,\hat{x}_{g_\lambda}] \\ 1, & x\ge \hat{x}_{g_\lambda}. \end{cases}$$

4.1.2. *Verification theorem.* The next result verifies that the function u constructed in the previous subsection is the value function.

Theorem 4.2. *The following claims hold true:*

(1) The candidate value function u coincides with the true value function V^λ;

(2) The reflection policy $\xi_t^\lambda:=\sup_{0\le s\le t}\big(y-g_\lambda(X_s^x)\big)^+$ is optimal for the initial condition (x,y), where g_λ is defined in (4.5).

Proof. The proof is divided in three steps.

Step 1. In this step we show the C^2-regularity of u. We split $(0,\infty)\times[0,1]$ into the three sets $\mathcal{E}_0:=\{(x,y)\,|\,y<y^\lambda\}$, $\mathcal{E}:=\{(x,y)\,|\,y\ge y^\lambda,\ y<g_\lambda(x)\}$ and $\mathcal{S}:=\{(x,y)\,|\,y\geqslant g_\lambda(x),\ y\geqslant y^\lambda\}$. Clearly, u is regular in the interior of $\mathcal{E}_0$, $\mathcal{E}$ and $\mathcal{S}$. In order to show the C^2-regularity across $\mathcal{E}_0$ and $\mathcal{E}$, it is sufficient to notice that $A_2(y^\lambda)=A_2'(y^\lambda)=A_2''(y^\lambda)=0$.

We next study the behavior across $\mathcal{S}$ and $\mathcal{E}$. To this end, recall the definition of F and notice that

$$\begin{cases} u(x,y)=F(x,y), & y<g_\lambda(x), \\ u(x,y)=F\left(x,g_\lambda(x)\right), & y\geqslant g_\lambda(x). \end{cases}$$

Clearly we have

$$\lim_{y\uparrow g_\lambda(x)}u(x,y)=F\left(x,g_\lambda(x)\right)=\lim_{y\downarrow g_\lambda(x)}u(x,y),$$

so that u is continuous across the boundary. Furthermore, after straightforward calculations we find

$$\begin{aligned} &u_x(x,y)=F_x(x,y), \quad y<b_\lambda(x), \\ &u_x(x,y)=F_x\left(x,g_\lambda(x)\right)+\underbrace{F_y\left(x,g_\lambda(x)\right)(g_\lambda)'(x)}_{=0}, \quad y>g_\lambda(y), \end{aligned}$$

so that

$$\lim_{y\uparrow g_\lambda(x)}u_x(x,y)=F_x\left(x,g_\lambda(x)\right)=\lim_{y\downarrow g_\lambda(x)}u_x(x,y),$$

which in turn implies the continuity of u_x. Finally,

$$\begin{aligned} &u_{xx}(x,y)=F_{xx}(x,y), \quad y<g_\lambda(x), \\ &u_{xx}(x,y)=F_{xx}\left(x,g_\lambda(x)\right)+\underbrace{F_{xy}\left(x,g_\lambda(x)\right)(g_\lambda)'(x)}_{=0}, \quad y>g_\lambda(x), \end{aligned}$$

from which we conclude that

$$\lim_{y\downarrow g_\lambda(x)} u_{xx}(x,y) = F_{xx}\left(x, g_\lambda(x)\right) = \lim_{y\uparrow g_\lambda(x)} u_{xx}(x,y), \tag{4.9}$$

which is the continuity of u_{xx}. For the derivatives involving y, by using the definitions of A_2 and g_λ in (4.5), a direct computation leads to

$$\begin{aligned}
\lim_{x\downarrow g_\lambda^{-1}(y)} u_y(x,y) &= F_y\left(g_\lambda^{-1}(y), y\right) = 0 = \partial_y F\left(x, g_\lambda(x)\right), \\
\lim_{x\downarrow g_\lambda^{-1}(y)} u_{yy}(x,y) &= F_{yy}\left(g_\lambda^{-1}(y), y\right) = 0 = \partial^2_{yy} F\left(x, g_\lambda(x)\right), \\
\lim_{x\downarrow g_\lambda^{-1}(y)} u_{yx}(x,y) &= F_{yx}\left(g_\lambda^{-1}(y), y\right) = 0 = \partial^2_{yx} F\left(x, g_\lambda(x)\right),
\end{aligned}$$

which gives us the desired regularity.

Step 2. In this step we show that u solves the HJB equation. Let $\mathcal{E}_0$, $\mathcal{E}$ and $\mathcal{S}$ be as in the previous step. First, on the set $\mathcal{E}_0$, u satisfies the HJB equation by (4.8). Secondly, on the set $\mathcal{E}$, by construction we have $(\mathcal{L}_x - \rho)u + (\pi(x) - \rho\kappa)y - \lambda y \log y = 0$, and we need to verify that $-u_y \le 0$. To this end, use the definition of A_2 and b_λ to compute the derivative

$$\begin{aligned}
-u_{yx}(x,y) &= -\alpha_- A_2'(y) x^{\alpha_- - 1} - H_\pi'(x) \\
&= -\alpha_- \frac{\kappa + \frac{\lambda}{\rho}\log(y) + \frac{\lambda}{\rho} - H_\pi(g_\lambda^{-1}(y))}{(g_\lambda^{-1}(y))^{\alpha_-}} x^{\alpha_- - 1} - H_\pi'(x) \\
&= -\alpha_- \frac{\left(\kappa + \frac{\lambda}{\rho}\log(y) + \frac{\lambda}{\rho} - H_\pi(g_\lambda^{-1}(y))\right)x^{\alpha_- - 1} + \frac{1}{\alpha_-} H_\pi'(x)(g_\lambda^{-1}(y))^{\alpha_-}}{(g_\lambda^{-1}(y))^{\alpha_-}} x^{\alpha_- - 1} \\
&= -\alpha_- \frac{\frac{H_\pi'(g_\lambda^{-1}(y))}{-\alpha_-} g_\lambda^{-1}(y) g_\lambda^{-1}(y)\left(x^{\alpha_- - 1} - (g_\lambda^{-1}(y))^{\alpha_- - 1}\right)}{(g_\lambda^{-1}(y))^{\alpha_-}} x^{\alpha_- - 1} \le 0
\end{aligned}$$

where the last inequality holds since $x^{\alpha_- - 1} - (g_\lambda^{-1}(y))^{\alpha_- - 1} < 0$, which is a direct consequence of $g_\lambda^{-1}(y) < x$. Thus, since $-u_y(x, g_\lambda(x)) = 0$, we have that $-u_y \le 0$ on $\mathcal{E}$.

We next study the behavior on $\mathcal{S}$. By construction we have $-u_y = 0$ in $\mathcal{S}$. Moreover, using the expression for u, u_x and u_{xx} computed in the previous step, for any x such that $y > g_\lambda^{-1}(x)$ we find

$$\begin{aligned}
&(\mathcal{L}_x - \rho)u + (\pi(x) - \rho\kappa)y - \lambda y \log y \\
&\quad = \frac{1}{2}\sigma^2 x^2 \partial_{xx}\left(F\left(x, b_\lambda^{-1}(x)\right)\right) + \mu x \partial_x\left(F\left(x, b_\lambda^{-1}(x)\right)\right) - \rho F\left(x, b_\lambda^{-1}(x)\right) \\
&\qquad + (\pi(x) - \rho\kappa)y - \lambda y \log y \\
&\quad = \frac{1}{2}\sigma^2 x^2 F_{xx}\left(x, b_\lambda^{-1}(x)\right) + \mu x F_x\left(x, b^{-1}(x)\right) - \rho F\left(x, b^{-1}(x)\right) \\
&\qquad + (\pi(x) - \rho\kappa)y - \lambda y \log y.
\end{aligned}$$

Hence, using the ordinary differential equation for $y < g_\lambda(x)$, in the limit as $y \uparrow g_\lambda(x)$, we proceed with

$$\begin{aligned}(\mathcal{L}_x - \rho)u &+ (\pi(x) - \rho\kappa)y - \lambda y \log y \\ &= -(\pi(x) - \rho\kappa)g_\lambda(x) + \lambda g_\lambda(x)\log(g_\lambda(x)) + (\pi(x) - \rho\kappa)y - \lambda y \log y \\ &= \int_{g_\lambda(x)}^{y} \left(\pi(x) - \rho\left(k + \frac{\lambda}{\rho}(1 + \log z)\right)\right) dz \\ &< (y - g_\lambda(x))\left(\pi(x) - \rho\left(k + \frac{\lambda}{\rho}(1 + \log(g_\lambda(x)))\right)\right) \\ &= \underbrace{(y - g_\lambda(x))}_{\geq 0}\left(\pi(x) - \rho\left(\frac{H'_\pi(x)}{-\alpha}x + H_\pi(x)\right)\right).\end{aligned}$$

Note that H_π satisfies the following PDE:

$$\mu x\, H'_\pi(x) - \rho H_\pi(x) + \pi(x) = -\frac{1}{2}\sigma^2 x^2 H''_\pi(x). \tag{4.10}$$

Therefore

$$\pi(x) - \rho\left(\frac{H'_\pi(x)}{-\alpha}x + H_\pi(x)\right) = -\left(\mu + \frac{\rho}{-\alpha_-}\right)x\, H'_\pi(x) - \frac{1}{2}\sigma^2 x^2 H''_\pi(x) \leq 0, \tag{4.11}$$

where the last inequality holds since $\frac{1}{2}\sigma^2\alpha_-(\alpha_- - 1) + \mu\alpha_- - \rho = 0$ and $\alpha_- < 0$. Hence $(\mathcal{L}_x - \rho)u + (\pi(x) - \rho\kappa)y - \lambda y \log y \leq 0$ on $\mathcal{S}$, so that u solves the HJB equation in the region $\mathcal{S}$.

Step 3. The last step consists of constructing a candidate optimal control ξ^λ as in the statement of the theorem and verifying its optimality by Itô's formula. The argument is analogous to that in Subsections 3.3 and 3.4, so that we omit the details to avoid repetitions. ☐

4.1.3. *Vanishing exploration and free boundary limits.* In the case of no exploration (i.e., $\lambda = 0$), the standard OS problem (4.1) can be solved semi-explicitly. In particular, the optimal stopping time is given by

$$\tau^* = \inf\{t \geq 0 : X_t^x \geq b^*\}$$

with b^* solves the equation: $-\frac{H'_\pi(x)}{\alpha_-}x + H_\pi(x) = \kappa$. As a special case of $\pi(x) = x^\theta$ with $\theta \in (0,1)$, we have $b^* = \left[-\frac{1}{P}\left(\frac{\alpha_-}{\theta - \alpha_-}\right)\kappa\right]^{1/\theta}$ with $P = \frac{1}{\rho + \frac{1}{2}\sigma^2\theta(1-\theta) - \theta\mu}$.

It is important to notice that, if $b_\lambda := g_\lambda^{-1} : [y^\lambda, 1] \to [0, \infty)$ denotes the free boundary of the entropy regularized OS problem with $\lambda > 0$ (determined in (4.5)) and with y^λ as in (4.7), we have the limits $\lim_{\lambda \downarrow 0} y^\lambda = 0$. Let $\lambda \downarrow 0$ in the following equation for any $y \in (0, 1]$

$$\frac{\lambda}{\rho}\log(y) + \frac{\lambda}{\rho} = \frac{H'_\pi(b_\lambda(y))}{-\alpha_-}b_\lambda(y) \;+\; H_\pi(b_\lambda(y)) - \kappa.$$

We observe that for a fixed $y \in (0, 1]$, $b_\lambda(y)$ is monotone increasing in λ, as we have $(H'_\pi(x)x)' > 0$ and $H'_\pi \geq 0$ according to Assumption 4.1. Hence $b_\lambda(y) \downarrow$ as $\lambda \downarrow 0$. Given that $b_\lambda(y) \geq 0$ for all $\lambda \geq 0$, there must exist a limit. By continuity argument, it is easy to show that the limit $\lim_{\lambda \downarrow 0} b_\lambda(y)$, which is independent of y, solves

$$0 = -\frac{H'_\pi(x)}{\alpha_-}x \;+\; H_\pi(x) - \kappa.$$

Hence we have $b^* \equiv \lim_{\lambda \downarrow 0} b_\lambda(y)$.

Thus, the boundary of the entropy regularized OS problem approximates the boundary of the (standard) OS problem (see also Propositions 2.9 and 2.10).

4.2. **Reinforcement learning algorithm.** In this section, we propose an RL framework to learn the reflection boundary g^λ and the corresponding value function V^λ, without prior knowledge on the form of g^λ. Our framework consists of two steps: a model-free step (see Subsection 4.2.1) and a model-based step (see Subsection 4.2.2). In the first step, where all model parameters are known, we design a Policy Iteration algorithm to learn g_λ. In the second step, where the model parameters are unknown, we combine the Policy Iteration with a Policy Evaluation procedure.

Note that the RL framework is currently formulated in terms of the real option example. We believe the results can be extended to a more general setting and in higher dimensions.

4.2.1. *Model-based analysis.* Motivated by the analysis in Subsection 4.1.1, here we focus on a subclass of control policies that can be fully characterized by some reflection boundary g that satisfies the following assumption.

Assumption 4.3. *Assume $g : \mathbb{R}_+ \to (0, 1]$ is a non-decreasing function such that*

$$g \in C^1([0, \hat{x}_g]), \quad \textit{where} \quad \hat{x}_g = \inf_{x \in \mathbb{R}_+} \{g(x) = 1\}.$$

For a function g that satisfies Assumption 4.3, we define the associated policy ξ^g of the following form:

$$\xi_t^g = \sup_{s \le t} \big(y - g(X_s^x)\big)^+, t \ge 0. \tag{4.12}$$

Define the associated value function as:

$$\begin{aligned} V_g^\lambda(x, y) &:= J^\lambda(x, y; \xi^g) \\ &= \mathbb{E}\left[\int_0^\infty e^{-\rho t}\Big(\Big(\pi(X_t^x) - \rho\kappa\Big)Y_t^{y,\xi^g} - \lambda Y_t^{y,\xi^g}\log(Y_t^{y,\xi^g})\Big)\mathrm{d}t\right] \end{aligned} \tag{4.13}$$

subject to $Y_t^{y,\xi^g} = y - \xi_t^g$ with ξ_t^g defined in (4.12).

The policy in (4.12) defines two areas: the exploration area and the stopping area, both associated with g:

$$\mathcal{E}(g) = \Big\{(x, y) \;\Big|\; y \le g(x)\Big\}, \tag{4.14}$$

$$\mathcal{S}(g) = \Big\{(x, y) \;\Big|\; y > g(x)\Big\}. \tag{4.15}$$

Then we have the following results for the value function V_g^λ associated with policy ξ^g.

Theorem 4.4. *Assume Assumptions 4.1-4.3 hold. Then the following HJB equation has a unique $C^1(\mathbb{R}_+ \times [0, 1]) \cap C^2\big(\overline{\mathcal{E}(g)}\big)$ solution:*

$$(\mathcal{L}_x - \rho)u + \big(\pi(x) - \rho\kappa\big)y - \lambda y \log y = 0 \quad \textit{on} \quad \mathcal{E}(g), \tag{4.16}$$

$$-u_y = 0, \quad \textit{on} \quad \mathcal{S}(g). \tag{4.17}$$

In addition, the solution satisfies $u_{xx} \in \mathbb{L}^\infty_{loc}(\mathbb{R}_+ \times [0, 1])$. Finally, we have the verification theorem holds, namely

$$V_g^\lambda(x, y) \equiv u(x, y). \tag{4.18}$$

Proof. Note thatf g^{-1} is well defined under Assumption 4.3 with $g^{-1}(x) > 0$ for all $x \in \mathbb{R}_+$. We seek for a semi-explicit solution to

$$(\mathcal{L}_x - \rho)u + \Big(\pi(x) - \rho\kappa\Big)y - \lambda y \log y = 0 \quad \text{on} \quad \mathcal{E}(g), \tag{4.19}$$

$$-u_y = 0, \quad \text{on} \quad \mathcal{S}(g). \tag{4.20}$$

Similar to the analysis in Section 4 and following the idea in [20, 21], take

$$u(x,y) = A(y)x^{\alpha_-} + H_\pi(x)\,y - \kappa y - \frac{\lambda}{\rho} y \log y, \tag{4.21}$$

with α_- defined in (4.4) and $H_\pi(x)$ defined as

$$H_\pi(x) := \mathbb{E}\Big[\int_0^\infty e^{-\rho t}\pi(X_t^x)\mathrm{d}t\Big]. \tag{4.22}$$

Under Assumption 4.1, we have $H_\pi(x) \in C^2(\mathbb{R}_+)$ with sublinear growth [42, Eqn (2.11)].

Then,

$$u_y(x,y) = A'(y)x^{\alpha_-} + H_\pi(x) - \kappa - \frac{\lambda}{\rho}\log y - \frac{\lambda}{\rho}. \tag{4.23}$$

Setting $u_y(g^{-1}(y), y) = 0$, we have

$$A'(y)(g^{-1}(y))^{\alpha_-} + H_\pi(g^{-1}(y)) - \kappa - \frac{\lambda}{\rho}\log y - \frac{\lambda}{\rho} = 0. \tag{4.24}$$

Hence, $A'(y) = \frac{\kappa + \frac{\lambda}{\rho}\log y + \frac{\lambda}{\rho} - H_\pi(g^{-1}(y))}{(g^{-1}(y))^{\alpha_-}}$ and

$$A(y) = \int_{g(0)}^{y} \frac{\kappa + \frac{\lambda}{\rho}\log(u) + \frac{\lambda}{\rho} - H_\pi(g^{-1}(u))}{(g^{-1}(u))^{\alpha_-}}\mathrm{d}u. \tag{4.25}$$

It is easy to check that $u(x,y)$ is C^1 in y on $\mathbb{R}_+\times[0,1]$ and that, under condition $g \in C^1([0,\hat{x}_g])$, we have $u(x,y) \in C^2(\overline{\mathcal{E}(g)})$.

Now, on the curve $y = g(x)$,

$$u_x^+(x,y) = A(g(x))\alpha_- x^{\alpha_- - 1} + R_\pi'(x)g(x). \tag{4.26}$$

On the other hand, we have $u(x,y) = u(x,g(x))$ for $(x,y) \in \mathcal{S}(g)$. Hence

$$\begin{aligned}
&\frac{u\big(x,g(x)\big) - u\big(x-\delta, g(x)\big)}{\delta} = \frac{u\big(x,g(x)\big) - u\Big(x-\delta, g(x-\delta)\Big)}{\delta}\\
= \;& \frac{1}{\delta}\Big(A(g(x))x^{\alpha_-} - A(g(x-\delta))((x-\delta))^{\alpha_-}\Big) + \frac{1}{\delta}\Big(H_\pi(x)g(x) - H_\pi(x-\delta)g(x-\delta)\Big)\\
&-\frac{\kappa}{\delta}\Big(g(x) - g(x-\delta)\Big) - \frac{\lambda}{\rho\delta}\Big(g(x)\log\big(g(x)\big) - g(x-\delta)\log\big(g(x-\delta)\big)\Big),
\end{aligned}$$

and

$$\begin{aligned}
u_x^-(x,y) &= \lim_{\delta\downarrow 0}\frac{u\big(x,g(x)\big) - u\big(x-\delta,g(x)\big)}{\delta}\\
&= A(g(x))\alpha_- x^{\alpha_- - 1} + R_\pi'(x)g(x) + u_y(x,g(x))\\
&= A(g(x))\alpha_- x^{\alpha_- - 1} + R_\pi'(x)g(x)g(x) = u_x^+(x,y),
\end{aligned}$$

where the last equality holds since $u_y(x,g(x)) = 0$. Hence $u \in C^1(\mathbb{R}_+ \times [0,1])$.

It should be noted that u_{xx} fails to be continuous across the boundary although it remains bounded on any compact subset of $\mathbb{R}_+ \times [0,1]$.

Recall Y_t^{y,ξ^g} is the process under control ξ^g, defined in (4.12). Fix the initial condition $(X_0, Y_0^{y,\xi^g}) = (x,y)$ and take $R > 0$. Set $\tau_R := \inf\{t \geq 0 : X_t \geq R\}$. Applying Ito's formula in the weak version to u (see [23, Chapter 8, Section VIII.4, Theorem 4.1] and [11, Theorem 4.2]), up to the stopping time $\tau_R \wedge T$, for some $T > 0$, to obtain

$$\mathbb{E}\Big[e^{-\rho(\tau_R\wedge T)}u(X^x_{\tau_R\wedge T}, Y^{y,\xi^g}_{\tau_R\wedge T})\Big] \quad = \quad u(x,y) + \mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\mathcal{L}u(X_t^x, Y_{t-}^{y,\xi^g}) - \rho V_g^\lambda(X_t^x, Y_{t-}^{y,\xi^g})\Big)\mathrm{d}t\Big]$$

$$+\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\,\partial_y u(X_t^x, Y_{t-}^{y,\xi^g})\mathrm{d}Y_t^{y,\xi^g}\Big]$$
$$+\mathbb{E}\Big[\sum_{0\le t\le\tau_R\wedge T}\Big(\Delta u(X_t^x, Y_t^{y,\xi^g}) - \partial_y u(X_t, Y_{t-}^{y,\xi^g})\Delta Y_t^{y,\xi^g}\Big)\Big],$$

with the notations $\Delta u(X_t^x, Y_t^{y,\xi^g}) = u(X_t^x, Y_t^{y,\xi^g}) - u(X_t^x, Y_{t-}^{y,\xi^g})$ and $\Delta Y_t^{y,\xi^g} = Y_t^{y,\xi^g} - Y_{t-}^{y,\xi^g}$.

Recall that for any admissible control ξ^g, it can be decomposed into the sum of its continuous part and its pure jump part, i.e., $\mathrm{d}\xi^g = \mathrm{d}(\xi^g)^{cont} + \Delta\xi^g$. Hence we have the decomposition $\mathrm{d}Y^{y,\xi^g} = \mathrm{d}(Y^{y,\xi^g})^{cont} + \Delta Y^{y,\xi^g}$, and therefore,

$$\begin{aligned}
\mathbb{E}\Big[e^{-\rho(\tau_R\wedge T)}u(X_s^x, Y_s^{y,\xi^g})\Big] &= u(x,y) + \mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\mathcal{L}_x\, u(X_t^x, Y_{t-}^{y,\xi^g}) - \rho V_g^\lambda(X_t^x, Y_{t-}^{y,\xi^g})\Big)\mathrm{d}t\Big]\\
&\quad +\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\,\partial_y u(X_t^x, Y_{t-}^{y,\xi^g})\mathrm{d}(Y^{y,\xi^g})_t^{cont}\Big]\\
&\quad +\mathbb{E}\Big[\sum_{0\le t\le\tau_R\wedge T}\Delta u(X_t^x, Y_t^{y,\xi^g}))\Big] \qquad (4.27)\\
&= u(x,y) - \mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\Big(\pi(X_t^x) - \rho\kappa\Big)Y_t^{y,\xi^g} - \lambda Y_t^{y,\xi^g}\log\Big(Y_t^{y,\xi^g}\Big)\Big)\mathrm{d}t\Big]. \qquad (4.28)
\end{aligned}$$

The last equation holds due to the following facts:

(i) It holds that
$$\begin{aligned}
&\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\mathcal{L}_x\, u(X_t^x, Y_{t-}^{y,\xi^g}) - \rho V_g^\lambda(X_t^x, Y_{t-}^{y,\xi^g})\Big)\mathrm{d}t\Big]\\
=\ & -\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\Big(\pi(X_t^x) - \rho\kappa\Big)Y_t^{y,\xi^g} - \lambda Y_t^{y,\xi^g}\log\Big(Y_t^{y,\xi^g}\Big)\Big)\mathrm{d}t\Big],
\end{aligned}$$
as pure jump could only possibly happen at time 0, ξ^g keeps (X_t^x, Y_t^{y,ξ^g}) within $\overline{\mathcal{E}(g)}$ for $t>0$ and the fact (4.19);

(ii) $\mathbb{E}\Big[\sum_{0\le t\le\tau_R\wedge T}\Delta u(X_t^x, Y_t^{y,\xi^g}))\Big] = 0$ as $-u_y = 0$ on $\overline{\mathcal{S}(g)}$ according to the facts that (4.20) holds and $u\in C^1(\mathbb{R}_+\times[0,1])$;

(iii) $\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\,\partial_y u(X_t^x, Y_{t-}^{y,\xi^g})\mathrm{d}(Y^{y,\xi^g})_t^{cont}\Big] = 0$ as the infinitesimal control happens on the boundary $\overline{\mathcal{S}(g)}\cap\overline{\mathcal{E}(g)}$ where $-u_y = 0$.

When taking limits as $R\to\infty$ we have $\tau_R\wedge T\to T$, $\mathbb{P}$-a.s. By standard properties of Brownian motion, it is easy to prove that the integral terms in the last expression on the right-hand side of (4.28) are uniformly bounded in $L^2(\Omega,\mathbb{P})$, hence uniformly integrable. Moreover, u in (4.21) has sublinear growth by straightforward calculation. Then we also take limits as $T\to\infty$ and it follows that

$$u(x,y) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t}\Big(\Big(\pi(X_t^x) - \rho\kappa\Big)Y_t^{y,\xi^g} - \lambda Y_t^{y,\xi^g}\log\Big(Y_t^{y,\xi^g}\Big)\Big)\mathrm{d}t\Big].$$

Therefore $u\equiv V_g^\lambda$ and hence the verification theorem holds. □

With the results in Theorem 4.4, we update the new boundary following:

$$\widetilde{g}(x) = \begin{cases} \max\Big\{y^* < g(x)\,\Big|\,\partial_{xy}V_g^\lambda(x,y^*) = 0\Big\} & \text{if } \partial_{xy}^- V_g^\lambda(x,g(x)) < 0, \text{ and} \\ \widetilde{g}(x) = g(x) & \text{o.w,} \end{cases} \qquad (4.29)$$

where we define the notation $\partial_{xy}^- f(x,y) := \lim_{h\to 0-}\frac{\partial_x f(x,y+h) - \partial_x f(x,y)}{h}$ as the left y-derivative of $\partial_x f$ (if exists).

We can show that the updating rule (4.29) always improves in terms of the value function.

Theorem 4.5 (Policy improvement)**.** *Assume Assumptions 4.1-4.3 hold and $\widetilde{g}$ is updated according to (4.29), it holds that*

$$V^\lambda_{\widetilde{g}}(x,y) \geq V^\lambda_g(x,y), \tag{4.30}$$

where $V^\lambda_g, V^\lambda_{\widetilde{g}}$ are value functions associated with policies $\xi^g, \xi^{\widetilde{g}}$, see (4.12).

Proof. First of all, $V^\lambda_g \in C^2\left(\overline{\mathcal{E}(g)}\right)$ according to Theorem 4.4. Therefore, by the design of the algorithm, $\widetilde{g}$ satisfies the conditions in Assumption 4.3. Hence Theorem 4.4 also applies to $\widetilde{g}$.

Denote $Y^{y,\xi^{\widetilde{g}}}_t$ as the process under control $\xi^{\widetilde{g}}$, with $\widetilde{g}$ defined in (4.29). Fix the initial condition (x,y) and take $R>0$. Set $\tau_R := \inf\{t \geq 0 : X^x_t \geq R\}$. Similar to the proof of Theorem 4.21, apply Ito's formula in the weak version to V^λ_g (see [23, Chapter 8, Section VIII.4, Theorem 4.1] and [11, Theorem 4.2]) up to the stopping time $\tau_R \wedge T$ for some $T>0$. We then obtain

$$\begin{aligned}
&\mathbb{E}\Big[e^{-\rho(\tau_R\wedge T)}V^\lambda_g(X^x_{\tau_R\wedge T}, Y^{y,\xi^{\widetilde{g}}}_{\tau_R\wedge T})\Big]\\
=\;& V^\lambda_g(x,y) + \mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\mathcal{L}_x V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-}) - \rho V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-})\Big)\mathrm{d}t\Big]\\
&+\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\,\partial_y V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-})\mathrm{d}Y^{\widetilde{g}}_t\Big]\\
&+\mathbb{E}\Big[\sum_{0\leq t\leq \tau_R\wedge T} e^{-\rho t}\Big(\Delta V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_t) - \partial_y V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-})\Delta Y^{y,\xi^{\widetilde{g}}}_t\Big)\Big]\\
=\;& V^\lambda_g(x,y) - \mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\Big(\pi(X^x_t)-\rho\kappa\Big)Y^{y,\xi^{\widetilde{g}}}_t - \lambda Y^{y,\xi^{\widetilde{g}}}_t \log\Big(Y^{y,\xi^{\widetilde{g}}}_t\Big)\Big)\mathrm{d}t\Big]\\
&+\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\,\partial_y V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-})\mathrm{d}(Y^{y,\xi^{\widetilde{g}}})^{\mathrm{cont}}_t\Big]\\
&+\mathbb{E}\Big[\sum_{0\leq t\leq \tau_R\wedge T} e^{-\rho t}\Big(\Delta V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-})\Big)\Big]\\
\geq\;& V^\lambda_g(x,y) - \mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\Big(\pi(X^x_t)-\rho\kappa\Big)Y^{y,\xi^{\widetilde{g}}}_t - \lambda Y^{y,\xi^{\widetilde{g}}}_t \log\Big(Y^{y,\xi^{\widetilde{g}}}_t\Big)\Big)\mathrm{d}t\Big],
\end{aligned}$$

with the notations $\Delta V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_t) = V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_t) - V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-})$ and $\Delta Y^{y,\xi^{\widetilde{g}}}_t = Y^{y,\xi^{\widetilde{g}}}_t - Y^{y,\xi^{\widetilde{g}}}_{t-}$. The second equation holds because

$$\begin{aligned}
&\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\mathcal{L}_x\, u(X^x_t, Y^{y,\xi^g}_{t-}) - \rho\, V^\lambda_g(X^x_t, Y^{y,\xi^g}_{t-})\Big)\mathrm{d}t\Big]\\
=\;& -\mathbb{E}\Big[\int_0^{\tau_R\wedge T} e^{-\rho t}\Big(\Big(\pi(X^x_t)-\rho\kappa\Big)Y^{y,\xi^g}_t - \lambda Y^{y,\xi^g}_t \log\Big(Y^{y,\xi^g}_t\Big)\Big)\mathrm{d}t\Big]
\end{aligned}$$

as pure jump could only possibly happen at time 0, $\xi^{\widetilde{g}}$ keeps $(X^x_t, Y^{y,\xi^{\widetilde{g}}}_t)$ within $\overline{\mathcal{E}(g)}$ for $t>0$ and the fact (4.19). The last inequality holds by the design of (4.29). In particular, for $(x,y)\in\mathcal{S}(\widetilde{g})\cap\mathcal{E}(g)$,

$$\partial_y V^\lambda_g(x,y) = \partial_y V^\lambda_g(g^{-1}(y),y) + \int_{g^{-1}(y)}^x \partial_{xy}V^\lambda_g(u,y)\mathrm{d}u \leq 0,$$

and

$$\Delta V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_t) \;=\; V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_t) - V^\lambda_g(X^x_t, Y^{y,\xi^{\widetilde{g}}}_{t-})$$

$$= \int_0^{\Delta Y_t^{y,\xi^{\widetilde{g}}}} \partial_y V(X_t^x, Y_{t-}^{y,\xi^{\widetilde{g}}} + u)\mathrm{d}u \geq 0,$$

as $\Delta Y_t^{y,\xi^{\widetilde{g}}} \leq 0$.

When taking limits as $R \to \infty$ we have $\tau_R \wedge T \to T$, $\mathbb{P}$-a.s. By standard properties of Brownian motion it is easy to prove that the integral terms in the last expression on the right-hand side of (4.28) are uniformly bounded in $\mathbb{L}^2(\Omega, \mathbb{P})$, hence uniformly integrable. Moreover, V_g^λ (taking the form as in (4.21)) has sublinear growth by straightforward calculation. Then we also take limits as $T \to \infty$ and it follows that

$$V_{\widetilde{g}}^\lambda(x,y) \geq V_g^\lambda(x,y).$$

□

Repeat the updating procedure (4.29) iteratively, we have the general heuristic described in Algorithm 1. The idea is that we start with an initial $g_0(x)$ that has a sufficiently large exploration region $\mathcal{E}(g_0)$ (*i.e.*, Assumption 4.6). Then in each iteration k, we update g_k function "downwards" into $\mathcal{E}(g_k)$ according to the Hessian information (*cf.* line 3 of Algorithm 1), in which u_k has a better regularity (*cf.* line 3 of Algorithm 1). See Figure 4.2.1 for a demonstration.

Algorithm 1 Policy Iteration for Optimal Stopping (PI-OS)

1: Initialize $g_0(x)$ for $x \in [0, \infty)$ according to Assumption 4.6.
2: **for** $k = 0, 1, \cdots, K-1$ **do**
3: Find $u_k(x,y)$ a $C^1(\mathbb{R}_+ \times [0,1]) \cap C^2\left(\overline{\mathcal{E}(g_k)}\right)$ solution to the following equations:

$$(\mathcal{L}_x - \rho)u + \Big(\pi(x) - \rho\kappa\Big)y - \lambda y \log y = 0 \quad \text{on} \quad \mathcal{E}(g_k), \tag{4.31}$$

$$-u_y = 0, \quad \text{on} \quad \mathcal{S}(g_k). \tag{4.32}$$

4: Update the strategy

$$g_{k+1}(x) = \begin{cases} \max\left\{y^* < g_k(x) \,\middle|\, \partial_{xy} u_k(x, y^*) = 0\right\} & \text{if } \partial_{xy}^- u_k(x, g_k(x)) < 0, \\ g_{k+1}(x) = g_k(x) & \text{o.w.} \end{cases}$$

5: **end for**

This iterative scheme is inspired by [38] for a one-dimensional singular control problem. However, the convergence result in that work requires a second-order condition to hold throughout the entire iteration process, which is difficult to verify.

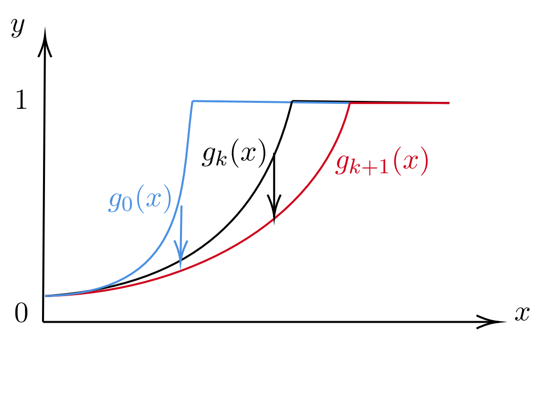

FIGURE 1. Demonstration of the Policy Iteration Algorithm.

In addition to the policy improvement result in Proposition 4.5, we provide the following policy convergence result under additional conditions on the initialization.

Assumption 4.6 (Initial policy)**.** *Assume the initial policy g_0 satisfies the following conditions:*

- **(a)** $g_0 \in C^1([0, \hat{x}_{g_0}])$ *is strictly increasing on* $[0, \hat{x}_{g_0}]$,
- **(b)** $g_0(0) = e^{-(1+\frac{\kappa\rho}{\lambda})}$, *and*
- **(c)** $-\alpha_-\Big(\kappa + \frac{\lambda}{\rho}\log(g_0(x)) + \frac{\lambda}{\rho}\Big) + \alpha_- H_\pi(x) - H_\pi'(x)\cdot x \geq 0$ *on* $[0, \hat{x}_{g_0}]$.

Note that **(c)** implies that $\hat{x}_{g_0} \leq \hat{x}_{g_\lambda}$.

Remark 4.7 (Justification of Assumption 4.6)**.** *Assumption 4.6-(a) is easy to satisfy and Assumption 4.6-(c) requires $\mathcal{E}(g_0)$ to be sufficiently large so that $\mathcal{E}(g_0) \supseteq \mathcal{E}(g_\lambda)$. Assumption 4.6-(b) seems to be the most restrictive one, as it requires the knowledge of $g_\lambda(0) = \exp(-(1+\frac{\kappa\rho}{\lambda}))$. In the model-free setting (such as in Section 4.2.2) and when $g_\lambda(0)$ is not available, the following algorithm can be applied to learn $g_\lambda(0)$ at a fast convergence rate.*

Algorithm 2 *Learning initial value $g_0(0)$*

1: *Initialize $y_0 \in (0,1)$, $c_0 > 0$ and $\{\eta_i\}_{i\geq 1}$.*
2: ***for*** $i = 1, \cdots,$ ***do***
3: *Acquire the value functions $u_i^+ = J^\lambda(0, y_i + \varepsilon_i; \xi = 0)$ and $u_i^- = J^\lambda(0, y_i - \varepsilon_i; \xi = 0)$, with $\varepsilon_i = \min\{y_i, 1-y_i, c_0/i\}$.*
4: *Update y_i using the following zeroth order gradient descent with a two-point estimator:*

$$y_{i+1} = y_i - \eta_i \frac{u_i^+ - u_i^-}{2\varepsilon_i}. \tag{4.33}$$

5: ***end for***

We know that $J^\lambda(0, y; \xi = 0) = -\kappa y - \frac{\lambda y}{\rho}\log y$ is convex in y with the minimizer taking value $y^* := e^{-(1+\frac{\kappa\rho}{\lambda})}$. When the functional form (hence the minimizer) is unknown to the decision-maker, Algorithm 2 converges linearly to the minimizer [18] when taking $\eta_i = \frac{\eta}{\sqrt{i}}$ with some $\eta > 0$. Mathematically,

$$|y_i - y^*|^2 \sim \mathcal{O}(\delta^i), \tag{4.34}$$

for some $\delta \in (0,1)$ depending on model parameters.

Remark 4.8 (Examples that satisfy Assumption 4.6)**.** *Assumption 4.6 is easy to satisfy and we provide two examples under the special case $\pi(x) = x^\theta$ with $0 < \theta < 1$: one is linear initialization and the other one is exponential initialization. For linear initialization, we can take*

$$g_0(x) = \min\left\{ e^{-(1+\frac{\kappa\rho}{\lambda})} + 2(1 - e^{-(1+\frac{\kappa\rho}{\lambda})})\Big(\frac{-\alpha_-\big(\kappa + \frac{\lambda}{\rho}\big)}{C(\theta - \alpha_-)}\Big)^\theta x, 1 \right\}, \tag{4.35}$$

which interpolates (x_0, y_0) and (x_1, y_1) with $x_0 = 0$, $y_0 = e^{-(1+\frac{\kappa\rho}{\lambda})}$, $x_1 = \frac{1}{2}\Big(\frac{-\alpha_-\big(\kappa+\frac{\lambda}{\rho}\big)}{C(\theta-\alpha_-)}\Big)^{-\theta}$ and $y_1 = 1$.

For exponential initialization, we can take

$$g_0(x) = \min\left\{ \exp\Big(\frac{\rho C(\zeta - \alpha_-)(x)^\zeta}{-\alpha_-\lambda} - \frac{\rho}{\lambda}(\kappa + \frac{\lambda}{\rho})\Big), 1 \right\}, \tag{4.36}$$

for some $\zeta \in (\theta, 1)$*. This is inspired by the form of* (4.5) *that under* $\pi(x) = x^\theta$

$$g_\lambda(x) = \min\left\{\exp\Big(\frac{\rho C(\theta-\alpha_-)(x)^\theta}{-\alpha_-\lambda} - \frac{\rho}{\lambda}(\kappa + \frac{\lambda}{\rho})\Big), 1\right\}.$$

Theorem 4.9 (Policy convergence)**.** *Assume Assumptions 4.1 and 4.6 hold. Then we have the following results*

$$\lim_{k\to\infty} g_k = g_\lambda, \tag{4.37}$$

$$\lim_{k\to\infty} V^\lambda_{g_k} = V_\lambda. \tag{4.38}$$

with g_λ *defined in* (4.5)*. In particular, we have* g_k *is is strictly increasing on* $[0, \hat{x}_{g_k}]$ *and* $g_k \in C^1([0, \hat{x}_{g_k}])$ *for all* $k \in \mathbb{N}_+$*.*

Note that the monotonicity and smoothness of the boundary g_k ensures the existence and uniqueness of the solution to (4.31)-(4.32), which coincides with the value function under policy ξ_{g_k} (see Theorem 4.4).

Remark 4.10. *It is worth noting that Theorem 4.9 presents the first policy improvement result for RL algorithms related to free boundary problems. Existing literature primarily focuses on the development of continuous-time regular controls. Compared to policy improvement, showing policy convergence remains significantly more challenging, even for regular controls. To the best of our knowledge, the only results on policy convergence for regular controls are found in* [3, 30, 40, 60]*, and all of these results are technically intricate.*

Proof. Step 1. By the construction of the sequence in Algorithm 1, we have $0 \le g_{k+1} \le g_k \le g_0$ for all $k \in \mathbb{N}_+$. Hence, by monotone convergence theorem, there exists a limit $\bar{g} \ge 0$ such that

$$\bar{g}(x) := \lim_{k\to\infty} g_k(x). \tag{4.39}$$

The main goal is to show that $\bar{g}(x) = g_\lambda(x)$ for $x \in \mathbb{R}_+$.

Step 2. Recall $V^\lambda_{g_k}(x, y)$ as the value function associated with strategy ξ^{g_k}. We use induction to prove the following iteratively:

- **(a)** g_k is is strictly increasing on $[0, \hat{x}_{g_k}]$,
- **(b)** $g_k(0)=e^{-(1+\frac{\kappa\rho}{\lambda})}$ for $k \in \mathbb{N}_+$,
- **(c)** $-\alpha_-\big(\kappa + \frac{\lambda}{\rho}\log(g_k(x)) + \frac{\lambda}{\rho}\big) + \alpha_- H_\pi(x) - H'_\pi(x)\cdot x \ge 0$ for $x \in [0, \hat{x}_{g_k}]$, namely, $g_k \ge g_\lambda$ on $[0, \hat{x}_{g_k}]$,
- **(d)** $\partial^-_{xy} u_k(x, g_k(x)) \le 0$ for $x \in [0, \hat{x}_{g_k}]$.
- **(e)** $g_k \in C^1([0, \hat{x}_{g_k}])$,

The conditions (a)-(c), (e) hold for $k = 0$ by the design of the initial reflection boundary g_0 (see Assumption 4.6). Therefore, Theorems 4.4 and 4.5 apply to g_0. Hence we have $u_0(x, y) = A_0(y)x^{\alpha_-} + H_\pi(x)\,y - \kappa y - \frac{\lambda}{\rho} y \log y$, with $A_0(y) = \int_{g_0(0)}^y \frac{\kappa + \frac{\lambda}{\rho}\log(u) + \frac{\lambda}{\rho} - H_\pi(g_0^{-1}(u))}{(g_0^{-1}(u))^{\alpha_-}} \mathrm{d}u$. Combined with condition **(c)**, we have for $x \in [0, \hat{x}_{g_0}]$,

$$\partial_{yx} u_0(x, g_0(x)) = \frac{\alpha_-}{x}\Big(\kappa + \frac{\lambda}{\rho}\log(g_0(x)) + \frac{\lambda}{\rho} - H_\pi(x)\Big) + H'_\pi(x) \le 0.$$

Therefore condition **(d)** holds.

Assume the above property holds for k. Then by Theorem 4.4, the following HJB equation

$$(\mathcal{L}_x - \rho)u + \Big(\pi(x) - \rho\kappa\Big)y - \lambda y \log y = 0 \quad \text{on} \quad \mathcal{E}(g_k), \tag{4.40}$$

$$-u_y = 0, \quad \text{on} \quad \mathcal{S}(g_k). \tag{4.41}$$

has a unique $C^1(\mathbb{R}_+ \times [0,1]) \cap C^2\left(\overline{\mathcal{E}(g_k)}\right)$ solution, which we denoted as u_k. Also, $V^\lambda_{g_k}(x,y) \equiv u_k(x,y)$. Then for $y \le g_k(x)$ the value function u_k satisfies the form:

$$u_k(x,y) = A_k(y)x^{\alpha_-} + H_\pi(x)y - \kappa y - \frac{\lambda}{\rho} y \log y, \tag{4.42}$$

with

$$A_k(y) = \int_{g_k(0)}^{y} \frac{\kappa + \frac{\lambda}{\rho}\log(u) + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(u))}{(g_k^{-1}(u))^{\alpha_-}} \mathrm{d}u. \tag{4.43}$$

In this region, we also have,

$$\begin{aligned}
\partial_y u_k(x,y) &= A_k'(y)x^{\alpha_-} + H_\pi(x) - k - \frac{\lambda}{\rho} - \frac{\lambda}{\rho}\log(y) \\
&= \frac{x^{\alpha_-}}{(g_k^{-1}(y)^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log y + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(y))\Big) + H_\pi(x) - k - \frac{\lambda}{\rho} - \frac{\lambda}{\rho}\log(y) \\
&= \frac{x^{\alpha_-} - (g_k^{-1}(y)^{\alpha_-}}{(g_k^{-1}(y)^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log y \\
&\qquad + \frac{\lambda}{\rho} - H_\pi(x)\Big) + \frac{x^{\alpha_-}}{(g_k^{-1}(y)^{\alpha_-}}\Big(H_\pi(x) - H_\pi(g_k^{-1}(y))\Big),
\end{aligned}$$

and

$$\begin{aligned}
\partial_{yx} u_k(x,y) &= \alpha_-\, A_k'(y)x^{\alpha_- - 1} + H_\pi'(x) \\
&= \frac{\alpha_- x^{\alpha_- -1}}{(g_k^{-1}(y)^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log y + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(y))\Big) + H_\pi'(x).
\end{aligned}$$

In the kth iteration there are two scenarios for each point $(x_0, g_k(x_0))$: $\partial^-_{xy}u_k(x_0,g_k(x_0)) = 0$ or $\partial^-_{xy}u_k(x_0,g_k(x_0)) < 0$. When $\partial^-_{xy}u_k(x_0,g_k(x_0)) = 0$, we have $g_{k+1}(x_0) = g_k(x_0)$. Therefore, for both scenarios, $\partial^-_{xy}u_k(x,y) = 0$ leads to $y = g_{k+1}(x)$, or equivalently, we solve $y = g_{k+1}(x) \le g_k(x)$ from $\partial^-_{xy}u_k\big(x, g_{k+1}(x)\big) = 0$. Then we have

$$\begin{aligned}
0 &= -\frac{\alpha_- x^{\alpha_-}}{(g_k^{-1}(g_{k+1}(x))^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log(g_{k+1}(x)) + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(g_{k+1}(x)))\Big) - H_\pi'(x)\,x \\
&\le -\alpha_-\Big(\kappa + \frac{\lambda}{\rho}\log(g_{k+1}(x)) + \frac{\lambda}{\rho} - H_\pi(x)\Big) - H_\pi'(x)\,x. \qquad (4.44)
\end{aligned}$$

The last inequality holds since for $y \ge g_\lambda(x)$,

$$\frac{k + \frac{\lambda}{\rho}\log(y) + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(g_{k+1}(x)))}{g_k^{-1}(g_{k+1}(x))^{\alpha_-}} \le \frac{k + \frac{\lambda}{\rho}\log(y) + \frac{\lambda}{\rho} - H_\pi(x)}{x^{\alpha_-}}. \tag{4.45}$$

Equation (4.45) holds since $g_k^{-1}(g_{k+1}(x)) \le x$ and

$$\begin{aligned}
&\partial_x\left(\frac{k + \frac{\lambda}{\rho}\log(y) + \frac{\lambda}{\rho} - H_\pi(x)}{x^{\alpha_-}}\right) \\
&= \frac{-H_\pi'(x)x^{\alpha_-} - \alpha_-\Big(k + \frac{\lambda}{\rho}\log(y) + \frac{\lambda}{\rho} - H_\pi(x)\Big)x^{\alpha_- - 1}}{x^{2\alpha_-}} \ge 0. \qquad (4.46)
\end{aligned}$$

And finally (4.46) holds for $y \ge g_\lambda(x)$.

Hence (4.44) suggests that **(c)** is satisfied in the $k+1$ iteration.

Next we show that $g_{k+1}(0) = e^{-(1+\frac{\kappa\rho}{\lambda})}$. For any points x_0 such that **(c)** holds with equality sign in the kth iteration, namely,

$$g_k(x_0) = \exp\left(\frac{\frac{-\alpha_- H_\pi(x_0)+H'_\pi(x_0)x_0}{-\alpha_-} - (\kappa + \frac{\lambda}{\rho}))\rho}{\lambda}\right),$$

it holds that $\partial_{xy} u_k^-(x_0, g_k(x_0)) = 0$ by simple calculation. Hence it holds that $g_{k+1}(x_0) = g_k(x_0)$. Take $x_0 = 0$, we have $g_{k+1}(0) = g_k(0) = e^{-(1+\frac{\kappa\rho}{\lambda})}$.

We now show condition **(a)** by contradiction. Suppose there exists $0 < x_1 < x_2$ such that $\bar{y} = g_{k+1}(x_1) = g_{k+1}(x_2)$. Then we have

$$\begin{aligned}
0 &= x_1 \partial_{xy} u_k(x_1, \bar{y}) - x_2 \partial_{xy} u_k(x_2, \bar{y}) \\
&= \frac{\alpha_- x_1^{\alpha_-}}{(g_k^{-1}(\bar{y})^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log(\bar{y}) + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(\bar{y}))\Big) + H'_\pi(x_1)x_1 \\
&\quad - \frac{\alpha_- x_2^{\alpha_-}}{(g_k^{-1}(\bar{y})^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log(\bar{y}) + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(\bar{y}))\Big) - H'_\pi(x_2)_2 \\
&= \frac{\alpha_-(x_1^{\alpha_-} - x_1^{\alpha_-})}{(g_k^{-1}(\bar{y})^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log(\bar{y}) + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(\bar{y}))\Big) + H'_\pi(x_1)x_1 - H'_\pi(x_2)x_2. \qquad (4.47)
\end{aligned}$$

Note that $x_1^{\alpha_-} > x_2^{\alpha_-}$ due to the facts that $x_1 < x_2$ and $\alpha_- < 0$. In addition, we have $x_1 H'_\pi(x_1) < x_2 H'_\pi(x_2)$ as $x\pi'(x)$ is strictly increasing according to Assumption 4.1. Hence we get a contradiction to (4.47). This suggests that g_{k+1} is either strictly increasing or decreasing. Given that $0 < g_\lambda \le g_{k+1}$ on $[0, \hat{x}_{g_{k+1}}]$, then g_{k+1} must be strictly increasing.

(d) holds since, according to **(c)**

$$\partial_{yx} u_{k+1}(x, g_{k+1}(x)) = \frac{\alpha_-}{x}\Big(\kappa + \frac{\lambda}{\rho}\log(g_{k+1}(x)) + \frac{\lambda}{\rho} - H_\pi(x)\Big) + H'_\pi(x) \le 0.$$

Finally, given that $u_k \in C^2\left(\overline{\mathcal{E}(g_k)}\right)$, $\partial_{xy} u_k(x, g_{k+1}(x)) = 0$ for $x \in [0, \hat{x}_{g_{k+1}}]$ and $(x, g_{k+1}(x)) \in \overline{\mathcal{E}(g_k)}$, we have $g_{k+1} \in C^1([0, \hat{x}_{g_{k+1}}])$. Hence we have **(e)** holds.

Step 3. We study the property of the limiting function $\bar{g} \ge g_\lambda$ as $g_k \ge g_\lambda$ for all $k \in \mathbb{N}_+$.

According to Step 2, **(a)**-**(e)** holds iteratively. Therefore it holds for all $k \in \mathbb{N}_+$ that

$$\frac{-\alpha_- x^{\alpha_-}}{(g_k^{-1}(g_{k+1}(x))^{\alpha_-}}\Big(\kappa + \frac{\lambda}{\rho}\log(g_{k+1}(x)) + \frac{\lambda}{\rho} - H_\pi(g_k^{-1}(g_{k+1}(x)))\Big) - H'_\pi(x)\,x = 0. \qquad (4.48)$$

For each $x \in \mathbb{R}_+$ fixed, we have $\lim_{k\to\infty} g_k^{-1}(g_{k+1}(x)) = x$ hold. Combined with (4.48), we have

$$0 \;=\; -\alpha_-\Big(\kappa + \frac{\lambda}{\rho}\log(\bar{g}(x)) + \frac{\lambda}{\rho} - H_\pi(x)\Big) - H'_\pi(x)\,x.$$

Therefore, $\bar{g} = g_\lambda$ for $x \in [0, x_{g_\lambda}]$. Hence (4.37) holds.

In addition, according to Step 2, for $(x, y) \in \overline{\mathcal{E}(g_k)}$, we have u_k follows (4.42) with A_k defined in (4.42). As g_k converges to g_λ uniformly, we have A_k converges to A, which is defined in (4.25). Hence u_k converges to u^λ in $\overline{\mathcal{E}(g^\lambda)}$. Hence $u_k = V^\lambda_{g_k}$ converges to $u^\lambda = V^\lambda$ on $\mathbb{R}_+ \times [0, 1]$. Finally, (4.38) holds. □

4.2.2. *Model-free implementation.* When the model parameters are unknown, we are not able to calculate the solution u_k of (4.31)-(4.32). In this case, we need to approximate the value function by acquiring instantaneous rewards along multiple trajectories and take the average (see lines 7-9 in Algorithm 3). This is referred to as the Policy Evaluation in the RL literature [53, 58].

Algorithm 3 Policy Iteration and Policy Evaluation for Optimal Stopping (PIPE-OS)

1: Initialize $g_0(x)$ according to Assumption 4.6. Specify a grid size δ_x for partitioning the x-axis and a grid size δ_y for partitioning the y-axis. Also, specify an upper bound $\bar{x} := N\delta$.
2: **for** $k = 0, 1, \cdots, K-1$ **do**
3: **for** $x \in \{0, \delta_x, 2\delta_x, \cdots, N\delta_x\}$ and $y \in \{0, \delta_y, 2\delta_y, \cdots, \lfloor 1/\delta_y \rfloor\}$ **do**
4: **for** $m = 1, \cdots, M$ **do**
5: Simulate the m-th path $(X^{(m),x}, Y^{(m),y,\xi^{g_k}})$ under policy ξ^{g_k} (defined in (4.12))
6: ▷ independent randomness across paths
7: Calculate the instantaneous value function $u_k^{(m)}(x, y)$:

$$u_k^{(m)}(x,y) = \int_0^\infty e^{-\rho t}\Big(\Big(\pi(X_t^{(m),x}) - \rho\kappa\Big)Y_t^{(m),y,\xi^{g_k}} - \lambda Y_t^{(m),y,\xi^{g_k}} \log(Y_t^{(m),y,\xi^{g_k}})\Big)\mathrm{d}t \tag{4.49}$$

8: **end for**
9: Calculate the approximated value function $\bar{u}_k(x,y) = \frac{\sum_{m=1}^M u_k^{(m)}(x,y)}{M}$
10: Update the strategy

$$g_{k+1}(x) = \begin{cases} \max\Big\{y^* < g_k(x) \,\Big|\, \partial_{xy}\bar{u}_k(x,y^*) = 0\Big\} & \text{if } \partial_{xy}^-\bar{u}_k(x, g_k(x)) < 0, \\ g_{k+1}(x) = g_k(x) & \text{o.w.} \end{cases}$$

11: **end for**
12: **end for**

Notably, we do not directly estimate model parameters in Algorithm 3, which enhances robustness against model misspecification and environmental shifts [1, 19, 28].

4.2.3. *Numerical performance.* We test the performance of Algorithm 3 on a few examples.

In the experiment, we set $\mu = 0.2$, $\sigma = 0.2$, $\rho = 0.5$, $k = 5$, and $\beta = 0.5$ (assumed to be unknown to the learner).

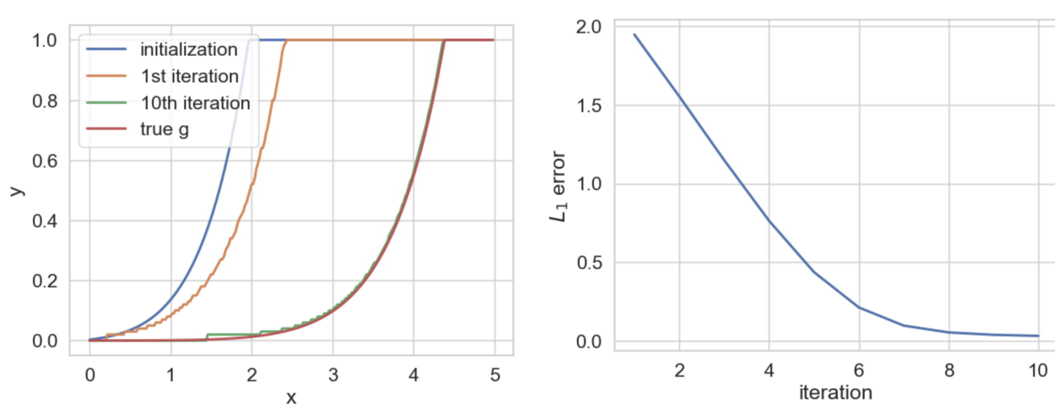

FIGURE 2. Exponential initialization. **Left**: Ground-truth and learned g function in selected iterations . **Right**: Convergence to the ground truth in L_1 norm
.

We evaluate the performance of the algorithm using two different initial policies: one with exponential initialization and the other with linear initialization (see Remark 4.8). As shown in Figure 2, Algorithm 3 converges within 10 outer iterations when using exponential initialization, highlighting its effectiveness. In addition, it takes approximately 20 outer iterations for the algorithm to converge with linear initialization (see Figure 3). This suggests that learning the boundary is more challenging when x is smaller.

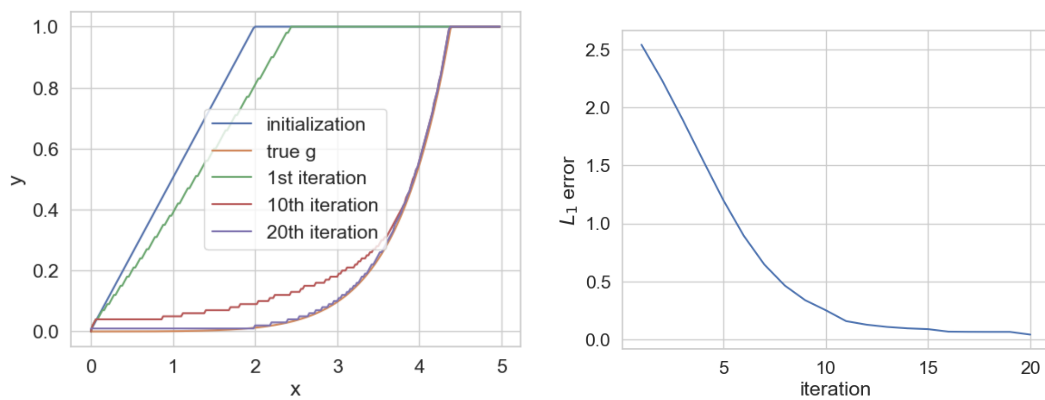

FIGURE 3. Linear initialization. **Left**: Ground-truth and learned g function in selected iterations . **Right**: Convergence to the ground truth in L_1 norm .

APPENDIX A. PROOF OF LEMMA 2.3

Setting $\hat{V} := V - G$, by Itô's formula, problem (2.1) reduces to the OS problem $\hat{V}(x) = \sup_{\tau \in \mathcal{T}} \mathbb{E}\big[\int_0^\tau e^{-\rho t}\hat{\pi}(X_t^x)dt\big]$, for which the value function $\hat{V}$ is the unique $W^{2,2}_{loc}(\mathbb{R})$-a.e. solution to the HJB variational inequality (see. e.g., Section 5.2 in [47])

$$\max\big\{\,(\mathcal{L}_x - \rho)\,\hat{V} + \hat{\pi}, -\hat{V}\,\big\} = 0. \tag{A.1}$$

By Theorem 2.3.5 in [39], the minimal stopping time τ^* is the first hitting time to the region $\hat{\mathcal{S}} := \{\hat{V} = 0\}$. Also, since $\hat{\pi}$ is increasing, $\hat{V}$ is nondecreasing so that $\hat{\mathcal{S}}$ is a closed interval (possibly empty). If $\hat{\mathcal{S}}$ is empty, then $\tau^* = \infty$ is the unique optimal stopping time, by minimality of τ^*. On the contrary, if $\hat{\mathcal{S}}$ is nonempty, then we can find $x^* \in \mathbb{R}$ such that $\hat{\mathcal{S}} := \{\hat{V} = 0\} = (-\infty, x^*]$. Notice moreover that necessarily we have $x^* \leq \bar{x}$, as otherwise we would have $(\mathcal{L}_x - \rho)\hat{V} + \hat{\pi} = \hat{\pi} > 0$ in the interval $(\bar{x}, x^*)$, thus contradicting the HJB equation. In other words,

$$(\mathcal{L}_x - \rho)\hat{V} + \hat{\pi} = \hat{\pi} < 0 \text{ in } (-\infty, x^*] \text{ and } (\mathcal{L}_x - \rho)\hat{V} + \hat{\pi} < 0, \text{ a.e. in } \mathbb{R}. \tag{A.2}$$

Next, assume by contradiction that there exists another optimal stopping time τ. By minimality of τ^* we have $\tau^* \leq \tau$. Moreover, by the martingale optimality principle (see e.g. Theorem 2.2 at p. 29 in [46]), the process $M = (M_t)_t$ defined by

$$M_t := e^{-\rho t \wedge \tau}\hat{V}(X^x_{t\wedge\tau}) + \int_0^{t\wedge\tau} e^{-\rho s}\hat{\pi}(X_s^x)\mathrm{d}s,$$

is an $\mathbb{F}^W$-martingale and, using Itô's formula (thanks to the regularity of $\hat{V}$) we have

$$\mathbb{E}[M_t] := \hat{V}(x) + \mathbb{E}\bigg[\int_0^{t\wedge\tau} e^{-\rho s}\big((\mathcal{L}_x - \rho)\hat{V} + \hat{\pi}\big)(X_s^x)\mathrm{d}s\bigg].$$

Notice that, since $\sigma^2(x) \geq c > 0$, after the time τ^* the Brownian motion pushes the state X_t^x in the region $(-\infty, x^*)$ with probability 1. Thus, by (A.2) and the characterization of τ^* we have

$$\begin{aligned}\mathbb{E}[M_\tau] &= \hat{V}(x) + \mathbb{E}\left[\int_0^\tau e^{-\rho s}\big((\mathcal{L}_x - \rho)\hat{V} + \hat{\pi}\big)(X_s^x)\mathrm{d}s\right] \\ &= \hat{V}(x) + \mathbb{E}\left[\mathbb{1}_{\{\tau>\tau^*\}}\int_{\tau^*}^\tau e^{-\rho s}\big((\mathcal{L}_x - \rho)\hat{V} + \hat{\pi}\big)(X_s^x)\mathrm{d}s\right] < \hat{V}(x) = \mathbb{E}[M_0],\end{aligned}$$

where the last inequality follows from assuming that the event $\tau > \tau^*$ has positive probability. This contradicts the fact that M is a martingale, and thus show that the optimal stopping time τ^* is unique.

Appendix B. Proof of the auxiliary estimates

We prove each of the estimate (3.5)–(3.10) separately.

Proof of (3.5). Noticing that $|x|(1+|x|) \le \frac{3}{2}(1+|x|^2)$, we have

$$xb(x) + \frac{q-1}{2}|\sigma(x)|^2 \le \left(\frac{3}{2}L + (q-1)L^2\right)\left(1+|x|^2\right).$$

Hence, the estimate (3.5) follows from Theorem 4.1 at p. 60 in [41].

Proof of (3.6). We adopt the rationals in the proof of Theorem 4.1 at p. 60 in [41]. Set $\Delta_t := X_t^{\bar{x}} - X_t^x$. For any $\varepsilon > 0$, using Itô's formula together with elementary estimates, we find

$$\begin{aligned}&\left(\varepsilon + |\Delta_t|^2\right)^{\frac{q}{2}} \\ &= \left(\varepsilon + |\bar{x}-x|^2\right)^{\frac{q}{2}} + q\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q-2}{2}} \Delta_t^*\left(\sigma\left(X_t^{\bar{x}}\right) - \sigma\left(X_t^x\right)\right)\mathrm{d}W_t \\ &\quad + q\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q-2}{2}} \Delta_t^*\left(b\left(X_t^{\bar{x}}\right) - b\left(X_t^x\right)\right)\mathrm{d}t \\ &\quad + q\frac{q-2}{2}\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q-4}{2}} \left|\Delta_t^*\left(\sigma\left(X_t^{\bar{x}}\right) - \sigma\left(X_t^x\right)\right)\right|^2\mathrm{d}t \\ &\quad + \frac{q}{2}\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q-2}{2}} \left|\sigma\left(X_t^{\bar{x}}\right) - \sigma\left(X_t^x\right)\right|^2\mathrm{d}t \\ &\le \left(\varepsilon + |\bar{x}-x|^2\right)^{\frac{q}{2}} + q\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q-2}{2}} \Delta_t^*\left(\sigma\left(X_t^{\bar{x}}\right) - \sigma\left(X_t^x\right)\right)\mathrm{d}W_t \\ &\quad + q\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q-2}{2}} \left[\Delta_t^*\left(b\left(X_t^{\bar{x}}\right) - b\left(X_t^x\right)\right) + \frac{q-1}{2}\left|\sigma\left(X_t^{\bar{x}}\right) - \sigma\left(X_t^x\right)\right|^2\right]\mathrm{d}t \\ &\le \left(\varepsilon + |\bar{x}-x|^2\right)^{\frac{q}{2}} + q\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q-2}{2}} \Delta_t^*\left(\sigma\left(X_t^{\bar{x}}\right) - \sigma\left(X_t^x\right)\right)\mathrm{d}W_t \\ &\quad + q\left(L + \frac{q-1}{2}L^2\right)\int_0^t \left(\varepsilon + |\Delta_t|^2\right)^{\frac{q}{2}}\mathrm{d}t.\end{aligned} \tag{B.1}$$

Taking expectations and using Grönwall inequality, we obtain

$$\mathbb{E}\left[\left(\varepsilon + |\Delta_t|^2\right)^{\frac{q}{2}}\right] \le \left(\varepsilon + |\bar{x}-x|^2\right)^{\frac{q}{2}} e^{q\left(L+\frac{q-1}{2}L^2\right)t},$$

which completes the proof sending $\varepsilon \to 0$.

Proof of (3.7). Set

$$\begin{aligned}&\hat{X}_t := \delta X_t^{\bar{x}} + (1-\delta)X_t^x, \quad X_t^\delta := X_t^{\delta\bar{x}+(1-\delta)x}, \quad \Gamma_t := \hat{X}_t - X_t^\delta, \\ &\hat{b}_t := \delta b\left(X_t^{\bar{x}}\right) + (1-\delta)b\left(X_t^x\right), \quad \hat{\sigma}_t := \delta\sigma\left(X_t^{\bar{x}}\right) + (1-\delta)\sigma\left(X_t^x\right).\end{aligned}$$

By Itô's formula, with elementary estimates we obtain

$$\begin{aligned}\left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q}{2}} &= \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^* \left[\left(\hat{b}_t - b\left(X_t^\delta\right)\right)\mathrm{d}t + \left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right)\mathrm{d}W_t\right] \\ &\quad + q\frac{q-2}{2}\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-4}{2}} \left|\Gamma_t^*\left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right)\right|^2 \mathrm{d}t \\ &\quad + \frac{q}{2}\int_0^t \left(\varepsilon+|\Gamma_t|^{q-2}\right)^{\frac{q-2}{2}} \left|\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right|^2 \mathrm{d}t \\ &\leq \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \left[\Gamma_t^*\left(\hat{b}_t - b\left(X_t^\delta\right)\right) + \frac{q-1}{2}\left|\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right|^2\right]\mathrm{d}t \\ &\quad + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^*\left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right)\mathrm{d}W_t.\end{aligned}$$

Using the properties of b and σ in Condition 1 in Assumption 3.1, we continue with

$$\begin{aligned}&\left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q}{2}} \\ &\leq \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^*\left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right)\mathrm{d}W_t \\ &\quad + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \left[\Gamma_t^*\left(b\left(\hat{X}_t\right) - b\left(X_t^\delta\right)\right) + (q-1)\left|\sigma\left(\hat{X}_t\right) - \sigma\left(X_t^\delta\right)\right|^2\right]\mathrm{d}t \\ &\quad + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \left[\Gamma_t^*\left(\hat{b}_t - b\left(\hat{X}_t\right)\right) + (q-1)\left|\hat{\sigma}_t - \sigma\left(\hat{X}_t\right)\right|^2\right]\mathrm{d}t \\ &\leq \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^*\left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right)\mathrm{d}W_t \\ &\quad + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \left(L+(q-1)L^2\right)|\Gamma_t|^2\,\mathrm{d}t \\ &\quad + q\int_0^t \left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}} \Big(|\Gamma_t^*|\,\delta(1-\delta)L\left|X_t^{\bar{x}} - X_t^x\right|^2 \\ &\qquad\qquad + (q-1)\delta^2(1-\delta)^2L^2\left|X_t^{\bar{x}} - X_t^x\right|^4\Big)\mathrm{d}t.\end{aligned}$$

Next, since $y\bar{y} \le \frac{1}{2}(y^2 + \bar{y}^2)$, after rearranging some of the terms we arrive at the expression

$$\begin{aligned}\left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q}{2}} \le & \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^* \left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right) \mathrm{d}W_t \\ & + q\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q-2}{2}} \left(L + (q-1)L^2\right)|\Gamma_t|^2 \,\mathrm{d}t \\ & + q\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q-2}{2}} \left(\frac{L}{2}\left(|\Gamma_t|^2 + \delta^2(1-\delta)^2\left|X_t^{\bar{x}} - X_t^x\right|^4\right)\right. \\ & \qquad \left. + (q-1)L^2\delta^2(1-\delta)^2\left|X_t^{\bar{x}} - X_t^x\right|^4\right)\mathrm{d}t \\ \le & \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^* \left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right) \mathrm{d}W_t \\ & + q\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q-2}{2}} \left[\left(\frac{3}{2}L + (q-1)L^2\right)|\Gamma_t|^2\right. \\ & \qquad \left. + \delta^2(1-\delta)^2\left(\frac{L}{2} + (q-1)L^2\right)\left|X_t^{\bar{x}} - X_t^x\right|^4\right]\mathrm{d}t.\end{aligned}$$

Next, using Young inequality with exponent $\frac{q}{q-2}$ and its conjugate $\frac{q}{2}$, we obtain

$$\begin{aligned}&\left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q}{2}} \\ &\le \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^* \left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right) \mathrm{d}W_t \\ &\quad + q\left(\frac{3}{2}L + (q-1)L^2\right)\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q}{2}} \mathrm{d}t \\ &\quad + \delta^2(1-\delta)^2 q\left(\frac{L}{2} + (q-1)L^2\right)\int_0^t \left(\frac{q-2}{q}\left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q}{2}} + \frac{2}{q}\left|X_t^{\bar{x}} - X_t^x\right|^{2q}\right)\mathrm{d}t \\ &\le \varepsilon^{\frac{q}{2}} + q\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q-2}{2}} \Gamma_t^* \left(\hat{\sigma}_t - \sigma\left(X_t^\delta\right)\right) \mathrm{d}W_t \\ &\quad + \left((2q-1)L + 2(q-1)^2L^2\right)\int_0^t \left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q}{2}} \mathrm{d}t \\ &\qquad + \delta^2(1-\delta)^2\left(L + 2(q-1)L^2\right)\int_0^t \left|X_t^{\bar{x}} - X_t\right|^{2q} \mathrm{d}t.\end{aligned} \tag{B.2}$$

Taking expectations and using Grönwall, (3.6) implies that

$$\mathbb{E}\left[\left(\varepsilon + |\Gamma_t|^2\right)^{\frac{q}{2}}\right] \le \left(\varepsilon^{\frac{q}{2}} + C\delta^2(1-\delta)^2|\bar{x} - x|^{2q} e^{2q\left(L + \frac{2q-1}{2}L^2\right)t}\right) e^{\left((2q-1)L + 2(q-1)^2L^2\right)t},$$

which in turn gives (3.7) by sending $\varepsilon \to 0$.

Proof of (3.8). First, by Itô's formula we have

$$
\begin{aligned}
&e^{-\rho t}\left(1+|X_t^x|^2\right)^{\frac{q}{2}}\\
&=\left(1+|x|^2\right)^{\frac{q}{2}}+q\int_0^t e^{-\rho s}\left(1+|X_t^x|^2\right)^{\frac{q}{2}-1}(X_t^x)^*\sigma(X_t^x)\mathrm{d}W_t\\
&\quad+\int_0^t e^{-\rho t}\left(1+|X_t^x|^2\right)^{\frac{q}{2}}\Bigg[-\rho+q\bigg(\frac{(X_t^x)^*}{1+|X_t^x|^2}b\left(X_s^x\right)\\
&\qquad\qquad+\left(\frac{q}{2}-1\right)\frac{|(X_t^x)^*\sigma(X_t^x)|^2}{(1+|X_t^x|^2)^2}+\frac{1}{2}\frac{|\sigma(X_t^x)|^2}{1+|X_t^x|^2}\bigg)\Bigg]\mathrm{d}t\\
&\le(1+|x|^2)^{\frac{q}{2}}+q\int_0^t e^{-\rho t}\left(1+|X_t^x|^2\right)^{\frac{q}{2}-1}(X_t^x)^*\sigma\left(X_t^x\right)\mathrm{d}W_t\\
&\quad+C\int_0^t e^{-\rho t}\left(1+|X_t^x|^2\right)^{\frac{q}{2}}\mathrm{d}t.
\end{aligned}
$$

Next, for $\tau\in\mathcal{T}$ and $T>0$, Burkholder-Davis-Gundy inequality gives

$$
\begin{aligned}
\mathbb{E}\left[\sup_{t\le\tau\wedge T}e^{-\rho t}\left(1+|X_t^x|^2\right)^{\frac{q}{2}}\right]&\le\left(1+|x|^2\right)^{\frac{q}{2}}+C\int_0^T e^{-\rho t}\mathbb{E}\left[\left(1+|X_t^x|^2\right)^{\frac{q}{2}}\right]\mathrm{d}t\\
&\quad+C\mathbb{E}\left[\left(\int_0^T e^{-2\rho t}\left(1+|X_t^x|^2\right)^{q-2}|X_t^x|^2\,|\sigma(X_t^x)|^2\mathrm{d}t\right)^{\frac{1}{2}}\right].
\end{aligned}
$$

Thus, using (3.5) and Jensen inequality we find

$$
\begin{aligned}
\mathbb{E}\left[\sup_{t\le\tau\wedge T}e^{-\rho t}\left(1+|X_t^x|^2\right)^{\frac{q}{2}}\right]&\le C\left(1+|x|^q\right)+C\left(1+|x|^q\right)\int_0^\infty e^{(-\rho+\hat{c}_0(q))t}\mathrm{d}t\\
&\quad+C\left(\int_0^\infty e^{-2\rho t}\mathbb{E}\left[\left(1+|X_t^x|^2\right)^q\right]\mathrm{d}t\right)^{\frac{1}{2}}\\
&\le C\left(1+|x|^q\right)+C\left(1+|x|^q\right)\left(\int_0^\infty e^{(-2\rho+\hat{c}_0(2q))t}\mathrm{d}t\right)^{\frac{1}{2}}\\
&\le C\left(1+|x|^q\right),
\end{aligned}
$$

where the finiteness of the integrals follows from the conditions $\rho>\hat{c}_0(q)$ and $2\rho>\hat{c}_0(2q)$ and the particular choice of q in (3.8) (which are part of Condition 4 in Assumption 3.1). The estimate (3.8) follows taking limits as $T\to\infty$ and using the dominated convergence theorem.

Proof of (3.9). With the notation $\Delta_t:=X_t^{\bar{x}}-X_t^x$, similarly to (B.1) we find

$$
\begin{aligned}
\left(\varepsilon+|\Delta_t|^2\right)^{\frac{q}{2}}\varepsilon\left(\varepsilon+|\bar{x}-x|^2\right)^{\frac{q}{2}}+\left(-\rho+q\left(L+\frac{q-1}{2}L^2\right)\right)\int_0^t e^{-\rho t}\left(\varepsilon+|\Delta_t|^2\right)^{\frac{q}{2}}\mathrm{d}t\\
+q\int_0^t e^{-\rho t}\left(\varepsilon+|\Delta_t|^2\right)^{\frac{q-2}{2}}\Delta_t^*\left(\sigma\left(X_t^{\bar{x}}\right)-\sigma\left(X_t^x\right)\right)\mathrm{d}W_t.
\end{aligned}
$$

For $\tau \in \mathcal{T}$ and $T > 0$, by using Burkholder-Davis-Gundy inequality and Jensen inequality, thanks to (3.6) we obtain

$$
\begin{aligned}
\mathbb{E}\left[\sup_{t\leq\tau\wedge T}\left(\varepsilon+|\Delta_t|^2\right)^{\frac{q}{2}}\right] \leq & C(\varepsilon^{\frac{q}{2}}+|\bar{x}-x|^q)+C\int_0^T e^{-\rho t}\mathbb{E}\left[|\Delta_t|^q\right]\mathrm{d}t \\
& + C\left(\int_0^T e^{-2\rho t}\mathbb{E}\left[|\Delta_t|^{2q}\right]\mathrm{d}t\right)^{\frac{1}{2}} \\
\leq & C\left(\varepsilon^{\frac{q}{2}}+|\bar{x}-x|^q\left(1+\int_0^\infty\left(e^{(-\rho+\hat{c}_1(q))t}+e^{(-2\rho+\hat{c}_1(2q))t}\right)\mathrm{d}t\right)\right) \\
\leq & C\left(\varepsilon^{\frac{q}{2}}+|\bar{x}-x|^q\right),
\end{aligned}
$$

where the finiteness of the integrals follows from the conditions $\rho > \hat{c}_1(q)$ and $2\rho > \hat{c}_1(2q)$ and the choice of q in (3.9) (which are part of Condition 4 in Assumption 3.1).

From the latter estimate, take limits as $\varepsilon \to 0$ and $T \to \infty$ and use monotone convergence theorem to derive (3.9).

Proof of (3.10). With the same notations as in the proof of (3.7) above, by repeating the steps leading to (B.2), we find

$$
\begin{aligned}
e^{-\rho t}\left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q}{2}} \leqslant \varepsilon^{\frac{q}{2}} + C\int_0^t e^{-\rho t}\left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q}{2}}\mathrm{d}t + C\delta^2(1-\delta)^2\int_0^t e^{-\rho t}\left|X_t^{\bar{x}}-X_t^x\right|^{2q}\mathrm{d}t \\
+ q\int_0^t e^{-\rho t}\left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q-2}{2}}\Gamma_t^*\left(\hat{\sigma}_t-\sigma\left(X_t^\delta\right)\right)\mathrm{d}W_t
\end{aligned}
$$

Hence, for $\tau \in \mathcal{T}$ and $T > 0$, by using Burkholder-Davis-Gundy inequality and Jensen inequality, thanks to (3.6) and (3.7) we get

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{t\leq\tau\wedge T} e^{-\rho t}\left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q}{2}}\right] \\
&\quad\leq C\varepsilon^{\frac{q}{2}} + C\int_0^T e^{-\rho t}\mathbb{E}\left[|\Gamma_t|^q\right]\mathrm{d}t \\
&\quad\quad + C\delta^2(1-\delta)^2\int_0^T e^{-\rho t}\mathbb{E}\left[\left|X_t^{\bar{x}}-X_t^x\right|^{2q}\right]\mathrm{d}t \\
&\quad\quad + C\left(\int_0^T e^{-2\rho t}\mathbb{E}\left[\left(\varepsilon+|\Gamma_t|^2\right)^{q-2}|\Gamma_t|^2\left|\hat{\sigma}_t-\sigma\left(X_t^\delta\right)\right|^2\right]\mathrm{d}t\right)^{\frac{1}{2}}.
\end{aligned}
\tag{B.3}
$$

To estimate the last term in the right-hand side, we first use the regularity of σ to obtain

$$
\begin{aligned}
&\int_0^T e^{-2\rho t}\mathbb{E}\left[\left(\varepsilon+|\Gamma_t|^2\right)^{q-2}|\Gamma_t|^2\left|\hat{\sigma}_t-\sigma\left(X_t^\delta\right)\right|^2\right]\mathrm{d}t \\
&\quad\leq C\int_0^T e^{-2\rho t}\mathbb{E}\left[\left(\varepsilon+|\Gamma_t|^2\right)^{q-1}\left(\left|\hat{\sigma}_t-\sigma\left(\hat{X}_t\right)\right|^2+\left|\sigma\left(\hat{X}_t\right)-\sigma\left(X_t^\delta\right)\right|^2\right)\right]\mathrm{d}t \\
&\quad\leq C\int_0^T e^{-2\rho t}\mathbb{E}\left[\left(\varepsilon+|\Gamma_t|^2\right)^{q-1}\left(\delta^2(1-\delta)^2\left|\hat{X}_t-X_t^\delta\right|^4+|\Gamma_t|^2\right)\right]\mathrm{d}t,
\end{aligned}
$$

and then employ Young inequality with exponent $\frac{q}{q-1}$ and its conjugate q to conclude that

$$\int_0^T e^{-2\rho t}\mathbb{E}\left[\left(\varepsilon+|\Gamma_t|^2\right)^{q-2}|\Gamma_t|^2\left|\hat{\sigma}_t-\sigma\left(X_t^\delta\right)\right|^2\right]\mathrm{d}t$$
$$\leq C\left(\varepsilon^{\frac{q}{2}}+\int_0^T e^{-2\rho t}\mathbb{E}\left[|\Gamma_t|^{2q}\right]\mathrm{d}t+\delta^2(1-\delta)^2\int_0^T e^{-2\rho t}\mathbb{E}\left[\left|\hat{X}_t-X_t^\delta\right|^{4q}\right]\mathrm{d}t\right).$$

By plugging the latter inequality into (B.3), and then using (3.6) and (3.7), we find

$$\mathbb{E}\left[\sup_{t\leq\tau\wedge T}e^{-\rho t}\left(\varepsilon+|\Gamma_t|^2\right)^{\frac{q}{2}}\right]$$
$$\leq C\left(\varepsilon^{\frac{q}{2}}+\int_0^T e^{-\rho t}\mathbb{E}\left[|\Gamma_t|^q\right]dt+\delta^2(1-\delta)^2\int_0^T e^{-\rho t}\mathbb{E}\left[\left|X_t^{\bar{x}}-X_t^x\right|^{2q}\right]\mathrm{d}t\right.$$
$$\left.+\left(\varepsilon^{\frac{q}{2}}+\int_0^T e^{-2\rho t}\mathbb{E}\left[|\Gamma_t|^{2q}\right]\mathrm{d}t+\delta^2(1-\delta)^2\int_0^T e^{-2\rho t}\mathbb{E}\left[\left|\hat{X}_t-X_t^\delta\right|^{4q}\right]\mathrm{d}t\right)^{\frac{1}{2}}\right)$$
$$\leq C\left(\varepsilon^{\frac{q}{2}}+\delta^2(1-\delta)^2|\bar{x}-x|^{2q}\int_0^\infty\left(e^{(-\rho+\hat{c}_2(q))t}+e^{(-\rho+\hat{c}_1(2q))t}\right)\mathrm{d}t\right.$$
$$\left.+\left(\varepsilon^{\frac{q}{2}}+\delta^2(1-\delta)^2|\bar{x}-x|^{4q}\int_0^\infty\left(e^{(-2\rho+\hat{c}_2(2q))t}+e^{(-2\rho+\hat{c}_1(4q))t}\right)\mathrm{d}t\right)^{\frac{1}{2}}\right).$$

Finally, by sending $\varepsilon\to 0$, we conclude that

$$\mathbb{E}\left[\sup_{t\leq\tau\wedge T}e^{-\rho t}|\Gamma_t|^q\right]\leq C\delta(1-\delta)|\bar{x}-x|^{2q}\int_0^\infty\left(e^{(-\rho+\hat{c}_2(q))t}+e^{(-\rho+\hat{c}_1(2q))t}\right)\mathrm{d}t$$
$$+C\delta(1-\delta)|\bar{x}-x|^{2q}\left(\int_0^\infty\left(e^{(-2\rho+\hat{c}_2(2q))t}+e^{(-2\rho+\hat{c}_1(4q))t}\right)\mathrm{d}t\right)^{\frac{1}{2}}$$
$$\leq C\delta(1-\delta)|\bar{x}-x|^{2q},$$

where the finiteness of the integrals follows from taking $q=2$ and from the conditions $\rho>\hat{c}_1(4q)$ and $2\rho>\hat{c}_2(2q)$ (which are part of Condition 4 in Assumption 3.1). The estimate (3.9) follows taking limits as $T\to\infty$ and using the dominated convergence theorem.

Acknowledgements. Supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 317210226 - SFB 1283 and the NSF CAREER award DMS-2339240.

References

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, *On the theory of policy gradient methods: Optimality, approximation, and distribution shift*, J. Mach. Learn. Res., 22 (2021), pp. 1–76.

[2] B. Ata, J. M. Harrison, and N. Si, *Singular control of (reflected) Brownian motion: A computational method suitable for queueing applications*, arXiv preprint arXiv:2312.11823, (2023).

[3] L. Bai, T. Gamage, J. Ma, and P. Xie, *Reinforcement learning for optimal dividend problem under diffusion model*, arXiv preprint arXiv:2309.10242, (2023).

[4] P. Bank and N. El Karoui, *A stochastic representation theorem with applications to optimization and obstacle problems*, Ann. Probab., 32 (2004), pp. 1030–1067.

[5] P. Bank and F. Riedel, *Optimal consumption choice with intertemporal substitution*, Ann. Appl. Probab., 11 (2001), pp. 750–788.

[6] M. Basei, X. Guo, A. Hu, and Y. Zhang, *Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon*, J. Mach. Learn. Res., 23 (2022), pp. 1–34.

[7] J. Bather, *A continuous time inventory model*, J. Appl. Probab., 3 (1966), pp. 538–549.

[8] F. Boetius and M. Kohlmann, *Connections between optimal stopping and singular stochastic control*, Stochastic Process. Appl., 77 (1998), pp. 253–281.
[9] T. De Angelis, S. Federico, and G. Ferrari, *Optimal boundary surface for irreversible investment with stochastic costs*, Math. Oper. Res., 42 (2017), pp. 1135–1161.
[10] T. De Angelis and G. Ferrari, *Stochastic nonzero-sum games: a new connection between singular control and optimal stopping*, Adv. in Appl. Probab., 50 (2018), pp. 347–372.
[11] T. De Angelis, G. Ferrari, and J. Moriarty, *A solvable two-dimensional degenerate singular stochastic control problem with nonconvex costs*, Math. Oper. Res., 44 (2019), pp. 512–531.
[12] T. De Angelis, N. Merkulov, and J. Palczewski, *On the value of non-Markovian Dynkin games with partial and asymmetric information*, Ann. Appl. Probab., 32 (2022), pp. 1774–1813.
[13] H. Denkert, Robert Pham and X. Warin, *Control randomisation approach for policy gradient and application to reinforcement learning in optimal switching*, arXiv preprint arXiv:2404.17939, (2024).
[14] R. Devidze, P. Kamalaruban, and A. Singla, *Exploration-guided reward shaping for reinforcement learning under sparse rewards*, Adv. Neural Inf. Process. Syst., 35 (2022), pp. 5829–5842.
[15] J. Dianetti and G. Ferrari, *Multidimensional singular control and related Skorokhod problem: sufficient conditions for the characterization of optimal controls*, Stochastic Process. Appl., 162 (2023), pp. 547–592.
[16] Y. Dong, *Randomized optimal stopping problem in continuous time and reinforcement learning algorithm*, SIAM J. Control Optim., 62 (2024), pp. 1590–1614.
[17] K. Doya, *Reinforcement learning in continuous time and space*, Neural Comput., 12 (2000), pp. 219–245.
[18] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, *Optimal rates for zero-order convex optimization: The power of two function evaluations*, IEEE Trans. Inf. Theory, 61 (2015), pp. 2788–2806.
[19] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, *Global convergence of policy gradient methods for the linear quadratic regulator*, in International conference on machine learning, PMLR, 2018, pp. 1467–1476.
[20] G. Ferrari, *On an integral equation for the free-boundary of stochastic, irreversible investment problems*, Ann. Appl. Probab., 25 (2015), pp. 150 – 176.
[21] G. Ferrari and P. Salminen, *Irreversible investment under lévy uncertainty: An equation for the optimal boundary*, Adv. in Appl. Probab., 48 (2016), pp. 298–314.
[22] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer Science & Business Media, 2006.
[23] ———, *Controlled Markov processes and viscosity solutions*, vol. 25, Springer Science & Business Media, 2006.
[24] M. Giegrich, C. Reisinger, and Y. Zhang, *Convergence of policy gradient methods for finite-horizon exploratory linear-quadratic control problems*, SIAM J. Control Optim., 62 (2024), pp. 1060–1092.
[25] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, vol. 224, Springer Science & Business Media, 2001.
[26] X. Guo, A. Hu, and Y. Zhang, *Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls*, SIAM J. Control Optim., 61 (2023), pp. 755–787.
[27] X. Guo and P. Tomecek, *A class of singular control problems and the smooth fit principle*, SIAM J. Control Optim., 47 (2009), pp. 3076–3099.
[28] B. Hambly, R. Xu, and H. Yang, *Policy gradient methods for the noisy linear quadratic regulator over a finite horizon*, SIAM J. Control Optim., 59 (2021), pp. 3359–3391.
[29] J. Hare, *Dealing with sparse rewards in reinforcement learning*, arXiv preprint arXiv:1910.09281, (2019).
[30] Y.-J. Huang, Z. Wang, and Z. Zhou, *Convergence of policy improvement for entropy-regularized stochastic control problems*, arXiv preprint arXiv:2209.07059, (2022).
[31] Y. Jia and X. Y. Zhou, *Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms*, J. Mach. Learn. Res., 23 (2022), pp. 1–50.
[32] ———, *q-learning in continuous time*, J. Mach. Learn. Res., 24 (2023), pp. 1–61.
[33] Y. M. Kabanov, *Hedging and liquidation under transaction costs in currency markets*, Finance Stoch., 3 (1999), pp. 237–248.
[34] I. Karatzas, D. Ocone, H. Wang, and M. Zervos, *Finite-fuel singular control with discretionary stopping*, Stochastics, 71 (2000), pp. 1–50.
[35] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control I. Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
[36] N. E. Karoui and I. Karatzas, *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11 (1988), pp. 223–258.
[37] Y. Kifer, *Dynkin's games and Israeli options*, International Scholarly Research Notices, 2013 (2013), p. 856458.
[38] S. Kumar and K. Muthuraman, *A numerical method for solving singular stochastic control problems*, Oper. Res., 52 (2004), pp. 563–582.

[39] D. Lamberton et al., *Optimal stopping and American options*, Ljubljana Summer School on Financial Mathematics, 134 (2009).
[40] J. Ma, G. Wang, and J. Zhang, *On convergence and rate of convergence of policy improvement algorithms*, arXiv preprint arXiv:2406.10959, (2024).
[41] X. Mao, *Stochastic Differential Equations and Applications*, Elsevier, 2007.
[42] J.-L. Menaldi and M. Robin, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.
[43] J. L. Menaldi and M. I. Taksar, *Optimal correction problem of a multidimensional stochastic system*, Automatica J. IFAC, 25 (1989), pp. 223–232.
[44] E. Neuman, W. Stockinger, and Y. Zhang, *An offline learning approach to propagator models*, arXiv preprint arXiv:2309.02994, (2023).
[45] M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb, *Continuous-time q-learning for infinite-horizon discounted cost linear quadratic regulator problems*, IEEE Trans. Cybern., 45 (2014), pp. 165–176.
[46] G. Peskir and A. Shiryaev, *Optimal stopping and free-boundary problems*, Springer, 2006.
[47] H. Pham, *Continuous-time Stochastic Control and Optimization with Financial Applications*, vol. 61, Springer Science & Business Media, 2009.
[48] M. Rao, Y. Chen, B. C. Vemuri, and F. Wang, *Cumulative residual entropy: a new measure of information*, IEEE Trans. Inform. Theory, 50 (2004), pp. 1220–1228.
[49] C. Reisinger, W. Stockinger, and Y. Zhang, *Linear convergence of a policy gradient method for some finite horizon continuous time control problems*, SIAM J. Control Optim., 61 (2023), pp. 3526–3558.
[50] A. M. Reppen, H. M. Soner, and V. Tissot-Daguette, *Neural optimal stopping boundary*, arXiv preprint arXiv:2205.04595, (2022).
[51] E. Solan, B. Tsirelson, and N. Vieille, *Random stopping times in stopping problems and stopping games*, arXiv preprint arXiv:1211.5802, (2012).
[52] H. M. Soner and V. Tissot-Daguette, *Stopping times of boundaries: Relaxation and continuity*, arXiv preprint arXiv:2305.09766, (2023).
[53] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.
[54] L. Szpruch, T. Treetanthiploet, and Y. Zhang, *Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models*, arXiv preprint arXiv:2112.10264, (2021).
[55] ———, *Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning*, SIAM J. Control Optim., 62 (2024), pp. 135–166.
[56] C. Tallec, L. Blier, and Y. Ollivier, *Making deep q-learning methods robust to time discretization*, in International Conference on Machine Learning, PMLR, 2019, pp. 6096–6104.
[57] W. Tang, Y. P. Zhang, and X. Y. Zhou, *Exploratory HJB equations and their convergence*, SIAM J. Control Optim., 60 (2022), pp. 3191–3216.
[58] P. Thomas and E. Brunskill, *Data-efficient off-policy policy evaluation for reinforcement learning*, in International Conference on Machine Learning, PMLR, 2016, pp. 2139–2148.
[59] N. Touzi and N. Vieille, *Continuous-time Dynkin games with mixed strategies*, SIAM J. Control Optim., 41 (2002), pp. 1073–1088.
[60] H. V. Tran, Z. Wang, and Y. P. Zhang, *Policy iteration for exploratory Hamilton–Jacobi–Bellman equations*, arXiv preprint arXiv:2406.00612, (2024).
[61] H. Wang, T. Zariphopoulou, and X. Y. Zhou, *Reinforcement learning in continuous time and space: A stochastic control approach*, J. Mach. Learn. Res., 21 (2020), pp. 1–34.
[62] S. A. Williams, P.-L. Chow, and J.-L. Menaldi, *Regularity of the free boundary in singular stochastic control*, J. Differential Equations, 111 (1994), pp. 175–201.