Regularity and stability of feedback relaxed controls

Christoph Reisinger* Yufei Zhang*

Abstract. This paper proposes a relaxed control regularization with general exploration rewards to design robust feedback controls for multi-dimensional continuous-time stochastic exit time problems. We establish that the regularized control problem admits a Hölder continuous optimal feedback control, and demonstrate that both the value function and the feedback control of the regularized control problem are Lipschitz stable with respect to parameter perturbations. Moreover, we show that a pre-computed feedback relaxed control gives a robust performance in a perturbed system, and derive a first-order sensitivity equation for both the value function and optimal feedback relaxed control. These stability results provide a theoretical justification for recent reinforcement learning heuristics that including an exploration reward in the optimization objective leads to more robust decision making. We finally prove first-order monotone convergence of the value functions for relaxed control problems with vanishing exploration parameters, which subsequently enables us to construct the pure exploitation strategy of the original control problem based on the feedback relaxed controls.

Key words. exploration and exploitation, feedback relaxed control, Lipschitz stability, sensitivity equation, reinforcement learning, Hamilton-Jacobi-Bellman equation.

AMS subject classifications. 93B52, 93B35, 93E20, 68Q32

1 Introduction

In this paper, we propose a relaxed control regularization with a class of exploration rewards to design robust feedback controls for multi-dimensional stochastic control problems in a continuous setting. In particular, we shall rigorously demonstrate that the constructed optimal feedback control is Lipschitz stable with respect to perturbations in the underlying model.

Since parameter uncertainty in a given model is practically inevitable, it is essential but challenging to *a priori* evaluate the performance of a pre-computed feedback control in a perturbed system, and to design feedback policies capable of handling model uncertainty. For instance, let us consider the following infinite-horizon stochastic control problem. Suppose $(\alpha_t)_{t\geq 0}$ is an admissible control process taking values in a *finite* action space $\mathbf{A}$, and the underlying state dynamics follows a controlled stochastic differential equation (SDE) defined as follows: $X_0^{\alpha,x} = x \in \mathbb{R}^n$, and

$$dX_t^{\alpha,x} = b(X_t^{\alpha,x}, \alpha_t)\, dt + \sigma(X_t^{\alpha,x}, \alpha_t)\, dW_t, \quad t \geq 0,$$

where $b : \mathbb{R}^n \times \mathbf{A} \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \times \mathbf{A} \to \mathbb{R}^{n\times n}$ are given coefficients. The aim of the controller is to maximize the total expected discounted reward over all admissible strategies. It is well-known that (see e.g. [19, Corollary 5.1 on p. 167] and Theorem 2.2 for more precise statements), under certain regularity assumptions, the optimal control strategy can be represented as a deterministic function $\alpha^u : \mathbb{R}^n \to \mathbf{A}$, called the optimal feedback control, which maps the current state space into the action space. Moreover, one can construct such an optimal feedback control α^u via a verification argument, which consists of solving a nonlinear Hamilton–Jacobi–Bellman (HJB) partial differential equation (PDE) arising from the dynamic programming principle for the optimal reward function u, and then performing a pointwise maximization of the associated Hamiltonian involving the function u and its derivatives $(\partial_i u, \partial_{ij} u)_{i,j=1}^n$ as follows: for

*Mathematical Institute, University of Oxford, United Kingdom (christoph.reisinger@maths.ox.ac.uk, yufei.zhang@maths.ox.ac.uk)

any given $x \in \mathbb{R}^n$,

$$\alpha^u(x) \in \arg\max_{\alpha \in \mathbf{A}} \left[\sum_{i,j=1}^{n} a^{ij}(x,\alpha)\partial_{ij}u(x) + \sum_{i=1}^{n} b^i(x,\alpha)\partial_i u(x) - c(x,\alpha)u(x) + f(x,\alpha) \right], \tag{1.1}$$

where $a(x,\alpha) = \sigma(x,\alpha)\sigma^T(x,\alpha)/2$, the functions c and f denote the discount rate and the instantaneous reward, respectively. We refer the reader to Theorems 2.2 and 3.5 for rigorous arguments of the above procedure for control problems of our interest, and to [19, Theorem 5.1 on p. 166] for a general statement.

We observe, however, that the control strategy α^u satisfying (1.1) in general is difficult to implement and unstable to parameter perturbations, which in practice would result in numerical instability of learning algorithms. Due to the finiteness of the action space $\mathbf{A}$ and the fact that arg max is a set-valued mapping, a function $\alpha^u : \mathbb{R}^n \to \mathbf{A}$ satisfying (1.1) in general is non-unique and merely measurable, and hence it is hard to follow such an irregular strategy in practice. More importantly, the discreteness of the set $\mathbf{A}$ implies that the arg max mapping is not continuous (in the sup-norm), which makes the feedback control α^u very sensitive to perturbations of the coefficients (b, σ, c, f). In other words, a slight change of the model parameters will result in a significant change of the feedback control, especially in the regions where two or more actions lead to similar performances based on the current model. Since it is difficult to determine the occurance of such regions *a priori*, it is unclear how well the control strategy α^u will perform in a real system with the perturbed coefficients $(\tilde{b}, \tilde{\sigma}, \tilde{c}, \tilde{f})$, even if $(\tilde{b}, \tilde{\sigma}, \tilde{c}, \tilde{f})$ is very close to (b, σ, c, f). See the last paragraph of Section 2 for more details on the instability of feedback controls and its practical impact on learning algorithms.

A tremendous amount of effort has been made to overcome the above difficulties, particularly in the (discrete-time) Reinforcement Learning (RL) setting (see e.g. [39]), where the agent seeks (nearly) optimal decisions in a random environment with incomplete information. Generally speaking, the controller must balance between greedily exploiting the available information to choose actions that maximize short-term rewards, and continuously exploring the environment to acquire more knowledge for long-term benefits. In particular, an entropy-regularized formulation has been proposed for solving (discrete-time) RL problems in [46, 33, 21], where the authors incorporate explorations by explicitly including the entropy of the exploration strategy in the optimization objective as a reward function, and balance exploitation and exploration by adjusting a weight imposed on this regularization term. Empirical studies (e.g. [46, 25, 33, 21]) show that such a regularized formulation leads to more robust decision making. Recently, the authors in [42, 43] extended this entropy-regularized formulation to continuous-time RL problems by using the relaxed control framework, and study the exploration/exploitation trade-off for one-dimensional linear-quadratic (LQ) control problems via explicit solutions. The relaxed control approach has then been extended to (discrete-time) RL problems with mean-field controls in [23].

In this work, we propose an exploratory framework with general exploration rewards to design robust feedback controls for continuous-time stochastic exit time problems with continuous state space and discrete action space. Our formulation extends the relaxed control approach in [42, 43] to multi-dimensional state dynamics and general exploration rewards, including Shannon's differential entropy and other commonly used regularization functions in the optimization literature (see e.g. [15, 45]); see the remark at the end of Section 3 for a detailed comparison among different exploration reward functions.

A major theoretical contribution of this work is a rigorous stability analysis of the regularized control problem and its associated feedback control strategy. Although the entropy-regularized RL formulation has demonstrated remarkable robustness in various empirical studies (e.g. [46, 25, 33, 21, 23, 43]), to the best of our knowledge, there is no published theoretical work on the Lipschitz stability of *feedback relaxed controls* with respect to parameter uncertainty (even in a discrete-time setting) nor on the Lipschitz stability of the value functions for regularized continuous-time stochastic control problems with general multi-dimensional nonlinear state dynamics. In fact, most existing results on the Lipschitz stability of feedback controls are for LQ control problems with linear state dynamics and quadratic cost functions (see e.g. [31] for discrete-time LQ problems in an ergodic setting and [6] for finite-horizon continuous-time LQ problems). The stability analysis of such problems relies heavily on the linearity of optimal feedback controls and the associated Riccati equations, and hence cannot be directly extended to general nonlinear control problems. We refer the reader also to [2, 30, 3, 4, 7, 8, 27] for the continuity of various stochastic optimization problems, including stochastic control problems and optimal stopping problems, in the underlying processes with respect to the (extended) weak topology.

In this work, we shall close the gap by providing a theoretical justification for recent RL heuristics that including an exploration reward in the optimization objective leads to more robust decision making. In particular, we shall demonstrate that the change in value functions of the regularized control problems (in the $C^{2,\beta}$-norm) depends Lipschitz-continuously on the perturbations of the model parameters, including the coefficients of the state dynamics and reward functions in the optimization objective. We shall also prove that the regularized control problem admits a Hölder continuous feedback control (cf. the original control α^u in (1.1) is merely measurable), which is Lipschitz stable (in the C^β-norm) with respect to parameter perturbations; see Theorem 4.2.

Moreover, this is the first paper which precisely quantifies the performance of a feedback control pre-computed based on a given model in a new multi-dimensional controlled dynamics with perturbed coefficients. We will prove that the gap between the suboptimal reward function achieved by the pre-computed feedback relaxed control and the optimal reward function of the perturbed relaxed control problem depends Lipschitz-continuously on the magnitude of perturbations in the coefficients (see Theorem 4.4). We also establish a first-order sensitivity equation for the value function and feedback control of the perturbed relaxed control problem (see Theorem 5.2 and Remark 5.1), which enables us to quantify the explicit dependence of the Lipschitz stability of feedback controls on the exploration parameter ε (see Theorem 5.4).

Let us briefly comment on the two main difficulties encountered in the stability analysis of feedback relaxed controls beyond those encountered in the finite-dimensional RL setting (see e.g. [20, 12, 21]) and the LQ setting (see e.g. [31, 6]). As we shall see in (3.6), the feedback relaxed control (in the present continuous setting) is defined as the pointwise maximizer of the associated Hamiltonian, which in general involves not only the value function of the regularized control problem, but also its first and second order derivatives. Hence, besides estimating the sup-norm of the value functions as in the finite-dimensional RL setting, we also need to quantify the impact of parameter uncertainty on the (first and second order) derivatives of the value functions, which are solutions to a fully nonlinear HJB PDEs. For continuous-time LQ problems, such an analysis can be greatly simplified by taking advantage of the quadratic structure of the value function, which reduces the study of HJB PDEs to that of Riccati ordinary differential equations. Such a simplification is not possible for general nonlinear control problems, which requires us to derive a precise *a priori* estimate for the derivatives of solutions to the associated fully nonlinear HJB equations.

Moreover, the Lipschitz stability and the first-order sensitivity analysis of the feedback relaxed controls also require us to establish the regularity of the HJB operator and the arg max-mapping between suitable function spaces for regularized control problems. As already pointed out in [40, 26], the fact that the HJB operator is fully nonlinear (since we allow the diffusion coefficients to be controlled) poses a significant challenge for choosing proper function spaces to simultaneously ensure the differentiability of the fully nonlinear HJB operator and the bounded invertibility of its (Fréchet) derivative, which are essential for deriving the sensitivity equations of the value functions and feedback controls (see Theorem 5.2 and Remark 5.1). Here, by taking advantage of the exploration reward functions, we demonstrate that the HJB operator and the arg max-mapping for the regularized control problem are sufficiently smooth between suitable Hölder spaces, which together with an elliptic regularity estimate leads us to the desired sensitivity results for the feedback relaxed controls; see Remark 4.1 for more details.

Finally, we establish that, as the exploration parameter tends to zero, the value function of the relaxed control problem converges monotonically to that of the classical stochastic control problem with a first-order accuracy (see Theorem 6.1). The convergence of value functions (in the $C^{2,\beta}$-norm) subsequently enables us to deduce a novel uniform result (on compact sets) for the feedback relaxed control to a pure exploitation strategy of the original control problem. We further prove an exact regularization property for a class of reward functions, which allows us to recover the pure exploitation strategy based on the feedback relaxed control *without* sending the exploration parameter to 0 (see Theorem 6.4).

We organize this paper as follows. Section 2 introduces the stochastic exit control problem, and establishes its connection to HJB equations. In Section 3, we propose a relaxed control regularization involving general exploration reward functions for the stochastic control problem, and establish the Hölder regularity of the feedback relaxed control strategy. Then, for a fixed positive exploration parameter, we prove the Lipschitz stability of the value function and feedback relaxed control with respect to parameter perturbations in Section 4, and derive their first-order sensitivity equations in Section 5. We establish the convergence of value functions and relaxed control strategies for vanishing exploration parameters in Section 6. Appendix A is devoted to the proofs of some technical results.

2 Stochastic exit time problem and HJB equation

In this section, we introduce the stochastic exit time problem of our interest, state the main assumptions on its coefficients, and recall its connection with HJB equations. We start with some useful notation which is needed frequently throughout this work.

For any given multi-index $\beta = (\beta_1, \dots, \beta_n)$ with $\beta_i \in \mathbb{N} \cup \{0\}$, $i = 1, \dots, n$, we define $|\beta| = \sum_{i=1}^n \beta_i$ and $D^\beta \phi = \frac{\partial^{|\beta|}\phi}{\partial x_1^{\beta_1} \dots \partial x_n^{\beta_n}}$. For any given open subset $\mathcal{O} \subset \mathbb{R}^n$, $k \in \mathbb{N} \cup \{0\}$, $\theta \in (0, 1]$, and function $\phi : \overline{\mathcal{O}} \to \mathbb{R}$, we define the following semi-norms:

$$[\phi]_{0;\overline{\mathcal{O}}} = \sup_{x \in \overline{\mathcal{O}}} |\phi(x)|, \quad [\phi]_{\theta;\overline{\mathcal{O}}} = \sup_{x,y \in \overline{\mathcal{O}}, x \neq y} \frac{|\phi(x) - \phi(y)|}{|x - y|^\theta}, \quad [\phi]_{k,0;\overline{\mathcal{O}}} = \sum_{|\beta|=k} [D^\beta \phi]_{0;\overline{\mathcal{O}}}, \quad [\phi]_{k,\theta;\overline{\mathcal{O}}} = \sum_{|\beta|=k} [D^\beta \phi]_{\theta;\overline{\mathcal{O}}}.$$

Then we shall denote by $C^k(\overline{\mathcal{O}})$ the space of k-times continuously differentiable functions in $\overline{\mathcal{O}}$ equipped with the norm $|\phi|_{k;\overline{\mathcal{O}}} = \sum_{m=0}^k [\phi]_{m,0;\overline{\mathcal{O}}}$, and by $C^{k,\theta}(\overline{\mathcal{O}})$ the space consisting of all functions in $C^k(\overline{\mathcal{O}})$ satisfying $[\phi]_{k,\theta;\overline{\mathcal{O}}} < \infty$, equipped with the norm $|\phi|_{k,\theta;\overline{\mathcal{O}}} = |\phi|_{k;\overline{\mathcal{O}}} + [\phi]_{k,\theta;\overline{\mathcal{O}}}$. When $k = 0$, we use $C^\theta(\overline{\mathcal{O}})$ to denote $C^{0,\theta}(\overline{\mathcal{O}})$, and use $|\cdot|_{\theta;\overline{\mathcal{O}}}$ to denote $|\cdot|_{0,\theta;\overline{\mathcal{O}}}$. We shall omit the subscript $\overline{\mathcal{O}}$ in the (semi-)norms if no confusion appears.

Finally, we shall denote by $[a^{ij}]$ the $n \times n$ matrix whose ijth-entries are given by a^{ij}, by $\mathbb{S}^n$, $\mathbb{S}^n_0$ and $\mathbb{S}^n_>$, respectively, the set of $n \times n$ symmetric, symmetric positive semi-definite and symmetric positive definite matrices, by $X \geq Y$ in $\mathbb{S}^n$ the fact that $X - Y$ is positive semi-definite. For any given $K \in \mathbb{N}$, we denote by Δ_K the probability simplex in $\mathbb{R}^K$, i.e.,

$$\Delta_K = \left\{ \lambda \in \mathbb{R}^K \;\middle|\; \textstyle\sum_{k=1}^K \lambda_k = 1, \lambda_k \geq 0, \, \forall k = 1, \dots, K \right\}. \tag{2.1}$$

Now we are ready to introduce the control problem of interest. In order to allow irregular feedback control strategies, we consider the following weak formulation of a control problem, which includes the underlying probability space as part of control strategies (see e.g. [44, 19]). See Remark 2.2 for possible extensions to stochastic control problems under strong formulation, for which the underlying probability reference system is fixed.

Definition 2.1. A 5-tuple $\pi = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P}, W)$ is said to be a reference probability system if $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ is a filtered probability space satisfying the usual condition[1], and $W = (W_t)_{t \geq 0}$ is an $\{\mathcal{F}_t\}_{t \geq 0}$-adapted n-dimensional Brownian motion. We denote by Π_{ref} the set of all reference probability systems.

Now let $\mathcal{O}$ be a given bounded domain in $\mathbb{R}^n$, i.e., a bounded connected open subset of $\mathbb{R}^n$. The aim of the controller is to maximize the expected discounted reward up to the first exit time of a controlled dynamics from the domain $\mathcal{O}$. More precisely, let $\pi = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P}, W) \in \Pi_{\text{ref}}$ be a given reference probability system, and $\mathcal{A}_\pi$ be the set of $\{\mathcal{F}_t\}_{t \geq 0}$-progressively measurable processes α taking values in a finite set $\mathbf{A}$. For any given initial state $x \in \mathbb{R}^n$, and control $\alpha \in \mathcal{A}_\pi$, we consider the controlled dynamics $X^{\alpha,x}$ satisfying the following SDE: $X_0^{\alpha,x} = x$ and

$$dX_t^{\alpha,x} = b(X_t^{\alpha,x}, \alpha_t)\, dt + \sigma(X_t^{\alpha,x}, \alpha_t)\, dW_t, \quad t \in (0, \infty), \tag{2.2}$$

where $b : \mathbb{R}^n \times \mathbf{A} \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \times \mathbf{A} \to \mathbb{R}^{n \times n}$ are given Lipschitz continuous functions (see (H.1) for precise conditions), and denote by $\tau^{\alpha,x} := \inf\{t \geq 0 \mid X_t^{\alpha,x} \notin \mathcal{O}\}$ the first exit time of the dynamics $X^{\alpha,x}$ from the domain $\mathcal{O}$,[2] and by $(\Gamma_t^{\alpha,x})_{t \in [0, \tau^{\alpha,x}]}$ the controlled discount factor: $\Gamma_t^{\alpha,x} := \exp\left(-\int_0^t c(X_s^{\alpha,x}, \alpha_s)\, ds\right)$ for all $t \in [0, \tau^{\alpha,x}]$. Then, for each given $x \in \overline{\mathcal{O}}$, we shall consider the following value function:

$$v(x) = \sup_{\pi \in \Pi_{\text{ref}}} \sup_{\alpha \in \mathcal{A}_\pi} \mathbb{E}^{\mathbb{P}}\left[\int_0^{\tau^{\alpha,x}} \Gamma_s^{\alpha,x} f(X_s^{\alpha,x}, \alpha_s)\, ds + g(X_{\tau^{\alpha,x}}^{\alpha,x}) \Gamma_{\tau^{\alpha,x}}^{\alpha,x} \right], \tag{2.3}$$

[1] We say $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ satisfies the usual condition if $(\Omega, \mathcal{F}, \mathbb{P})$ is complete, $\mathcal{F}_0$ contains all the $\mathbb{P}$-null sets in $\mathcal{F}$, and $\{\mathcal{F}_t\}_{t \geq 0}$ is right continuous.

[2] Note that, if $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ is a filtered probability space satisfying the usual condition, $(X_t)_{t \geq 0}$ is an $\{\mathcal{F}_t\}_{t \geq 0}$-progressively measurable continuous process, and $\mathcal{O}$ is an open subset of $\mathbb{R}^n$, then the first exit time $\tau = \inf\{t \geq 0 \mid X_t^{\alpha,x} \notin \mathcal{O}\}$ is an $\{\mathcal{F}_t\}_{t \geq 0}$-stopping time; see [44, Example 3.3 on p. 24].

where the functions f and g denote, respectively, the running reward and the exit reward.

Throughout this work, we shall perform the analysis under the following assumptions on the coefficients:

H.1. *Let $n, K \in \mathbb{N}$, $\mathcal{K} = \{1, \dots, K\}$, $\boldsymbol{A}$ is a set of cardinality K, i.e., $\boldsymbol{A} = \{\boldsymbol{a}_k\}_{k\in\mathcal{K}}$, and $\mathcal{O}$ be a bounded domain in $\mathbb{R}^n$. There exist constants $\nu, \Lambda > 0$, $\theta \in (0,1]$ such that the boundary $\partial\mathcal{O}$ of $\mathcal{O}$ is of class $C^{2,\theta}$, $g \in C^{2,\theta}(\overline{\mathcal{O}})$, and the functions $b : \mathbb{R}^n \times \boldsymbol{A} \to \mathbb{R}^n$, $\sigma : \mathbb{R}^n \times \boldsymbol{A} \to \mathbb{R}^{n\times n}$, $c : \overline{\mathcal{O}} \times \boldsymbol{A} \to [0,\infty)$ and $f : \overline{\mathcal{O}} \times \boldsymbol{A} \to \mathbb{R}$ satisfy the following conditions: for each $k \in \mathcal{K}$,*

$$\sigma(x, \boldsymbol{a}_k)\sigma^T(x, \boldsymbol{a}_k) \geq \nu I_n, \quad \text{for all } x \in \mathbb{R}^n, \tag{2.4}$$

$$\sum_{i,j} |\sigma^{ij}(\cdot, \boldsymbol{a}_k)|_{0,1;\mathbb{R}^n} + \sum_i |b^i(\cdot, \boldsymbol{a}_k)|_{0,1;\mathbb{R}^n} + |c(\cdot, \boldsymbol{a}_k)|_{\theta;\overline{\mathcal{O}}} + |f(\cdot, \boldsymbol{a}_k)|_{\theta;\overline{\mathcal{O}}} \leq \Lambda. \tag{2.5}$$

Remark 2.1. The Lipschitz continuity of b and σ on $\mathbb{R}^n$ ensures that, for any given $\pi \in \Pi_{\text{ref}}$, $\alpha \in \mathcal{A}_\pi$ and $x \in \mathbb{R}^n$, the controlled SDE (2.2) admits a unique strong solution. Moreover, the non-degeneracy of σ on $\mathbb{R}^n$ ensures that SDEs with non-Lipschitz feedback controls admit a weak solution (cf. Theorems 2.2 and 3.5); see also Lemma 3.1.

As shown in [22, Lemma 6.38], the fact that $\partial\mathcal{O}$ is of class $C^{2,\theta}$ ensures that a function in $C^{2,\theta}(\overline{\mathcal{O}})$ has boundary values in $C^{2,\theta}(\partial\mathcal{O})$, and conversely, any function $\phi \in C^{2,\theta}(\partial\mathcal{O})$ can be extended to a function in $C^{2,\theta}(\overline{\mathcal{O}})$. Hence, one can introduce a boundary norm $|\cdot|_{2,\theta;\partial\mathcal{O}}$ for the space $C^{2,\theta}(\partial\mathcal{O})$, such that for any given $\phi \in C^{2,\theta}(\partial\mathcal{O})$, $|\phi|_{2,\theta,\partial\mathcal{O}} = \inf_\Phi |\Phi|_{2,\theta;\overline{\mathcal{O}}}$, where $\Phi \in C^{2,\theta}(\overline{\mathcal{O}})$ is a global extension of ϕ to $\overline{\mathcal{O}}$. The space $C^{2,\theta}(\partial\mathcal{O})$ equipped with the norm $|\cdot|_{2,\theta;\partial\mathcal{O}}$ is a Banach space (see e.g. the discussions on page 94 in [22]).

To simplify the presentation, we study exit time control problems with Hölder continuous coefficients in this work and analyze classical solutions of associated elliptic HJB equations. Similar results, including the characterization and Lipchitz stability of feedback relaxed controls in Sections 3 and 4, can be obtained for finite horizon control problems with measurable coefficients, whose corresponding parabolic HJB equations admit weak solutions in suitable Sobolev spaces (see [41] for the well-posedness of weak solutions to parabolic HJB equations and [29, Theorem 1 on p. 122] for a generalized Itô's formula). The first-order sensitivity analysis in Section 5 in general can only be performed for classical solutions in Hölder spaces; see Remark 4.1 for details.

The rest of this section is devoted to the connection between the stochastic exit time problem and a Hamilton-Jacobi-Bellman (HJB) boundary value problem, which plays an essential role in the construction of feedback control strategies. More precisely, we now consider the following HJB equation with inhomogeneous Dirichlet boundary data:

$$F_0[u] \coloneqq H_0(\mathbf{L}u + \mathbf{f}) = 0 \quad \text{in } \mathcal{O}, \qquad u = g \quad \text{on } \partial\mathcal{O}, \tag{2.6}$$

where $H_0 : \mathbb{R}^K \to \mathbb{R}$ is the pointwise maximum function, i.e., $H_0(x) = \max_{k\in\mathcal{K}} x_k$ for all $x = (x_1, \dots, x_K)^T \in \mathbb{R}^K$, $\mathbf{f} : \overline{\mathcal{O}} \to \mathbb{R}^K$ is the function satisfying $\mathbf{f}(x) = (f(x, \mathbf{a}_k))_{k\in\mathcal{K}}$ for all $x \in \overline{\mathcal{O}}$, and $\mathbf{L} = (\mathcal{L}_k)_{k\in\mathcal{K}}$ is a family of elliptic operators satisfying for all $k \in \mathcal{K}$, $\phi \in C^2(\mathcal{O})$, $x \in \mathcal{O}$ that

$$\mathcal{L}_k\phi(x) \coloneqq a_k^{ij}(x)\partial_{ij}\phi(x) + b_k^i(x)\partial_i\phi(x) - c_k(x)\phi(x), \quad \text{with } a_k \coloneqq \tfrac{1}{2}\sigma_k\sigma_k^T. \tag{2.7}$$

Above and hereafter, when there is no ambiguity, we shall denote by $\phi_k(\cdot)$ a generic function $\phi(\cdot, \mathbf{a}_k)$ for all $k \in \mathcal{K}$, and adopt the summation convention as in [22, 16], i.e., repeated equal dummy indices indicate summation from 1 to n.

Throughout this paper, we shall focus on the classical solution $u \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ to (2.6) established in the following theorem, which subsequently enables us to characterize optimal feedback controls for (2.3).

Theorem 2.1. *Suppose (H.1) holds, and let $M = \sup_{i,j,k} |\sigma_k^{ij}|_{0;\overline{\mathcal{O}}}$. Then the Dirichlet problem (2.6) admits a unique solution $u \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$. Moreover, there exists a constant $\beta_0 = \beta_0(n, \nu, M) \in (0,1)$ and a Borel measurable function $\alpha^u : \overline{\mathcal{O}} \to \boldsymbol{A}$ such that $u \in C^{2,\min(\beta_0,\theta)}(\overline{\mathcal{O}})$ and*

$$\alpha^u(x) \in \operatorname*{arg\,max}_{\boldsymbol{a}_k\in\boldsymbol{A}} \big(\mathcal{L}_k u(x) + f_k(x)\big) \quad \forall x \in \overline{\mathcal{O}}. \tag{2.8}$$

Proof. We shall only prove the uniqueness of solutions in $C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$, since the existence of classical solutions in $C^{2,\min(\beta_0,\theta)}(\overline{\mathcal{O}})$ will be established constructively based on the relaxed control approximation

in Theorem 6.1 (see also [16, Theorem 7.5] for a proof of existence based on the method of continuity), and the existence of a Borel measurable function satisfying (2.8) follows directly from the measurable selection theorem (see [1, Theorem 18.19]).

Let $u_1, u_2 \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ be solutions to (2.6). Then for all $x \in \mathcal{O}$, we can deduce from the fundamental theorem of calculus that

$$0 = H_0(\mathbf{L}u_1(x) + \mathbf{f}(x)) - H_0(\mathbf{L}u_2(x) + \mathbf{f}(x)) = \int_0^1 h(s,x)^T \mathbf{L}(u_1 - u_2)(x)\, ds = \tilde{L}(u_1 - u_2)(x),$$

where $h : [0,T] \times \mathcal{O} \to \Delta_K$ is a measurable function, and $\tilde{L}$ denotes the elliptic operator satisfying for all $\phi \in C^2(\mathcal{O})$ and $x \in \mathcal{O}$ that $\tilde{L}\phi(x) = \eta^T(x)\mathbf{L}\phi(x)$ with $\eta(x) = \int_0^1 h(s,x)\, ds$. In particular, the function h can be chosen as the weak limit of the functions $([0,T] \times \mathcal{O} \ni (s,x) \mapsto (\nabla H_0^\varepsilon)(\mathbf{L}u_2(x) + \mathbf{f}(x) + s\mathbf{L}(u_1 - u_2)(x)) \in \Delta_K)_{\varepsilon > 0}$ in $L^2([0,T] \times \mathcal{O})$, where $(H_0^\varepsilon)_{\varepsilon>0}$ is a sequence of smooth approximations of H_0 obtained by using the standard mollification argument. Then we can easily show that $\eta(x) \in \Delta_K$ for all $x \in \mathcal{O}$, $\tilde{L}$ is a uniform elliptic operator, and $\sum_{k=1}^K \eta_k(x) c_k(x) \geq 0$ for all $x \in \mathcal{O}$. Hence the classical maximum principle (see e.g. [22, Theorem 3.7]) and $u_1 = u_2$ on $\partial\mathcal{O}$ imply that $u_1 = u_2$ on $\overline{\mathcal{O}}$, which shows that the Dirichlet problem (2.6) admits at most one solution in $C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$. □

We now present a verification result, i.e., Theorem 2.2, which shows that the classical solution to the HJB equation (2.6) is the value function (2.3), and the Borel measurable function α^u defined as in (2.8) is a feedback control of (2.3). The proof will be postponed to Appendix A, which essentially follows from Itô's formula and the existence result of weak solutions to SDEs with non-degenerate diffusion coefficients (see [32, Theorem 1]).

We first recall the definition of optimal feedback control (see e.g. [44, Definition 6.1]).

Definition 2.2. A Borel measurable function $h : \overline{\mathcal{O}} \to \mathbf{A}$ is said to be a feedback control of (2.3) if for all $x \in \overline{\mathcal{O}}$, there exists $\pi^x = (\Omega^x, \mathcal{F}^x, \{\mathcal{F}_t^x\}_{t\geq 0}, \mathbb{P}^x, W) \in \Pi_{\text{ref}}$, and an $\{\mathcal{F}_t^x\}_{t\geq 0}$-progressively measurable continuous process $(X_t^x)_{t\geq 0}$, such that $X_0^x = x$, and for $\mathbb{P}^x$-a.s. that

$$dX_t^{h,x} = b(X_t^{h,x}, h(X_t^{h,x}))\, dt + \sigma(X_t^{h,x}, h(X_t^{h,x}))\, dW_t \quad \forall t \in [0, \tau^x], \tag{2.9}$$

and $\int_0^{\tau^{h,x}} \big(|b(X_s^{h,x}, h(X_s^{h,x}))| + |\sigma(X_s^{h,x}, h(X_s^{h,x}))|^2\big)\, ds < \infty$, where $\tau^{h,x} := \inf\{t \geq 0 \mid X_t^{h,x} \notin \mathcal{O}\}$. A feedback control h is said to be optimal if we have for all $x \in \overline{\mathcal{O}}$ that $v(x) = J(x,h)$, where

$$J(x,h) := \mathbb{E}^{\mathbb{P}^x}\bigg[\int_0^{\tau^{h,x}} \Gamma_s^{h,x} f(X_s^{h,x}, h(X_s^{h,x}))\, ds + g(X_{\tau^{h,x}}^{h,x}) \Gamma_{\tau^{h,x}}^{h,x}\bigg], \tag{2.10}$$

and $\Gamma_t^{h,x} = \exp\big(-\int_0^t c(X_s^{h,x}, h(X_s^{h,x}))\, ds\big)$ for all $t \in [0, \tau^{h,x}]$.

Theorem 2.2. *Suppose (H.1) holds. Let $v : \overline{\mathcal{O}} \to \mathbb{R}$ be the value function defined as in (2.3), $u \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ be the solution to the Dirichlet problem (2.6), and $\alpha^u : \overline{\mathcal{O}} \to \boldsymbol{A}$ be a Borel measurable function satisfying (2.8). Then we have $u(x) = v(x)$ for all $x \in \overline{\mathcal{O}}$, and α^u is an optimal feedback control of (2.3).*

Remark 2.2. As shown in Theorem 2.2, by considering a weak formulation of the stochastic control problem (2.3) with reference probability systems varying in Π_{ref}, we can rigorously demonstrate that a measurable function α^u satisfying (2.8) is indeed an optimal feedback control strategy.

One can also consider stochastic exit time problems under a strong formulation, for which we first fix a reference probability system $\pi = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P}, W)$, and the agent only maximizes the reward functional over all admissible control processes in $\mathcal{A}_\pi$. It has been shown in [14, Theorem 2.1] that, if we assume (H.1) and $c > 0$ on $\bar{\mathcal{O}} \times \mathbf{A}$, then (2.6) satisfies the strong comparison principle i.e., a comparison result for semicontinuous viscosity solutions. In particular, (H4) in [14] is satisfied since $\partial\mathcal{O} \in C^{2,\theta}$ enjoys the exterior ball condition, and (H5) in [14] is satisfied with $\Gamma_{\text{out}} = \partial\mathcal{O}$ due to the uniform ellipticity condition (2.4). The strong comparison principle further enables us to show that the value function of the stochastic control problem (under the strong formulation) is the unique continuous viscosity solution to (2.6); see [5, Theorem 3.1]. Since the classical solution u is a viscosity solution of (2.6), we see it is the value function of the stochastic control problem (under the strong formulation), and the strategy α^u defined in (2.8) will lead to the optimal reward. Hence, we can still view the function α^u as an optimal feedback control.

We reiterate that, due to the fact that arg max is a set-valued mapping, the feedback control strategy (2.8) in general is non-unique, discontinuous, and sensitive to the perturbation of the coefficients. For instance, let $K = 2$, and consider the set $G = \{x \in \mathcal{O} \mid (\mathcal{L}_1 - \mathcal{L}_2)u(x) + (f_1 - f_2)(x) = 0\}$ at whose boundary the optimal control α^u in (2.8) could have a jump discontinuity. Except for the trivial case where α^u is a constant on $\mathcal{O}$, one can easily deduce from the connectedness of $\mathcal{O}$, the fact that $u \in C^2(\mathcal{O})$, and the continuity of the coefficients that the set G is non-empty. Since the boundary of the level set G can have poor regularity, we see the feedback control α^u in general is merely Borel measurable, which introduces a substantial difficulty to follow the optimal control in practice. Moreover, the discontinuity of α^u also implies that a small perturbation of the coefficients could lead to a significant difference of α^u in the sup-norm, especially near the boundary of the set G. It is well-known (see e.g. [9, Section 6.4.2] and [24, Figure 4]) that such an instability of feedback controls would result in a numerical instability of the learning process, i.e., the approximate policies generated by an iterative learning algorithm may change subsequently from one iteration to the next, and eventually oscillate among several far-from-optimal policies.

3 Relaxation of stochastic exit time problem

In this section, we propose a relaxation of the stochastic exit time problem (2.3), which extends the ideas used in [42] to control problems with multi-dimensional controlled dynamics and general exploration reward functions. As we shall see shortly, the relaxed control problem has a Hölder continuous feedback control strategy, and enjoys better stability with respect to perturbation of the coefficients.

The following technical lemma is essential for the formulation of relaxed control problems with multi-dimensional dynamics, whose proof is included in Appendix A.

Lemma 3.1. *Suppose (H.1) holds. Then there exist unique functions $\tilde{b} : \mathbb{R}^n \times \Delta_K \to \mathbb{R}^n$ and $\tilde{\sigma} : \mathbb{R}^n \times \Delta_K \to \mathbb{S}^n_>$ such that it holds for all $x \in \mathbb{R}^n$, $\lambda \in \Delta_K$ that*

$$\tilde{b}(x,\lambda) = \sum_{k=1}^K b(x, \boldsymbol{a}_k)\lambda_k, \quad \tilde{\sigma}(x,\lambda)\tilde{\sigma}(x,\lambda)^T = \sum_{k=1}^K \sigma(x, \boldsymbol{a}_k)\sigma(x, \boldsymbol{a}_k)^T\lambda_k.$$

Moreover, it holds for all $x \in \mathbb{R}^n$, $\lambda \in \Delta_K$ that $\tilde{\sigma}(x,\lambda) \geq \sqrt{\nu} I_n$ and $\sum_{i,j} |\tilde{\sigma}^{ij}(\cdot,\lambda)|_{0,1} + \sum_i |\tilde{b}^i(\cdot,\lambda)|_{0,1} < \infty$.

We now proceed to introduce the relaxation of the exit time problem (2.3). Roughly speaking, instead of seeking the optimal feedback action, which maps the current state to *a specific action* in the space $\mathbf{A}$, we seek the optimal feedback control distribution, which is a deterministic mapping from the current state to *a probability measure* over the space $\mathbf{A}$, i.e., $\lambda^* : \mathcal{O} \to \mathcal{P}(\mathbf{A})$. Once such a mapping is determined, at each given state, the agent will execute the control by sampling a control action based on the distribution $\lambda^*(x)$. We refer the reader to [42] for a more detailed derivation of the following regularized control problem (3.6) in a one-dimensional setting. Note that the fact that $\mathbf{A}$ has cardinality $K < \infty$ enables us to identify the space of probability measures over $\mathbf{A}$ as the probability simplex Δ_K.

More precisely, let $\pi = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P}, W) \in \Pi_{\text{ref}}$ be a given reference probability system, and $\mathcal{M}_\pi$ be the set of $\{\mathcal{F}_t\}_{t\geq 0}$-progressively measurable processes λ taking values in the set Δ_K. Suppose that (H.1) holds, for any given initial state $x \in \mathbb{R}^n$, and control $\lambda \in \mathcal{M}_\pi$, we consider the controlled diffusion process $X^{\lambda,x}$ satisfying the following SDE: $X_0^{\lambda,x} = x$ and

$$dX_t^{\lambda,x} = \tilde{b}(X_t^{\lambda,x}, \lambda_t)\, dt + \tilde{\sigma}(X_t^{\lambda,x}, \lambda_t)\, dW_t, \quad t \in (0,\infty), \tag{3.1}$$

where $\tilde{b} : \mathbb{R}^n \times \Delta_K \to \mathbb{R}^n$ and $\tilde{\sigma} : \mathbb{R}^n \times \Delta_K \to \mathbb{S}^n_>$ are the functions defined in Lemma 3.1. We further introduce the first exit time of $X^{\lambda,x}$ from the domain $\mathcal{O}$ defined as $\tau^{\lambda,x} := \inf\{t \geq 0 \mid X_t^{\lambda,x} \notin \mathcal{O}\}$, and the controlled discount factor $\Gamma_t^{\lambda,x} := \exp\left(-\int_0^t \sum_{k=1}^K c(X_s^{\lambda,x}, \mathbf{a}_k)\lambda_s^k\, ds\right)$ for all $t \in [0, \tau^{\lambda,x}]$.

Now let $\rho : \mathbb{R}^K \to \mathbb{R} \cup \{\infty\}$ be a given exploration reward function satisfying $\rho < \infty$ on Δ_K (precise conditions will be specified in (H.2)). For any given relaxation parameter $\varepsilon > 0$, we consider the following value function: for each $x \in \overline{\mathcal{O}}$,

$$v^\varepsilon(x) = \sup_{\pi\in\Pi_{\text{ref}}} \sup_{\lambda\in\mathcal{M}_\pi} \mathbb{E}^{\mathbb{P}}\left[\int_0^{\tau^{\lambda,x}} \Gamma_s^{\lambda,x}\left(\sum_{k=1}^K f(X_s^{\lambda,x}, \mathbf{a}_k)\lambda_s^k - \varepsilon\rho(\lambda_s)\right) ds + g(X_{\tau^{\lambda,x}}^{\lambda,x})\Gamma_{\tau^{\lambda,x}}^{\lambda,x}\right]. \tag{3.2}$$

Note that the exploration reward function ρ plays a crucial role in the above relaxed control regularization. If we set the exploration reward function $\rho \equiv 0$ or the relaxation parameter $\varepsilon = 0$, then one can show that Dirac measures supported on the optimal strategies of the original control problem (2.8) (see α^u defined as in (2.8)) are optimal control distributions of the relaxed control problem (3.2), and the value function v in (2.3) will be equal to the value function v^ε in (3.2) (see Theorems 6.1 and 6.4). Hence, to achieve the stability of the optimal control strategy for the relaxed control problem (3.2), we shall impose the following condition on the reward function ρ:

H.2. *There exists a convex function $H \in C^2(\mathbb{R}^K)$ and a constant $c_0 > 0$, depending on K, such that for all $x, y \in \mathbb{R}^K$, we have $H(x) - c_0 \leq \max_{k\in\mathcal{K}} x_k \leq H(x)$ and $\rho(y) = \sup_{z\in\mathbb{R}^K} \big(z^T y - H(z)\big)$.*

We remark that (H.2) is satisfied by most commonly used reward functions, including Shannon's differential entropy proposed in [46, 25, 33, 21, 42]. We refer the reader to the discussion at the end of this section for a detailed comparison of different reward functions.

Given a function $H : \mathbb{R}^K \to \mathbb{R}$, we define for each $\varepsilon \geq 0$ the function $H_\varepsilon : \mathbb{R}^K \to \mathbb{R}$ such that for all $x = (x_1, \ldots, x_K)^T \in \mathbb{R}^K$,

$$H_\varepsilon(x) = \begin{cases} \varepsilon H(\varepsilon^{-1} x), & \varepsilon > 0, \\ \max\{x_1, \ldots, x_K\}, & \varepsilon = 0. \end{cases} \tag{3.3}$$

Note that $(H_\varepsilon)_{\varepsilon\geq 0}$ are convex functions if H is a convex function. The next lemma follows directly from (H.2) and standard arguments in convex analysis, whose proof will be given in Appendix A for completeness.

Lemma 3.2. *Suppose (H.2) holds, and let $(H_\varepsilon)_{\varepsilon\geq 0}$ be defined as in (3.3). Then we have that*

(1) the function $\rho : \mathbb{R}^K \to \mathbb{R} \cup \{\infty\}$ is convex on $\mathbb{R}^K$, continuous relative to Δ_K, and satisfies that $\rho(y) \in [-c_0, 0]$ for all $y \in \Delta_K$ and $\rho(y) = \infty$ for all $y \in (\Delta_K)^c$,

(2) it holds for all $x \in \mathbb{R}^K$ and $\varepsilon > 0$ that $H_\varepsilon(x) - \varepsilon c_0 \leq H_0(x) \leq H_\varepsilon(x)$, $H_\varepsilon(x) = \max_{y\in\Delta_K} \big(y^T x - \varepsilon\rho(y)\big)$, and $(\nabla H_\varepsilon)(x) = \arg\max_{y\in\Delta_K} \big(y^T x - \varepsilon\rho(y)\big)$. Consequently, we have for all $x, y \in \mathbb{R}^K$ and $\varepsilon > 0$ that $|H_\varepsilon(x) - H_\varepsilon(y)| \leq |x - y|$.

We proceed to study the corresponding HJB equation of the relaxed control problem (3.2), which plays a crucial role in our subsequent analysis. For each $\lambda = (\lambda^1, \ldots, \lambda^K)^T \in \Delta_K$, let $f^\lambda : \mathcal{O} \to \mathbb{R}$ be the function satisfying for all $x \in \mathcal{O}$ that $f^\lambda(x) = \sum_{k=1}^K f(x, \mathbf{a}_k)\lambda^k = \lambda^T \mathbf{f}(x)$ (with $\mathbf{f}$ defined as in (2.6)), and $\mathcal{L}^\lambda$ be the elliptic operator satisfying for all $\phi \in C^2(\mathcal{O})$ and $x \in \mathcal{O}$ that

$$\begin{aligned} \mathcal{L}^\lambda \phi(x) &= \frac{1}{2}\big(\tilde\sigma(x,\lambda)\tilde\sigma^T(x,\lambda)\big)^{ij}\partial_{ij}\phi(x) + \tilde b^i(x,\lambda)\partial_i\phi(x) - \left(\sum_{k=1}^K c(x,\mathbf{a}_k)\lambda^k\right)\phi(x) \\ &= \sum_{k=1}^K \left(\frac{1}{2}\big(\sigma(x,\mathbf{a}_k)\sigma^T(x,\mathbf{a}_k)\big)^{ij}\partial_{ij}\phi(x) + b^i(x,\mathbf{a}_k)\partial_i\phi(x) - c(x,\mathbf{a}_k)\phi(x)\right)\lambda^k = \lambda^T \mathbf{L}\phi(x), \end{aligned} \tag{3.4}$$

where we have used the definition of the elliptic operators $\mathbf{L} = (\mathcal{L}_k)_{k\in\mathcal{K}}$ (cf. (2.7)), and the definition of the functions $\tilde b$ and $\tilde\sigma$ (cf. Lemma 3.1).

Since the diffusion coefficient of SDE (3.1) is non-degenerate (see Lemma 3.1) and all coefficients of the relaxed control problem (3.2) are continuous on $\overline{\mathcal{O}} \times \Delta_K$, a formal application of the dynamic programming principle (see e.g. [19, 13] and references within) enables us to associate the relaxed control problem (3.2) with the following HJB equation:

$$\max_{\lambda\in\Delta_K} \big[\mathcal{L}^\lambda u^\varepsilon + f^\lambda - \varepsilon\rho(\lambda)\big] = 0 \quad \text{in } \mathcal{O}, \qquad u^\varepsilon = g \quad \text{on } \partial\mathcal{O}.$$

Moreover, (3.4) and Lemma 3.2(2) imply that the above Dirichlet problem is equivalent to

$$F_\varepsilon[u^\varepsilon] := H_\varepsilon(\mathbf{L}u^\varepsilon + \mathbf{f}) = 0 \quad \text{in } \mathcal{O}, \qquad u^\varepsilon = g \quad \text{on } \partial\mathcal{O}, \tag{3.5}$$

where the function H_ε is defined as in (3.3), and $\mathbf{L}$, $\mathbf{f}$ are defined as those in (2.6).

In order to rigorously justify the connection between (3.2) and (3.5), we establish the well-posedness of classical solutions to (3.5) in Theorem 3.4, and then prove a verification result in Theorem 3.5.

We need the following proposition, which gives an *a priori* estimate of classical solutions to (3.5). We postpone the proof to Appendix A, which adapts the technique in [16, Theorem 7.5 on p. 127] to HJB equations with compact control sets, and reduces the problem to an *a priori* estimate for HJB equations involving only principal terms.

Proposition 3.3. *Suppose (H.1) and (H.2) hold, and let* $M = \sup_{i,j,k} |\sigma_k^{ij}|_{0;\overline{\mathcal{O}}}$. *Then there exists a constant* $\beta_0 = \beta_0(n, \nu, M) \in (0, 1)$, *such that it holds for all* $\beta \in (0, \min(\beta_0, \theta)]$ *that, if* $u^\varepsilon \in C^{2,\beta}(\overline{\mathcal{O}})$ *is a solution to the Dirichlet problem (3.5) with parameter* $\varepsilon > 0$, *then* u^ε *satisfies the estimate that* $|u^\varepsilon|_{2,\beta} \le C(|g|_{2,\beta} + \varepsilon c_0 + 1)$, *where the constant* C *depends only on* n, ν, Λ, β *and* $\mathcal{O}$.

Theorem 3.4. *Suppose (H.1) and (H.2) hold, let* $\varepsilon > 0$ *and* $M = \sup_{i,j,k} |\sigma_k^{ij}|_{0;\overline{\mathcal{O}}}$. *Then the Dirichlet problem (3.5) admits a unique solution* $u^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$. *Moreover, there exists a constant* $\beta_0 = \beta_0(n, \nu, M) \in (0, 1)$ *and a unique function* $\lambda^{u^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$ *such that* $u^\varepsilon \in C^{2,\min(\beta_0,\theta)}(\overline{\mathcal{O}})$, $\lambda^{u^\varepsilon} \in C^{\min(\beta_0,\theta)}(\overline{\mathcal{O}}, \mathbb{R}^K)$ *and*

$$\lambda^{u^\varepsilon}(x) = \arg\max_{\lambda \in \Delta_K} \left(\mathcal{L}^\lambda u^\varepsilon(x) + f^\lambda(x) - \varepsilon \rho(\lambda) \right) = (\nabla H_\varepsilon)(\boldsymbol{L} u^\varepsilon(x) + \boldsymbol{f}(x)) \quad \forall x \in \overline{\mathcal{O}}. \tag{3.6}$$

In particular, one can take the same constant β_0 *as in Proposition 3.3.*

Proof. One can deduce by similar arguments as those for Theorem 2.1 and the classical maximum principle that (3.5) admits a unique classical solution in $C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$. Moreover, by using the *a priori* bound of classical solutions in Proposition 3.3, we can establish the existence and regularity of the classical solution u^ε to (3.5) based on the method of continuity; see [16, Theorem 5.1 on p. 116].

Now let $u^\varepsilon \in C^{2,\beta}(\overline{\mathcal{O}})$ be the solution to (3.5) with some $\beta \in (0, \theta]$. The continuity of $\mathcal{L}^\lambda, f^\lambda$ and ρ on Δ_K, and Lemma 3.2(2) ensure that the function λ^{u^ε} is well-defined on $\overline{\mathcal{O}}$, and has the expression $\lambda^{u^\varepsilon} = (\nabla H_\varepsilon)(\mathbf{L}u^\varepsilon + \mathbf{f})$. Note that, it holds for any given $\phi_1, \phi_2 \in C^\beta(\overline{\mathcal{O}})$ that $\phi_1\phi_2 \in C^\beta(\overline{\mathcal{O}})$. Hence the Hölder continuity of the coefficients (see (H.1)) implies that $\mathbf{L}u^\varepsilon + \mathbf{f} \in C^\beta(\overline{\mathcal{O}}, \mathbb{R}^K)$. We can then easily deduce from the local Lipschitz continuity of $\nabla H_\varepsilon : \mathbb{R}^K \to \mathbb{R}^K$ that $\lambda^{u^\varepsilon} \in C^\beta(\overline{\mathcal{O}}, \mathbb{R}^K)$. □

The next theorem shows that the function (3.6) is an optimal feedback control of (3.2), which is defined similarly to Definition 2.2. The proof of this statement is similar to that of Theorem 2.2 and hence omitted.

Theorem 3.5. *Suppose (H.1) and (H.2) hold. Let* $\varepsilon > 0$, $v^\varepsilon : \overline{\mathcal{O}} \to \mathbb{R}$ *be the value function defined as in (3.2),* $u^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ *be the solution to the Dirichlet problem (3.5), and* $\lambda^{u^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$ *be the function defined as in (3.6). Then* $u^\varepsilon(x) = v^\varepsilon(x)$ *for all* $x \in \overline{\mathcal{O}}$, *and* λ^{u^ε} *is an optimal feedback control of (3.2).*

Remark 3.1. Theorem 3.4 shows that the feedback control λ^{u^ε} is uniquely defined and Hölder continuous. This improved regularity makes it easier to implement the relaxed control λ^{u^ε} in practice, compared to the original (merely measurable) feedback control α^u (cf. Theorem 2.1).

We end this section with a remark about possible choices of reward functions. Generally speaking, we shall choose a reward function ρ whose generating function H and its gradient ∇H can be efficiently evaluated, such that one can design an efficient algorithm to solve the relaxed control problem (3.2) (see e.g. [46, 25, 33, 21, 26]). A common choice of reward functions in the literature is the following entropy-type reward function (see e.g. [28, 35, 36, 42]):

$$\rho_{\text{en}}(y) = \begin{cases} \sum_{k=1}^K y_k \ln(y_k), & y \in \Delta_K, \\ \infty, & y \in (\Delta_K)^c, \end{cases}$$

whose generating function is $H_{\text{en}}(x) = \ln \sum_{k=1}^K \exp(x_k)$, $x \in \mathbb{R}^K$. One can show that $H_{\text{en}} \in C^\infty(\mathbb{R}^K) \cap C^{2,1}(\mathbb{R}^K)$, and it satisfies (H.2) with $c_0 = \ln K$ (see e.g. [36]).

The advantage of the entropy reward function is that both H_{en} and ∇H_{en} are given in closed form, and they can be naturally extended to continuous action spaces $\mathbf{A}$ (see e.g. [42]). However, it is important to notice that the evaluation of H_{en} and ∇H_{en} involves exponentials. Hence, when the relaxation parameter ε is small, a naive implementation of iterative algorithms for solving (3.5), which in general involves evaluating

the value and inverse of H_{en} and ∇H_{en} at a large argument $z = (\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x))/\varepsilon \in \mathbb{R}^K$ with $x \in \mathcal{O}$, may lead to unreliable results due to unstable floating-point arithmetic; see [10, Example 4.2] and [11] for more details. Moreover, since $\nabla H_{\text{en}}(x) \in (0,1)^K$ for all $x \in \mathbb{R}^K$, the optimal relaxed control of (3.2) may converge to the optimal control of (2.3) with a very slow rate as the relaxation parameter ε tends to zero.

Alternatively, by virtue of the fact that only the generating function H and its gradient are involved in the HJB equation (3.5) and the feedback control (3.6), we can also obtain a reward function ρ by directly constructing a K-dimensional function H based on a recursive application of smoothing functions for the two-dimensional max function. For instance, we can start with the following two-dimensional smoothing functions (see e.g. [15, 45]): for $x = (x_1, x_2)^T \in \mathbb{R}^2$,

$$H_{\text{chks}}(x) = \frac{\sqrt{(x_1 - x_2)^2 + 1} + x_1 + x_2}{2}, \tag{3.7}$$

$$H_{\text{zang}}(x) = \begin{cases} x_1, & x_2 - x_1 < -1/2, \\ -\frac{1}{2}(x_1 - x_2)^4 + \frac{3}{4}(x_1 - x_2)^2 + \frac{x_1+x_2}{2} + \frac{3}{32}, & |x_1 - x_2| \le 1/2, \\ x_2, & x_2 - x_1 > 1/2. \end{cases} \tag{3.8}$$

Then, for any given $K \ge 3$, by using the fact that $\max_{k\in\mathcal{K}} x_k = \max(\max_{i\in\mathcal{K}_1} x_i, \max_{j\in\mathcal{K}_2} x_j)$, with $\mathcal{K}_1 = \{1, \dots, K_0\}$, $\mathcal{K}_2 = \{K_0 + 1, \dots, K\}$ and $K_0 = \lfloor (K+1)/2 \rfloor$, we can express the K-dimensional max function as a nested application of the two-dimensional max function and one-dimensional identity function. Hence, by replacing the two-dimensional max function with the two-dimensional smoothing function (3.7) (resp. (3.8)) in the recursive expression, we can obtain the K-dimensional smoothing function $H_{\text{chks}} \in C^\infty(\mathbb{R}^K) \cap C^{2,1}(\mathbb{R}^K)$ (resp. $H_{\text{zang}} \in C^{2,1}(\mathbb{R}^K)$). It has been shown in [10, Lemma 3.3] that for any given $K \ge 2$, both functions H_{chks} and H_{zang} satisfy (H.2) with $c_0 = (\log_2(K-1) + 1)/2$ for H_{chks}, and $c_0 = 3(\log_2(K-1) + 1)/32$ for H_{zang}.

Note that, the evaluation of H_{chks}, H_{zang} and their gradients only involves square-roots and multiplications, hence they are numerically more stable than the entropy-type smoothing H_{en} (see [10]). More importantly, since H_{zang} only modifies the function H_0 locally near the non-differentiable points, we can determine the optimal control of (2.3) precisely from the optimal control of (3.2) without sending the relaxation parameter ε to zero (see Theorem 6.4 and Remark 6.2 for details).

Figure 1 compares the functions $H_{\text{en}}, H_{\text{zang}} : \mathbb{R}^3 \to \mathbb{R}$ and the reward functions generated by them. One can clearly see from Figure 1 (left) that H_{en} substantially modifies the pointwise maximum function H_0 everywhere, while H_{zang} only performs a modification of H_0 locally near the kinks. For both functions, the difference from H_0 peaks around the the points where $\arg\max_{k\in\mathcal{K}} x_k$ is not a singleton. Such points correspond to the regions where the agent of the control problem (2.3) cannot make a clear decision based on the current model, since two or more different actions would result in a very similar reward.

Figure 1 (right) depicts the reward functions $\rho_{\text{en}}(y_1, y_2, y_3)$ and $\rho_{\text{zang}}(y_1, y_2, y_3)$ with $y_3 = 1 - y_1 - y_2$, for all $(y_1, y_2) \in \mathcal{C} := \{(y_1, y_2) \in \mathbb{R}^2 \mid 0 \le y_1, y_2 \le 1, y_1 + y_2 \le 1\}$. The point $(1/3, 1/3, 1/3)$ corresponds to the pure exploration strategy, i.e., the uniform distribution on the action space $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$, while the vertices of $\mathcal{C}$ corresponds to the pure exploitation strategy, i.e., the Dirac measures supported on some $\mathbf{a}_i \in \mathbf{A}$. Both functions achieve their minimum around the point $(1/3, 1/3, 1/3)$, which indicates that the exploration reward functions encourage the controller of the relaxed control problem to explore further, especially when it is difficult to choose a unique optimal action based on the current model.

Note that, by comparing the values of the reward functions near the point $(1/3, 1/3, 1/3)$ and near the vertices of $\mathcal{C}$, we see that ρ_{en} in general gives more rewards for exploration than ρ_{zang}. Consequently, to recover the value function and optima control of (2.3), we have to take a smaller relaxation parameter for (2.3) with ρ_{en} than that for (2.3) with ρ_{zang}, which could cause a numerical instability issue due to the exponentials in H_{en} and ∇H_{en} (see e.g. [10]).

4 Lipschitz stability of optimal feedback relaxed control

In this section, we shall fix a relaxation parameter $\varepsilon > 0$ and study the robustness of the feedback control strategy (3.6) for a relaxed control problem associated with a perturbed model. In particular, we shall show that the control strategy (3.6) admits a (locally) Lipschitz continuous dependence on the

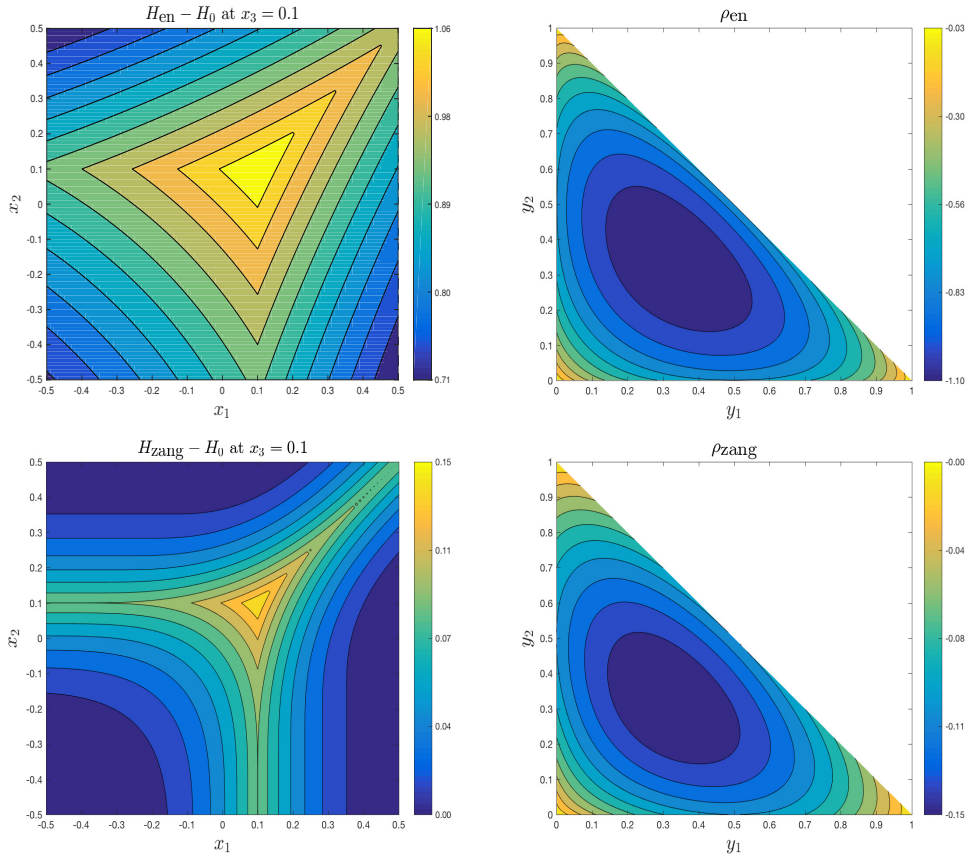

Figure 1: Comparison of H_{en} and H_{zang} and their corresponding reward functions for $K = 3$.

perturbation of the coefficients, if the reward function is generated by a function H with locally Lipschitz continuous Hessian.

We start by presenting two technical results, which are essential for our subsequent analysis. The first one is due to Nugari [34], which establishes the regularity of Nemytskij operators in Hölder spaces.

Lemma 4.1. *Let $n, K \in \mathbb{N}$, $\alpha \in (0, 1]$, $\mathcal{O} \subset \mathbb{R}^n$ be an open bounded set, $\phi : \mathbb{R}^K \to \mathbb{R}$ be a continuously differentiable function, and $\Phi : u \in C^\alpha(\overline{\mathcal{O}}, \mathbb{R}^K) \mapsto \Phi[u] \in C^\alpha(\overline{\mathcal{O}})$ be the Nemytskij operator satisfying for all $u = (u_1, \dots u_K)$ that $\Phi[u](x) = \phi(u(x))$, $x \in \overline{\mathcal{O}}$. Then Φ is well-defined, continuous and bounded. Moreover, if we further suppose $\nabla\phi$ is locally Lipschitz continuous (resp. ϕ is twice continuously differentiable), then Φ is locally Lipschitz continuous (resp. continuously differentiable with the Fréchet derivative $\Phi'[u] = (\nabla\phi)^T(u)$ for all $u \in C^\alpha(\overline{\mathcal{O}}, \mathbb{R}^K)$).*

Remark 4.1. Lemma 4.1 enables us to view the fully nonlinear HJB operator F_ε in (3.5) and the value-to-action map $u^\varepsilon \mapsto \lambda^{u^\varepsilon}$ defined in (3.6) as differentiable maps between suitable Hölder spaces, which is essential for the sensitivity analysis on the value functions and feedback relaxed controls in Section 5.

Note that in general it is not possible to perform the same first-order sensitivity analysis by interpreting the HJB operator F_ε as a map between the Sobolev space $W^{2,p}(\mathcal{O})$ and the Lebesgue space $L^q(\mathcal{O})$. In fact, since the operator $F_\varepsilon : W^{2,p}(\mathcal{O}) \to L^q(\mathcal{O})$ in general is only differentiable with $p > q$ (see [40, Theorem 13]), we see the derivative of F_ε, which is a second-order linear elliptic operator, is not bijective between $W^{2,p}(\mathcal{O})$ and $L^q(\mathcal{O})$. Consequently, we cannot apply the implicit function theorem to derive the sensitivity equation for the value function (3.2) as in Theorem 5.2.

If the operator F_ε is only semilinear, i.e., the diffusion coefficient of (2.2) is uncontrolled, then one can show that F_ε is differentiable between $W^{2,p}(\mathcal{O})$ and $L^p(\mathcal{O})$ for $1 < p < \infty$, and its derivative is a bijection between the same spaces (see [26] for the case with $p = 2$). In this case, we can extend Theorem 5.2 and study L^p-perturbation of the coefficients in (3.2).

Now we proceed to introduce a relaxed control problem with a set of perturbed coefficients satisfying the following conditions:

H.3. *Let $\nu > 0$, $\theta \in (0,1]$ be the constants in (H.1), and $\Lambda' > 0$ be a constant. The functions $\hat{b} : \mathbb{R}^n \times \boldsymbol{A} \to \mathbb{R}^n$, $\hat{\sigma} : \mathbb{R}^n \times \boldsymbol{A} \to \mathbb{R}^{n\times n}$, $\hat{c} : \overline{\mathcal{O}} \times \boldsymbol{A} \to [0,\infty)$, $\hat{f} : \overline{\mathcal{O}} \times \boldsymbol{A} \to \mathbb{R}$, and $\hat{g} : \overline{\mathcal{O}} \to \mathbb{R}$ satisfy that $\hat{g} \in C^{2,\theta}(\overline{\mathcal{O}})$, $\hat{\sigma}(x,\boldsymbol{a}_k)\hat{\sigma}^T(x,\boldsymbol{a}_k) \geq \nu I_n$ for all $(x,\boldsymbol{a}_k) \in \mathbb{R}^n \times \boldsymbol{A}$, and for all $\boldsymbol{a}_k \in \boldsymbol{A}$ that*

$$\sum_{i,j} |\hat{\sigma}^{ij}(\cdot,\boldsymbol{a}_k)|_{0,1;\mathbb{R}^n} + \sum_i |\hat{b}^i(\cdot,\boldsymbol{a}_k)|_{0,1;\mathbb{R}^n} + |\hat{c}(\cdot,\boldsymbol{a}_k)|_{\theta;\overline{\mathcal{O}}} + |\hat{f}(\cdot,\boldsymbol{a}_k)|_{\theta;\overline{\mathcal{O}}} \leq \Lambda'.$$

Let $\varepsilon > 0$ be a fixed relaxation parameter. We shall consider a perturbed control problem (2.3) with the coefficients $(\hat{b},\hat{\sigma},\hat{c},\hat{f},\hat{g})$, and its relaxation (see (3.2)) with parameter ε, whose value function is denoted as $\hat{v}^\varepsilon$. Then, by using Lemma 3.2, Theorems 3.4 and 3.5, one can verify that, under (H.2) and (H.3), the value function $\hat{v}^\varepsilon$ is the classical solution $\hat{u}^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ of the following Dirichlet problem:

$$\max_{\lambda\in\Delta_K} \big(\lambda^T(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}) - \varepsilon\rho(\lambda)\big) = H_\varepsilon\big(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}\big) = 0 \quad \text{in } \mathcal{O}, \qquad \hat{u}^\varepsilon = \hat{g} \quad \text{on } \partial\mathcal{O}, \tag{4.1}$$

where the function H_ε is defined as in (3.3), $\hat{\mathbf{f}} : \overline{\mathcal{O}} \to \mathbb{R}^K$ is the function satisfying $\hat{\mathbf{f}}(x) = (\hat{f}(x,\mathbf{a}_k))_{k\in\mathcal{K}}$ for all $x \in \overline{\mathcal{O}}$, and $\hat{\mathbf{L}} = (\hat{\mathcal{L}}_k)_{k\in\mathcal{K}}$ is a family of elliptic operators satisfying for all $k \in \mathcal{K}$, $\phi \in C^2(\mathcal{O})$, $x \in \mathcal{O}$ that

$$\hat{\mathcal{L}}_k\phi(x) := \hat{a}_k^{ij}(x)\partial_{ij}\phi(x) + \hat{b}_k^i(x)\partial_i\phi(x) - \hat{c}_k(x)\phi(x), \quad \text{with } \hat{a}_k = \tfrac{1}{2}\hat{\sigma}_k\hat{\sigma}_k^T.$$

Moreover, we can deduce from (3.6) that, the optimal feedback control of the perturbed relaxed control problem is given by

$$\hat{\lambda}^{\hat{u}^\varepsilon}(x) = \arg\max_{\lambda\in\Delta_K} \big(\lambda^T(\hat{\mathbf{L}}\hat{u}^\varepsilon(x) + \hat{\mathbf{f}}(x)) - \varepsilon\rho(\lambda)\big) = (\nabla H_\varepsilon)(\hat{\mathbf{L}}\hat{u}^\varepsilon(x) + \hat{\mathbf{f}}(x)) \quad \forall x \in \overline{\mathcal{O}}. \tag{4.2}$$

Note that Theorem 3.4 shows that the classical solution $\hat{u}^\varepsilon$ of (4.1) is in $C^{2,\beta}(\overline{\mathcal{O}})$ for some $\beta > 0$, so the above function $\hat{\lambda}^{\hat{u}^\varepsilon}$ is well-defined on $\partial\mathcal{O}$.

The following result shows the (local) Lipschitz dependence of $\hat{u}^\varepsilon - u^\varepsilon$ and $\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon}$ on perturbation of the coefficients, which demonstrates the robustness of the relaxed control problem. For notational simplicity, given the functions (b,σ,c,f,g) and $(\hat{b},\hat{\sigma},\hat{c},\hat{f},\hat{g})$ satisfying (H.1) and (H.3) respectively, we shall introduce for each $\beta \in (0,\theta]$ the following measurement of perturbations:

$$\mathcal{E}_{\mathrm{per},\beta} := \sup_{i,j,k}(|a_k^{ij} - \hat{a}_k^{ij}|_{\beta;\overline{\mathcal{O}}} + |b_k^i - \hat{b}_k^i|_{\beta;\overline{\mathcal{O}}} + |c_k - \hat{c}_k|_{\beta;\overline{\mathcal{O}}} + |f_k - \hat{f}_k|_{\beta;\overline{\mathcal{O}}}) + |g - \hat{g}|_{2,\beta;\overline{\mathcal{O}}}, \tag{4.3}$$

where $a_k = \sigma_k\sigma_k^T$, $\hat{a}_k = \hat{\sigma}_k\hat{\sigma}_k^T$ for each $k \in \mathcal{K}$.

Theorem 4.2. *Suppose (H.1), (H.2) and (H.3) hold. Let $\varepsilon > 0$, $M = \sup_{i,j,k}\max(|\sigma_k^{ij}|_{0;\overline{\mathcal{O}}}, |\hat{\sigma}_k^{ij}|_{0;\overline{\mathcal{O}}})$, $M_g = \max(|g|_{2,\theta}, |\hat{g}|_{2,\theta})$, $u^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ (resp. $\hat{u}^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$) be the solution to the Dirichlet problem (3.5) (resp. (4.1)), and $\lambda^{u^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$ (reps. $\hat{\lambda}^{\hat{u}^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$) be the function defined as in (3.6) (resp. (4.2)). Then there exists $\beta_0 = \beta_0(n,\nu,M) \in (0,1)$, such that it holds for all $\beta \in (0,\min(\beta_0,\theta)]$ that,*

$$|\hat{u}^\varepsilon - u^\varepsilon|_{2,\beta} \leq C\mathcal{E}_{\mathrm{per},\beta} \tag{4.4}$$

with the constant $\mathcal{E}_{\mathrm{per},\beta}$ defined as in (4.3), and a constant $C = C(\varepsilon,n,K,\nu,\Lambda,\Lambda',\beta,c_0,M_g,\mathcal{O})$.

If we further suppose the function $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) has a locally Lipschitz continuous Hessian, then it also holds that $|\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon}|_\beta \leq C\mathcal{E}_{\mathrm{per},\beta}$.

Proof. Throughout this proof, we shall denote by C a generic constant, which depends only on ε, n, K, ν, Λ, Λ', β, c_0, M_g and $\mathcal{O}$, and may take a different value at each occurrence.

The *a priori* estimate in Proposition 3.3 shows that there exists a constant $\beta_0 = \beta_0(n,\nu,M) \in (0,1)$, such that we have for all $\beta \in (0,\min(\beta_0,\theta)]$ the estimates $|u^\varepsilon|_{2,\beta}, |\hat{u}^\varepsilon|_{2,\beta} \leq C$. Moreover, we have by the fundamental theorem of calculus that

$$0 = H_\varepsilon\big(\mathbf{L}u^\varepsilon + \mathbf{f}\big) - H_\varepsilon\big(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}\big) = \eta^T\big(\mathbf{L}u^\varepsilon + \mathbf{f} - \hat{\mathbf{L}}\hat{u}^\varepsilon - \hat{\mathbf{f}}\big) = \eta^T\big(\mathbf{L}(u^\varepsilon - \hat{u}^\varepsilon) + (\mathbf{L} - \hat{\mathbf{L}})\hat{u}^\varepsilon + \mathbf{f} - \hat{\mathbf{f}}\big) \tag{4.5}$$

in $\mathcal{O}$, where $\eta : \overline{\mathcal{O}} \to \Delta_K$ is the function defined as $\eta := \int_0^1 (\nabla H_\varepsilon)\big(s(\mathbf{L}u^\varepsilon + \mathbf{f}) + (1-s)(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}})\big)\, ds$.

Now let $\beta \in (0, \min(\beta_0, \theta)]$ be a fixed constant. The fact that $\nabla H_\varepsilon \in C^1(\mathbb{R}^K, \Delta_K)$ (see (H.2)), the Hölder continuity of coefficients (see (H.1) and (H.3)), and the *a priori* estimates of $|u^\varepsilon|_{2,\beta}$ and $|\hat{u}^\varepsilon|_{2,\beta}$ yield the estimate that $|\eta|_\beta \le C$ (see Lemma 4.1). Then, by setting $w = u^\varepsilon - \hat{u}^\varepsilon \in C^{2,\beta}(\overline{\mathcal{O}})$, we can deduce from (4.5) that w is the classical solution to the following Dirichlet problem:

$$\eta^T \mathbf{L} w = -\eta^T\big((\mathbf{L} - \hat{\mathbf{L}})\hat{u}^\varepsilon + \mathbf{f} - \hat{\mathbf{f}}\big) \quad \text{in } \mathcal{O}, \qquad w = g - \hat{g} \quad \text{on } \partial\mathcal{O}.$$

Hence the fact that $\eta \in C^\beta(\overline{\mathcal{O}}, \Delta_K)$ and the global Schauder estimate in [22, Theorem 6.6] lead us to the estimate that

$$|w|_{2,\beta} \le C(|w|_0 + |g - \hat{g}|_{2,\beta} + |\eta^T\big((\mathbf{L} - \hat{\mathbf{L}})\hat{u}^\varepsilon + \mathbf{f} - \hat{\mathbf{f}}\big)|_\beta),$$

which, together with the maximum principle (see [22, Theorem 3.7]) and the *a priori* estimate of $|\hat{u}^\varepsilon|_{2,\beta}$, enables us to conclude that:

$$|u^\varepsilon - \hat{u}^\varepsilon|_{2,\beta} = |w|_{2,\beta} \le C(|g - \hat{g}|_{2,\beta} + |\eta^T\big((\mathbf{L} - \hat{\mathbf{L}})\hat{u}^\varepsilon + \mathbf{f} - \hat{\mathbf{f}}\big)|_\beta) \le C\mathcal{E}_{\mathrm{per},\beta},$$

with the constant $\mathcal{E}_{\mathrm{per},\beta}$ defined as in (4.3).

Now we show the stability of feedback controls. Note that (4.4) implies that

$$|\big(\mathbf{L}u^\varepsilon + \mathbf{f}\big) - \big(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}\big)|_\beta = |\mathbf{L}(u^\varepsilon - \hat{u}^\varepsilon) + (\mathbf{L} - \hat{\mathbf{L}})\hat{u}^\varepsilon + \mathbf{f} - \hat{\mathbf{f}}|_\beta \le C\mathcal{E}_{\mathrm{per},\beta}.$$

The additional assumption that $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) has a locally Lipschitz continuous Hessian implies that ∇H_ε is differentiable with locally Lipschitz continuous derivatives, which along with Lemma 4.1 shows that the Nemytskij operator $\nabla H_\varepsilon : C^\beta(\overline{\mathcal{O}}, \mathbb{R}^K) \to C^\beta(\overline{\mathcal{O}}, \mathbb{R}^K)$ is locally Lipschitz continuous. Hence there exists a constant C, such that for all perturbed coefficients $(\hat{b}, \hat{\sigma}, \hat{c}, \hat{f}, \hat{g})$ satisfying (H.3), we have

$$\begin{aligned}
&|\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon}|_\beta = |(\nabla H_\varepsilon)(\hat{\mathbf{L}}\hat{u}^\varepsilon(x) + \hat{\mathbf{f}}(x)) - (\nabla H_\varepsilon)(\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x))| \\
&\le C|\big(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}\big) - \big(\mathbf{L}u^\varepsilon + \mathbf{f}\big)|_\beta \le C\mathcal{E}_{\mathrm{per},\beta},
\end{aligned}$$

which finishes the desired (local) Lipschitz estimate. □

Remark 4.2. The assumption that $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) has a locally Lipschitz continuous Hessian is satisfied by most commonly used functions, including H_{en}, H_{chks} and H_{zang} given in Section 3. In general, if H is merely twice continuously differentiable as in (H.2), we can follow a similar argument and establish that the Hölder norm of the difference between two relaxed control strategies is continuously dependent on the Hölder norms of the perturbations in the coefficients.

Note that the Lipschitz stability result (4.4) in general does not hold for the original control problem (2.3) (or equivalently, $\varepsilon = 0$ in (3.2)). In fact, for any given $\beta \in (0,1)$, [18, Theorem 2] shows that the Nemytskij operator $\mathbf{f} \in (C^\beta(\overline{\mathcal{O}}))^K \mapsto H_0(\mathbf{f}) \in C^\beta(\overline{\mathcal{O}})$ is not continuous, which implies that there exists $(\mathbf{f}_m)_{m\in\mathbb{N}\cup\{\infty\}} \subset (C^\beta(\overline{\mathcal{O}}))^K$ such that $\lim_{m\to\infty} |\mathbf{f}_m - \mathbf{f}_\infty|_\beta = 0$ and $|H_0(\mathbf{f}_m) - H_0(\mathbf{f}_\infty)|_\beta \ge 1$ for all $m \in \mathbb{N}$. Now for each $m \in \mathbb{N}\cup\{\infty\}$, we consider the following simple HJB equation (2.6): $\Delta u_m + H_0(\mathbf{f}_m) = 0$ in $\mathcal{O}$ and $u_m = 0$ on $\partial\mathcal{O}$. Hence we have $|\Delta(u_m - u_\infty)|_\beta = |H_0(\mathbf{f}_m) - H_0(\mathbf{f}_\infty)|_\beta \ge 1$ for all $m \in \mathbb{N}$, which implies that the $C^{2,\beta}$-norm of the value function (2.3) does not depend continuously on the C^β-perturbation of the model parameters. See Theorem 5.4 for a precise quantification of ε-dependence in (4.4).

The remaining part of this section is devoted to an important application of Theorem 4.2, where we shall examine the performance of the control strategy λ^{u^ε}, computed based on the relaxed control problem with the original coefficients (b, σ, c, f, g) (see (3.6)), on a new relaxed control problem with perturbed coefficients satisfying (H.3).

We first observe that, if there exists a classical solution $\underline{u}^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ to the following problem:

$$(\lambda^{u^\varepsilon})^T(\hat{\mathbf{L}}\underline{u}^\varepsilon + \hat{\mathbf{f}}) - \varepsilon\rho(\lambda^{u^\varepsilon}) = 0 \quad \text{in } \mathcal{O}, \qquad \underline{u}^\varepsilon = \hat{g} \quad \text{on } \partial\mathcal{O}, \tag{4.6}$$

with $\hat{\mathbf{L}}$ and $\hat{\mathbf{f}}$ defined as in (4.1), then by using Itô's formula, one can easily show that the reward function $\underline{v}^\varepsilon$, resulting by implementing the Hölder continous feedback control λ^{u^ε} to the relaxed control problem

with the coefficients $(\hat{b}, \hat{\sigma}, \hat{c}, \hat{f}, \hat{g})$, coincides with the function $\underline{u}^\varepsilon$ (see e.g. Theorems 2.2 and 3.5). On the other hand, we have seen that the (optimal) value function $\hat{v}^\varepsilon$ of the perturbed relaxed control problem is the classical solution $\hat{u}^\varepsilon$ to (4.1). Hence it suffices to compare the classical solutions to (4.6) and (4.1).

The following proposition shows that (4.6) indeed admits an unique classical solution.

Proposition 4.3. *Suppose (H.1), (H.2) and (H.3) hold. Let $\varepsilon > 0$, $M = \sup_{i,j,k} |\sigma_k^{ij}|_{0;\overline{\mathcal{O}}}$, $\lambda^{u^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$ be the function defined as in (3.6), and $\beta_0 = \beta_0(n, \nu, M) \in (0,1)$ be the constant in Proposition 3.3. Then the Dirichlet problem (4.6) admits a unique solution $\underline{u}^\varepsilon \in C^{2,\min(\beta_0,\theta)}(\overline{\mathcal{O}})$.*

Proof. Let us denote $\bar{\beta}_0 = \min(\beta_0, \theta)$ throughout this proof. The uniqueness of classical solutions to (4.6) follows directly from the classical maximum principle (see [22, Theorem 3.7]). Hence we shall focus on establishing the existence and regularity of solutions to (4.6). Note that, Theorem 3.4 shows that $\lambda^{u^\varepsilon} \in C^{\bar{\beta}_0}(\overline{\mathcal{O}}, \Delta_K)$ and $u^\varepsilon \in C^{2,\bar{\beta}_0}(\overline{\mathcal{O}})$, where u^ε is the classical solution to (3.5).

We now study the function $\rho(\nabla H_\varepsilon) : x \in \mathbb{R}^K \mapsto \rho(\nabla H_\varepsilon(x)) \in \mathbb{R}$. Note that, (H.2) and (3.3) imply that $H_\varepsilon : \mathbb{R}^K \to \mathbb{R}$ is convex and differentiable. Moreover, Lemma 3.2(2) shows that the convex conjugate of H_ε, denoted by $(H_\varepsilon)^*$, is given by $(H_\varepsilon)^*(y) = \sup_{x\in\mathbb{R}^K}(x^T y - H_\varepsilon(x)) = \varepsilon\rho(y)$, for all $y \in \mathbb{R}^K$. Hence, we can deduce from [38, Theorem 23.5] (by setting $f = H_\varepsilon$ in the statement) that

$$(\varepsilon\rho)((\nabla H_\varepsilon)(x)) = (H_\varepsilon)^*((\nabla H_\varepsilon)(x)) = x^T(\nabla H_\varepsilon)(x) - H_\varepsilon(x), \quad x \in \mathbb{R}^K, \tag{4.7}$$

which implies that $(\varepsilon\rho)(\nabla H_\varepsilon) \in C^1(\mathbb{R}^K)$. Then, by using the representation $\lambda^{u^\varepsilon} = (\nabla H_\varepsilon)(\mathbf{L}u^\varepsilon + \mathbf{f})$, we can conclude that $\varepsilon\rho(\lambda^{u^\varepsilon}) = (\varepsilon\rho)\big((\nabla H_\varepsilon)(\mathbf{L}u^\varepsilon + \mathbf{f})\big) \in C^{\bar{\beta}_0}(\overline{\mathcal{O}})$.

Therefore, we can deduce from the fact that all coefficients of (4.6) are in $C^{\bar{\beta}_0}(\overline{\mathcal{O}})$, $\sum_{k=1}^K \lambda^{u^\varepsilon,k} c_k \geq 0$, and [22, Theorem 6.14] that (4.6) admits a unique solution in $C^{2,\bar{\beta}_0}(\overline{\mathcal{O}})$. □

We are ready to show that, the difference between this suboptimal reward function $\underline{v}^\varepsilon$ and the (optimal) value function $\hat{v}^\varepsilon$ of the perturbed relaxed control problem depends Lipschitz-continuously on the magnitude of perturbations in the coefficients.

Theorem 4.4. *Suppose (H.1), (H.2) and (H.3) hold. Let $\varepsilon > 0$, $M = \sup_{i,j,k} \max(|\sigma_k^{ij}|_{0;\overline{\mathcal{O}}}, |\hat{\sigma}_k^{ij}|_{0;\overline{\mathcal{O}}})$, $M_g = \max(|g|_{2,\theta}, |\hat{g}|_{2,\theta})$, and $\underline{u}^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ (resp. $\hat{u}^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$) be the solution to the Dirichlet problem (4.6) (resp. (4.1)). Then we have $\hat{u}^\varepsilon \geq \underline{u}^\varepsilon$ on $\overline{\mathcal{O}}$.*

If we further suppose the function $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) has a locally Lipschitz continuous Hessian, then there exists $\beta_0 = \beta_0(n, \nu, M) \in (0,1)$, such that for all $\beta \in (0, \min(\beta_0, \theta)]$, we have the estimate $|\hat{u}^\varepsilon - \underline{u}^\varepsilon|_{2,\beta} \leq C\mathcal{E}_{\mathrm{per},\beta}$, with the constant $\mathcal{E}_{\mathrm{per},\beta}$ defined as in (4.3), and a constant $C = C(\varepsilon, n, K, \nu, \Lambda, \Lambda', \beta, c_0, M_g, \mathcal{O})$.

Proof. Let $\lambda^{u^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$ (reps. $\hat{\lambda}^{\hat{u}^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$) be the function defined as in (3.6) (resp. (4.2)), and C be a generic constant, which depends only on ε, n, K, ν, Λ, Λ', β, c_0, M_g and $\mathcal{O}$, and may take a different value at each occurrence.

The fact $\hat{u}^\varepsilon$ solves (4.1) implies that, for all $x \in \mathcal{O}$,

$$\begin{aligned} 0 = H_\varepsilon\big(\hat{\mathbf{L}}\hat{u}^\varepsilon(x) + \hat{\mathbf{f}}(x)\big) &= \max_{\lambda\in\Delta_K} \big(\lambda^T(\hat{\mathbf{L}}\hat{u}^\varepsilon(x) + \hat{\mathbf{f}}(x)) - \varepsilon\rho(\lambda)\big) \\ &\geq (\lambda^{u^\varepsilon}(x))^T(\hat{\mathbf{L}}\hat{u}^\varepsilon(x) + \hat{\mathbf{f}}(x)) - \varepsilon\rho(\lambda^{u^\varepsilon}(x)), \end{aligned}$$

which, together with the fact that $\hat{u}^\varepsilon = \underline{u}^\varepsilon = \hat{g}$ and the classical maximum principle (see [22, Theorem 3.7]), shows that $\hat{u}^\varepsilon \geq \underline{u}^\varepsilon$ on $\overline{\mathcal{O}}$.

We now estimate $\hat{u}^\varepsilon - \underline{u}^\varepsilon$ by assuming the function $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) has a locally Lipschitz continuous Hessian. By using the definition of the optimal control $\hat{\lambda}^{\hat{u}^\varepsilon}$, we have that

$$(\hat{\lambda}^{\hat{u}^\varepsilon})^T(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}) - \varepsilon\rho(\hat{\lambda}^{\hat{u}^\varepsilon}) = 0, \quad \text{in } \mathcal{O}.$$

By subtracting (4.6) from the above equation, we have

$$\begin{aligned} 0 &= \big[(\hat{\lambda}^{\hat{u}^\varepsilon})^T(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}) - \varepsilon\rho(\hat{\lambda}^{\hat{u}^\varepsilon})\big] - \big[(\lambda^{u^\varepsilon})^T(\hat{\mathbf{L}}\underline{u}^\varepsilon + \hat{\mathbf{f}}) - \varepsilon\rho(\lambda^{u^\varepsilon})\big] \\ &= (\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon})^T(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}) + (\lambda^{u^\varepsilon})^T\hat{\mathbf{L}}(\hat{u}^\varepsilon - \underline{u}^\varepsilon) - \big(\varepsilon\rho(\hat{\lambda}^{\hat{u}^\varepsilon}) - \varepsilon\rho(\lambda^{u^\varepsilon})\big), \quad \text{in } \mathcal{O}. \end{aligned}$$

Note that, the *a priori* estimate in Proposition 3.3 shows that, under (H.1), (H.2) and (H.3), there exists a constant $\beta_0 = \beta_0(n,\nu,M) \in (0,1)$, such that we have for all $\beta \in (0,\min(\beta_0,\theta)]$ the estimates $|u^\varepsilon|_{2,\beta}, |\hat{u}^\varepsilon|_{2,\beta} \leq C$, which, along with the fact that $\nabla H_\varepsilon \in C^1(\mathbb{R}^K)$ and Lemma 4.1, implies the *a priori* bounds $|\hat{\lambda}^{\hat{u}^\varepsilon}|_\beta, |\lambda^{u^\varepsilon}|_\beta \leq C$. Hence, from any given $\beta \in (0,\min(\beta_0,\theta)]$, we can deduce from the Schauder theory in [22, Theorem 6.6] and the maximum principle in [22, Theorem 3.7] that

$$\begin{aligned} |\hat{u}^\varepsilon - \underline{u}^\varepsilon|_{2,\beta} &\leq C\big(|(\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon})^T(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}})|_\beta + |\varepsilon\rho(\hat{\lambda}^{\hat{u}^\varepsilon}) - \varepsilon\rho(\lambda^{u^\varepsilon})|_\beta\big) \\ &\leq C\big(|\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon}|_\beta + |\varepsilon\rho(\hat{\lambda}^{\hat{u}^\varepsilon}) - \varepsilon\rho(\lambda^{u^\varepsilon})|_\beta\big). \end{aligned} \tag{4.8}$$

By using the additional assumption that H has a locally Lipschitz continuous Hessian, and the identity (4.7), we can deduce that $\rho(\nabla H_\varepsilon) : \mathbb{R}^K \to \mathbb{R}$ is continuously differentiable with a locally Lipschitz continuous gradient, from which, we can obtain from Lemma 4.1 that for any $\alpha \in (0,1]$, the corresponding Nemytskij operator $(\varepsilon\rho)(\nabla H_\varepsilon) : C^\alpha(\overline{\mathcal{O}},\mathbb{R}^K) \to C^\alpha(\overline{\mathcal{O}},\mathbb{R})$ is locally Lipschitz continuous. Hence, we can obtain from (4.8) and the definitions of λ^{u^ε} and $\hat{\lambda}^{\hat{u}^\varepsilon}$ (see (3.6) and (4.2)) that

$$\begin{aligned} |\hat{u}^\varepsilon - \underline{u}^\varepsilon|_{2,\beta} &\leq C\big(|\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon}|_\beta + |(\varepsilon\rho)((\nabla H_\varepsilon)(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}})) - (\varepsilon\rho)((\nabla H_\varepsilon)(\mathbf{L}u^\varepsilon + \mathbf{f}))|_\beta\big) \\ &\leq C\big(|\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon}|_\beta + |(\hat{\mathbf{L}}\hat{u}^\varepsilon + \hat{\mathbf{f}}) - (\mathbf{L}u^\varepsilon + \mathbf{f})|_\beta\big) \\ &\leq C\big(|\hat{\lambda}^{\hat{u}^\varepsilon} - \lambda^{u^\varepsilon}|_\beta + |(\hat{\mathbf{L}} - \mathbf{L})\hat{u}^\varepsilon + \mathbf{L}(\hat{u}^\varepsilon - u^\varepsilon) + \hat{\mathbf{f}} - \mathbf{f}|_\beta\big), \end{aligned}$$

from which, we can conclude from the *a priori* bound of $|\hat{u}^\varepsilon|_{2,\beta}$ and Theorem 4.2 the desired estimate $|\hat{u}^\varepsilon - \underline{u}^\varepsilon|_{2,\beta} \leq C\mathcal{E}_{\mathrm{per},\beta}$. □

5 First-order sensitivity equations for relaxed control problems

In this section, we proceed to derive a first-order Taylor expansion for the value function and the optimal control of the relaxed control problem (3.2) with perturbed coefficients, which subsequently leads us to a first-order approximation of the optimal strategy for the perturbed problem based on the pre-computed optimal control. The sensitivity equation further enables us to quantify the explicit dependence of the Lipschitz stability result (4.4) on the relaxation parameter ε.

The following proposition establishes the Fréchet differentiability of the fully nonlinear HJB operator with inhomogeneous boundary conditions. For notational simplicity, for any given $\beta \in (0,1]$, and bounded open subset $\mathcal{O} \subset \mathbb{R}^n$ with $C^{2,\beta}$ boundary, we shall introduce the Banach space Θ^β for the coefficients:

$$\Theta^\beta = \big(C^\beta(\overline{\mathcal{O}},\mathbb{R}^{n\times n}) \times C^\beta(\overline{\mathcal{O}},\mathbb{R}^n) \times C^\beta(\overline{\mathcal{O}}) \times C^\beta(\overline{\mathcal{O}})\big)^K \times C^{2,\beta}(\overline{\mathcal{O}}) \tag{5.1}$$

equipped with the product norm $|\cdot|_{\Theta^\beta}$, and denote by $\boldsymbol{\vartheta} = ((a_k,b_k,c_k,f_k)_{k\in\mathcal{K}}, g)$ a generic element in Θ^β. We also denote by $C^{2,\beta}(\partial\mathcal{O})$ the Banach space of $C^{2,\beta}$ functions defined on $\partial\mathcal{O}$ (see Remark 2.1), and by $\tau_D : C^{2,\beta}(\overline{\mathcal{O}}) \to C^{2,\beta}(\partial\mathcal{O})$ the restriction operator on $\partial\mathcal{O}$. Furthermore, for any given Banach spaces X and Y, we denote by $B(X,Y)$ the Banach space containing all continuous linear mappings from X into Y, equipped with the operator norm.

Proposition 5.1. *Suppose (H.2) holds. Let $\varepsilon > 0$, $\beta \in (0,1]$, $\mathcal{O}$ be a bounded domain in $\mathbb{R}^n$ with $C^{2,\beta}$ boundary, $H_\varepsilon : \mathbb{R}^K \to \mathbb{R}$ be the function defined as in (3.3), Θ^β be the Banach space defined as in (5.1), and $F^\beta : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O})$ be the following HJB operator:*

$$F^\beta : (\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \mapsto F^\beta[\boldsymbol{\vartheta}, u] := (H_\varepsilon(\boldsymbol{L}^{\boldsymbol{\vartheta}}u + \boldsymbol{f}^{\boldsymbol{\vartheta}}), \tau_D(u - g^{\boldsymbol{\vartheta}})) \in C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O}),$$

where for any given $\boldsymbol{\vartheta} = ((a_k,b_k,c_k,f_k)_{k\in\mathcal{K}}, g) \in \Theta^\beta$, $\boldsymbol{f}^{\boldsymbol{\vartheta}} = (f_k)_{k\in\mathcal{K}} \in C^\beta(\overline{\mathcal{O}})^K$, $g^{\boldsymbol{\vartheta}} = g$ and $\boldsymbol{L}^{\boldsymbol{\vartheta}} = (\mathcal{L}_k^{\boldsymbol{\vartheta}})_{k\in\mathcal{K}}$ is the elliptic operators satisfying $\mathcal{L}_k^{\boldsymbol{\vartheta}}\phi = a_k^{ij}\partial_{ij}\phi + b_k^i\partial_i\phi - c_k\phi$ for all $k \in \mathcal{K}$, $\phi \in C^2(\mathcal{O})$.

Then F^β is continuously differentiable with the derivative $F^\beta : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to B(\Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}), C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O}))$ satisfying for all $(\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}})$, $\tilde{\boldsymbol{\vartheta}} \in \Theta^\beta$ and $v \in C^{2,\beta}(\overline{\mathcal{O}})$ that

$$(F^\beta)'[\boldsymbol{\vartheta}, u](\tilde{\boldsymbol{\vartheta}}, v) = \big((\nabla H_\varepsilon)^T(\boldsymbol{L}^{\boldsymbol{\vartheta}}u + \boldsymbol{f}^{\boldsymbol{\vartheta}})(\boldsymbol{L}^{\boldsymbol{\vartheta}}v + \boldsymbol{L}^{\tilde{\boldsymbol{\vartheta}}}u + \boldsymbol{f}^{\tilde{\boldsymbol{\vartheta}}}), \tau_D(v - g^{\tilde{\boldsymbol{\vartheta}}})\big).$$

Proof. We first write the HJB operator as $F^\beta = (F_1, F_2)$, where $F_1 : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}})$ is the composition of the Nemytskij operator $H_\varepsilon : C^\beta(\overline{\mathcal{O}})^K \to C^\beta(\overline{\mathcal{O}})$ and the mapping $G : (\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \mapsto G[\boldsymbol{\vartheta}, u] := \mathbf{L}^{\boldsymbol{\vartheta}} u + \mathbf{f}^{\boldsymbol{\vartheta}} \in C^\beta(\overline{\mathcal{O}})^K$, and $F_2 : (\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \mapsto F_2[\boldsymbol{\vartheta}, u] := \tau_D(u - g^{\boldsymbol{\vartheta}}) \in C^{2,\beta}(\partial\mathcal{O})$ is the linear boundary operator.

Since the function H_ε is in $C^2(\mathbb{R}^K)$, we can deduce from Lemma 4.1 that the Nemytskij operator $H_\varepsilon : C^\beta(\overline{\mathcal{O}})^K \to C^\beta(\overline{\mathcal{O}})$ is well-defined and continuously differentiable with the Fréchet derivative $(H_\varepsilon)'[u] = (\nabla H_\varepsilon)^T(u) \in B(C^\beta(\overline{\mathcal{O}})^K, C^\beta(\overline{\mathcal{O}}))$ for all $u \in C^\beta(\overline{\mathcal{O}})^K$.

Moreover, since for any given $(\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}})$, $G[\cdot, u] : \Theta^\beta \to C^\beta(\overline{\mathcal{O}})^K$ and $G[\boldsymbol{\vartheta}, \cdot] : C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}})^K$ are affine mappings, one can easily compute the partial derivatives $\partial_u G : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to B(C^{2,\beta}(\overline{\mathcal{O}}), C^\beta(\overline{\mathcal{O}})^K)$ and $\partial_{\boldsymbol{\vartheta}} G : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to B(\Theta^\beta, C^\beta(\overline{\mathcal{O}})^K)$ of G as follows: $(\partial_u G)[\boldsymbol{\vartheta}, u](v) = \mathbf{L}^{\boldsymbol{\vartheta}} v$ and $(\partial_{\boldsymbol{\vartheta}} G)[\boldsymbol{\vartheta}, u](\tilde{\boldsymbol{\vartheta}}) = \mathbf{L}^{\tilde{\boldsymbol{\vartheta}}} u + \mathbf{f}^{\tilde{\boldsymbol{\vartheta}}}$ for all $(\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}})$, $\tilde{\boldsymbol{\vartheta}} \in \Theta^\beta$ and $v \in C^{2,\beta}(\overline{\mathcal{O}})$. Moreover, it is clear that $\partial_u G$ and $\partial_{\boldsymbol{\vartheta}} G$ are both continuous, which implies that $G : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}})^K$ is continuously differentiable with derivative

$$G'[\boldsymbol{\vartheta}, u](v, \tilde{\boldsymbol{\vartheta}}) = (\partial_u G)[\boldsymbol{\vartheta}, u](v) + (\partial_{\boldsymbol{\vartheta}} G)[\boldsymbol{\vartheta}, u](\tilde{\boldsymbol{\vartheta}}) = \mathbf{L}^{\boldsymbol{\vartheta}} v + \mathbf{L}^{\tilde{\boldsymbol{\vartheta}}} u + \mathbf{f}^{\tilde{\boldsymbol{\vartheta}}}$$

for all $(\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}})$, $\tilde{\boldsymbol{\vartheta}} \in \Theta^\beta$ and $v \in C^{2,\beta}(\overline{\mathcal{O}})$ (see [17, Theorem 7.2-3]).

Therefore, by using the chain rule (see [17, Theorem 7.1-3]), we see the composite mapping $F_1 : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}})$ is also continuously differentiable with the derivative $F_1'[\boldsymbol{\vartheta}, u] = (H_\varepsilon)'[G[\boldsymbol{\vartheta}, u]]G'[\boldsymbol{\vartheta}, u]$ for all $(\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}})$. This, along with the fact that $F_2 : C^{2,\beta}(\overline{\mathcal{O}}) \times \Theta^\beta \to C^{2,\beta}(\partial\mathcal{O})$ is a linear operator, enables us to conclude the desired differentiability of the operator $F^\beta = (F_1, F_2)$. □

With the above proposition in hand, we are ready to derive the first-order sensitivity equation for the value function of the relaxed control problem with respect to the parameter perturbations.

Theorem 5.2. *Suppose (H.1) and (H.2) hold. Let $\varepsilon > 0$, $(\Theta^\beta)_{\beta \in (0,1]}$ be the Banach spaces defined as in (5.1), $\boldsymbol{\vartheta}_0 = ((\sigma_k \sigma_k^T/2, b_k, c_k, f_k)_{k \in \mathcal{K}}, g)$, $u^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ be the solution to the Dirichlet problem (3.5) (with the coefficients $\boldsymbol{\vartheta}_0$), and $\beta_0 \in (0,1)$ be the constant in Proposition 3.3.*

Then it holds for each $\beta \in (0, \min(\beta_0, \theta)]$ that, there exists a neighborhood $\mathcal{V}$ of $\boldsymbol{\vartheta}_0$ in Θ^β, a neighborhood $\mathcal{W}$ of u^ε in $C^{2,\beta}(\overline{\mathcal{O}})$, and a mapping $\mathcal{S} : \mathcal{V} \to \mathcal{W}$ satisfying the following properties:

(1) for each $\tilde{\boldsymbol{\vartheta}} \in \mathcal{V}$, $\mathcal{S}[\tilde{\boldsymbol{\vartheta}}]$ is the classical solution to the following Dirichlet problem:

$$H_\varepsilon(\boldsymbol{L}^{\tilde{\boldsymbol{\vartheta}}} u + \boldsymbol{f}^{\tilde{\boldsymbol{\vartheta}}}) = 0 \quad \text{in } \mathcal{O}, \qquad u = g^{\tilde{\boldsymbol{\vartheta}}} \quad \text{on } \partial\mathcal{O},$$

where $(\boldsymbol{L}^{\boldsymbol{\vartheta}}, \boldsymbol{f}^{\boldsymbol{\vartheta}}, g^{\boldsymbol{\vartheta}})$ are defined as in Proposition 5.1 for each $\boldsymbol{\vartheta} \in \Theta^\beta$,

(2) $\mathcal{S} : \mathcal{V} \to \mathcal{W}$ is continuously differentiable with $\mathcal{S}[\boldsymbol{\vartheta}_0 + \delta\boldsymbol{\vartheta}] = u^\varepsilon + \mathcal{S}'[\boldsymbol{\vartheta}_0]\delta\boldsymbol{\vartheta} + o(|\delta\boldsymbol{\vartheta}|_{\Theta^\beta})$ as $|\delta\boldsymbol{\vartheta}|_{\Theta^\beta} \to 0$, and for each $\delta\boldsymbol{\vartheta} \in \Theta^\beta$, $\delta u = \mathcal{S}'[\boldsymbol{\vartheta}_0]\delta\boldsymbol{\vartheta} \in C^{2,\beta}(\overline{\mathcal{O}})$ is the solution to the following Dirichlet problem:

$$(\lambda^{u^\varepsilon})^T(\boldsymbol{L}^{\boldsymbol{\vartheta}_0}\delta u + \boldsymbol{L}^{\delta\boldsymbol{\vartheta}} u^\varepsilon + \boldsymbol{f}^{\delta\boldsymbol{\vartheta}}) = 0 \quad \text{in } \mathcal{O}, \qquad \delta u = g^{\delta\boldsymbol{\vartheta}} \quad \text{on } \partial\mathcal{O}, \tag{5.2}$$

where $\lambda^{u^\varepsilon} : \overline{\mathcal{O}} \to \Delta_K$ is the function defined as in (3.6).

Proof. The desired result comes from a direct application of the implicit function theorem (see [17, Theorem 7.13-1]). Theorem 3.4 shows that the Dirichlet problem (3.5) with the coefficients $\boldsymbol{\vartheta}_0$ admits a solution $u^\varepsilon \in C^{2,\beta}(\overline{\mathcal{O}})$ for each $\beta \in (0, \min(\beta_0, \theta)]$.

Let $\beta \in (0, \min(\beta_0, \theta)]$ be a fixed constant. We shall consider the mapping $F^\beta : \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O})$ defined as follows:

$$F^\beta : (\boldsymbol{\vartheta}, u) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}}) \mapsto F^\beta[\boldsymbol{\vartheta}, u] := (H_\varepsilon(\mathbf{L}^{\boldsymbol{\vartheta}} u + \mathbf{f}^{\boldsymbol{\vartheta}}), \tau_D(u - g^{\boldsymbol{\vartheta}})) \in C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O}).$$

Due to the fact that $u^\varepsilon \in C^{2,\beta}(\overline{\mathcal{O}})$ satisfies (3.5) with the coefficients $\boldsymbol{\vartheta}_0$, we have $H_\varepsilon(\mathbf{L}^{\boldsymbol{\vartheta}_0} u^\varepsilon + \mathbf{f}^{\boldsymbol{\vartheta}_0}) = 0$ in $\mathcal{O}$ and $H_\varepsilon(\mathbf{L}^{\boldsymbol{\vartheta}_0} u^\varepsilon + \mathbf{f}^{\boldsymbol{\vartheta}_0}) \in C^\beta(\overline{\mathcal{O}})$, which subsequently implies that $H_\varepsilon(\mathbf{L}^{\boldsymbol{\vartheta}_0} u^\varepsilon + \mathbf{f}^{\boldsymbol{\vartheta}_0}) = 0$ on $\overline{\mathcal{O}}$. The boundary condition of (3.5) implies that $\tau_D(u^\varepsilon - g^{\boldsymbol{\vartheta}_0}) = 0$ in $C^{2,\beta}(\partial\mathcal{O})$. Hence $F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon] = 0$.

Proposition 5.1 shows that F^β is continuously differentiable on $\Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}})$, and for each $(\tilde{\boldsymbol{\vartheta}}, v) \in \Theta^\beta \times C^{2,\beta}(\overline{\mathcal{O}})$,

$$\partial_u F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon](v) = \big((\nabla H_\varepsilon)^T(\mathbf{L}^{\boldsymbol{\vartheta}_0} u^\varepsilon + \mathbf{f}^{\boldsymbol{\vartheta}_0})\mathbf{L}^{\boldsymbol{\vartheta}_0} v, \tau_D v\big) = \big((\lambda^{u^\varepsilon})^T \mathbf{L}^{\boldsymbol{\vartheta}_0} v, \tau_D v\big),$$
$$\partial_{\boldsymbol{\vartheta}} F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon](\tilde{\boldsymbol{\vartheta}}) = \big((\nabla H_\varepsilon)^T(\mathbf{L}^{\boldsymbol{\vartheta}_0} u^\varepsilon + \mathbf{f}^{\boldsymbol{\vartheta}_0})(\mathbf{L}^{\tilde{\boldsymbol{\vartheta}}} u^\varepsilon + \mathbf{f}^{\tilde{\boldsymbol{\vartheta}}}), -\tau_D g^{\tilde{\boldsymbol{\vartheta}}}\big) = \big((\lambda^{u^\varepsilon})^T(\mathbf{L}^{\tilde{\boldsymbol{\vartheta}}} u^\varepsilon + \mathbf{f}^{\tilde{\boldsymbol{\vartheta}}}), -\tau_D g^{\tilde{\boldsymbol{\vartheta}}}\big),$$

where we have used the definition of $\lambda^{u^\varepsilon} \in C^\beta(\overline{\mathcal{O}}, \Delta_K)$ (see (3.6)). The classical maximum principle (see e.g. [22, Theorem 3.7]) implies that the map $\partial_u F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon](\cdot) : C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O})$ is an injection. We now show it is also a surjection. Let $(\hat{f}, \hat{g}) \in C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O})$ be given. Then the assumption that $\partial\mathcal{O} \in C^{2,\beta}$ enables us to apply [22, Lemma 6.38] and extend $\hat{g}$ to a function in $C^{2,\beta}(\overline{\mathcal{O}})$, which is still denoted by $\hat{g}$. The fact that $\lambda^{u^\varepsilon} \in C^\beta(\overline{\mathcal{O}}, \Delta_K)$ (see Theorem 3.4) and the elliptic regularity theory (see [22, Theorem 6.14]) ensure that the Dirichlet problem $\partial_u F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon](v) = (\hat{f}, \hat{g})$ admits a unique solution $v \in C^{2,\beta}(\overline{\mathcal{O}})$. Hence we see $\partial_u F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon] : C^{2,\beta}(\overline{\mathcal{O}}) \to C^\beta(\overline{\mathcal{O}}) \times C^{2,\beta}(\partial\mathcal{O})$ is a bijection.

Therefore, the implicit function theorem (see [17, Theorem 7.13-1]) ensures the existence of $\mathcal{S} \in C^1(\mathcal{V}, \mathcal{W})$ with derivative $\mathcal{S}'[\boldsymbol{\vartheta}_0] = -(\partial_u F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon])^{-1}\partial_{\boldsymbol{\vartheta}} F^\beta[\boldsymbol{\vartheta}_0, u^\varepsilon] \in B(\Theta^\beta, C^{2,\beta}(\overline{\mathcal{O}}))$. Hence we have $\mathcal{S}[\boldsymbol{\vartheta}_0 + \delta\boldsymbol{\vartheta}] = u^\varepsilon + \mathcal{S}'[\boldsymbol{\vartheta}_0]\delta\boldsymbol{\vartheta} + o(|\delta\boldsymbol{\vartheta}|_{\Theta^\beta})$ as $|\delta\boldsymbol{\vartheta}|_{\Theta^\beta} \to 0$. Let $\delta\boldsymbol{\vartheta} \in \Theta^\beta$ and $\delta u = \mathcal{S}'[\boldsymbol{\vartheta}_0]\delta\boldsymbol{\vartheta}$, the characterization of partial derivatives of F^β enables us to conclude that δu satisfies (5.2). □

Remark 5.1. We can further obtain a first-order expansion of the optimal control λ^{u^ε} in terms of the perturbations of the coefficients. If $\varepsilon > 0$ and the function H in (H.2) is in $C^3(\mathbb{R}^K)$ (c.f. H_{en} and H_{chks} in Section 3), then Lemma 4.1 shows that $\nabla H_\varepsilon : C^\alpha(\overline{\mathcal{O}}, \mathbb{R}^K) \to C^\alpha(\overline{\mathcal{O}}, \mathbb{R}^K)$, $\alpha \in (0, 1]$, is continuously differentiable with derivative $(\nabla H_\varepsilon)'[u]h = (\nabla^2 H_\varepsilon)(u)h$ for all $u, h \in C^\alpha(\overline{\mathcal{O}}, \mathbb{R}^K)$, where $\nabla^2 H_\varepsilon$ is the Hessian of H_ε. Hence, by using the chain rule and Theorem 5.2, we have for all $\beta \in (0, \min(\beta_0, \theta)]$ that

$$\lambda^{\mathcal{S}[\boldsymbol{\vartheta}_0 + \delta\boldsymbol{\vartheta}]} = \lambda^{u^\varepsilon} + \big((\nabla^2 H_\varepsilon)(\mathbf{L}^{\boldsymbol{\vartheta}_0} u^\varepsilon + \mathbf{f}^{\boldsymbol{\vartheta}_0})\big)\big(\mathbf{L}^{\boldsymbol{\vartheta}_0}\delta u + \mathbf{L}^{\delta\boldsymbol{\vartheta}} u^\varepsilon + \mathbf{f}^{\delta\boldsymbol{\vartheta}}\big) + o(|\delta\boldsymbol{\vartheta}|_{\Theta^\beta})$$

as $|\delta\boldsymbol{\vartheta}|_{\Theta^\beta} \to 0$, where $\lambda^{\mathcal{S}[\boldsymbol{\vartheta}_0 + \delta\boldsymbol{\vartheta}]}$ is the optimal feedback control of the relaxed control problem with the perturbed coefficients $\boldsymbol{\vartheta}_0 + \delta\boldsymbol{\vartheta}$, and δu is the classical solution to (5.2).

With the sensitivity equation (5.2) in hand, we now estimate the precise dependence of δu on the relaxation parameter ε, which strengthens the Lipschitz stability result (4.4) by quantifying the explicit ε-dependence of the (local) Lipschitz constant. Note that Remark 4.2 shows that the value function (2.3) (in the $C^{2,\beta}$-norm) does not depend continuously on the C^β-perturbation of the parameters, which suggests that for a fixed $\delta\boldsymbol{\vartheta} \in \Theta^\beta$, the $|\cdot|_{2,\beta}$-norm of δu will blow up as the parameter ε tends to 0.

Since the Hölder norm of the function λ^{u^ε} in (5.2) tends to infinity as $\varepsilon \to 0$, we first present a precise *a priori* estimate for the classical solutions to linear elliptic equations with ε-dependent coefficients. The proof will be postponed to Appendix A, where we first reduce the equation to a constant coefficient equation involving only second-order terms, and then apply the classical Schauder estimate.

Proposition 5.3. *Let $\alpha \in [0, 1]$, $\beta \in (0, 1)$, $\nu, \Lambda > 0$, and $\mathcal{O}$ be a bounded domain in $\mathbb{R}^n$ with $C^{2,\beta}$ boundary. For every $\varepsilon \in (0, 1]$, let $a_\varepsilon : \overline{\mathcal{O}} \to \mathbb{R}^{n\times n}$, $b_\varepsilon : \overline{\mathcal{O}} \to \mathbb{R}^n$ and $c_\varepsilon : \overline{\mathcal{O}} \to [0, \infty)$ be given functions satisfying $a_\varepsilon \geq \nu I_n$ on $\overline{\mathcal{O}}$. Suppose that $[a_\varepsilon^{ij}]_0, [b_\varepsilon^i]_0, [c_\varepsilon]_0 \leq \Lambda$ and $[a_\varepsilon^{ij}]_\beta, [b_\varepsilon^i]_\beta, [c_\varepsilon]_\beta \leq \Lambda\varepsilon^{-\alpha}$ for all $\varepsilon \in (0, 1]$ and $i, j = 1, \ldots, n$. Then for every $\varepsilon \in (0, 1]$, $f \in C^\beta(\overline{\mathcal{O}})$ and $g \in C^{2,\beta}(\overline{\mathcal{O}})$, the Dirichlet problem*

$$a_\varepsilon^{ij}\partial_{ij} w + b_\varepsilon^i \partial_i w - c_\varepsilon w + f = 0, \quad \text{in } \mathcal{O}, \qquad w = g \quad \text{on } \partial\mathcal{O}$$

admits a unique solution $w^\varepsilon \in C^{2,\beta}(\overline{\mathcal{O}})$ satisfying the following estimate with a constant $C = C(n, \beta, \nu, \Lambda, \mathcal{O})$:

$$|w^\varepsilon|_{2,\beta} \leq C\big(\varepsilon^{-\alpha(\beta+2)/\beta}|f|_0 + [f]_\beta + \varepsilon^{-\alpha(\beta+2)/\beta}|g|_{2,\beta}\big).$$

Now we present the *a priori* estimate of δu exhibiting its explicit ε-dependence (cf. (4.4)), which applies to relaxed control problems with reward functions generated by H_{en}, H_{chks} and H_{zang}.

Theorem 5.4. *Assume the setting of Theorem 5.2 and in addition that the function $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) has a Lipschitz continuous gradient. Let $\beta_0 \in (0, 1)$ be the constant in Proposition 3.3 and $\bar{\beta}_0 = \min(\beta_0, \theta)$. Then it holds for all $\varepsilon \in (0, 1]$, $\beta \in (0, \bar{\beta}_0]$ and $\delta\boldsymbol{\vartheta} \in \Theta^\beta$ that, the classical solution δu to the Dirichlet problem (5.2) satisfies the estimate $|\delta u|_{2,\beta} \leq C\varepsilon^{-(\beta+2)/\bar{\beta}_0}|\delta\boldsymbol{\vartheta}|_{\Theta^\beta}$, where C is a constant independent of ε and $\delta\boldsymbol{\vartheta}$.*

Proof. Throughout this proof, let C be a generic constant depending possibly on $\boldsymbol{\vartheta}_0$ and β, but independent of ε and $\delta\boldsymbol{\vartheta}$. Proposition 3.3 shows that $|u^\varepsilon|_{2,\bar\beta_0} \le C$ for all $\varepsilon \in (0,1]$, which together with (3.6), the fact that $\nabla H_\varepsilon(x) = \nabla H(\varepsilon^{-1}x)$ for all $x \in \mathbb{R}^K$ (see (3.3)) and the Lipschitz continuity of ∇H implies that $|\lambda^{u^\varepsilon}|_0 \le C$ and $|\lambda^{u^\varepsilon}|_{\bar\beta_0} \le C\varepsilon^{-1}$ for all $\varepsilon \in (0,1]$. Consequently, we have for all $\beta \in (0,\bar\beta_0]$ and $\varepsilon \in (0,1]$ that $|\lambda^{u^\varepsilon}|_\beta \le C|\lambda^{u^\varepsilon}|_{\bar\beta_0}^{\beta/\bar\beta_0}|\lambda^{u^\varepsilon}|_0^{(\bar\beta_0-\beta)/\bar\beta_0} \le C\varepsilon^{-\beta/\bar\beta_0}$.

Now let us fix $\beta \in (0,\bar\beta_0]$ and $\delta\boldsymbol{\vartheta} \in \Theta^\beta$. Since $\lambda^{u^\varepsilon} \in \Delta_K$ on $\overline{\mathcal{O}}$, we can apply Proposition 5.3 (with $\alpha = \beta/\bar\beta_0$) to (5.2) and conclude the desired estimate from the following inequality:

$$\begin{aligned}|\delta u|_{2,\beta} &\le C\big(\varepsilon^{-(\beta+2)/\bar\beta_0}|(\lambda^{u^\varepsilon})^T(\mathbf{L}^{\,\delta\boldsymbol{\vartheta}}u^\varepsilon + \mathbf{f}^{\delta\boldsymbol{\vartheta}})|_0 + |(\lambda^{u^\varepsilon})^T(\mathbf{L}^{\,\delta\boldsymbol{\vartheta}}u^\varepsilon + \mathbf{f}^{\delta\boldsymbol{\vartheta}})|_\beta + \varepsilon^{-(\beta+2)/\bar\beta_0}|g^{\,\delta\boldsymbol{\vartheta}}|_{2,\beta}\big)\\ &\le C\big(\varepsilon^{-(\beta+2)/\bar\beta_0}|\delta\boldsymbol{\vartheta}|_{\Theta^\beta} + |\lambda^{u^\varepsilon}|_\beta|\mathbf{L}^{\,\delta\boldsymbol{\vartheta}}u^\varepsilon + \mathbf{f}^{\delta\boldsymbol{\vartheta}}|_\beta\big) \le C\big(\varepsilon^{-(\beta+2)/\bar\beta_0}|\delta\boldsymbol{\vartheta}|_{\Theta^\beta} + \varepsilon^{-\beta/\bar\beta_0}|\delta\boldsymbol{\vartheta}|_{\Theta^\beta}\big).\end{aligned}$$ □

6 Convergence analysis for vanishing relaxation parameter

In this section, we analyze the convergence of the relaxed control problem (3.2) to the original control problem (2.3) as the relaxation parameter tends to zero. In particular, with the help of the HJB equations (2.6) and (3.5), we shall establish first-order monotone convergence of the value functions, and also uniform convergence of the feedback controls (in regions where a strict complementary condition is satisfied).

We first study the convergence of the value functions of the relaxed control problems. The following theorem shows that, as the relaxation parameter ε tends to zero, the value function (3.2) converges monotonically to the value function (2.3) in $C^{2,\beta}(\overline{\mathcal{O}})$ with first order.

Theorem 6.1. *Suppose (H.1) and (H.2) hold. Let $\beta_0 \in (0,1)$ be the constant in Proposition 3.3, and $u \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ (resp. $u^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$) be the solution to (2.6) (resp. (3.5) with parameter $\varepsilon > 0$). Then we have $u^{\varepsilon_1} \ge u^{\varepsilon_2}$ for all $\varepsilon_1 \ge \varepsilon_2 > 0$. Moreover, it holds for any $\beta \in (0, \min(\beta_0,\theta))$ that $(u^\varepsilon)_{\varepsilon>0}$ converges to u in $C^{2,\beta}(\overline{\mathcal{O}})$ as $\varepsilon \to 0$, and satisfies the estimate:*

$$0 \le u^\varepsilon - u \le \left(\exp\left[\left(\frac{\max_{k\in\mathcal{K}}\sum_{i=1}^n |b_k^i|_0}{\nu/2} + 1\right)\operatorname{diam}(\mathcal{O})\right] - 1\right)\frac{2\varepsilon c_0}{\nu}. \tag{6.1}$$

Proof. Let $(F_\varepsilon)_{\varepsilon\ge0}$ be defined as in (2.6) and (3.5), and $\varepsilon_1 \ge \varepsilon_2 > 0$ be given constants. Lemma 3.2 shows that $\rho \le 0$ on Δ_K, and $H_\varepsilon(x) = \max_{y\in\Delta_K}\big(y^Tx - \varepsilon\rho(y)\big)$ for all $x \in \mathbb{R}^K$. Hence, we have $H_{\varepsilon_1} \ge H_{\varepsilon_2}$, and

$$0 = F_{\varepsilon_1}[u^{\varepsilon_1}] - F_{\varepsilon_2}[u^{\varepsilon_2}] \ge F_{\varepsilon_2}[u^{\varepsilon_1}] - F_{\varepsilon_2}[u^{\varepsilon_2}] = \eta^T\mathbf{L}(u^{\varepsilon_1} - u^{\varepsilon_2}),$$

where we write $\eta := \int_0^1 (\nabla H_{\varepsilon_2})(\mathbf{L}u^{\varepsilon_2} + \mathbf{f} + s\mathbf{L}(u^{\varepsilon_1} - u^{\varepsilon_2}))\,ds$. Since $\eta(x) \in \Delta_K$ for all $x \in \mathcal{O}$, we can deduce from the classical maximum principle (see e.g. [22, Theorem 3.7]) that $\inf_{x\in\overline{\mathcal{O}}}(u^{\varepsilon_1} - u^{\varepsilon_2})(x) \ge \inf_{x\in\partial\mathcal{O}}(u^{\varepsilon_1} - u^{\varepsilon_2})^-(x) = 0$.

Similarly, for any given $\varepsilon > 0$, we can obtain from Lemma 3.2(2) that

$$\begin{aligned}0 = F_\varepsilon[u^\varepsilon] - F_0[u] &\le F_\varepsilon[u^\varepsilon] - (F_\varepsilon[u] - \varepsilon c_0) = \tilde\eta^T\mathbf{L}(u^\varepsilon - u) + \varepsilon c_0,\\ 0 = F_\varepsilon[u^\varepsilon] - F_0[u] &\ge F_\varepsilon[u^\varepsilon] - F_\varepsilon[u] = \tilde\eta^T\mathbf{L}(u^\varepsilon - u),\end{aligned}$$

where we have $\tilde\eta := \int_0^1 (\nabla H_\varepsilon)(\mathbf{L}u + \mathbf{f} + s\mathbf{L}(u^\varepsilon - u))\,ds$. By using $a_k = \sigma_k(\sigma_k)^T/2$, (2.4) in (H.1), and the fact that $\tilde\eta \in \Delta_K$ on $\overline{\mathcal{O}}$, we deduce that $\sum_{k=1}^K \tilde\eta_k c_k \ge 0$ and $\sum_{k=1}^K \tilde\eta_k a_k \ge (\nu/2)I_n$. Hence the classical maximum principle (see e.g. [22, Theorem 3.7]) and the fact that $u^\varepsilon = u$ on $\partial\mathcal{O}$ give us the estimate (6.1).

Finally, the *a priori* bound in Proposition 3.3 and the Arzelà–Ascoli theorem ensure that for any given $\beta \in (0,\min(\beta_0,\theta))$, there exists a subsequence $(u^{\varepsilon_m})_{m\in\mathbb{N}}$ with $\lim_{m\to\infty}\varepsilon_m = 0$, such that $(u^{\varepsilon_m})_{m\in\mathbb{N}}$ converges in $C^{2,\beta}(\overline{\mathcal{O}})$ to some function $\bar u$ and $\bar u \in C^{2,\min(\beta_0,\theta)}(\overline{\mathcal{O}})$. Since the entire sequence $(u^\varepsilon)_{\varepsilon>0}$ converges monotonically to u, we have $u = \bar u$ and $(u^\varepsilon)_{\varepsilon>0}$ converges to u in $C^{2,\beta}(\overline{\mathcal{O}})$ for all $\beta \in (0,\min(\beta_0,\theta))$. □

Remark 6.1. The estimate (6.1) depends on $\varepsilon, c_0, \nu, b_k^i$ and $\mathcal{O}$ in a rather intuitive way. Note that, compared with the original control problem (2.3), the relaxed control problem (3.2) introduces additional randomness for exploration to achieve more robust decisions, especially at regions where two or more strategies lead to

similar performances based on the given model (the points at which arg max in (2.8) is not a singleton). The relation (2.8) between feedback controls and the derivatives of value functions further suggests that such regions usually correspond to a sign change of derivatives of value functions.

The exploration surplus in the value functions clearly increases as ε or c_0 increase (see Lemma 3.2(1) and Figure 1), since the same level of exploration will bring more rewards. It will also increase with $\mathrm{diam}(\mathcal{O})$ as the dynamics will stay in $\mathcal{O}$ longer. Furthermore, due to the lack of regularization from the Laplacian operator, a small volatility or a large drift-to-volalitly ratio of the underlying model usually leads to a more rapidly changing value function, which increases the occurrence of the uncertain regions and makes the relaxation approach more beneficial.

Now we turn to investigate the convergence of the feedback relaxed control (3.6). To distinguish different convergence behaviours related to reward functions generated by H_{en} and H_{zang}, we first introduce the following concept for functions which only modify the pointwise maximum function locally near the kinks.

Definition 6.1. Let $n \in \mathbb{N}$, we say a function $\phi : \mathbb{R}^n \to \mathbb{R}$ satisfies (S_{loc}) with constant $\vartheta \geq 0$, if it holds for all $k = 1, \ldots, n$ and $x \in \mathbb{R}^n$ with $x_k \geq x_j + \vartheta$, $\forall j \neq k$, that $\phi(x) = x_k$.

It is clear that the pointwise maximum function on $\mathbb{R}^n$ satisfies (S_{loc}) with $\vartheta = 0$, and the two-dimensional function H_{zang} defined in (3.8) satisfies (S_{loc}) with $\vartheta = 1/2$. The following lemma shows that property (S_{loc}) is preserved under function composition and scaling, which consequently implies that the recursively constructed K-dimensional H_{zang} and its corresponding scaled function $(H_{\mathrm{zang}})_\varepsilon$ (cf. (3.3)) satisfy (S_{loc}). The proof follows directly from Definition 6.1, and is included in Appendix A.

Lemma 6.2. *(1) For each $n \in \mathbb{N}$, let $H_0^{(n)} : \mathbb{R}^n \to \mathbb{R}$ be the n-dimensional pointwise maximum function (see (3.3)). Let $n_1 = 2$, $n_2, n_3 \in \mathbb{N}$, $(\vartheta_i, c_i)_{i=1}^3 \subset [0, \infty)$, $\phi_i : \mathbb{R}^{n_i} \to \mathbb{R}$, $i = 1, 2, 3$, be given functions, and $\phi : \mathbb{R}^{n_2+n_3} \to \mathbb{R}$ be the function satisfying for all $x = (x_1, \ldots, x_{n_2+n_3})^T \in \mathbb{R}^{n_2+n_3}$ that $\phi(x) = \phi_1(\phi_2(x^{(1)}), \phi_3(x^{(2)}))$ with $x^{(1)} = (x_1, \ldots, x_{n_2})$ and $x^{(2)} = (x_{n_2+1}, \ldots, x_{n_2+n_3})$. Suppose that for each $i = 1, 2, 3$, the function ϕ_i satisfies (S_{loc}) with constant ϑ_i, and $\phi_i(x) \leq H_0^{(n_i)}(x) + c_i$ for all $x \in \mathbb{R}^{n_i}$. Then the function ϕ satisfies (S_{loc}) with constant $\max(\vartheta_2, \vartheta_3, c_2 + \vartheta_1, c_3 + \vartheta_1)$, and it holds for all $x \in \mathbb{R}^{n_2+n_3}$ that $\phi(x) \leq H_0^{(n_2+n_3)}(x) + c_1 + \max(c_2, c_3)$.*

(2) If $\phi : \mathbb{R}^n \to \mathbb{R}$ satisfies (S_{loc}) with constant $\vartheta \geq 0$, then for each $\varepsilon > 0$, the scaled function $\phi_\varepsilon : x \in \mathbb{R}^n \mapsto \varepsilon\phi(\varepsilon^{-1}x) \in \mathbb{R}$ satisfies (S_{loc}) with constant $\varepsilon\vartheta$.

The following proposition presents several important convergence properties of the functions $(\nabla H_\varepsilon)_{\varepsilon>0}$. In the sequel, we shall denote by $e_k \in \mathbb{R}^K$, $k \in \mathcal{K}$, the unit vector from the k-th column of the identify matrix I_K, and by $\mathrm{conv}(S)$ the convex hull of a given set $S \subset \mathbb{R}^K$.

Proposition 6.3. *Suppose (H.2) holds. Let $(H_\varepsilon)_{\varepsilon\geq 0}$ be defined as in (3.3), $(\partial H_0)(x) = \mathrm{conv}(\{e_k \in \mathbb{R}^K \mid x_k = H_0(x), k \in \mathcal{K}\})$ for all $x \in \mathbb{R}^K$, and $U = \{x \in \mathbb{R}^K \mid (\partial H_0)(x) \text{ is a singleton}\}$. Then it holds for all $x \in \mathbb{R}^K$ and compact subset $\mathcal{C} \subset U$ that*

(1) $\lim_{k\to\infty} \mathrm{dist}((\nabla H_{\varepsilon_k})(x_k), (\partial H_0)(x)) = 0$ provided that $\lim_{k\to\infty} x_k = x$ and $\lim_{k\to\infty} \varepsilon_k = 0^+$,

(2) $(\nabla H_\varepsilon)_{\varepsilon>0}$ converges uniformly to ∂H_0 on $\mathcal{C}$ as $\varepsilon \to 0$. If we further suppose the function $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) satisfies (S_{loc}) with constant $\vartheta \geq 0$, then there exists $\varepsilon_0 > 0$ such that $(\nabla H_\varepsilon)(x) = (\partial H_0)(x)$ for all $x \in \mathcal{C}$ and $\varepsilon \in (0, \varepsilon_0]$.

Proof. We first establish Property (1) by considering the following function:

$$\phi : (x, \varepsilon, y) \in \mathbb{R}^K \times [0, 1] \times \Delta_K \mapsto y^T x - \varepsilon\rho(y) \in \mathbb{R}.$$

Note that Lemma 3.2(1) shows that the restriction of ρ on Δ_K is continuous, which subsequently implies that ϕ is a continuous function. Then we can deduce from [1, Theorem 17.31] that the set-valued mapping $\Xi : (x, \varepsilon) \in \mathbb{R}^K \times [0, 1] \rightrightarrows \arg\max_{y\in\Delta_K} \phi(x, \varepsilon, y) \subset \Delta_K$ is upper hemicontinuous, which along with the fact that $\Xi(x, \varepsilon) = (\nabla H_\varepsilon)(x)$ for all $(x, \varepsilon) \in \mathbb{R}^K \times (0, 1]$ (see Lemma 3.2(2)) enables us to deduce $\lim_{k\to\infty} \mathrm{dist}((\nabla H_{\varepsilon_k})(x_k), \Xi(x, 0)) = 0$ for any given $\lim_{k\to\infty} x_k = x$ and $\lim_{k\to\infty} \varepsilon_k = 0^+$. Property (1) now follows from the fact that $\Xi(x, 0) = (\partial H_0)(x)$ (see e.g. [37, Theorem 2]).

Now we shall prove Property (2). We first define the set $U_k = \{x \in \mathbb{R}^K \mid x_k > x_j, \forall j \neq k\}$ for each $k \in \mathcal{K}$. It is clear that $(U_k)_{k\in\mathcal{K}}$ are disjoint open convex sets, $U = \cup_{k\in\mathcal{K}} U_k$, and it holds for all $k \in \mathcal{K}$ and $x \in U_k$ that H_0 is differentiable at x with $(\nabla H_0)(x) = e_k = (\partial H_0)(x)$.

Let $\mathcal{C} \subset U$ be a compact set, then we have $\mathcal{C} = \cup_{k\in\mathcal{K}}(\mathcal{C} \cap U_k)$ due to $U = \cup_{k\in\mathcal{K}} U_k$. Let us fix an arbitrary index $k \in \mathcal{K}$. By using the fact that $(U_k)_{k\in\mathcal{K}}$ are disjoint open sets, we can deduce that $\mathcal{C} \cap U_k$ is also compact. Since $(H_\varepsilon)_{\varepsilon\geq 0}$ are convex and differentiable on U_k and $\lim_{\varepsilon\to 0} H_\varepsilon(x) = H_0(x)$ for all $x \in U_k$, we can deduce from the convexity of U_k and [38, Theorem 25.7] that $(\nabla H_\varepsilon)_{\varepsilon>0}$ converges uniformly to $\nabla H_0 = \partial H_0$ on $\mathcal{C} \cap U_k$. Since $\mathcal{K}$ is a finite set, we have shown the desired uniform convergence on $\mathcal{C}$.

Moreover, for each $k \in \mathcal{K}$, the compactness of $\mathcal{C} \cap U_k$ implies that there exists $\varepsilon_{0,k} > 0$ such that $\mathcal{C} \cap U_k \subset \{x \in \mathbb{R}^K \mid x_k > x_j + \varepsilon_{0,k}, \forall j \neq k\}$. Then, if H satisfies $(S_{\rm loc})$ with constant $\vartheta \geq 0$, then Lemma 6.2(2) shows that for all $\varepsilon > 0$ satisfying $\varepsilon\vartheta \leq \varepsilon_{0,k}$, we have $H_\varepsilon = H_0$ (and hence $\nabla H_\varepsilon = \nabla H_0$) on $\mathcal{C} \cap U_k$. Hence, by setting $\varepsilon_0 > 0$ to be a constant satisfying $\varepsilon_0\vartheta \leq \min_{k\in\mathcal{K}} \varepsilon_{0,k}$, we can conclude for all $\varepsilon \in (0, \varepsilon_0]$ that $\nabla H_\varepsilon = \nabla H_0 = \partial H_0$ on $\mathcal{C}$. □

Now we are ready to present the convergence of the feedback relaxed control (3.6). Note that the Hölder continuity of the relaxed controls (3.6) and the possible discontinuity of the feedback control (2.8) suggest that the sequence $(\lambda^{u^\varepsilon})_{\varepsilon>0}$ in general does not converge uniformly to α^u on $\mathcal{O}$ as $\varepsilon \to 0$. Thus we shall show that the relaxed controls converge in terms of the Hausdorff metric everywhere in $\mathcal{O}$, and converge uniformly on compact subsets of the following region:

$$\mathcal{O}_{\rm st} = \left\{ x \in \mathcal{O} \;\middle|\; \operatorname*{arg\,max}_{k\in\mathcal{K}} \big(\mathcal{L}_k u(x) + f_k(x)\big) \text{ is a singleton} \right\}, \tag{6.2}$$

where $u \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ is the solution to (2.6) (or equivalently the value function (2.3) if the function $\sigma \in \mathbb{S}_0^n$; see Theorem 2.2), and $(\mathcal{L}_k)_{k\in\mathcal{K}}$ are the elliptic operators defined as in (2.7). Note that $\mathcal{O}_{\rm st}$ contains the points at which a strict complementary condition is satisfied, i.e., the optimal feedback control strategy of (2.3) is uniquely determined.

Theorem 6.4. *Suppose (H.1) and (H.2) hold. Let $(\lambda^{u^\varepsilon})_{\varepsilon>0}$ be the functions defined as in (3.6) for each $\varepsilon > 0$, $u \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ be the solution to (2.6), and $\mathcal{O}_{\rm st}$ be the set defined as in (6.2). Then we have for all $x \in \mathcal{O}$ and $(x_\varepsilon)_{\varepsilon>0} \subset \mathcal{O}$ with $\lim_{\varepsilon\to 0} x_\varepsilon = x$ that*

$$\lim_{\varepsilon\to 0} \operatorname{dist}\left(\lambda^{u^\varepsilon}(x_\varepsilon), \operatorname{conv}\left(\left\{ e_k \in \mathbb{R}^K \;\middle|\; k \in \operatorname*{arg\,max}_{k\in\mathcal{K}} \big(\mathcal{L}_k u(x) + f_k(x)\big) \right\}\right)\right) = 0. \tag{6.3}$$

Moreover, it holds for all compact subset $\mathcal{C} \subset \mathcal{O}_{\rm st}$ that $(\lambda^{u^\varepsilon})_{\varepsilon>0}$ converges uniformly to the function $\lambda^* : x \in \mathcal{O}_{\rm st} \to e_{\kappa^u(x)} \in \Delta_K$ on $\mathcal{C}$ as $\varepsilon \to 0$, where $\kappa^u(x) = \arg\max_{k\in\mathcal{K}} \big(\mathcal{L}_k u(x) + f_k(x)\big)$ for all $x \in \mathcal{O}_{\rm st}$. If we further suppose the function $H : \mathbb{R}^K \to \mathbb{R}$ in (H.2) satisfies $(S_{\rm loc})$ with constant $\vartheta > 0$, then there exists $\varepsilon_0 > 0$ such that it holds for all $\varepsilon \in (0, \varepsilon_0]$ that $\lambda^{u^\varepsilon} \equiv \lambda^*$ on $\mathcal{C}$.

Proof. For any give $\varepsilon > 0$, let $u^\varepsilon \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ be the solution to (3.5). We first prove (6.3) by fixing an arbitrary point $x \in \mathcal{O}$. By using (3.6) and Proposition 6.3(1), we see it suffices to show $\lim_{\varepsilon\to 0}(\mathbf{L}u^\varepsilon(x_\varepsilon) + \mathbf{f}(x_\varepsilon)) = \mathbf{L}u(x) + \mathbf{f}(x)$, where $\mathbf{L}$, $\mathbf{f}$ are defined as those in (2.6). Then the fact that $(u^\varepsilon)_{\varepsilon>0}$ converges to u uniformly in $C^2(\overline{\mathcal{O}})$ (see Theorem 6.1) and the continuity of coefficients enable us to conclude (6.3).

We now proceed to demonstrate the uniform convergence of $(\lambda^{u^\varepsilon})_{\varepsilon>0}$ in $\mathcal{O}_{\rm st}$. Note that for all $x \in \mathcal{O}_{\rm st}$, we have $e_{\kappa^u(x)} = (\partial H_0)\big(\mathbf{L}u(x) + \mathbf{f}(x)\big)$, where the set-valued mapping $\partial H_0 : \mathbb{R}^K \rightrightarrows \Delta_K$ is defined as in Proposition 6.3. We further define for any given $k \in \mathcal{K}$ the set

$$\mathcal{O}_{{\rm st},k} = \{x \in \mathcal{O} \mid \mathcal{L}_k u(x) + f_k(x) > \mathcal{L}_j u(x) + f_j(x),\ \forall j \neq k\},$$

where $u \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ is the solution to (2.6), and $(\mathcal{L}_k)_{k\in\mathcal{K}}$ are the elliptic operators defined as in (2.7). The continuity of the coefficients in $(\mathcal{L}_k)_{k\in\mathcal{K}}$ (see (H.1)) implies that $(\mathcal{O}_{{\rm st},k})_{k\in\mathcal{K}}$ are disjoint open sets satisfying $\mathcal{O}_{\rm st} = \cup_{k\in\mathcal{K}} \mathcal{O}_{{\rm st},k}$.

Now let $\mathcal{C} \subset \mathcal{O}_{\rm st} \subseteq \mathcal{O}$ be a given compact subset. Then we have $\mathcal{C} = \cup_{k\in\mathcal{K}}(\mathcal{C} \cap \mathcal{O}_{{\rm st},k})$, and $\mathcal{C} \cap \mathcal{O}_{{\rm st},k}$ is a compact set for each $k \in \mathcal{K}$. Let $k \in \mathcal{K}$ be a fixed index. Then the continuity of the coefficients in $(\mathcal{L}_k)_{k\in\mathcal{K}}$,

the fact that $u \in C^2(\overline{\mathcal{O}})$, and the compactness of $\mathcal{C} \cap \mathcal{O}_{\mathrm{st},k}$ imply that, there exist constants $C_1, C_2 \in (0,\infty)$ such that we have for all $x \in \mathcal{C} \cap \mathcal{O}_{\mathrm{st},k}$ and $j \in \mathcal{K}$ that, $|\mathcal{L}_j u(x) + f_j(x)| \leq C_2$ and

$$0 < C_1 \leq \mathcal{L}_k u(x) + f_k(x) - \max_{j \neq k} \big(\mathcal{L}_j u(x) + f_j(x)\big) \leq C_2 < \infty.$$

Now by using the fact that $(u^\varepsilon)_{\varepsilon>0}$ converges to u uniformly in $C^2(\overline{\mathcal{O}})$, we can deduce that there exist $\varepsilon_0, C_1, C_2 > 0$ such that the same estimates hold for all $(u^\varepsilon)_{\varepsilon \in (0,\varepsilon_0]}$. In other words, let U be the set defined as in Proposition 6.3, we can introduce the compact set

$$G_k \coloneqq \{x \in \mathbb{R}^K \mid 0 < C_1 \leq x_k - \max_{j \neq k} x_j \leq C_2,\ |x_j| \leq C_2, \forall j \in \mathcal{K}\} \subset U,$$

and conclude for all $\varepsilon \in (0, \varepsilon_0]$, $x \in \mathcal{C} \cap \mathcal{O}_{\mathrm{st},k}$ that $\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x) \in G_k$ and $\mathbf{L}u(x) + \mathbf{f}(x) \in G_k$.

For any given $\eta > 0$, the uniform convergence of $(\nabla H_\varepsilon)_{\varepsilon>0}$ to ∂H_0 on G_k (see Proposition 6.3(2)) ensures that there exists $\delta_k > 0$, such that we have for all $y \in G_k$ and $\varepsilon < \delta_k$ that $|(\nabla H_\varepsilon)(y) - (\partial H_0)(y)| \leq \eta$. Hence, by using the fact that $\partial H_0 = \{e_k\}$ on G_k, we have for all $\varepsilon < \min(\delta_k, \varepsilon_0)$ and $x \in \mathcal{C} \cap \mathcal{O}_{\mathrm{st},k}$ that

$$\begin{aligned} |\lambda^{u^\varepsilon}(x) - \lambda^*(x)| &= |(\nabla H_\varepsilon)\big(\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x)\big) - (\partial H_0)\big(\mathbf{L}u(x) + \mathbf{f}(x)\big)| \\ &= |(\nabla H_\varepsilon)\big(\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x)\big) - (\partial H_0)\big(\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x)\big)| \leq \eta, \end{aligned}$$

which shows the uniform convergence of $(\lambda^{u^\varepsilon})_{\varepsilon>0}$ to λ^* on $\mathcal{C} \cap \mathcal{O}_{\mathrm{st},k}$. Since $\mathcal{C} = \cup_{k \in \mathcal{K}}(\mathcal{C} \cap \mathcal{O}_{\mathrm{st},k})$ and $\mathcal{K}$ is a finite set, we can conclude the desired uniform convergence on $\mathcal{C}$.

Finally, if we further suppose H satisfies (S_{loc}) with constant $\vartheta \geq 0$, Proposition 6.3(2) ensures that $\nabla H_\varepsilon \equiv \partial H_0$ on G_k for all small enough $\varepsilon > 0$, which leads to the fact that $\lambda^{u^\varepsilon} \equiv \lambda^*$ for all small enough $\varepsilon > 0$ on $\mathcal{C}$ and finishes our proof. □

Remark 6.2. One can identify the unit vector $e_k \in \Delta_K$, $k \in \mathcal{K}$, as the Dirac measure supported on $\{\mathbf{a}_k\}$, which shows that, as the relaxation parameter tends to zero, the agent of the relaxed control problem will emphasize more on exploitation, and the relaxed control distribution will collapse to a pure exploitation strategy for the classical control problem.

Note that Theorem 6.4 demonstrates an exact regularization feature of the reward function ρ_{zang} generated by H_{zang}, which means that we can recover the original control strategy in the region $\mathcal{O}_{\mathrm{st}}$ based on the feedback relaxed control *without* sending the relaxation parameter ε to 0. The main intuition of the proof is that the region $\mathcal{O}_{\mathrm{st}}$ can be mapped into a finite number of convex sets (i.e., the sets $(U_k)_{k \in \mathcal{K}}$ in the proof of Proposition 6.3). Hence, if a reward function only modifies the pointwise maximum function locally near the kinks, then one can employ the local compactness and local convexity structure of $\mathcal{O}_{\mathrm{st}}$ and the finiteness of the action set $\mathbf{A}$, and deduce the local exact regularization property in the region $\mathcal{O}_{\mathrm{st}}$.

The exact regularization feature of ρ_{zang} helps avoid the possible numerical instability for solving the relaxed control problem (3.2) with an extremely small relaxation parameter. In contrast, the feedback relaxed control λ^{u^ε} based on the entropy reward function ρ_{en} is always in $(0,1)^K$, and the convergence rate to the original control strategy can be arbitrarily slow.

7 Conclusions

To the best of our knowledge, this is the first paper which constructs Lipschitz stable feedback control strategies for general multi-dimensional continuous-time stochastic control problems, and rigorously analyzes the performance of a pre-computed feedback control for a perturbed problem in a continuous setting. We also perform a novel first-order sensitivity analysis for the value function and feedback relaxed control with respect to perturbations in the model parameters, and quantify the explicit dependence of the Lipschitz stability of feedback controls on the exploration parameter. These stability results provide a theoretical justification for recent reinforcement learning heuristics that including an exploration reward in the optimization objective leads to more robust decision making.

A natural next step would be to extend the stability analysis to finite horizon stochastic control problems and mean-field control problems with continuous action spaces (see e.g. [23, 42]). The infinite cardinality of action spaces implies that the corresponding relaxed controls take values in an infinite-dimensional space of

probability measures, which poses additional challenges for the analysis of the regularized control problems. For example, infinite-dimensional convex analysis on spaces of measures must be employed to analyze the regularity of the modified Hamiltonians and the well-posedness of the associated HJB equations. Moreover, one must endow the action space of relaxed controls with a suitable metric structure (such as the Wasserstein metric) in order to study the spatial regularity and Lipschitz stability of feedback relaxed controls.

Another interesting direction is to design efficient numerical algorithms for solving the regularized control problems in a continuous setting.

A Proofs of technical results

Proof of Theorem 2.2. Let $\pi = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P}, W) \in \Pi_{\rm ref}$, $\alpha \in \mathcal{A}_\pi$, $x \in \overline{\mathcal{O}}$, let $X^{\alpha,x} = (X_t^{\alpha,x})_{t\geq 0}$ be the strong solution to (2.2) with control α, and for all $t \geq 0$, let $Z_t^{\alpha,x} = \int_0^t c(X_s^{\alpha,x}, \alpha_s)\, ds$. It is shown in [13, Lemma 3.1] that $\mathbb{E}[\exp(\mu\tau^{\alpha,x})] < \infty$ for some constant $\mu > 0$, which implies that $\tau^{\alpha,x} < \infty$ with probability 1. Applying Itô's formula to the function $\phi(y,z) = u(y)\exp(-z)$, $(y,z) \in \mathbb{R}^n \times \mathbb{R}$, gives us that

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}[\phi(X_{\tau^{\alpha,x}}^{\alpha,x}, Z_{\tau^{\alpha,x}}^{\alpha,x})] &= \phi(x,0) + \mathbb{E}^{\mathbb{P}}\left[\int_0^{\tau^{\alpha,x}} (\mathcal{L}_{X^{\alpha,x}}\phi + c_\alpha \partial_z \phi)(X_t^{\alpha,x}, Z_t^{\alpha,x})\, dt\right] \\ &= u(x) + \mathbb{E}^{\mathbb{P}}\left[\int_0^{\tau^{\alpha,x}} (\mathcal{L}_{X^{\alpha,x}} u - c_\alpha u)(X_t^{\alpha,x})\Gamma_t^{\alpha,x}\, dt\right],\end{aligned} \tag{A.1}$$

where $\mathcal{L}_{X^{\alpha,x}}$ is the generator of the controlled dynamics $X^{\alpha,x}$, and $\Gamma_t^{\alpha,x} = \exp\big(-\int_0^t c(X_s^{\alpha,x}, \alpha_s)\, ds\big)$ for all $t \in [0, \tau^{\alpha,x}]$. The fact that u is a solution to (2.6) implies that for $\mathbb{P}$-a.s. $\omega \in \Omega$, and $t \in [0, \tau^{\alpha,x}(\omega)]$,

$$\begin{aligned}&(\mathcal{L}_{X^{\alpha,x}} u - c_\alpha u)(X_t^{\alpha,x}(\omega)) + f(X_t^{\alpha,x}(\omega), \alpha_t(\omega)) \\ &\leq \max_{k\in\mathcal{K}} \Big(a^{ij}(\cdot, \mathbf{a}_k)\partial_{ij} u(\cdot) + b^i(\cdot, \mathbf{a}_k)\partial_i u(\cdot) - c(\cdot, \mathbf{a}_k)u(\cdot) + f(\cdot, \mathbf{a}_k)\Big)(X_t^{\alpha,x}(\omega)) = 0,\end{aligned} \tag{A.2}$$

Then, by rearranging the terms, using the fact that $\phi(X_{\tau^{\alpha,x}}^{\alpha,x}, Z_{\tau^{\alpha,x}}^{\alpha,x}) = g(X_{\tau^{\alpha,x}}^{\alpha,x})\Gamma_{\tau^{\alpha,x}}^{\alpha,x}$ and taking the supremum over all $\alpha \in \mathcal{A}_\pi$ and $\pi \in \Pi_{\rm ref}$, we can deduce that $u(x) \geq v(x)$ for all $x \in \overline{\mathcal{O}}$.

We proceed to show α^u is a feedback control of (2.3) (cf. Definition 2.2). Let $\alpha^u : \overline{\mathcal{O}} \to \mathbf{A}$ be a Borel measurable function satisfying (2.8), and $\tilde{\alpha}^u : \mathbb{R}^n \to \mathbf{A}$ be an extension of α^u such that $\tilde{\alpha}^u = \alpha^u$ on $\overline{\mathcal{O}}$ and $\tilde{\alpha}^u = \mathbf{a}_1$ on $\overline{\mathcal{O}}^c$. We shall consider the functions $b_\alpha : \mathbb{R}^n \to \mathbb{R}^n$, $\sigma_\alpha : \mathbb{R}^n \to \mathbb{S}_0^n$ such that $b_\alpha(x) = b(x, \tilde{\alpha}^u(x))$, $\sigma_\alpha(x) = \sigma(x, \tilde{\alpha}^u(x))$ for all $x \in \mathbb{R}^n$. The measurability of α^u and the continuity of b, σ imply that b_α, σ_α and $\tilde{\alpha}^u$ are Borel measurable. Then, for any given $x \in \mathbb{R}^n$, by using the boundedness of functions b_α, σ_α, and [32, Theorem 1], we can deduce that there exists $\pi^x = (\Omega^x, \mathcal{F}^x, \{\mathcal{F}_t^x\}_{t\geq 0}, \mathbb{P}^x, W) \in \Pi_{\rm ref}$, and an $\{\mathcal{F}_t^x\}_{t\geq 0}$-progressively measurable continuous process $(X_t^x)_{t\geq 0}$, such that $X_0^x = x$, and

$$dX_t^x = b(X_t^x, \tilde{\alpha}^u(X_t^x))\, dt + \sigma(X_t^x, \tilde{\alpha}^u(X_t^x))\, dW_t \quad \text{for all } t \geq 0 \text{ and } \mathbb{P}^x\text{-a.s.} \tag{A.3}$$

Thus we can obtain from the definition of $\tilde{\alpha}^u$ that $(X_t^x)_{t\geq 0}$ satisfies (2.9) with $h = \alpha^u$. Moreover, [29, Theorem 2.2.4 on p. 54] implies that $\mathbb{E}^{\mathbb{P}^x}[\int_0^{\tau^{\alpha^u,x}} \big(|b(X_s^x, \alpha^u(X_s^x))| + |\sigma(X_s^x, \alpha^u(X_s^x))|^2\big)\, ds] < \infty$, which shows that α^u is a feedback control of (2.3).

It remains to show α^u is an optimal feedback control. If $x \in \partial\mathcal{O}$, we can deduce from the definition that $\tau^{\tilde{\alpha}^u,x} = 0$, which shows that $g(x) = g(X_{\tau^{\tilde{\alpha}^u,x}}^x) = J(x, \alpha^u)$, where $J(x,\alpha^u)$ is defined as in (2.10). Similarly, we have for all $\pi \in \Pi_{\rm ref}$, $\alpha \in \mathcal{A}_\pi$, $x \in \partial\mathcal{O}$ that the first exit time of $X^{\alpha,x}$ from $\mathcal{O}$ is 0, i.e., $\tau^{\alpha,x} = 0$, which implies that $v(x) = g(x)$. Hence, we can deduce from the fact that u satisfies the boundary condition of (2.6) that $u(x) = g(x) = v(x) = J(x,\alpha^u)$ for all $x \in \partial\mathcal{O}$.

For each $x \in \mathcal{O}$, let X^x be a progressively measurable continuous process satisfying the SDE (A.3), defined on the reference probability system $\pi^x \in \Pi_{\rm ref}$. The assumption that α^u satisfies (2.8) ensures that $\tilde{\alpha}^u(X^x)$ and X^x obtain the equality in (A.2) for $\mathbb{P}$-a.s. $\omega \in \Omega$, and $t \in [0, \tau^{\tilde{\alpha}^u,x}(\omega)]$, from which, by using similar arguments as (A.1), we can obtain that $u(x) = J(x,\alpha^u)$ (c.f. (2.10)). On the other hand, owing to the fact that $\tilde{\alpha}^u(X^x) \in \mathcal{A}_{\pi^x}$, we have by the definition of v that $u(x) \leq v(x)$ for all $x \in \mathcal{O}$. Combining this with the fact that $u(x) \geq v(x)$ for all $x \in \mathcal{O}$, we can conclude that $u(x) = v(x) = J(x,\alpha^u)$ in $\mathcal{O}$, which shows that α^u is an optimal feedback control and $u \equiv v$ on $\overline{\mathcal{O}}$. □

Proof of Lemma 3.1. The definition of Δ_K and (H.1) clearly imply that the function $\tilde{b}$ is well-defined and enjoys the desired estimates. Hence we shall focus on establishing the properties of the function $\tilde{\sigma}$.

It has been shown in [17, Theorem 7.14-3] that for any given $A \in \mathbb{S}^n_>$, there exists a unique matrix $A^{1/2} \in \mathbb{S}^n_>$ such that $A^{1/2}(A^{1/2})^T = A$, $A^{1/2} \geq \sqrt{\mu} I_n$ if $A \geq \mu I_n$, and the mapping $\Phi : A \in \mathbb{S}^n_> \mapsto \Phi(A) = A^{1/2} \in \mathbb{S}^n_>$ is infinitely differentiable. Note that (2.4) and (2.5) in (H.1) ensure that there exists a constant $C \in (0, \infty)$, such that it holds for all $x \in \mathbb{R}^n$, $\lambda \in \Delta_K$ that

$$\sum_{k=1}^K \sigma(x, \mathbf{a}_k)\sigma(x, \mathbf{a}_k)^T \lambda_k \in G := \left\{ A = [a_{ij}] \in \mathbb{S}^n_> \;\middle|\; A \geq \nu I_n, \sum_{i,j=1}^n |a_{ij}| \leq C \right\} \subset \mathbb{S}^n_>. \tag{A.4}$$

We now define the function $\tilde{\sigma} : \mathbb{R}^n \times \Delta_K \to \mathbb{S}^n_>$ by $\tilde{\sigma}(x, \lambda) = \Phi\left(\sum_{k=1}^K \sigma(x, \mathbf{a}_k)\sigma(x, \mathbf{a}_k)^T \lambda_k\right)$ for all $x \in \mathbb{R}^n, \lambda \in \Delta_K$. The facts that Φ is a smooth function and G is a compact subset of $\mathbb{S}^n_>$ imply that Φ is bounded and Lipschitz continuous on G. Therefore, we can conclude from (2.4), (2.5), (A.4) and the definition of $\tilde{\sigma}$ that it holds for all $x \in \mathbb{R}^n$, $\lambda \in \Delta_K$ that $\tilde{\sigma}(x, \lambda) \geq \sqrt{\nu} I_n$ and $\sum_{i,j} |\tilde{\sigma}^{ij}(\cdot, \lambda)|_{0,1} < \infty$. □

Proof of Lemma 3.2. We start by establishing Property (1). Since $H : \mathbb{R}^K \to \mathbb{R}$ is a continuous convex function, the representation of ρ in (H.2) and [38, Theorem 12.2] ensure that ρ is a closed convex proper function satisfying

$$H(x) = \sup_{y \in \mathbb{R}^K} \left(y^T x - \rho(y) \right) \quad \forall x \in \mathbb{R}^K. \tag{A.5}$$

The assumption that $H(x) - c_0 \leq \max_{k \in \mathcal{K}} x_k \leq H(x)$ for all $x \in \mathbb{R}^K$ implies that for all $y \in \mathbb{R}^K$,

$$\sup_{x \in \mathbb{R}^K} \left(x^T y - \left[\max_{k \in \mathcal{K}} x_k + c_0 \right] \right) \leq \sup_{x \in \mathbb{R}^K} \left(x^T y - H(x) \right) = \rho(y) \leq \sup_{x \in \mathbb{R}^K} \left(x^T y - \max_{k \in \mathcal{K}} x_k \right),$$

which together with the fact that

$$\sup_{x \in \mathbb{R}^K} \left(x^T y - \max_{k \in \mathcal{K}} x_k \right) = \begin{cases} 0, & y \in \Delta_K, \\ \infty, & y \notin \Delta_K, \end{cases}$$

shows that $\rho(y) \in [-c_0, 0]$ for all $y \in \Delta_K$ and $\rho(y) = \infty$ for all $y \in (\Delta_K)^c$. Finally, since ρ is a closed convex function satisfying $\{y \in \mathbb{R}^K \mid \rho(y) < \infty\} = \Delta_K$, we can deduce from [38, Theorem 10.2] (Δ_K is the standard simplex and hence locally simplicial) that the restriction of ρ to Δ_K is a continuous function.

We now show Property (2). It is clear from (H.2) and (3.3) that $H_\varepsilon(x) - c_0\varepsilon \leq H_0(x) \leq H_\varepsilon(x)$ for all $x \in \mathbb{R}^K$. Note that (A.5) and the fact that $\rho = \infty$ on Δ_K imply that for all $\varepsilon > 0$ we have

$$\varepsilon H(\varepsilon^{-1} x) = \varepsilon \max_{y \in \mathbb{R}^K} \left(y^T \varepsilon^{-1} x - \rho(y) \right) = \max_{y \in \Delta_K} \left(y^T x - \varepsilon \rho(y) \right) \quad \forall x \in \mathbb{R}^K,$$

which shows the function $\varepsilon\rho$ is the convex conjugate of H_ε, i.e., $(H_\varepsilon)^* = \varepsilon\rho$. Hence, we can further deduce from [38, Theorem 23.5], the differentiability and convexity of H_ε that

$$(\nabla H_\varepsilon)(x) = \arg\max_{y \in \mathbb{R}^K} \left(y^T x - (H_\varepsilon)^*(y) \right) = \arg\max_{y \in \Delta_K} \left(y^T x - \varepsilon \rho(y) \right) \in \Delta_K \quad \forall x \in \mathbb{R}^K.$$

Consequently, we can obtain from the fundamental theorem of calculus and the Cauchy-Schwarz inequality that H_ε is Lipschitz continuous with constant $L_{H_\varepsilon} = \sup_{x \in \mathbb{R}^K} |(\nabla H_\varepsilon)(x)| \leq \max_{y \in \Delta_K} |y|$. Note that Δ_K is the convex hull of $\{e_1, \ldots, e_K\}$, where e_k is the unit vector from the k-th column of the identify matrix I_K. Hence [38, Theorem 32.2] ensures that $\max_{y \in \Delta_K} |y|$ is attained at $\{e_1, \ldots, e_K\}$, which implies that $L_{H_\varepsilon} \leq 1$, and finishes the proof of Lemma 3.2. □

Before establishing Proposition 3.3, we first present an *a priori* estimate for solutions of fully nonlinear equations involving only the second order term.

Lemma A.1. *[16, Theorem 7.2 on p. 125] Let $\mathcal{O}$ be a bounded connected open subset of $\mathbb{R}^n$, and $F : \mathcal{O} \times \mathbb{S}^n \to \mathbb{R}$ be a given function. Suppose the function F is differentiable and convex in its second component, and there exist constants $\lambda, \Lambda > 0$ such that $\lambda I_n \le \left[\frac{\partial F}{\partial r_{ij}}(x,r)\right] \le \Lambda I_n$ for all $(x,r) \in \mathcal{O} \times \mathbb{S}^n$. Then there exists a constant $\alpha = \alpha(n, \Lambda/\lambda) \in (0,1)$ such that for any $\beta \in (0,\alpha)$, if we have in addition that $\partial\mathcal{O} \in C^{2,\beta}$, $g \in C^{2,\beta}(\overline{\mathcal{O}})$, and there exist constants $\gamma, \mu > 0$ such that it holds for all $x, y \in \mathcal{O}$, $r \in \mathbb{S}^n$ that $|F(x,r) - F(y,r)| \le \gamma(\mu + |r|)|x-y|^\beta$, then the Dirichlet problem*

$$F(x, D^2u) = 0 \quad \text{in } \mathcal{O}, \qquad u = g \quad \text{on } \partial\mathcal{O},$$

admits a unique solution $u \in C^{2,\beta}(\overline{\mathcal{O}})$ satisfying the estimate $[u]_{2,\beta} \le C\big[|u|_0 + |g|_{2,\beta} + \mu\big]$, where the constant C depends only on n, Λ/λ, γ, $(\alpha-\beta)^{-1}$ and the $C^{2,\beta}$-norm of $\partial\mathcal{O}$.

Now we proceed to prove the *a priori* estimate for solutions to (3.5).

Proof of Proposition 3.3. Throughout this proof, we shall denote by C a generic constant, which may take a different value at each occurrence.

Let $\phi \in C(\overline{\mathcal{O}}) \cap C^2(\mathcal{O})$ be a given function, we consider the Dirichlet problem

$$F_\phi(x, D^2u(x)) = 0 \quad \text{in } \mathcal{O}, \qquad u = g \quad \text{on } \partial\mathcal{O}, \tag{A.6}$$

where we define $D^2u(x) = [\partial_{ij}u(x)] \in \mathbb{S}^n$, and the function $F_\phi : \mathcal{O} \times \mathbb{S}^n \to \mathbb{R}$ such that for all $x \in \mathcal{O}$ and $r = [r_{ij}] \in \mathbb{S}^n$,

$$F_\phi(x,r) \coloneqq H_\varepsilon\left(\big(a_k^{ij}(x)r_{ij} + b_k^i(x)\partial_i\phi(x) - c_k(x)\phi(x) + f_k(x)\big)_{k\in\mathcal{K}}\right).$$

It follows from (H.2) that F_ϕ is differentiable and convex in r. Moreover, a straightforward computation shows for all $(x,r) \in \mathcal{O} \times \mathbb{S}^n$ that $\left[\frac{\partial F_\phi}{\partial r_{ij}}(x,r)\right] = \sum_{k=1}^K \eta_k(x,r)a_k(x)$, where we have

$$\eta_l(x,r) \coloneqq \partial_l H_\varepsilon\left(\big(a_k^{ij}(x)r_{ij} + b_k^i(x)\partial_i\phi(x) - c_k(x)\phi(x) + f_k(x)\big)_{k\in\mathcal{K}}\right), \quad l = 1, \dots, K.$$

Note that for each $k \in \mathcal{K}$, the fact that $a_k = \sigma\sigma^T/2$ and (H.1) (see (2.4)-(2.5)) imply that there exists a constant C, depending only on n, such that for all $x \in \mathcal{O}$,

$$\frac{\nu}{2}I_n \le a_k(x) \le \operatorname{tr}\big(a_k^T(x)a_k(x)\big)I_n \le C\left(\sup_{i,j,k}|\sigma_k^{ij}|_{0;\overline{\mathcal{O}}}\right)^4 I_n,$$

which, along with the fact that $(\eta_1(x,r), \dots, \eta_K(x,r))^T \in \Delta_K$ for all $(x,r) \in \mathcal{O} \times \mathbb{S}^n$ (see Lemma 3.2(2)), shows that $\frac{\nu}{2}I_n \le \left[\frac{\partial F_\phi}{\partial r_{ij}}(x,r)\right] \le CI_n$, for some constant C depending only on n and the constant M defined in the statement of Proposition 3.3.

The regularity of the coefficients in (H.1) and the Lipschitz continuity of H_ε (see Lemma 3.2(2)) imply that, if the function $\phi \in C^{2,\eta}(\overline{\mathcal{O}})$, $0 < \eta \le \theta$, then the function F_ϕ satisfies for all $x, y \in \mathcal{O}$, $r \in \mathbb{S}^n$ that

$$|F_\phi(x,r) - F_\phi(y,r)| \le C\Lambda(|r| + |\phi|_{1,\eta} + 1)|x-y|^\eta,$$

for some constant C depending only on n. Consequently, we can deduce from Lemma A.1 that, there exists a constant $\beta_0 = \beta_0(n,\nu,M) \in (0,1)$, such that for all $\beta \in (0, \min(\beta_0,\theta)]$ and $\phi \in C^{2,\beta}(\overline{\mathcal{O}})$, the Dirichlet problem (A.6) admits a unique solution $u^\phi \in C^{2,\beta}(\overline{\mathcal{O}})$, and satisfies $[u^\phi]_{2,\beta} \le C\big[|u^\phi|_0 + |g|_{2,\beta} + |\phi|_{1,\beta} + 1\big]$, where the constant C depends only on n, ν, Λ, β, and $\mathcal{O}$.

Now let $u^\varepsilon \in C^{2,\beta}(\overline{\mathcal{O}})$, $\beta \in (0, \min(\beta_0,\theta)]$ be a solution to (3.5). Then it is clear that u^ε is a solution to the Dirichlet problem: $F_{u^\varepsilon}(x, D^2u(x)) = 0$ in $\mathcal{O}$ and $u = g$ on $\partial\mathcal{O}$. We can then deduce from the above arguments that, there exists a constant C, depending only on n, ν, Λ, β and $\mathcal{O}$, such that $[u^\varepsilon]_{2,\beta} \le C\big[|g|_{2,\beta} + |u^\varepsilon|_{1,\beta} + 1\big]$. Hence by using the interpolation inequality (see [16, Theorem 1.2 on p. 18]), we have $|u^\varepsilon|_{2,\beta} \le C\big[|g|_{2,\beta} + |u^\varepsilon|_0 + 1\big]$.

It remains to estimate $|u^\varepsilon|_0$. By using the fundamental theorem of calculus, we have for all $x \in \mathcal{O}$ that

$$-H_\varepsilon(\mathbf{f}(x)) = H_\varepsilon(\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x)) - H_\varepsilon(\mathbf{f}(x)) = \int_0^1 (\nabla H_\varepsilon)^T(s\mathbf{L}u^\varepsilon(x) + \mathbf{f}(x))\mathbf{L}u^\varepsilon(x)\,ds,$$

from which, by using the classical maximum principle (see e.g. [22, Theorem 3.7]) and the fact that $\nabla H_\varepsilon \in \Delta_K$ (see Lemma 3.2(2)), we can deduce that, there exists a constant $C = C(n, \Lambda, \mathcal{O}) > 0$ that

$$|u^\varepsilon|_0 \le C\Big(\sup_{x\in\partial\mathcal{O}} |g(x)| + |H_\varepsilon(\mathbf{f})|_0 \Big) \le C\big(|g|_{0;\overline{\mathcal{O}}} + |H_0(\mathbf{f})|_0 + \varepsilon c_0\big) \le C\big(|g|_{0;\overline{\mathcal{O}}} + 1 + \varepsilon c_0\big),$$

which together with the fact that $|u^\varepsilon|_{2,\beta} \le C\big[|g|_{2,\beta} + |u^\varepsilon|_0 + 1\big]$ leads to the desired estimate. □

Proof of Proposition 5.3. The well-posedness of the classical solution w^ε follows from the standard elliptic regularity theory (see [22, Theorem 6.14]), hence it suffices to prove the *a priori* estimate for a fixed $\varepsilon > 0$.

Let $\rho > 0$ be a constant whose value will be specified later, and $(\xi_m)_{m=1}^M$ be a partition of unity in a domain containing $\overline{\mathcal{O}}$ such that the following properties hold: (1) the support of each function ξ_m is contained in a ball $B_\rho(x_m)$ for some $x_m \in \mathbb{R}^n$; (2) $\xi_m \in C^\infty(\mathbb{R}^n)$ satisfies for all $\gamma \ge 0$ that $|\xi_m|_{\lfloor\gamma\rfloor,\gamma-\lfloor\gamma\rfloor} \le C_\gamma \rho^{-\gamma}$, where $\lfloor\gamma\rfloor$ is the integer part of γ and C_γ is a constant independent of m and γ; (3) for each $x \in \overline{\mathcal{O}}$, $\sum_{m=1}^M \xi_m(x) = 1$ and the number of intersected supports of $(\xi_m)_{m=1}^M$ at x is bounded by a constant M_n depending only on the dimension n. In the following, we shall denote by w the solution w^ε, and by C a generic constant independent of α, m and ε.

For each $m = 1, \dots, M$, we define the function $w_m = w\xi_m$, which satisfies $w_m = g\xi_m$ on $\partial\mathcal{O}$ and

$$a_\varepsilon^{ij}(x_m)\partial_{ij} w_m = (a_\varepsilon^{ij}(x_m) - a_\varepsilon^{ij})\partial_{ij} w_m + a_\varepsilon^{ij}(\partial_j w \partial_i \xi_m + \partial_i w \partial_j \xi_m + w\partial_{ij}\xi_m) + \tilde{f}, \quad \text{in } \mathcal{O},$$

where $\tilde{f} \coloneqq (-b_\varepsilon^i \partial_i w + c_\varepsilon w - f)\xi_m$. Hence applying the classical Schauder estimate yields that

$$|w_m|_{2,\beta} \le C\big(|(a_\varepsilon^{ij}(x_m) - a_\varepsilon^{ij})\partial_{ij} w_m|_\beta + |a_\varepsilon^{ij}(\partial_j w \partial_i \xi_m + \partial_i w \partial_j \xi_m + w\partial_{ij}\xi_m)|_\beta + |\tilde{f}|_\beta + |g\xi_m|_{2,\beta}\big) \tag{A.7}$$

for some constant $C = C(n, \beta, \nu, \Lambda, \mathcal{O})$.

Note that by choosing $\rho = (2C\Lambda\varepsilon^{-\alpha})^{-1/\beta}$, we have for all $x \in B_\rho(x_m)$ and $i, j = 1, \dots, n$ that

$$|(a_\varepsilon^{ij}(x_m) - a_\varepsilon^{ij}(x)| \le [\alpha_\varepsilon^{ij}]_\beta |x - x_m|^\beta \le [\alpha_\varepsilon^{ij}]_\beta \rho^\beta \le 1/(2C),$$

which together with the fact that $\partial_{ij} w_m = 0$ on $\overline{\mathcal{O}} \setminus B_\rho(x_m)$ implies that

$$\begin{aligned}
&|(a_\varepsilon^{ij}(x_m) - a_\varepsilon^{ij})\partial_{ij} w_m|_{\beta;\overline{\mathcal{O}}} = |(a_\varepsilon^{ij}(x_m) - a_\varepsilon^{ij})\partial_{ij} w_m|_{\beta;\overline{\mathcal{O}}\cap B_\rho(x_m)} \\
&\le |a_\varepsilon^{ij}(x_m) - a_\varepsilon^{ij}|_{0;\overline{\mathcal{O}}\cap B_\rho(x_m)} |\partial_{ij} w_m|_{\beta;\overline{\mathcal{O}}\cap B_\rho(x_m)} + |a_\varepsilon^{ij}(x_m) - a_\varepsilon^{ij}|_{\beta;\overline{\mathcal{O}}\cap B_\rho(x_m)} |\partial_{ij} w_m|_{0;\overline{\mathcal{O}}\cap B_\rho(x_m)} \\
&\le |w_m|_{2,\beta;\overline{\mathcal{O}}}/(2C) + [a^{ij}]_\beta |w_m|_{2;\overline{\mathcal{O}}} \le |w_m|_{2,\beta;\overline{\mathcal{O}}}/(2C) + \Lambda\varepsilon^{-\alpha}|w_m|_{2;\overline{\mathcal{O}}}.
\end{aligned}$$

Then we can deduce from the interpolation inequality (see [16, Theorem 1.3 on p. 19]) and (A.7) that

$$|w_m|_{2,\beta} \le C\big(\varepsilon^{-\alpha(\beta+2)/\beta}|w_m|_0 + |a_\varepsilon^{ij}(\partial_j w \partial_i \xi_m + \partial_i w \partial_j \xi_m + w\partial_{ij}\xi_m)|_\beta + |\tilde{f}|_\beta + |g\xi_m|_{2,\beta}\big). \tag{A.8}$$

Note that for all $\gamma \ge 0$, we can obtain from property (2) of $(\xi_m)_{m=1}^M$ that $|\xi_m|_{\lfloor\gamma\rfloor,\gamma-\lfloor\gamma\rfloor} \le C_\gamma(2C\Lambda\varepsilon^{-\alpha})^{\gamma/\beta}$. Hence by repeatedly applying interpolation inequalities, we can simplify (A.8) into

$$|w_m|_{2,\beta} \le C\big(\varepsilon^{-\alpha(\beta+2)/\beta}|w_m|_0 + \varepsilon^{-\alpha}|f|_0 + [f]_\beta + \varepsilon^{-\alpha(\beta+2)/\beta}|g|_{2,\beta}\big),$$

which along with properties (2) and (3) of $(\xi_m)_{m=1}^M$ leads to the estimate that

$$|w|_{2,\beta} \le 2M_n \max_m |w_m|_{2,\beta} \le C\big(\varepsilon^{-\alpha(\beta+2)/\beta}|w|_0 + \varepsilon^{-\alpha}|f|_0 + [f]_\beta + \varepsilon^{-\alpha(\beta+2)/\beta}|g|_{2,\beta}\big).$$

Finally, we can conclude from the classical maximum principle (see e.g. [22, Theorem 3.7]) that $|w|_0 \le C(|f|_0 + |g|_0)$, which finishes the proof of the desired *a priori* estimate. □

Proof of Lemma 6.2. We first establish Property (1). For any given $x = (x_1, \dots, x_{n_2+n_3})^T \in \mathbb{R}^{n_2+n_3}$, we write $x^{(1)} = (x_1, \dots, x_{n_2}) \in \mathbb{R}^{n_2}$ and $x^{(2)} = (x_{n_2+1}, \dots, x_{n_2+n_3}) \in \mathbb{R}^{n_3}$.

Let $x \in \mathbb{R}^{n_2+n_3}$ satisfy for some $k \in \{1, \dots, n_2+n_3\}$ that $x_k \ge \max_{j\ne k} x_j + c$ with $c = \max(\vartheta_2, \vartheta_3, c_2 + \vartheta_1, c_3 + \vartheta_1)$. We assume without loss of generality that $k \le n_2$. Then since ϕ_2 satisfies (S_{loc}) with ϑ_2 and $c \ge \vartheta_2$, we have that $\phi_2(x^{(1)}) = x_k$ and $\phi_3(x^{(2)}) \le H_0^{(n_3)}(x^{(2)}) + c_3$. Moreover, since $x_k \ge H_0^{(n_3)}(x^{(2)}) + c$

and $c \geq c_3 + \vartheta_1$, we see $\phi_2(x^{(1)}) \geq \phi_3(x^{(2)}) + \vartheta_1$, which, along with the assumption that ϕ_1 satisfies (S_{loc}) with ϑ_1, implies $\phi(x) = \phi_2(x^{(1)}) = x_k$. Similar arguments show that the same conclusion holds if $k \geq n_2+1$, which enables us to conclude that ϕ satisfies (S_{loc}) with c.

Now let $x \in \mathbb{R}^{n_2+n_3}$ be an arbitrary given point. We have by assumptions that $\phi_2(x^{(1)}) \leq H_0^{(n_2)}(x^{(1)})+c_2$ and $\phi_3(x^{(2)}) \leq H_0^{(n_3)}(x^{(2)}) + c_3$. Hence, by using the fact that $H_0^{(2)}$ is component-wise increasing and subadditive on $\mathbb{R}^2$, we have

$$\begin{aligned}\phi(x) &= \phi_1(\phi_2(x^{(1)}), \phi_3(x^{(2)})) \leq H_0^{(n_1)}(\phi_2(x^{(1)}), \phi_3(x^{(2)})) + c_1 \\ &\leq H_0^{(n_1)}(H_0^{(n_2)}(x^{(1)}), H_0^{(n_3)}(x^{(2)})) + c_1 + \max(c_2, c_3) = H_0^{(n_2+n_3)}(x) + c_1 + \max(c_2, c_3),\end{aligned}$$

which finishes the proof of Property (1). Property (2) follows directly from the definition of ϕ_ε. □

References

[1] D. Aldous, *Weak convergence and the general theory of processes*, manuscript, 1981. Available online at https://www.stat.berkeley.edu/ aldous/Papers/weak-gtp.pdf

[2] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd ed., Springer-Verlag, Berlin, 2006.

[3] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder, *All adapted topologies are equal*, Probab. Theory Relat. Fields, 178 (2020), pp. 1125–1172.

[4] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and J. Wiesel, *Estimating processes in adapted Wasserstein distance*, preprint, arXiv:2002.07261, 2020.

[5] G. Barles and E. Rouy, *A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications*, Comm. Partial Differential Equations, 23 (1998), pp. 1945–2033.

[6] M. Basei, X. Guo, and A. Hu, *Linear quadratic reinforcement learning: Sublinear regret in the episodic continuous-time framework*, preprint, arXiv:2006.15316, 2020.

[7] E. Bayraktar, Y. Dolinsky, and J. Guo, *Continuity of utility maximization under weak convergence*, Math. Financ. Econ., 14 (2020), pp. 725–757.

[8] E. Bayraktar, L. Dolinskyi, and Y. Dolinsky, *Extended weak convergence and utility maximisation with proportional transaction costs*, Finance Stoch., 24 (2020), pp. 1013–1034.

[9] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[10] S. I. Birbil, S.-C. Fang, J. Frenk, and S. Zhang, *Recursive approximation of the high dimensional max function*, Oper. Res. Lett., 33 (2005), pp. 450–458.

[11] P. Blanchard, D. J. Higham, and N. J. Higham, *Accurate computation of the log-sum-exp and softmax functions*, preprint (2019) arXiv:1909.03469. Accepted in IMA J. Numer. Anal., https://doi.org/10.1093/imanum/draa038.

[12] O. Bokanowski, S. Maroso, and H. Zidani, *Some convergence results for Howard's algorithm*, SIAM J. Numer. Anal., 47 (2009), pp. 3001–3026.

[13] R. Buckdahn and T. Y. Nie, *Generalized Hamilton-Jacobi-Bellman equations with Dirichlet boundary condition and stochastic exit time optimal control problem*, SIAM J. Control Optim., 54 (2016), pp. 602–631.

[14] S. Chaumont, *Uniqueness to elliptic and parabolic Hamilton–Jacobi–Bellman equations with non-smooth boundary*, C.R. Math. Acad. Sci. Paris, 339 (2004), pp. 555–560.

[15] C. Chen and O. L. Mangasarian, *Smoothing methods for convex inequalities and linear complementarity problems*, Math. Program., 71 (1995), pp. 51–69.

[16] Y.-Z. Chen and L.-C. Wu, *Second Order Elliptic Equations and Elliptic Systems*, Transl. Math. Monogr. 174, AMS, Providence, RI, 1998.

[17] P. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications*, Appl. Math. 130, SIAM, Philadelphia, 2013.

[18] P. Drábek, *Continuity of Nemyckij's operator in Hölder spaces*, Comm. Math. Univ. Carolinae, 16 (1975), pp. 37–57.

[19] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Springer, New York, 2006.

[20] P. Forsyth and G. Labahn, *Numerical methods for controlled Hamilton-Jacobi-Bellman PDEs in finance*, J. Comput. Finance, 11 (2007/2008, Winter), pp. 1–43.

[21] M. Geist, B. Scherrer, and O. Pietquin, *A theory of regularized Markov decision processes*, preprint, arXiv:1901.11275, 2019.

[22] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd edition, Springer-Verlag, Berlin, New York, 1985.

[23] H. Gu, X. Guo, X. Wei, and R. Xu, *Dynamic programming principles for learning MFGs*, preprint, arXiv:1911.07314, 2019.

[24] X. Guo, A. Hu, R. Xu, and J. Zhang, *A general framework for learning mean-field games*, preprint, arXiv:2003.06069, 2020.

[25] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, *Reinforcement learning with deep energy-based policies*, preprint, arXiv:1702.08165, 2017.

[26] K. Ito, C. Reisinger, and Y. Zhang, *A neural network based policy iteration algorithm with global H^2-superlinear convergence for stochastic games on domains*, preprint (2019) arXiv:1906.02304. Accepted in Found. Comput. Math., https://doi.org/10.1007/s10208- 020-09460-1.

[27] A. D. Kara and S. Yüksel, *Robustness to incorrect system models in stochastic control*, SIAM J. Control Optim., 58 (2020), pp. 1144–1182.

[28] B. W. Kort and D. P. Bertsekas, *A new penalty function algorithm for constrained minimization*, in Proceedings of the 1972 IEEE Conference on Decision and Control, New Orleans, Louisiana, 1972.

[29] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, 1980.

[30] H.J. Langen, *Convergence of dynamic programming models*, Math. Oper. Res., 6 (1981), pp. 493–512.

[31] H. Mania, S. Tu, and B. Recht, *Certainty equivalence is efficient for linear quadratic control*, in Advances in Neural Information Processing Systems, 2019, pp. 10154–10164.

[32] Y. S. Mishura and A. Y. Veretennikov, *Existence and uniqueness theorems for solutions of McKean-Vlasov stochastic equations*, preprint, arXiv:1603.02212, 2016.

[33] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, *Bridging the gap between value and policy based reinforcement learning*, preprint, arXiv:1702.08892, 2017.

[34] R. Nugari, *Further remarks on the Nemitskii operator in Hölder spaces*, Comment. Math. Univ. Carolin. 34 (1993) pp. 89–95.

[35] J. M. Peng, *A smoothing function and its applications*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, ed., Kluwer, Dordrecht, 1998, pp. 293–316.

[36] J. Peng and Z. Lin, *A non-interior continuation method for generalized linear complementarity problems*, Math. Program., 86 (1999), pp. 533–563.

[37] R. A. Poliquin and R. T. Rockafellar, *Proto-derivative formulas for basic subgradient mappings in mathematical programming*, Set-Valued Anal., 2 (1994), pp. 275–290.

[38] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[39] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[40] I. Smears and E. Süli, *Discontinuous Galerkin finite element approximation of Hamilton-Jacobi-Bellman equations with Cordes coefficients*, SIAM J. Numer. Anal., 52 (2014), pp. 993–1016,

[41] I. Smears and E. Süli, *Discontinuous Galerkin finite element methods for time-dependent Hamilton-Jacobi-Bellman equations with Cordes coefficients*, Numer. Math., (2015), pp. 1–36.

[42] H. Wang, Z. T. Zariphopoulou, and X. Zhou, *Exploration versus exploitation in reinforcement learning: a stochastic control approach*, J. Mach. Learn. Res., 21(2020). pp. 1–34.

[43] H. Wang and X. Zhou, *Continuous-time mean-variance portfolio selection: A reinforcement learning framework*, Math. Finance, 30 (2020), pp. 1273–1308.

[44] J. Yong and X. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.

[45] I. Zang, *A smoothing-out technique for min-max optimization*, Math. Program., 19 (1980), pp. 61–77.

[46] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, *Maximum entropy inverse reinforcement learning*, In AAAI, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.