Convergence of Policy Iteration for Entropy-Regularized Stochastic Control Problems

Yu-Jui Huang∗ Zhenhua Wang† Zhou Zhou‡

June 18, 2024

Abstract

For a general entropy-regularized stochastic control problem on an infinite horizon, we prove that a policy iteration algorithm (PIA) converges to an optimal relaxed control. Contrary to the standard stochastic control literature, classical Hölder estimates of value functions do not ensure the convergence of the PIA, due to the added entropy-regularizing term. To circumvent this, we carry out a delicate estimation by moving back and forth between appropriate Hölder and Sobolev spaces. This requires new Sobolev estimates designed specifically for the purpose of policy iteration and a nontrivial technique to contain the entropy growth. Ultimately, we obtain a uniform Hölder bound for the sequence of value functions generated by the PIA, thereby achieving the desired convergence result. Characterization of the optimal value function as the unique solution to an exploratory Hamilton–Jacobi–Bellman equation comes as a by-product. The PIA is numerically implemented in an example of optimal consumption.

MSC (2010): 93E20, 60H10, 94A17

Keywords: entropy regularization, policy iteration algorithm, exploratory Hamilton-Jacobi-Bellman equations, reinforcement learning.

1 Introduction

Policy improvement refers to updating a current policy (i.e., control strategy) to a new one that improves the performance. In practice, one may repeat this procedure indefinitely (which forms a policy iteration algorithm (PIA)), with the hope that the recursive improvement will eventually lead to an optimal policy.

For standard stochastic control problems in a diffusion model, the PIA generally lives up to one’s hope. By classical estimates of partial differential equations (PDEs), Puterman [20] shows that the optimal value is recovered under the PIA, for a specific control problem with a compact domain of time and space. Jacka and Mijatović [10] put forth a long list of general assumptions under which the PIA converges to an optimal control, and a few examples that fulfill the assumptions are provided. Relying on backward stochastic differential equations, Kerimkulov et al. [16] establish the convergence of the PIA to an optimal control with an appropriate convergence rate, under the condition that only the drift coefficient of the state process is controlled. Overall, these developments share one underlying idea: for the desired convergence of the PIA, one needs a uniform bound for the value functions.

∗Department of Applied Mathematics, University of Colorado Boulder, email: yujui.huang@colorado.edu. Partially supported by National Science Foundation (DMS-2109002).
†Department of Mathematics, Iowa State University, email: zhenhuaW@iastate.edu.
‡School of Mathematics and Statistics, University of Sydney, email: zhou.zhou@sydney.edu.au.
generated by the PIA as well as their derivatives—that is, a uniform bound in a Hölder space is in demand. Finding such a bound is made drastically more complicated by entropy regularization.

Entropy regularization changes a standard stochastic control problem in two distinctive ways. At each time, it first randomizes the control action through a density function (or more generally, a probability measure) defined on the action space $U$; next, it adds the entropy of the density to the reward functional to be maximized. The ultimate goal is to find a density-valued stochastic process (also called a relaxed control) that maximizes the expected cumulative reward plus entropy.

This class of problems has drawn substantial attention, due to its close ties to reinforcement learning (RL). In selecting a distribution under certain constraint (e.g., a performance criterion), the principle of maximum entropy, dating back to Jaynes [11, 12], stipulates that the best choice is the distribution with largest possible entropy, among those that satisfy the constraint. As argued in [11], maximum entropy amounts to least commitment to (or, dependence on) any additional assumption beyond the known constraint—put differently, in the words of Shannon [22], it preserves the highest level of information of the unknown. By applying the principle of maximum entropy to a dynamic discrete-time setting, Ziebart et al. [28] perform sequential optimization under a “softmax” criterion, which is equivalent to solving an entropy-regularized stochastic control problem. In the RL literature, it is now well-known that a “softmax” criterion (or entropy regularization) encourages exploration of the unknown environment, thereby preventing early settlement to suboptimal strategies; see Ziebart et al. [29], Fox et al. [5], Haarnoja et al. [8], and Haarnoja et al. [9], among many others.

Lately, Wang et al. [26] extended the above discrete-time RL setting to a continuous-time diffusion framework, which opens the door to mathematical analysis based on the Hamilton–Jacobi–Bellman (HJB) equation. This continuous-time analysis has been applied to entropy-regularized stochastic control problems in various settings. For linear–quadratic problems, an optimal relaxed control is derived explicitly in [26] and applied to mean–variance portfolio selection in Wang and Zhou [27]. Firoozi and Jaimungal [4] and Guo et al. [7] generalize this single-agent setup to a mean field game and obtain the explicit form of a Nash equilibrium. For general control problems beyond the linear–quadratic case, when the action space is finite, Reisinger and Zhang [21] study the regularity of the optimal value function and the related optimal relaxed control, as well as their stability with respect to model parameters and convergence in case of diminishing regularization. Tang et al. [24] obtain corresponding regularity and convergence results for the case of a general action space.

Despite this vibrant development, the PIA in a diffusion framework remains largely unexplored under entropy regularization. To the best of our knowledge, the convergence of the PIA to an optimal relaxed control is established only in the linear–quadratic case. Specifically, the HJB analysis in [26] suggests that the PIA be performed among Gibbs-form relaxed control (see e.g., (2.6) below). As Gibbs-form distributions reduce to Gaussian densities in the linear–quadratic case, the PIA is highly tractable: as shown in [27], an optimal policy is reached within two iterations in the PIA. Even in a time-inconsistent case where one looks for an equilibrium policy (instead of a value-maximizing policy), Dai et al. [2] show that the linear–quadratic structure still ensures that equilibrium policies can take the form of Gaussian densities and the PIA converges to such a policy.\(^1\)

Let us stress the importance of establishing the PIA convergence for general entropy-regularized stochastic control problems. For a standard stochastic control problem, one can always find the optimal value function (and thus an optimal policy) by solving the associated HJB equation, even without considering the PIA at all. For an entropy-regularized stochastic control problem, as it provides a theoretic description of an RL problem (thanks to the exposition in [26]), the study of the PIA has significant value. Indeed, in an actual RL scenario, as the underlying model is not perfectly known, the HJB equation cannot be readily solved. One instead has to rely on the PIA to gradually

\(^1\)In such a PIA, an initial policy needs to be carefully selected based on the linear–quadratic structure to guarantee the convergence of algorithm. Also, the policy improvement property no longer holds, as the goal is no longer value maximization but attaining an equilibrium.
improve the policy at hand and ultimately approach an optimal one.

In this paper, we take up a general entropy-regularized stochastic control problem beyond the linear–quadratic case. Our goal is to establish a general convergence result that the PIA yields an optimal relaxed control. As aforementioned, this boils down to finding a uniform bound in a Hölder space for the value functions $\{v^n\}_{n \in \mathbb{N}}$ generated by the PIA. Intriguingly, we find that classical Schauder estimates are not enough to provide the desired uniform bound, due to the added entropy term; see Remark 4.7 for details. In response, we devise a delicate estimation procedure (i.e., the “grand plan” in Section 4.1) that goes back and forth between appropriate Hölder and Sobolev spaces.

To actually implement this plan, we need preparations in two important aspects. First, we develop new Sobolev estimates specifically for the purpose of policy iteration (Section 4.2). In particular, Lemma 4.2 carefully extends a classical interior Sobolev estimate to higher order, and further spells out explicitly how the estimate depends on the controlled drift of the underlying state process. This explicit dependence is normally overlooked in the standard PDE literature, but indispensable for our analysis of the PIA; see the discussion above Lemma 4.2 and Remark 4.3 (i) for details. In addition, we identify suitable conditions to contain the growth of the entropy term (Section 4.3). By assuming that the reward function and the drift coefficient of the state process are Lipschitz in the control variable (Assumption 4.1) and the action space $U \subset \mathbb{R}^\ell$ has a positive finite measure and fulfills a uniform cone condition (Assumption 4.2), Lemma 4.4 shows that any Gibbs-form distribution has polynomial growth, so that its entropy can grow only logarithmically; see Corollary 4.2.

Let us stress that containing the entropy growth is challenging in general. While the condition “$U \subset \mathbb{R}^\ell$ has a positive finite measure” naturally yields an upper bound for the entropy (see (2.8) below), the uniform cone condition of $U$ (Assumption 4.2) is subtle: it relates any Gibbs-form distribution to an integral in $\mathbb{R}^\ell$ over a sector of a fixed size. Evaluating this integral (under the $\ell$-dimensional polar coordinates on a sector) gives a polynomial, as desired.

Based on all this, Proposition 4.2 carries out the “grand plan” in detail, producing a desired uniform bound for $\{v^n\}_{n \in \mathbb{N}}$ in a Hölder space. This allows us to show $v^n \to V^*$ in the Hölder space and obtain an optimal relaxed control in terms of $V^*$; see Theorem 4.1, the main result of this paper. As a by-product of Theorem 4.1, $V^*$ is characterized as the unique classical solution to a so-called exploratory HJB equation (i.e., (2.4) below) with sufficient regularity. This complements the well-posedness result in [24] for exploratory HJB equations, and particularly answers an open question raised in [24, Section 6]; see Remark 4.8 for details.

Most of our results require that only the drift coefficient of the state process is controlled. This is not very restrictive in terms of the literature of PIAs. Even for standard stochastic control problems (without entropy regularization), [20] and [16] assume outright that only the drift coefficient is controlled; [10] does not assume this, but the explicit example provided there (see [10, Remark 4 (i)]) still allows only the drift to be controlled; [15] considers the general case with the diffusion coefficient also controlled, but the PIA only converges to a locally optimal control. Furthermore, from a technical point of view, our “grand plan” does not easily extend to the case of a controlled diffusion coefficient: as explained in Section 4.5, a crucial “order reduction” no longer holds in the more general setting.

Finally, let us point out several recent studies related to the PIA, mostly for deterministic control without entropy regularization. For a state process modeled by a controlled ordinary differential equation (ODE), Lee and Sutton [17] propose two versions of the PIA and analyze their performance for various RL problems in continuous time and space. When the state dynamics is additionally affine in the control action, Wallace and Si [25] provide a comprehensive review of four implementations of the PIA; Lutter et al. [18], on the other hand, propose a value iteration algorithm (which performs iterations on the value function directly). Under a general ODE dynamics, Tang et al. [23] prove a convergence rate for a semi-discrete version of the PIA, using viscosity solution techniques.

The rest of the paper is organized as follows. Section 1.1 collects the notation in the paper. Section 2 introduces the model, the PIA to consider, and two useful technical lemmas. Section 3
investigates properties of the value function associated with a Gibbs-form relaxed control, including its regularity and PDE characterization. In Section 4, we first observe that the PIA does improve a policy recursively. To show that the recursive improvement generates an optimal relaxed control, we propose a delicate estimation procedure in Section 4.1, whose implementation requires theoretic preparations in Sections 4.2 and 4.3. Section 4.4 performs the estimation procedure, which yields the desired result that the PIA yields an optimal relaxed control. Section 5 presents an optimal consumption problem where our theoretic result can be applied. We numerically implement the PIA and show that it does lead to an optimal relaxed control in this example.

1.1 Notation

Let \( \mathbb{N} \) be the set of positive integers and define \( \mathbb{N}_0 := \mathbb{N} \cup \{0\} \). Consider \( \mathbb{R}_+ := (0, \infty) \). For any \( a \in \mathbb{R} \), we denote by \( \lfloor a \rfloor \) the largest integer \( n \) such that \( n \leq a \).

Fix any \( m, n \in \mathbb{N} \). In \( \mathbb{R}^m \), we denote by \( | \cdot | \) the Euclidean norm and by \( B_\eta(x) \) the open ball centered at \( x \in \mathbb{R}^m \) with radius \( \eta > 0 \). For any \( E \subseteq \mathbb{R}^m \), let \( \text{Leb}(E) \) be the Lebesgue measure of \( E \) and \( \text{dim} E \) be the radius of the smallest closed ball that covers \( E \) if such a ball exists. Given another \( E' \subseteq \mathbb{R}^m \), we consider \( \text{dist}(E, E') := \inf_{x \in E, y \in E'} |x - y| \) and write \( E' \subseteq E \) if \( E' \) is compactly embedded in \( E \). For any \( x, y \in \mathbb{R}^m \) and \( M \in \mathbb{R}^{m \times n} \), we denote by \( x' \) and \( M' \) the transpose of \( x \) and \( M \), respectively, and by \( x \cdot y \) the inner product of \( x \) and \( y \).

Given \( f : \mathbb{R}^m \to \mathbb{R} \), let \( D_x f \) and \( D_x^2 f \) denote the gradient and Hessian matrix of \( f \), respectively. For a multi-index \( a = (a_1, \ldots, a_m) \), with \( a_i \in \mathbb{N}_0 \) for all \( i = 1, \ldots, m \), we consider \( |a|_{l_1} := \sum_{i=1}^m a_i \). For any \( k \in \mathbb{N}_0 \) and multi-index \( a \) with \( |a|_{l_1} = k \), \( D_x^a f := \frac{\partial^k f}{\partial x_1^{a_1} \cdots \partial x_m^{a_m}} \) is called a \( k \)th order derivative of \( f \), if it exists. Given a domain \( E \subseteq \mathbb{R}^m \), \( k \in \mathbb{N}_0 \), and \( 0 < \alpha \leq 1 \), we define

\[
[f]_{0, \alpha(E)} := \|f\|_{L^\infty(E)} + \sup_{x, y \in E, x \neq y} \frac{|f(x) - f(y)|}{|x - y|^\alpha},
\]

\[
[f]_{k(E)} := \sum_{|a|_{l_1} = k} \|D_x^a f\|_{L^\infty(E)}, \quad [f]_{k, \alpha(E)} := \sum_{|a|_{l_1} = k} \|D_x^a f\|_{0, \alpha(E)},
\]

\[
\|f\|_{C^k(E)} := \sum_{j=0}^k [f]_{j, \alpha(E)}, \quad \|f\|_{C^{k, \alpha}(E)} := \sum_{j=0}^k [f]_{j, \alpha(E)}.
\]

Let \( C^k(E) \) (resp. \( C^{k, \alpha}(E) \)) be the set of \( k \)-times continuously differentiable \( f : E \to \mathbb{R} \) with \( \|f\|_{C^k(E)} < \infty \) (resp. \( \|f\|_{C^{k, \alpha}(E)} < \infty \)). For \( E = \mathbb{R}^m \), we further consider \( C^{k, \alpha}(\mathbb{R}^m) \), the set of \( f \in C^{k, \alpha}(\mathbb{R}^m) \) with

\[
\|f\|_{C^{k, \alpha}(\mathbb{R}^m)} := \sup_{x \in \mathbb{R}^m} \|f\|_{C^{k, \alpha}(B_1(x))} < \infty.
\]

By definition, \( C^{k, \alpha}(\mathbb{R}^m) \subseteq C^k(\mathbb{R}^m) \). Given \( q \geq 1 \), we also consider \( W^{k, q}(E) \), the set of \( f : E \to \mathbb{R} \) whose weak derivatives up to the \( k \)th order exist and lie in \( L^q(E) \), i.e., \( \|f\|_{W^{k, q}(E)} := \sum_{|a|_{l_1} \leq k} \|D_x^a f\|_{L^q(E)} < \infty \), where \( D_x^a f \) now represents a weak derivative. For any vector-valued \( g = (g_i)_{1 \leq i \leq m} : E \to \mathbb{R}^m \) (resp. a matrix-valued \( g = (g_{ij})_{1 \leq i \leq m, 1 \leq j \leq n} : E \to \mathbb{R}^{m \times n} \)), we will write \( g \in C^k(E) \) whenever \( g_i \in C^k(E) \) for all \( 1 \leq i \leq m \) (resp. \( g_{ij} \in C^k(E) \) for all \( 1 \leq i \leq m, 1 \leq j \leq n \)). Accordingly, we define \( \|g\|_{C^k(E)} = \sum_{i=1}^m \|g_i\|_{C^k(E)} \) (resp. \( \|g\|_{C^{k, \alpha}(E)} = \sum_{1 \leq i \leq m, 1 \leq j \leq n} \|g_{ij}\|_{C^{k, \alpha}(E)} \)). Similar notation is used when we say a vector- or matrix-valued function belongs to \( C^{k, \alpha}(E) \), \( C^{k, \alpha}(\mathbb{R}^m) \), or \( W^{k, q}(E) \).

2 The Setup and Preliminaries

Fix \( d, \tilde{d}, \ell \in \mathbb{N} \). Consider a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) that supports a \( \tilde{d} \)-dimensional Brownian motion \((W_t)_{t \geq 0}\), adapted to a filtration \( \mathbb{F} = (\mathcal{F}_t)_{t \geq 0} \) with \( \mathcal{F}_t \subseteq \mathcal{F} \) for all \( t \geq 0 \). Let \( U \subseteq \mathbb{R}^\ell \) have a positive
finite measure, i.e., $0 < \text{Leb}(U) < \infty$. Denote by $\mathcal{P}(U)$ the set of all probability density functions on $U$. A process $\pi = (\pi_t)_{t \geq 0}$ is called a relaxed control if it is $\mathbb{F}$-adapted with $\pi_t \in \mathcal{P}(U)$ for all $t \geq 0$.

Given Borel measurable functions $b = (b_1, \ldots, b_d) : \mathbb{R}^d \times U \rightarrow \mathbb{R}^d$ and $\sigma = (\sigma_{ij})_{i,j} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, an agent considers, for any relaxed control $\pi$ and $x \in \mathbb{R}^d$, the diffusion process
\[
dX_s^\pi = \left( \int_U b(X_s^\pi, u) \pi_s(u) du \right) ds + \sigma(X_s^\pi) dW_s \quad \text{for } s \geq 0, \quad X_0 = x, \tag{2.1}\]
and the associated value function
\[
V^\pi(x) := \mathbb{E}_x \left[ \int_0^\infty e^{-\rho s} \left( \int_U r(X_s^\pi, u) \pi_s(u) du - \lambda \int U \pi_s(u) \ln \pi_s(u) du \right) ds \right], \tag{2.2}\]
where $\rho > 0$ is the agent’s discount rate, $r : \mathbb{R}^d \times U \rightarrow \mathbb{R}$ is a given reward function, assumed to be Borel measurable, and $\lambda > 0$ is the weight the agent prescribes for the entropy-regularizing term $-\int_U \pi_s(u) \ln \pi_s(u) du$. A relaxed control $\pi$ is said to be admissible if (2.1) admits a unique strong solution and $V^\pi(x)$ is finite for all $x \in \mathbb{R}^d$. We denote by $\mathcal{A}$ the set of all admissible relaxed controls. Furthermore, we call $\pi \in \mathcal{A}$ a Markov relaxed control if there exists a Borel measurable $\mu : \mathbb{R}^d \times U \rightarrow [0, \infty)$ such that $\pi_s(u) = \mu(X_s, u)$ for all $s \geq 0$ and $u \in U$, and denote by $\mathcal{A}_M$ the set of all Markov relaxed controls. Oftentimes, we will write “$\pi \in \mathcal{A}_M$” and “$\mu \in \mathcal{A}_M$” interchangeably.

The agent aims at achieving the optimal value
\[
V^*(x) := \sup_{\pi \in \mathcal{A}} V^\pi(x) \tag{2.3}\]
by finding an optimal relaxed control $\pi^* \in \mathcal{A}$. In the search for $\pi^*$, policy improvement refers to updating an initial guess $\pi^0 \in \mathcal{A}$ to a better relaxed control $\pi^1 \in \mathcal{A}$, in the sense of $V^{\pi^1} \geq V^{\pi^0}$. One may repeat this procedure indefinitely, hoping that $\pi^*$ can be uncovered in the limit.

### 2.1 Policy Iteration Algorithm

In view of the derivation in [26], the HJB equation\(^2\) associated with $V^*$ is
\[
\rho v(x) = \sup_{\varpi \in \mathcal{P}(U)} \int_U \left( b(x, u) \cdot D_x v(x) + r(x, u) - \lambda \ln \varpi(u) \right) \varpi(u) du + \frac{1}{2} \text{tr}(\sigma \sigma^T) D_x^2 v(x), \tag{2.4}\]
and the resulting candidate optimizer $\pi^* \in \mathcal{A}_M$ is given in the Gibbs form
\[
\pi^*(x, u) := \Gamma(x, D_x V^*(x), u), \tag{2.5}\]
where $\Gamma : \mathbb{R}^d \times \mathbb{R}^d \times U \rightarrow \mathbb{R}$ is defined by
\[
\begin{align*}
\Gamma(x, y, u) &:= \arg \max_{\varpi \in \mathcal{P}(U)} \int_U \left( b(x, u) \cdot y + r(x, u) \right) \varpi(u) du \\
&= \frac{\exp(\frac{1}{\lambda} [b(x, u) \cdot y + r(x, u)])}{\int_U \exp(\frac{1}{\lambda} [b(x, \tilde{u}) \cdot y + r(x, \tilde{u})]) d\tilde{u}}. \tag{2.6}\end{align*}
\]
In view of this, we propose the following policy iteration algorithm (PIA):

- Take an arbitrary $v^0 \in C^{1,1}_{\text{uni}}(\mathbb{R}^d)$. For each $n \in \mathbb{N}$, introduce
  \[
  \pi^n(x, u) := \Gamma(x, D_x v^{n-1}(x), u) \in \mathcal{A}_M \quad \text{and} \quad v^n := V^{\pi^n}. \tag{2.7}\]

\(^2\)The equation is also called an exploratory HJB equation, coined by Tang et al. [24], to emphasize its connection to exploration in reinforcement learning.
The goal of this paper is to show that this algorithm is well-defined and fulfills two key properties:

(i) It produces improved relaxed controls recursively, i.e. \( v^n \geq v^{n-1} \) for all \( n \in \mathbb{N} \) (Proposition 4.1);

(ii) It leads to the optimal value function \( V^* \) and yields an optimal relaxed control, i.e., \( v^n \to V^* \) and \( \pi^* \) given in (2.5) is indeed an optimizer for (2.3) (Theorem 4.1).

### 2.2 Two Fundamental Lemmas

Let us present two lemmas that will be useful in Sections 3 and 4 below. First, we observe that \( V^\pi \) has a uniform upper bound (independent of \( \pi \in \mathcal{A} \), as long as the reward function \( r \) is bounded.

**Lemma 2.1.** For any \( \pi \in \mathcal{A} \), \( V^\pi(x) \leq \frac{1}{\rho}(\|r\|_{C^0(\mathbb{R}^d)} + \lambda \ln(\text{Leb}(U))) \) for all \( x \in \mathbb{R}^d \).

**Proof.** Consider the uniform density \( \nu \in \mathcal{P}(U) \) given by \( \nu(u) := 1/\text{Leb}(U) \) for all \( u \in U \). For any \( f \in \mathcal{P}(U) \), we compute the Kullback–Leibler divergence

\[
0 \leq D_{KL}(f||\nu) := \int_U f \ln \left( \frac{f}{\nu} \right) du = \int_U f \ln f du + \ln(\text{Leb}(U)),
\]

which gives \( -\int_U f \ln f du \leq \ln(\text{Leb}(U)) \). In view of (2.2) and this inequality,

\[
V^\pi(x) \leq \int_0^\infty e^{-ps}(\|r\|_{C^0(\mathbb{R}^d)} + \lambda \ln \text{Leb}(U)) ds = \frac{1}{\rho}(\|r\|_{C^0(\mathbb{R}^d)} + \lambda \ln(\text{Leb}(U))) \forall x \in \mathbb{R}^d,
\]

which holds for any \( \pi \in \mathcal{A} \). \(\square\)

Next, we explore regularity of certain functions composed of \( r(x, u), b(x, u) \), and \( \Gamma(x, y, u) \). Assume that there exists \( k \in \mathbb{N}_0 \) such that

\[
\Lambda_k := \sup_{u \in U} \left\{ \|r(\cdot, u)\|_{C^k(\mathbb{R}^d)} + \|b(\cdot, u)\|_{C^k(\mathbb{R}^d)} + \|\sigma\|_{C^k(\mathbb{R}^d)} \right\}
\]

is a finite number. In addition, for any generic function \( f = f(x, y, u) : \mathbb{R}^d \times \mathbb{R}^d \times U \to \mathbb{R} \), we define

\[
\hat{f}(x, y) := \int_U f(x, y, u) \Gamma(x, y, u) du \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d.
\]

We also consider \( \mathcal{H} : \mathbb{R}^d \times \mathbb{R}^d \times U \to \mathbb{R} \) defined by

\[
\mathcal{H}(x, y, u) := \lambda \ln \Gamma(x, y, u)
\]

\[
= b(x, u) \cdot y + r(x, u) - \lambda \ln \left( \int_U \exp \left( \frac{1}{\lambda} [b(x, \tilde{u}) \cdot y + r(x, \tilde{u})] \right) d\tilde{u} \right).
\]

**Lemma 2.2.** Fix \( k \in \mathbb{N}_0 \) and \( p : \mathbb{R}^d \to \mathbb{R}^d \). Set \( f(x, y, u) := r(x, u) - \mathcal{H}(x, y, u) \) (or \( f(x, u) := b(x, u) \)).

(i) If \( \Lambda_k < \infty \) in (2.9) and \( p \in C^k(\mathbb{R}^d) \), then for any multi-index \( a = (a_1, \ldots, a_d) \) with \( |a|_{l_1} = k \),

\[
\left| D_x a \hat{f}(x, p(x)) \right| \leq C_k \left( 1 + |p(x)|^{(k+1)2^k} \right) \left( 1 + \sum_{|c|_{l_1} = 0, \ldots, k} |D_c p(x)| \right)^k \forall x \in \mathbb{R}^d,
\]

where \( C_k > 0 \) is a constant depending on only \( \Lambda_k, k, \lambda, \) and \( d \). Moreover, \( \hat{f}(\cdot, p(\cdot)) \in C^k(\mathbb{R}^d) \).

(ii) If \( \Lambda_{k+1} < \infty \) in (2.9) and \( p \in C^{k,\alpha}_{\text{unif}}(\mathbb{R}^d) \) with \( 0 < \alpha \leq 1 \), then \( \hat{f}(\cdot, p(\cdot)) \in C^{k,\alpha}_{\text{unif}}(\mathbb{R}^d) \).

The proof of Lemma 2.2 is relegated to Appendix A.
3 Value Functions for Gibbs-Form Relaxed Control

In view of (2.7), every \( \pi^n \) generated in the PIA is a Gibbs-form relaxed control given by

\[
\pi(x, u) = \Gamma(x, p(x), u) \quad \text{for some } p : \mathbb{R}^d \to \mathbb{R}^d.
\] (3.1)

In this section, we will investigate properties of the value functions \( V^\pi \) associated with \( \pi \) in the form of (3.1). The first question is whether (3.1) is admissible. The next result provides a sufficient condition, which also ensures that \( V^\pi \) is well-defined (i.e., finitely-valued) and continuous.

Proposition 3.1. Let \( \Lambda_1 < \infty \) in (2.9). For any \( p : \mathbb{R}^d \to \mathbb{R}^d \) of the class \( C^{0,1}_{\text{uni}}(\mathbb{R}^d) \), \( \pi \) given by (3.1) is an admissible relaxed control, whence \( \pi \in \mathcal{A}_M \). Moreover, \( V^\pi \) is bounded, with

\[
|V^\pi(x)| \leq 3\Lambda_0(1 + \|p\|_{C^0(\mathbb{R}^d)} + \lambda|\ln(\text{Leb}(U))|) \quad \forall x \in \mathbb{R}^d,
\]

and also continuous on \( \mathbb{R}^d \).

Proof. By \( \Lambda_1 < \infty \) in (2.9), \( r, b, \) and \( \sigma \) are bounded on \( \mathbb{R}^d \times U \) and globally Lipschitz in \( x \), uniformly in \( u \). In view of (2.6), the boundedness of \( r, b, \) and \( \sigma \), along with \( 0 < \text{Leb}(U) < \infty \), implies that \( \Gamma(x, p(x), u) \) is well-defined and \( u \to \Gamma(x, p(x), u) \) belongs to \( \mathcal{P}(U) \). Hence, for \( \pi(x, u) = \Gamma(x, p(x), u) \) to be an admissible relaxed control, it remains to show that there is a unique strong solution to (2.1) and \( |V^\pi(x)| < \infty \) for all \( x \in \mathbb{R}^d \). Note from (2.10) that (2.1) can be written as

\[
dX_t^\pi = \dot{b}(X_t^\pi, p(X_t^\pi))dt + \sigma(X_t^\pi)dW_t, \quad X_0^\pi = x.
\] (3.2)

By Lemma 2.2 (ii) (with \( k = 0 \)), \( \dot{b}(\cdot, p(\cdot)) \in C^{0,1}_{\text{uni}}(\mathbb{R}^d) \), i.e., \( \dot{b}(\cdot, p(\cdot)) \) is uniformly locally Lipschitz. This readily implies that \( \dot{b}(\cdot, p(\cdot)) \) is globally Lipschitz. Indeed, for any \( y, z \in \mathbb{R}^d \), we can take \( \{y_i\}_{i=0}^n \) on the line segment joining \( y \) and \( z \) such that \( y_0 = y, y_n = z, |y_i - y_{i-1}| \leq 1 \) for all \( i = 1, \ldots, n \), and \( \sum_{i=1}^n |y_i - y_{i-1}| = |y - z| \). Then, by taking \( M = \|\dot{b}(\cdot, p(\cdot))\|_{C^{0,1}_{\text{uni}}(\mathbb{R}^d)} \), we get

\[
|\dot{b}(y, p(y)) - \dot{b}(z, p(z))| \leq \sum_{i=1}^n |\dot{b}(y, p(y_i)) - \dot{b}(z, p(y_{i-1}))| \leq M \sum_{i=1}^n |y_i - y_{i-1}| = M|y - z|,
\]
i.e., \( \dot{b}(\cdot, p(\cdot)) \) is globally Lipschitz. This, together with \( \sigma \) being globally Lipschitz, ensures the existence of a unique strong solution to (2.1).

To prove the boundedness of \( V^\pi \), consider the space \( C([0, \infty); \mathbb{R}^d) \) of continuous paths \( \omega : [0, \infty) \to \mathbb{R}^d \), equipped with the metric \( d(\omega^1, \omega^2) = \sum_{n \in \mathbb{N}} \frac{1}{2^n} \sup_{s \in [0, n]} |\omega^1_s - \omega^2_s| \) for \( \omega^1, \omega^2 \in C([0, \infty); \mathbb{R}^d) \). Recall (2.10) and (2.11). Let us define \( \Delta : (C([0, \infty); \mathbb{R}^d), d) \to (\mathbb{R}, |\cdot|) \) by

\[
\Delta(\omega) := \int_0^\infty e^{-\rho s} \left( \dot{r}(\omega_s, p(\omega_s)) - \dot{\mathcal{H}}(\omega_s, p(\omega_s)) \right) ds
\] (3.3)

and observe from (2.2) that

\[
V^\pi(x) = \mathbb{E}[\Delta(X^\pi)]
\] (3.4)

In view of \( \mathcal{H} \) in (2.11) and \( \Lambda_0 \) in (2.9), we have

\[
\mathcal{H}(x, p(x), u) \leq \Lambda_0 \left( 1 + \|p\|_{C^0(\mathbb{R}^d)} + \lambda|\ln(\exp \left( \frac{1}{\Lambda_0} \left( 1 + \|p\|_{C^0(\mathbb{R}^d)} \right) \right) \text{Leb}(U)) \right)
\leq 2\Lambda_0 \left( 1 + \|p\|_{C^0(\mathbb{R}^d)} + \lambda|\ln(\text{Leb}(U)) \right)< \infty
\] (3.5)
so that $\hat{H}(x, p(x)) = \int_U H(x, p(x), u)\Gamma(x, p(x), u)du \leq 2\Lambda_0(1 + \|p\|_{C^0(\mathbb{R}^d)} + \lambda|\ln(\text{Leb}(U))|$. Similarly, as $|r(x, u)| \leq \Lambda_0$, we have $|\hat{r}(x, p(x))| \leq \Lambda_0$. It follows that $|\hat{r}(. , p(.)) - \hat{H}(., p(.))|_{C^0(\mathbb{R}^d)} \leq K$, with $K := 3\Lambda_0(1 + \|p\|_{C^0(\mathbb{R}^d)} + \lambda|\ln(\text{Leb}(U))|) < \infty$. By (3.4), this implies $|V^\pi(x)| \leq K/\rho < \infty$.

To prove that $V^\pi$ is continuous, take $x \in \{x_n\} \in \mathbb{R}^d$ with $x_n \to x$ and denote by $X^u$ the strong solution to (2.1) with $X_0 = y$, for $y \in \{x, x_1, x_2, \ldots\}$. For any $\varepsilon > 0$, take $T > 0$ such that

$$\int_T^\infty e^{-\rho s}\|\hat{r}(., p(.)) - \hat{H}(., p(.))\|_{C^0(\mathbb{R}^d)}ds \leq \frac{K}{\rho}e^{-\rho T} < \varepsilon. \tag{3.6}$$

Since $\hat{r}(., p(.)) - \hat{H}(., p(.))$ is Lipschitz (by a similar use of Lemma 2.2 as indicated below (3.2)), there exists $L > 0$ such that for any $\omega^1, \omega^2 \in C([0, \infty); \mathbb{R}^d)$,

$$\left|\left(\hat{r}(\omega^1, p(\omega^1)) - \hat{H}(\omega^1, p(\omega^1))\right) - \left(\hat{r}(\omega^2, p(\omega^2)) - \hat{H}(\omega^2, p(\omega^2))\right)\right| \leq L|\omega^1 - \omega^2| \quad \forall s \geq 0. \tag{3.7}$$

Also, by the Lipschitz continuity of $\hat{b}(., p(.))$ and $\sigma(.)$, a classical use of Gronwall’s inequality yields

$$\mathbb{E}\left[\sup_{0 \leq s \leq T} |X^x_s - X^x_s| \right] \leq C_T|x_n - x|, \tag{3.8}$$

where $C_T > 0$ is a constant depending on $T$. We then deduce from (3.3), (3.4) and (3.7) that

$$|V^\pi(x_n) - V^\pi(x)| \leq \mathbb{E}[|\Delta(X^x_n) - \Delta(X^x)|]$$

$$\leq \mathbb{E}\left[\int_0^T e^{-\rho s}|X^x_s - X^x_s|ds + \int_T^\infty e^{-\rho s}2\|\hat{r}(., p(.)) - \hat{H}(., p(.))\|_{C^0(\mathbb{R}^d)}ds\right]$$

$$\leq LC_T(1 - e^{-\rho T})|x_n - x|/\rho + 2\varepsilon,$$

where the last inequality follows from (3.8) and (3.6). This implies $\limsup_{n \to \infty}|V^\pi(x_n) - V^\pi(x)| \leq 2\varepsilon$. As $\varepsilon > 0$ is arbitrarily picked, we conclude $\lim_{n \to \infty}|V^\pi(x_n) - V^\pi(x)| = 0$, as desired. 

To further investigate the regularity of $V^\pi$, we will invoke classical regularity results for elliptic equations. On an open connected domain $E \subseteq \mathbb{R}^d$, given $\gamma > 0$, $\beta : E \to \mathbb{R}^d$, and $\delta : E \to \mathbb{R}^{d \times d}$, we consider the elliptic operator

$$L(\gamma, \beta, \delta)w := -\gamma w + \beta \cdot D_xw + \frac{1}{2} \text{tr}(\delta \delta' D_x^2w), \quad \forall w \in C^2(E). \tag{3.9}$$

The next lemma summarizes results from Theorem 6.13, Theorem 6.17, Problem 6.1 in [6].

**Lemma 3.1.** Fix $k \in \mathbb{N}_0$ and $\alpha \in (0, 1)$. Let $E \subseteq \mathbb{R}^d$ be a bounded connected open domain satisfying the exterior sphere condition. Suppose that $\beta : E \to \mathbb{R}^d$ and $\delta : E \to \mathbb{R}^{d \times d}$ satisfy

$$M_k := \|\beta\|_{C^{k,\alpha}(E)} + \|\delta\|_{C^{k,\alpha}(E)} < \infty \tag{3.10}$$

and there exists $K > 0$ such that

$$\text{tr}(\delta(x)\delta'(x)y'y) \geq K|y|^2 \quad \forall x \in E \text{ and } y \in \mathbb{R}^d. \tag{3.11}$$

Then, for any $\gamma > 0$, $f \in C^{k,\alpha}(E)$, and $\phi : \overline{E} \to \mathbb{R}$ that is continuous, the Dirichlet problem

$$L(\gamma, \beta, \delta)w = f \quad \text{in } E, \tag{3.12}$$

$$w = \phi \quad \text{on } \partial E. \tag{3.13}$$

\^3$E$ is said to satisfy the exterior sphere condition if for each $x \in \partial E$, there exist $y \in \mathbb{R}^d$ and $r > 0$ such that $B_r(y) \subseteq \mathbb{R}^d \setminus \overline{E}$ and $\overline{B_r(y)} \cap E = \{x\}$. In this paper, since all the domains we consider will be either the whole space $\mathbb{R}^d$ or small balls in $\mathbb{R}^d$, this condition always holds.
with $L^{(\gamma, \beta, \delta)}$ as in (3.9), admits a unique solution $w \in C^{k+2,\alpha}(E)$. Moreover, for any $E' \subset E$,

$$
\|w\|_{C^{k+2,\alpha}(E')} \leq C_k(\|w\|_{C^0(E)} + \|f\|_{C^{k,\alpha}(E)}),
$$

(3.14)

for some $C_k > 0$ that depends on $k$, $M_k$, $	ext{dist}(E', \partial E)$, $\dim(E)$, $\gamma$, $\alpha$, $K$, $d$, and $\tilde{d}$.

For any $\pi \in A_M$ given by (3.1), we recall (2.10) and define the elliptic operator

$$
L^\pi w := L^{(\rho, \beta, p(\cdot), \sigma)} w = -\rho w + \beta(\cdot, p(\cdot)) \cdot D_x w + \frac{1}{2} \text{tr}(\sigma(\sigma') D_x^2 w), \quad \forall w \in C^2(\mathbb{R}^d).
$$

(3.15)

**Proposition 3.2.** Fix $\alpha \in (0, 1)$ and let $\Lambda_{k+1} < \infty$ in (2.9) for some $k \in \mathbb{N}_0$. Suppose that $\sigma$ fulfills the ellipticity condition, i.e., there exists $\eta_0 > 0$ such that

$$
\text{tr}(\sigma(x)\sigma(x)') \geq \eta_0 |\xi|^2 \quad \forall \xi, x \in \mathbb{R}^d.
$$

(3.16)

For any $p : \mathbb{R}^d \to \mathbb{R}^d$ of the class $C^{0,1}_{\text{unif}}(\mathbb{R}^d)$, consider $\pi$ given by (3.1). Then, we have $\pi \in A_M$, $V^\pi \in C^{k+2,\alpha}_{\text{unif}}(\mathbb{R}^d)$, and that $V^\pi$ satisfies the elliptic equation

$$
L^\pi V^\pi(x) + \hat{\rho}(x, p(x)) - \hat{\mathcal{H}}(x, p(x)) = 0 \quad \forall x \in \mathbb{R}^d.
$$

(3.17)

**Proof.** By Proposition 3.1, $\pi \in A_M$ and $V^\pi$ is bounded continuous on $\mathbb{R}^d$. For a bounded connected open domain $E \subset \mathbb{R}^d$ satisfying the exterior sphere condition, consider the Dirichlet problem

$$
L^{(\rho, \beta, p(\cdot), \sigma)} w + \hat{\rho}(\cdot, p(\cdot)) - \hat{\mathcal{H}}(\cdot, p(\cdot)) = 0 \quad \text{in} \ E, \quad w = V^\pi \quad \text{on} \ \partial E.
$$

(3.18)

In view of (2.9), $\Lambda_{k+1} < \infty$ directly gives

$$
\sigma \in C^{k+1}(\mathbb{R}^d) \subset C^{k,\alpha}(\mathbb{R}^d) \subset C^{k,\alpha}_{\text{unif}}(\mathbb{R}^d).
$$

(3.19)

Also, with $\Lambda_{k+1} < \infty$ and $p \in C^{k,\alpha}_{\text{unif}}(\mathbb{R}^d)$, Lemma 2.2 (ii) entails

$$
\hat{b}(\cdot, p(\cdot)), \quad \hat{\rho}(\cdot, p(\cdot)) - \hat{\mathcal{H}}(\cdot, p(\cdot)) \in C^{k,\alpha}_{\text{unif}}(\mathbb{R}^d).
$$

(3.20)

As $E$ is a bounded domain, (3.19) and (3.20) readily imply that $\sigma$, $\hat{b}(\cdot, p(\cdot))$, and $\hat{\rho}(\cdot, p(\cdot)) - \hat{\mathcal{H}}(\cdot, p(\cdot))$ all belong to $C^{k,\alpha}(E)$. Such regularity, along with (3.16) and the continuity of $V^\pi$ on $\overline{E}$, allows us to apply Lemma 3.1 to conclude that there exists a unique solution $w \in C^{k+2,\alpha}(E)$ to (3.18). Now, we claim that $w = V^\pi$ on $E$. For any $x \in E$, let $X^\pi$ denote the unique strong solution to (2.1) with $X_0 = x$ and consider the stopping time $\tau_E := \inf\{t \geq 0 : X^\pi_t \not\in E\}$. As (3.16) ensures $\mathbb{E}[\tau_E] < \infty$ (see e.g., [14, Lemma 5.7.4 and Remark 5.7.5]), [14, Proposition 5.7.2] asserts that

$$
w(x) = \mathbb{E} \left[ \int_0^{\tau_E} e^{-\rho t} \left( \hat{\rho}(X^\pi_t, p(X^\pi_t)) - \hat{\mathcal{H}}(X^\pi_t, p(X^\pi_t)) \right) dt + e^{-\rho \tau_E} V^\pi(X^\pi_{\tau_E}) \right] = V^\pi(x),
$$

where the last equality follows from (3.4) and the strong Markov property of $X^\pi$. Hence, $w = V^\pi$ on $E$. As the domain $E$ is arbitrarily picked, this implies that $V^\pi$ satisfies (3.17).

It remains to prove $V^\pi \in C^{k+2,\alpha}_{\text{unif}}(\mathbb{R}^d)$, i.e., $\sup_{x \in \mathbb{R}^d} \|V^\pi\|_{C^{k+2,\alpha}(B_1(x))} < \infty$. By taking $E' = B_1(x)$ and $E = B_2(x)$ in (3.14) and considering the supremum over all $x \in \mathbb{R}^d$, we get

$$
\sup_{x \in \mathbb{R}^d} \|V^\pi\|_{C^{k+2,\alpha}(B_1(x))} \leq \sup_{x \in \mathbb{R}^d} C_k(x) \left( \|V^\pi\|_{C^0(B_2(x))} + \|\hat{\rho}(\cdot, p(\cdot)) - \hat{\mathcal{H}}(\cdot, p(\cdot))\|_{C^{k,\alpha}(B_2(x))} \right)
$$

$$
\leq \left( \sup_{x \in \mathbb{R}^d} C_k(x) \right) \left( \|V^\pi\|_{C^0(\mathbb{R}^d)} + 2\|\hat{\rho}(\cdot, p(\cdot)) - \hat{\mathcal{H}}(\cdot, p(\cdot))\|_{C^{k,\alpha}(\mathbb{R}^d)} \right),
$$

(3.21)
where $C_k(x) > 0$ is a nondecreasing function of $\|\hat{b}(\cdot, p(\cdot))\|_{C^{k,\alpha}(B_2(x))}$ and $\|\sigma\|_{C^{k,\alpha}(B_2(x))}$. Recall from (3.19)-(3.20) that $\sigma, \hat{b}(\cdot, p(\cdot)) \in C_{\text{unif}}^{k,\alpha}(\mathbb{R}^d)$, so that $\sup_{x \in \mathbb{R}^d} \|f\|_{C^{k,\alpha}(B_2(x))} \leq 2\|f\|_{C^{k,\alpha}(\mathbb{R}^d)} < \infty$, for $f = \sigma, \hat{b}(\cdot, p(\cdot))$. Hence, $\sup_{x \in \mathbb{R}^d} C_k(x) < \infty$. This, along with $V^\pi$ being bounded and $\hat{r}(\cdot, p(\cdot)) - \hat{\mathcal{H}}(\cdot, p(\cdot)) \in C_{\text{unif}}^{k,\alpha}(\mathbb{R}^d)$ (by (3.20)), ensures that the right-hand side of (3.21) is finite, as desired.

\[\square\]

**Remark 3.1.** Proposition 3.2 requires both $p \in C_{\text{unif}}^{0,1}(\mathbb{R}^d)$ and $p \in C_{\text{unif}}^{k,\alpha}(\mathbb{R}^d)$, $\alpha \in A$. The former ensures that $\pi$ given by (3.1) is an admissible relaxed control (Proposition 3.1); the latter allows Lemma 2.2 to be applied, so that classical results for elliptic equation (i.e., Lemma 3.1) can be imported for the study of $V^\pi$. In reality, only one of the two requirements takes effect (as it covers the other): when $k = 0$ (resp. $k \geq 1$), “$p \in C_{\text{unif}}^{0,1}(\mathbb{R}^d)$” (resp. “$p \in C_{\text{unif}}^{1,\alpha}(\mathbb{R}^d)$”) takes effect.

When put in the context of policy iteration, Proposition 3.2 reveals how the regularity of $v^n$ in (2.7) evolves. Simply put, the PIA upgrades the regularity of $v^n$ up to the regularity of $b, \sigma$, and $r$.

**Corollary 3.1.** Let $\Lambda_k < \infty$ in (2.9) for some $\bar{k} \in \mathbb{N}$ and assume (3.16). Then, in the PIA (2.7), $\pi^n \in A_M$ for all $n \in \mathbb{N}$ and the sequence $\{v^n\}_{n \geq 0}$ satisfy the following for all $0 < \alpha < 1$:

$$v^n \in C_{\text{unif}}^{n+1,\alpha}(\mathbb{R}^d), \quad \text{for } 1 \leq n \leq \bar{k}; \quad v^n \in C_{\text{unif}}^{k+1,\alpha}(\mathbb{R}^d), \quad \text{for } n \geq \bar{k} + 1.$$

Moreover, for each $n \in \mathbb{N}$,

$$\mathcal{L}^\pi v^n(x) + \hat{r}(x, D_x v^{n-1}(x)) - \hat{\mathcal{H}}(x, D_x v^{n-1}(x)) = 0 \quad \forall x \in \mathbb{R}^d. \tag{3.22}$$

**Proof.** As $v^0 \in C_{\text{unif}}^{1,1}(\mathbb{R}^d)$, we have $D_x v^0 \in C_{\text{unif}}^{0,1}(\mathbb{R}^d)$. By taking $p = D_x v^0$ and $k = 0$ in Proposition 3.2, we have $\pi^1 \in A_M, v^1 \in C_{\text{unif}}^{2,\alpha}(\mathbb{R}^d)$, and that (3.22) holds for $n = 1$. Similarly, by taking $p = D_x v^1 \in C_{\text{unif}}^{1,\alpha}(\mathbb{R}^d)$ and $k = 1$ in Proposition 3.2, we have $v^2 \in A_M, v^2 \in C_{\text{unif}}^{3,\alpha}(\mathbb{R}^d)$, and that (3.22) holds for $n = 2$. Under $\Lambda_k < \infty$, this procedure (i.e., taking $p = D_x v^{n-1}$ and $k = n - 1$ in Proposition 3.2) can be carried out for all $n = 1, 2, \ldots, \bar{k}$, which shows that $v^n \in A_M, v^n \in C_{\text{unif}}^{n+1,\alpha}(\mathbb{R}^d)$ and that (3.22) holds for all $n = 1, 2, \ldots, \bar{k}$. Now, by taking $p = D_x v^{\bar{k}} \in C_{\text{unif}}^{k,\alpha}(\mathbb{R}^d) \subset C_{\text{unif}}^{\bar{k}-1,\alpha}(\mathbb{R}^d)$ and $k = \bar{k} - 1$ in Proposition 3.2, we have $\pi^{\bar{k}+1} \in A_M, v^{\bar{k}+1} \in C_{\text{unif}}^{\bar{k}+1,\alpha}(\mathbb{R}^d)$, and that (3.22) holds for $n = \bar{k} + 1$. By repeating the same procedure (i.e., taking $p = D_x v^{n-1}$ and $k = \bar{k} - 1$ in Proposition 3.2) for all $n > \bar{k} + 1$, we see that $v^n \in A_M, v^n \in C_{\text{unif}}^{k+1,\alpha}(\mathbb{R}^d)$, and (3.22) holds for all $n > \bar{k} + 1$. \[\square\]

## 4 Policy Improvement and the Convergence of PIA

The goal of this section is to establish the convergence of the PIA (2.7) to the optimal value $V^*$ in (2.3). First, we will show that the algorithm does improve a policy recursively, i.e., $\{v^n\}_{n \in \mathbb{N}}$ in (2.7) satisfies $v^{n+1} \geq v^n$, and the limiting value $V^*$ is well-defined; see Proposition 4.1 and Corollary 4.1. Proving $V^* = V^*$ is much more challenging, due to the added entropy term in (2.2). To overcome the technical hurdles, we propose a plan (the “grand plan” in Section 4.1) to find a uniform estimate of $\{v^n\}_{n \in \mathbb{N}}$. The plan’s implementation requires new Sobolev estimates designed specifically for the purpose of policy iteration (Section 4.2) and a nontrivial containment of the entropy growth (Section 4.3). Based on this, Proposition 4.2 gives a uniform bound for $\{v^n\}_{n \in \mathbb{N}}$ and their derivatives, which allows us to prove $v^* = V^*$ and obtain an optimal relaxed control; see Theorem 4.1.

First, under mild regularity conditions, we find that for any relaxed control $\pi$ given by (3.1),

$$\tilde{\pi}(x, u) := \Gamma(x, D_x V^\pi(x), u), \quad \forall (x, u) \in \mathbb{R}^d \times U, \tag{4.1}$$

is again a well-defined relaxed control and it outperforms $\pi$, i.e., $V^\pi \geq V^\pi$. 

10
Proposition 4.1. Let $\Lambda_1 < \infty$ in (2.9) and assume (3.16). For any $p : \mathbb{R}^d \rightarrow \mathbb{R}^d$ of the class $C^{0,1}_{\text{unif}}(\mathbb{R}^d)$, consider $\pi$ given by (3.1) and recall that $\pi \in \mathcal{A}_M$ (Proposition 3.1) and $V^\pi \in C^{2,\alpha}_{\text{unif}}(\mathbb{R}^d)$ for all $0 < \alpha < 1$ (Proposition 3.2). Then, $\tilde{\pi}$ given by (4.1) is well-defined and satisfies $\tilde{\pi} \in \mathcal{A}_M$, $V^{\tilde{\pi}} \in C^{2,\alpha}_{\text{unif}}(\mathbb{R}^d)$ for all $0 < \alpha < 1$, and $V^{\tilde{\pi}}(x) \geq V^\pi(x)$ for all $x \in \mathbb{R}^d$.

Proof. With $D_x V^\pi \in C^{1,\alpha}_{\text{unif}}(\mathbb{R}^d) \subset C^{0,1}_{\text{unif}}(\mathbb{R}^d)$, Proposition 3.1 (with $p = D_x V^\pi$) asserts $\hat{\pi} \in \mathcal{A}_M$. Applying Proposition 3.2 (with $k = 0$ and $p = D_x V^\pi$) then yields $V^{\tilde{\pi}} \in C^{2,\alpha}_{\text{unif}}(\mathbb{R}^d)$ for all $0 < \alpha < 1$. For any $x \in \mathbb{R}^d$, by applying Itô’s formula to $e^{-\rho t} V^\pi(X_t^{\tilde{\pi}})$ and recalling (3.15), we get

$$V^\pi(x) = e^{-\rho T} V^\pi(X_T^{\tilde{\pi}}) - \int_0^T e^{-\rho t} L^\pi V^\pi(X_t^{\tilde{\pi}})dt - \int_0^T e^{-\rho t} D_x V^\pi(X_t^{\tilde{\pi}}) \cdot \sigma(X_t^{\tilde{\pi}})dW_t, \quad \forall T > 0.$$  

As $b, \sigma, V^\pi, D_x V^\pi,$ and $D_x^2 V^\pi$ are bounded (by $\Lambda_1 < \infty$ and $V^\pi \in C^2(\mathbb{R}^d)$), when we take expectation on both sides of the above equality and let $T \rightarrow \infty$, the dominated convergence theorem gives $V^\pi(x) = -\mathbb{E}\left[\int_0^\infty e^{-\rho t} L^\pi V^\pi(X_t^{\tilde{\pi}})dt\right]$. This, along with $V^{\tilde{\pi}}(x) = \mathbb{E}[\Delta(X_t^{\tilde{\pi}})]$ as in (3.4), yields

$$V^{\tilde{\pi}}(x) - V^\pi(x) = \mathbb{E}\left[\int_0^\infty e^{-\rho t} \left(L^\pi V^\pi(X_t^{\tilde{\pi}}) + \hat{\pi}(X_t^{\tilde{\pi}}, D_x V^\pi(X_t^{\tilde{\pi}})) - \hat{H}(X_t^{\tilde{\pi}}, D_x V^\pi(X_t^{\tilde{\pi}}))\right)dt\right]$$

$$= \mathbb{E}\left[\int_0^\infty e^{-\rho t} \left(-\rho V^\pi(X_t^{\tilde{\pi}}) + \frac{1}{2} \text{tr}(\sigma\sigma^T D_x^2 V^\pi(X_t^{\tilde{\pi}})) + \int_U \left(b(u, X_t^{\tilde{\pi}}) \cdot D_x V^\pi(X_t^{\tilde{\pi}}) + r(X_t^{\tilde{\pi}}, u) - \lambda \ln \hat{\pi}(X_t^{\tilde{\pi}}, u)\right)\hat{\pi}(X_t^{\tilde{\pi}}, u)du\right)dt\right]$$

where the second equality above follows from (3.15), (2.10), and (2.11). By the definition of $\tilde{\pi}(x, u) = \Gamma(x, D_x V^\pi(x), u)$ in (2.6), the integral in the last line of (4.2) is larger or equal to the same integral with $\hat{\pi}(X_t^{\tilde{\pi}}, u)$ therein replaced by $\pi(X_t^{\tilde{\pi}}, u)$. Hence, with $\pi(X_t^{\tilde{\pi}}, u)$ in place of $\hat{\pi}(X_t^{\tilde{\pi}}, u)$ in the last line of (4.2), a calculation similar to that in (4.2) gives

$$V^{\pi}(x) - V^\pi(x) \geq \mathbb{E}\left[\int_0^\infty e^{-\rho t} \left(L^\pi V^\pi(X_t^{\tilde{\pi}}) + \hat{\pi}(X_t^{\tilde{\pi}}, p(X_t^{\tilde{\pi}})) - \hat{H}(X_t^{\tilde{\pi}}, p(X_t^{\tilde{\pi}}))\right)dt\right] = 0,$$

where the last equality holds because $V^\pi$ satisfies (3.17) (by Proposition 3.2). \(\square\)

Corollary 4.1. Let $\Lambda_1 < \infty$ in (2.9) and assume (3.16). Then, in the PIA (2.7), we have $\pi^n \in \mathcal{A}_M$ and $v^{n+1} \geq v^n$ for all $n \in \mathbb{N}$. Hence,

$$v^*(x) := \lim_{n \to \infty} v^n(x) \quad \forall x \in \mathbb{R}^d$$

is well-defined (i.e., exists and is finitely-valued).

Proof. For each $n \in \mathbb{N}$, by Corollary 3.1 (with $\tilde{k} = 1$), we have $\pi^n \in \mathcal{A}_M$ and $v^n \in C^{2,\alpha}_{\text{unif}}(\mathbb{R}^d)$ for all $0 < \alpha < 1$. It then follows from Proposition 4.1 that $v^{n+1} \geq v^n$. As the sequence $\{v^n\}_{n \in \mathbb{N}}$ is monotone and has a uniform upper bound (by Lemma 2.1), $v^*$ in (4.3) is well-defined. \(\square\)

Corollary 4.1 shows that, as long as $\Lambda_1 < \infty$ and (3.16) holds, the PIA (2.7) functions properly: it gives rise to a better relaxed control (which generates a larger value) in each iteration, and ultimately points to the largest possible value $v^*$ that cannot be further improved by the algorithm. It is then natural to ask whether $v^*$ coincides with the optimal value $V^*$ in (2.3).
4.1 The Grand Plan

To show that the sequence \(\{u^n\}_{n \in \mathbb{N}}\) in the PIA (2.7) will ultimately lead to the optimal value (i.e., \(u^* = V^*\)), we will focus on proving
\[
\sup_{n \in \mathbb{N}} \|v^n\|_{C^2(\mathbb{R}^d)} < \infty. \tag{4.4}
\]

Once this is established (Proposition 4.2), the Arzela-Ascoli theorem and a modified verification argument will ensure \(u^* = V^*\) and provide an optimal relaxed control (Theorem 4.1).

While \(\|v^n\|_{C^0(\mathbb{R}^d)}\) is known to be uniformly bounded (by Lemma 2.1 and Corollary 4.1), it is difficult to estimate \(\|D_xv^n\|_{C^1(\mathbb{R}^d)} = \|D_xv^n\|_{C^0(\mathbb{R}^d)} + \|D_x^2v^n\|_{C^0(\mathbb{R}^d)}\), due to the added entropy term in (2.2). To circumvent the involved technical challenges, we propose the following plan.

Take \(s := \lfloor d/2 \rfloor + 1\). For \(n \in \mathbb{N}\) large enough, we perform the following estimations in order:

1. For each \(x \in \mathbb{R}^d\), bound \(|D_xv^n(x)| + |D_x^2v^n(x)|\) by \(\|v^n\|_{W^{s+2,2}(B_1(x))}\) (Lemma 4.1).

2. Bound \(\|v^n\|_{W^{s+2,2}(B_1(x))}\) by a power function of the sum of
\[
\|v^{n-i}\|_{L^q(B_\eta(x))}, \quad i = 0, 1, \ldots, s - 1, \quad \text{and} \quad \|\hat{H}(\cdot, D_xv^{n-s-1}(\cdot))\|_{L^q(B_\eta(x))} \tag{4.5}
\]
for some \(q \geq 1\) and \(\eta > 0\) (see (4.31) and (4.33), which reply on Lemmas 4.2 and 4.3).

3. Bound \(|\hat{H}(\cdot, D_xv^{n-s-1}(\cdot))|\) by \(\psi(|D_xv^{n-s-1}(\cdot)|)\), for a suitable function \(\psi\) of logarithmic growth (Lemma 4.4 and Corollary 4.2).

4. Combine the local estimates in Steps 1 and 2. By Step 3 and \(\sup_{N \in \mathbb{N}} \|v^n\|_{C^0(\mathbb{R}^d)} < \infty\), turn the local estimate of \(|D_xv^n(x)| + |D_x^2v^n(x)|\) into a global one (i.e., independent of \(x \in \mathbb{R}^d\)), thereby obtaining a uniform bound for \(\|v^n\|_{C^1(\mathbb{R}^d)}\).

To properly appreciate the above procedure, we point out two important observations.

**Remark 4.1.** The added entropy term in (2.2) (i.e., \(\hat{H}(\cdot, D_xv^{n-s-1}(\cdot))\) therein) shows up in Step 2. It is critical that it has logarithmic growth in \(\|D_xv^{n-s-1}(\cdot)\|_{C^1(\mathbb{R}^d)}\), as stated in Step 3. Since any power of a logarithmic function remains sublinear, by plugging the estimate in Step 3 into the power function of Step 2, we can see \(\|v^n\|_{C^1(\mathbb{R}^d)}\) dominated by \(\|v^{n-s-1}\|_{C^1(\mathbb{R}^d)}\). This readily suggests (4.4), and is made rigorous in the proof of Proposition 4.2; see (4.35) and the arguments below it.

**Remark 4.2.** At first glance, Steps 1–4 above seem a puzzling “detour”: to obtain a uniform bound for \(\|D_xv^n\|_{C^1(\mathbb{R}^d)} = \|D_xv^n\|_{C^0(\mathbb{R}^d)} + \|D_x^2v^n\|_{C^0(\mathbb{R}^d)}\), we look for a local \(W^{s+2,2}\) bound, transform it into a local \(L^q\) bound, and finally switch back to the original \(C^1(\mathbb{R}^d)\) space. It is natural to wonder if one can simply work within \(C^1(\mathbb{R}^d)\) and derive the desired uniform bound by the standard Schauder estimates for elliptic equations. As we will explain subsequently, such a detour may be necessary, precisely because of the presence of entropy regularization; see Remark 4.7 for details.

To begin with Steps 1 and 2 of our plan, we will derive some useful estimates in Sobolev spaces.

4.2 Sobolev Estimates

A standard result in Sobolev spaces (see e.g., [3, Exercise 18, page 292]) states that for any \(w \in W^{s,2}(\mathbb{R}^d)\) with \(s > d/2\), \(w\) is bounded and satisfies
\[
\|w\|_{L^\infty(\mathbb{R}^d)} \leq C\|w\|_{W^{s,2}(\mathbb{R}^d)}, \tag{4.6}
\]
where \(C > 0\) is a constant depending on only \(s\) and \(d\). For the case where the derivatives of \(w\) do not necessarily lie in \(L^2(\mathbb{R}^d)\), we can generalize (4.6) to the following localized version.
Lemma 4.1. For any \( w \in C^s(\mathbb{R}^d) \) with \( s > d/2 \), there exists \( C > 0 \) depending on only \( s \) and \( d \) such that 
\[
|w(x)| \leq C \|w\|_{W^{s,2}(B_1(x))}
\]
for all \( x \in \mathbb{R}^d \).

Proof. For any \( x \in \mathbb{R}^d \), consider the cut-off function \( \xi \in C^\infty(\mathbb{R}^d) \) defined by \( \xi(z) := \exp \left( \frac{|z-x|^2}{|z|^2} \right) \) for \( z \in B_1(x) \) and \( \xi = 0 \) on \( \mathbb{R}^d \setminus B_1(x) \). Note that \( \xi(x) = 1 \) and \( 0 < \xi \leq 1 \) on \( B_1(x) \). It follows that
\[
\|\xi w\|_{W^{s,2}(\mathbb{R}^d)} = \|\xi w\|_{W^{s,2}(B_1(x))} \leq K \|w\|_{W^{s,2}(B_1(x))},
\]
where \( K > 0 \) is a constant depending on the \( L^\infty(B_1(x)) \)-norm of the derivatives of \( \xi \) up to the \( s \)-th order. By the definition of \( \xi \), this \( L^\infty(B_1(x)) \)-norm is independent of \( x \in \mathbb{R}^d \). Hence, \( K > 0 \) is independent of \( x \in \mathbb{R}^d \) and depends on only \( d \) and \( s \). Finally, observe that 
\[
|w(x)| \leq \|\xi w\|_{L^\infty(\mathbb{R}^d)} \leq C \|\xi w\|_{W^{s,2}(\mathbb{R}^d)} \leq CK \|w\|_{W^{s,2}(B_1(x))},
\]
where the second inequality stems from (4.6) (applicable as (4.7) implies \( \xi w \in W^{s,2}(\mathbb{R}^d) \)) and the third inequality is due to (4.7). The desired result follows as \( C \) and \( K \) depend on only \( d \) and \( s \).

Lemma 4.1 provides a local Sobolev estimate of \( w \) with a potentially large order \( s > d/2 \). Our goal now is to reduce the order to a sufficiently smaller one. We first note that the reduction from order 2 to 0 (i.e., from \( W^{2,q} \) to \( W^{0,q} = L^q \)) is readily known, when \( w \) is a solution to an elliptic equation. Specifically, given a bounded connected open domain \( E \subset \mathbb{R}^d \), recall the elliptic operator \( L^{(\gamma,\beta,\delta)} \) in (3.9). As long as \( \beta \in C^0(E) \), \( \delta \in C^0(\mathring{E}) \), and (3.11) is satisfied, for any fixed \( q \geq 1 \), a classical interior \( W^{2,q} \) estimate holds for any \( w \in W^{2,q}(E) \) that satisfies (3.12) a.e.: for any \( E' \subset E \),
\[
\|w\|_{W^{2,q}(E')} \leq C \|w\|_{L^q(E)} + \|f\|_{L^q(E)},
\]
where \( C > 0 \) is a constant that depends on only \( d, q, K \) in (3.11), \( \text{dist}(E', \partial E) \), \( \gamma \), \( \|\beta\|_{C^0(E)} + \|\delta\|_{C^0(E)} \), and the modulus of continuity of \( \delta_{ij} \); see e.g., [1, Theorem 4.2, page 47].

In the following, we will derive a higher-order counterpart of (4.8): for any \( k \in \mathbb{N}_0 \), \( W^{2,q} \) and \( L^q \) in (4.8) will be replaced by \( W^{k+2,q} \) and \( W^{k,q} \) (or \( L^q \)), respectively, for some \( q' > q \). In doing so, we find that the constant \( C > 0 \) will depend on not only \( \|\beta\|_{C^0(E)} \), but also \( \|\beta\|_{W^{k,q'}(E)} \). This additional dependence on \( \|\beta\|_{W^{k,q'}(E)} \) is spelled out explicitly in (4.9) below, our ultimate higher-order interior estimate that reduces order from \( k + 2 \) to \( k \). On one hand, this sets us apart from standard interior estimates in the PDE literature, where dependence on the coefficients of a PDE (including \( \beta \)) is incorporated into a constant multiplier \( C > 0 \); on the other hand, such spelled-out explicit dependence is necessary for Proposition 4.2 below; see Remark 4.3 (i) for details.

Lemma 4.2. Consider the elliptic operator \( L^{(\gamma,\beta,\delta)} \) in (3.9) with \( E = \mathbb{R}^d \) and suppose that (3.11) holds. Then, there exist three nondecreasing functions
\[
A(k,q) : \mathbb{N}_0 \times [1, \infty) \to \mathbb{R}_+ \text{ satisfying } A(k,q) \geq q,
\]
\[
Q(k,q) : \mathbb{N}_0 \times [1, \infty) \to \mathbb{N}_0,
\]
\[
R(k,\eta) : \mathbb{N}_0 \times \mathbb{R}_+ \to \mathbb{R}_+ \text{ satisfying } R(k,\eta) \geq 2\eta,
\]
such that whenever \( \beta \in C^k(\mathbb{R}^d) \) and \( \delta \in C^{k+1}(\mathbb{R}^d) \) for some \( k \in \mathbb{N}_0 \), the following holds: given \( f \in C^k(\mathbb{R}^d) \), any solution \( w \in C^{k+2}(\mathbb{R}^d) \) to (3.12) (with \( E = \mathbb{R}^d \)) satisfies
\[
\|w\|_{W^{k+2,q}(B_\eta(x))} \leq H_k \left( 1 + \|\beta\|_{W^{k,A(k,\eta)}(B_{R(k,\eta)}(x))} \right)^Q(k,q)
\]
\[
\cdot \left( \|w\|_{L^{A(k,\eta)}(B_{R(k,\eta)}(x))} + \|f\|_{W^{k,A(k,\eta)}(B_{R(k,\eta)}(x))} \right),
\]
for all \( x \in \mathbb{R}^d \) and \( (q,\eta) \in [1, \infty) \times \mathbb{R}_+ \), where \( H_k > 0 \) is a constant that depends on only \( k \), \( K \) in (3.11), \( q, \eta, d, \gamma, \) and \( \Phi_k := \|\beta\|_{C^0(\mathbb{R}^d)} + \|\delta\|_{C^{k+1}(\mathbb{R}^d)} \).
Proof. We prove the desired result by induction on \( k \in \mathbb{N}_0 \). For the case \( k = 0 \), by taking
\[
A(0, q) = q, \quad Q(0, q) = 0, \quad \text{and } R(0, \eta) = 2\eta \quad \forall (q, \eta) \in [1, \infty) \times \mathbb{R}_+,
\] (4.10)
we observe from the classical estimate (4.8) that (4.9) holds, with \( H_0 > 0 \) a constant depending on only \( K, q, \eta, d, \gamma \), and \( \Phi_0 \). Now, suppose that for some \( s \in \mathbb{N} \) such that \( \beta, \delta \in C^s(\mathbb{R}^d) \), (4.9) holds for all \( k \leq s - 1 \), with \( A(k, q), Q(k, q), R(k, \eta), \) and \( H_k \) all known for \( k \leq s - 1 \) and \( (q, \eta) \in [1, \infty) \times \mathbb{R}_+ \). Our goal is to establish (4.9) for \( k = s \).

For any \( f \in C^s(\mathbb{R}^d) \), let \( w \in C^{s+2}(\mathbb{R}^d) \) be a solution to (3.12) (with \( E = \mathbb{R}^d \)). Fix \( x \in \mathbb{R}^d \) and \( (q, \eta) \in [1, \infty) \times \mathbb{R}_+ \) and let \( R(s, \eta) \geq 2\eta \) be a number whose precise value will be determined subsequently. For \( j = 1, \ldots, d \), a direct calculation shows that \( D_{x_j}w \in C^{s+1}(\mathbb{R}^d) \) satisfies
\[
L^{(\gamma, \beta, \delta)}(D_{x_j}w) = D_{x_j}(L^{(\gamma, \beta, \delta)}w) - \text{tr}(D_{x_j}(\delta \delta')D_xw) - D_{x_j} \beta \cdot D_xw \\
= D_{x_j}f - \text{tr}(D_{x_j}(\delta \delta')D_xw) - D_{x_j} \beta \cdot D_xw =: \tilde{f} \in C^{s+1}(\mathbb{R}^d).
\] (4.11)
Hence, by the induction hypothesis (particularly, (4.9) with \( k = s - 1 \)), we have
\[
\|D_{x_j}w\|_{W^{s+1, q}(B_\eta(x))} \leq \|w\|_{W^{s+1, q}(B_\eta(x))} \leq 2\eta \]
(4.12)
In the following, we focus on estimating the right-hand side of (4.12). For simplicity, we set
\[
\tilde{q} := A(s - 1, q) \geq q \geq 1 \quad \text{and} \quad \tilde{\eta} := R(s - 1, \eta) \geq 2\eta > 0.
\] (4.13)
First, by the induction hypothesis (particularly, (4.9) with \((k, q, \eta) = (0, \tilde{q}, \tilde{\eta})\)) and (4.10),
\[
\|D_{x_j}w\|_{L^{\tilde{q}}(B_\eta(x))} \leq \|w\|_{W^{2, \tilde{q}}(B_\eta(x))} \leq \tilde{H}_0\left(\|w\|_{L^{\tilde{q}}(B_{2\eta}(x))} + \|f\|_{L^{\tilde{q}}(B_{2\eta}(x))}\right),
\] (4.14)
where \( \tilde{H}_0 > 0 \) is a constant depending on only \( K, d, \gamma, \Phi_0, \) and \((\tilde{q}, \tilde{\eta})\), which in turn depends on \( s, q, \) and \( \eta \). On the other hand, by the definition of \( \tilde{f} \) in (4.11),
\[
\|\tilde{f}\|_{W^{s+1, \tilde{q}}(B_\eta(x))} \leq \|f\|_{W^{s+1, \tilde{q}}(B_\eta(x))} + C_\delta \|w\|_{W^{s+1, \tilde{q}}(B_\eta(x))} + \|D_{x_j} \beta \cdot D_x w\|_{W^{s+1, \tilde{q}}(B_\eta(x))},
\] (4.15)
where \( C_\delta > 0 \) is a constant depending on only \( \|\delta\|_{C^s(\mathbb{R}^d)} \). Moreover, by Hölder’s inequality, the last term on the right-hand side of (4.15) satisfies
\[
\|D_{x_j} \beta \cdot D_x w\|_{W^{s+1, \tilde{q}}(B_\eta(x))} \leq C\|\|w\|_{W^{s+2, \tilde{q}}(B_\eta(x))}\|_{W^{s+2, \tilde{q}}(B_\eta(x))},
\] (4.16)
where \( C > 0 \) is a constant depending on only \( s \) and \( d \). By plugging (4.14), (4.15), and (4.16) into (4.12) and using the fact
\[
\|\cdot\|_{W^{a, c}(E)} \leq \|\cdot\|_{W^{a, \tilde{c}}(E)}, \quad \text{if } a \leq \tilde{a}, \ c \leq \tilde{c}, \ \text{and } E \subset \mathbb{R}^d \text{ is bounded},
\] (4.17)
we obtain
\[
\|D_{x_j}w\|_{W^{s+1, q}(B_\eta(x))} \leq H^* \left(1 + \|\beta\|_{W^{s-1, \tilde{q}}(B_\eta(x))}\right)^{Q(s-1, q)} \left(\|w\|_{L^{\tilde{q}}(B_{2\eta}(x))} + \|f\|_{W^{s, \tilde{q}}(B_{2\eta}(x))}\right) + (1 + \|\beta\|_{W^{s, 2\tilde{q}}(B_\eta(x))}) \|w\|_{W^{s+1, 2\tilde{q}}(B_\eta(x))},
\] (4.18)
where $H^* := 2H_{s-1}(1 + \tilde{H}_0 + C\beta + C) > 0$ depends on only $s$, $K$, $q$, $\eta$, $d$, $\gamma$, and $\Phi_s$. Let us further estimate the last term in the above inequality, by noting that the induction hypothesis (particularly, (4.9) with $(k, q, \eta) = (s - 1, 2\bar{q}, \bar{\eta})$) implies

$$
\|w\|_{W^{s+1,2\bar{q}}(B_{\bar{q}}(x))} \leq \tilde{H}_{s-1} \left( 1 + \|\beta\|_{W^{s-1,2\bar{q}}(B_{R(s-1,\bar{q})}(x))} \right) Q(s-1,2\bar{q}) \cdot \left( \|\dot{w}\|_{L^A(s-1,2\bar{q}')(B_{R(s-1,\bar{q})}(x))} + \|f\|_{W^{s-1,2\bar{q}}(B_{R(s-1,\bar{q})}(x))} \right),
$$

(4.19)

where $\tilde{H}_{s-1} > 0$ is a constant depending on only $k = s - 1$, $K$, $d$, $\gamma$, $\Phi_{s-1}$, and $(2\bar{q}, \bar{\eta})$, which in turn depends on $s$, $q$, and $\eta$. Now, let us define

$$
A(s, q) := A(s - 1, 2\bar{q}) \geq 2\bar{q}, \quad R(s, \eta) := R(s - 1, \bar{\eta}) \geq 2\bar{\eta},
$$

$$
Q(s, q) := Q(s - 1, q) + Q(s - 1, 2\bar{q}) + 1,
$$

(4.20)

and consider the constant $H_s := 2H^*(1 + \tilde{H}_{s-1}) > 0$, which depends on only $s$, $K$, $q$, $\eta$, $d$, $\gamma$, and $\Phi_s$. Then, by plugging (4.19) into (4.18) and using (4.17), we get

$$
\|D_x w\|_{W^{s+1,q}(B_q(x))} \leq H^* \left( 1 + \|\beta\|_{W^{s-1,\bar{q}}(B_{\bar{q}}(x))} \right) Q(s-1,q) \cdot \left( \|\dot{w}\|_{L^\bar{q}(B_{2\bar{q}}(x))} + \|f\|_{W^{s,\bar{q}}(B_{2\bar{q}}(x))} + \tilde{H}_{s-1} \left( 1 + \|\beta\|_{W^{s,A(x,q)}(B_{R(s,q)}(x))} \right) Q(s-1,2\bar{q})+1 \cdot \left( \|\dot{w}\|_{L^{A(x,q)}(B_{R(s,q)}(x))} + \|f\|_{W^{s,A(x,q)}(B_{R(s,q)}(x))} \right) \right) \leq H_s \left( 1 + \|\beta\|_{W^{s,A(x,q)}(B_{R(s,q)}(x))} \right) Q(s,q) \left( \|\dot{w}\|_{L^{A(x,q)}(B_{R(s,q)}(x))} + \|f\|_{W^{s,A(x,q)}(B_{R(s,q)}(x))} \right).
$$

Summing this up over all $j = 1, 2, \ldots, d$ gives (4.9) for $k = s$.

Remark 4.3. In Proposition 4.2 below, we will apply (4.9) recursively to $\{v^n\}_{n \in \mathbb{N}}$ in the PIA (2.7). Specifically, $\beta$ and $f$ in (4.9) will respectively play the roles of $\tilde{b}(\cdot, D_x v^{n-1}(\cdot))$, the controlled drift, and $\tilde{r}(\cdot, D_x v^{n-1}(\cdot))$) $- \tilde{H}(\cdot, D_x v^{n-1}(\cdot))$, the reward plus entropy; see (4.26)-(4.27) below. Note that

(i) As $\beta = \tilde{b}(\cdot, D_x v^{n-1}(\cdot))$ changes across different iterations, it is important that (4.9) spells out the explicit dependence on $\beta$, so that the estimate of $v^n$ can be made precise, in terms of the lower-order Sobolev norms of $\{v^m\}_{m \leq n}$; see (4.31) below.

(ii) The explicit formulas of functions $A(k, q), Q(k, q), R(k, \eta)$ in (4.9) are not needed, although (4.13) and (4.20) above already provide a recursive formula to precisely define these functions.

In view of Remark 4.3, for the eventual application of (4.9) in Proposition 4.2, we still need to compute the $W^{k,q}$-norms of $\tilde{b}(\cdot, p(\cdot))$ and $\tilde{r}(\cdot, p(\cdot)) - \tilde{H}(\cdot, p(\cdot))$.

Lemma 4.3. Fix $q \in [1, \infty)$. Suppose that there exists $k \in \mathbb{N}_0$ such that $\Lambda_k < \infty$ in (2.9) and $p : \mathbb{R}^d \to \mathbb{R}^d$ is of the class $C^k(\mathbb{R}^d)$. Then, for any $x_0 \in \mathbb{R}^d$ and $\eta > 0$, the function $f(x, y, u) := r(x, y) - \mathcal{H}(x, y, u)$ (or, $f(x, u) := b(x, u)$) satisfies

$$
\|\hat{f}(\cdot, p(\cdot))\|_{W^{k,q}(B_q(x_0))} \leq C_k \left( 1 + \|p\|_{W^{k,(k+1)2^{k+1}q}(B_q(x_0))} \right)^{(k+1)2^{k+1}},
$$

where $C_k > 0$ is a constant that depends on only $k$, $\Lambda_k$, $\lambda$, $d$, $q$, and $\eta$. 

\[15\]
Proof. In the following, $C_k > 0$ is a generic constant that may change from line to line but depends on only $k$, $\Lambda_k$, $\lambda$, $d$, $q$, and $\eta$. By (2.12), for any multi-index $a = (a_1, \ldots, a_d)$ with $|a|_{l_1} \in \{0, 1, \ldots, k\}$,\[ ||D_x^a f(\cdot, p(\cdot))||_{L^q(B_n(x_0))} \]
\[ \leq C_k \left\{ \int_{B_n(x_0)} \left( 1 + |p(x)|^{(k+1)2^k} \right)^q \left( 1 + \sum_{|c|_{l_1} = 0, \ldots, k} |D_x^c p(x)| \right)^{kq} \right\} \frac{1}{q} \]
\[ \leq C_k \left\{ \int_{B_n(x_0)} \left( 1 + |p(x)|^{(k+1)2^k} \right)^q \left( 1 + \sum_{|c|_{l_1} = 0, \ldots, k} |D_x^c p(x)| \right)^{(k+1)2^kq} \right\} \frac{1}{q} \]
\[ \leq C_k \left\{ \left( \text{Leb}(B_n(x_0)) + ||p||_{W^{k,(k+1)2^k+1,q}(B_n(x_0))} \right)^{(k+1)2^{k+1}} \right\} \frac{1}{q} \]
where the third inequality follows from Hölder’s inequality and $(b_1 + b_2)^j \leq 2^{j-1}(b_1^j + b_2^j)$ for $b_1, b_2 \in \mathbb{R}$ and $j \geq 1$. Note that $C_k$ in the first inequality above (obtained from (2.12) in Lemma 2.2 (i)) depends on only $k$, $\Lambda_k$, $\lambda$, and $d$; it gets additional dependence on $q$ in the third line, and on $\eta$ in the last line. Summing up the above inequality over all $a$ with $|a|_{l_1} \in \{0, 1, \ldots, k\}$ yields the desired result. \hfill $\Box$

4.3 Logarithmic Growth of the Entropy

Now, we proceed to Step 3 of our plan specified in Section 4.1. While the entropy term in (2.2) has a natural upper bound (i.e., $\ln(\text{Leb}(U))$) thanks to (2.8), it is generally unclear how to control its behavior on the lower end. To obtain an appropriate lower bound that grows only logarithmically, we will impose additional regularity on the functions $r$ and $b$ as well as the action space $U$.

Assumption 4.1. The maps $u \mapsto r(x, u)$ and $u \mapsto b(x, u)$ are Lipschitz, uniformly in $x$; that is,\[ \Theta := \sup_{u_1, u_2 \in \hat{U}, x \in \mathbb{R}^d} \frac{|r(x, u_1) - r(x, u_2)| + |b(x, u_1) - b(x, u_2)|}{|u_1 - u_2|} < \infty. \] (4.21)

We will also assume that the action space $U \subset \mathbb{R}^\ell$ fulfills a uniform cone condition. To properly state the assumption, note that when $\ell > 1$, for any $\alpha \in [0, \pi/2]$,\[ \Delta_\alpha := \{ u = (u_1, \ldots, u_\ell) \in \mathbb{R}^\ell : u_1^2 + \ldots + u_{\ell-1}^2 \leq \tan^2(\alpha) u_\ell^2 \text{ and } u_\ell \geq 0 \} \]
is a cone with vertex $0 \in \mathbb{R}^\ell$, axis $u_1 = u_2 = \ldots = u_{\ell-1} = 0$, and angle $\alpha$. Given $u \in \mathbb{R}^\ell$, the region obtained by a rotation of $u + \Delta_\alpha$ in $\mathbb{R}^\ell$ about $u$ is called a cone with vertex $u$ and angle $\alpha$.\n
Assumption 4.2. When $\ell > 1$, there exist $\zeta > 0$ and $\alpha \in (0, \pi/2]$ such that for any $u \in U$, there is a cone with vertex $u$ and angle $\alpha$ (denoted by cone$(u, \alpha)$) that satisfies $(\text{cone}(u, \alpha) \cap B_\zeta(u)) \subseteq U$. When $\ell = 1$, there exists $\zeta > 0$ such that for any $u \in U$, either $[u - \zeta, u]$ or $[u, u + \zeta]$ is contained in $U$.

Remark 4.4. Assumption 4.2 states that a cone with a fixed size (determined by the fixed length $\zeta > 0$ and angle $\alpha \in (0, \pi/2]$) can be fitted inside $U$ at any $u \in U$. Any convex domain of $\mathbb{R}^\ell$ with a positive Lebesgue measure fulfills Assumption 4.2, and so does a finite disjoint union of such domains.

Remark 4.5. The standing condition “$0 < \text{Leb}(U) < \infty$” and Assumption 4.2 imply that $U$ is bounded. Indeed, if $U$ is not bounded, there exists $u_n \in U \setminus B_n(0)$ for all $n \in \mathbb{N}$. By Assumption 4.2, we have $(\text{cone}(u_n, \alpha) \cap B_{\zeta/12}(u_n)) \subseteq U$, where cone$(u_n, \alpha)$ is a cone with vertex $u_n$ and angle $\alpha$. As $(\text{cone}(u_n, \alpha) \cap B_{\zeta/12}(u_n))_{n \in \mathbb{N}}$ is an infinite collection of disjoint sets of a fixed size, we must have Leb$(U) = \infty$, a contradiction.
Under the above two assumptions, we find that the Gibbs-form function $\Gamma(x,y,u)$ in (2.6) has polynomial growth in $y$, uniformly in $x$ and $u$.

**Lemma 4.4.** Suppose that Assumptions 4.1 and 4.2 hold. Then, the function $\Gamma$ in (2.6) satisfies

\[
\Gamma(x,y,u) \leq C\ell(1 + |y|)^\ell \quad \forall (x,y,u) \in \mathbb{R}^d \times \mathbb{R}^d \times U,
\]

where $C > 0$ is a constant depending on only $\ell$, $\lambda$, $\Theta$, $\zeta$, and $\alpha$.

**Proof.** Fix $x, y \in \mathbb{R}^d$. For an arbitrary $u_0 \in U$, thanks to Assumption 4.1,

\[
| (r(x,u) + b(x,u) \cdot y) - (r(x,u_0) + b(x,u_0) \cdot y) | \leq \Theta(1 + |y|)|u - u_0|, \quad \forall u \in U. \tag{4.22}
\]

This readily implies

\[
\Gamma(x,y,u_0) = \frac{\exp \left( \frac{1}{\lambda} [r(x,u_0) + b(x,u_0) \cdot y] \right)}{\int_U \exp \left( \frac{1}{\lambda} [r(x,u) + b(x,u) \cdot y] \right) du} \leq \frac{1}{\int_U \exp \left( \frac{-\Theta}{\lambda}(1 + |y|)|u - u_0| \right) du}, \tag{4.23}
\]

where the inequality follows from dividing by $\exp(\frac{1}{\lambda}[r(x,u_0) + b(x,u_0) \cdot y])$ the numerator and the denominator and then using the estimate (4.22).

Now, let us prove the desired result for $\ell > 1$ first. By Assumption 4.2,

\[
\int_U e^{-\frac{\Theta}{\lambda}(1+|y|)|u-u_0|} du \geq \int_{\text{cone}(u_0, a) \cap B_\alpha(U)} e^{-\frac{\Theta}{\lambda}(1+|y|)|u-u_0|} du = \int_{\Delta_\alpha \cap B_\zeta(0)} e^{-\frac{\Theta}{\lambda}(1+|y|)|u|} du, \tag{4.24}
\]

where the equality follows from a translation from $u_0$ to $0 \in \mathbb{R}^\ell$ and a proper rotation about $0 \in \mathbb{R}^\ell$. Let us estimate the right-hand side of (4.24) in the next two cases. If $\frac{\Theta}{\lambda}(1 + |y|) \zeta \leq 1$,

\[
\int_{\Delta_\alpha \cap B_\zeta(0)} e^{-\frac{\Theta}{\lambda}(1+|y|)|u|} du \geq e^{-1} \text{Leb}(\Delta_\alpha \cap B_\zeta(0)) =: K_0.
\]

If $\frac{\Theta}{\lambda}(1 + |y|) \zeta > 1$, consider the two positive constants

\[
K_1 := \int_0^\pi \sin^{\ell-2}(\varphi_1) d\varphi_1 \cdots \int_0^\alpha \sin(\varphi_{\ell-2}) d\varphi_{\ell-2} \int_0^\alpha d\varphi_{\ell-1} \quad \text{and} \quad K_2 := \int_0^1 z^{\ell-1} e^{-z} dz.
\]

By using the $\ell$-dimensional spherical coordinates, we have

\[
\int_{\Delta_\alpha \cap B_\zeta(0)} e^{-\frac{\Theta}{\lambda}(1+|y|)|u|} du = K_1 \int_0^\zeta r^{\ell-1} e^{-\frac{\Theta}{\lambda}(1+|y|)r} dr = K_1 \left( \frac{\lambda}{\Theta(1 + |y|)} \right)^\ell \int_0^\zeta e^{-\frac{\Theta}{\lambda}(1+|y|)z} z^{\ell-1} e^{-z} dz \geq K_1 K_2 \left( \frac{\lambda}{\Theta(1 + |y|)} \right)^\ell,
\]

where the second equality follows from the change of variable $z = \frac{\Theta}{\lambda}(1 + |y|)r$. Combining the above two cases, we conclude from (4.23) and (4.24) that $\Gamma(x,y,u_0) \leq \max \left\{ \frac{1}{K_0}, \frac{1}{K_1}, \frac{1}{K_2} \left( \frac{\Theta}{\lambda}(1 + |y|) \right)^\ell \right\}$, which implies the desired result as $K_0, K_1$, and $K_2$ depend on only $\alpha$, $\zeta$, and $\ell$.

For $\ell = 1$, by following arguments similar to the above, with $\Delta_\alpha \cap B_\zeta(0)$, $K_0$, $K_1$, and $K_2$ replaced by $[0, \zeta]$, $e^{-1} \zeta$, 1, and $\int_0^1 e^{-z} dz = 1 - e^{-1}$, respectively, we can obtain the desired estimate. \hfill \Box

Now, recall $\mathcal{H}(x,y)$ defined in (2.10) and (2.11), which stands for the entropy of the density function $u \mapsto \Gamma(x,y,u)$. We observe that the polynomial growth of $y \mapsto \Gamma(x,y,u)$ in Lemma 4.4 swiftly turns into the logarithmic growth of $y \mapsto \mathcal{H}(x,y)$. 

17
Corollary 4.2. Suppose that Assumptions 4.1 and 4.2 hold. Then, there exists $\kappa > 0$, depending on only $\ell$, $\lambda$, $\text{Leb}(U)$, $\Theta$, $\zeta$, and $\alpha$, such that

$$\sup_{x \in \mathbb{R}^d} |\hat{H}(x,y)| \leq \kappa + \ell \ln(1 + |y|), \quad \forall y \in \mathbb{R}^d.$$ 

Proof. By Lemma 4.4, there exists a constant $C > 0$, depending on only $\ell$, $\lambda$, $\Theta$, $\zeta$, and $\alpha$, such that $\ln(\Gamma(x,y,u)) \leq \ell \ln C + \ell \ln(1 + |y|)$ for all $(x,y,u) \in \mathbb{R}^d \times \mathbb{R}^d \times U$. As $u \mapsto \Gamma(x,y,u)$ is by definition a density function (see (2.6)), this implies $\sup_{x \in \mathbb{R}^d} \int_U \ln(\Gamma(x,y,u)) \Gamma(x,y,u) du \leq \ell \ln C + \ell \ln(1 + |y|), \forall y \in \mathbb{R}^d$. On the other hand, (2.8) readily gives $\sup_{x \in \mathbb{R}^d} \int_U \ln(\Gamma(x,y,u)) \Gamma(x,y,u) du \geq -\ln(\text{Leb}(U))$. Hence, in view of (2.10) and (2.11), we conclude that for all $y \in \mathbb{R}^d$

$$\sup_{x \in \mathbb{R}^d} |\hat{H}(x,y)| = \lambda \sup_{x \in \mathbb{R}^d} \int_U \ln(\Gamma(x,y,u)) \Gamma(x,y,u) du \leq \lambda(\ell \ln(\text{Leb}(U))) + \ell \ln C + \ell \ln(1 + |y|),$$

which shows that the desired result holds. \qed

Remark 4.6. As pointed out in Remark 4.1, the logarithmic growth of $\|\hat{H}(\cdot, p(\cdot))\|_{C^0(\mathbb{R}^d)}$ in $\|p\|_{C^0(\mathbb{R}^d)}$ is indispensable for our eventual proof of (4.4). Notably, this modest growth is achieved under the $C^0(\mathbb{R}^d)$-norm, but not under any $C^k(\mathbb{R}^d)$-norm with $k \geq 1$. Indeed, Step 4 in the proof of Lemma 2.2 (i) in Appendix A indicates that $|D_x \hat{H}(\cdot, p(\cdot))|$ has at least linear growth in $\|p\|_{C^1(\mathbb{R}^d)}$.

4.4 Achieving Optimality

Based on the preparations in Sections 4.2 and 4.3, the next result carefully implements the “grand plan” set forth in Section 4.1 and ultimately establishes the desired uniform bound (4.4).

Proposition 4.2. Let $\Lambda_{|d/2|+2} < \infty$ in (2.9) and assume (3.16). Suppose additionally that Assumptions 4.1 and 4.2 hold. Then, the sequence $\{v^n\}_{n \in \mathbb{N}}$ in PIA (2.7) satisfies (4.4).

Proof. For convenience, we take $s := \lfloor d/2 \rfloor + 1$ throughout the proof.

Step 1. For any $x \in \mathbb{R}^d$ and $n \geq s + 2$, we will show that $|D_x v^n(x)| + |D_x^2 v^n(x)|$ is bounded from above by the $L^q$- and $W^{k,q}_c$-norms of $\{v^m\}_{m=n-s}^n$. By Corollary 3.1 (with $k = s + 1$), $v^n \in C^s_{\text{unif}}(\mathbb{R}^d) \subset C^{s+2}(\mathbb{R}^d)$. With $D_x v^n$, $D_x^2 v^n \in C^s(\mathbb{R}^d)$ and $s > d/2$, Lemma 4.1 implies

$$|D_x v^n(x)| + |D_x^2 v^n(x)| \leq C\|v^n\|_{W^{s+2,2}(B_1(x))},$$

where $C > 0$ is a constant that depends on only $s$ and $d$.

As Corollary 3.1 (with $k = s + 1$) also gives $v^m \in C^{(m+1)\wedge(s+2)}(\mathbb{R}^d)$ for all $1 \leq m \leq n$, by recalling the notation in (2.10) and (2.11), for all $2 \leq m \leq n$ we define

$$f^m(\cdot) := \hat{r}(\cdot, D_x v^{m-1}(\cdot)) - \hat{H}(\cdot, D_x v^{m-1}(\cdot)), \quad b^m(\cdot) := \hat{b}(\cdot, D_x v^{m-1}(\cdot)).$$

For each $2 \leq m \leq n$, as $D_x v^{m-1} \in C^{(m-1)\wedge(s+1)}(\mathbb{R}^d)$, by Lemma 2.2 (i) (with $k = (m-1) \wedge (s+1)$ and $p = D_x v^{m-1}$), we have $b^m \in C^{(m-1)\wedge(s+1)}(\mathbb{R}^d)$. Since $v^m$ satisfies (3.22) (recall Corollary 3.1), the conditions $b^m \in C^{(m-1)\wedge(s+1)}(\mathbb{R}^d)$, $\sigma \in C^{s+1}(\mathbb{R}^d)$ (as $\Lambda_{|d/2|+2} < \infty$), and (3.16) allow us to apply Lemma 4.2 to $v^m$. This implies that for any $k \leq (m-1) \wedge (s+1)$ and $(q, \eta) \in [1, \infty) \times \mathbb{R}_+$,

$$\|v^m\|_{W^{k+2,q}(B_{\eta}(x))} \leq H_k \left(1 + \|b^m\|_{W^{k,A(k,\eta)}(B_{R(k,\eta)}(x))}\right)^{Q(k,\eta)} \cdot \left(\|v^m\|_{L^{A(k,\eta)}(B_{R(k,\eta)}(x))} + \|f^m\|_{W^{k,A(k,\eta)}(B_{R(k,\eta)}(x))}\right),$$

where

$$H_k = \left(1 + \|b^m\|_{W^{k,A(k,\eta)}(B_{R(k,\eta)}(x))}\right)^{Q(k,\eta)} \cdot \left(\|v^m\|_{L^{A(k,\eta)}(B_{R(k,\eta)}(x))} + \|f^m\|_{W^{k,A(k,\eta)}(B_{R(k,\eta)}(x))}\right).$$
where \( H_k > 0 \) is a constant depending on only \( k, \Lambda_k, d, q, \eta, \) and \( \rho \). Furthermore, by taking
\[
\bar{q} := A(k, q) \geq q \quad \text{and} \quad \bar{\eta} := R(k, \eta) \geq 2\eta,
\]
we deduce from Lemma 4.3 (with \( p = D_xv^{m-1} \)) that
\[
\max \left\{ \| f^m \|_{W^{k,4}(B_0(x))}, \| b^m \|_{W^{k,4}(B_0(x))} \right\} \leq C_k \left( 1 + \| v^{m-1} \|_{W^{k+1,4}(B_0(x))} \right)^{ak},
\]
where we take \( a_k := (k+1)2^{k+1} \) and \( C_k > 0 \) is a constant depending on only \( k, \Lambda_k, \lambda, d, q, \eta \). By plugging (4.29) into (4.27), we obtain
\[
\| v^m \|_{W^{k+2,4}(B_0(x))} \leq H_k \left( 1 + C_k^{1/ak} \| \tilde{q} \|_{A(k,q)} \left( 1 + \| v^{m-1} \|_{W^{k+1,4}(A(k,q)(B_{R(k,\eta)}(x)))} \right)^{ak} \right)
\]
\[
\cdot \left( 1 + C_k^{1/ak} \| \tilde{q} \|_{L^1(A(k,q)(B_{R(k,\eta)}(x)))} + \| v^{m-1} \|_{W^{k+1,4}(A(k,q)(B_{R(k,\eta)}(x)))} \right)^{ak}
\]
\[
\leq \tilde{H}_k \left( 1 + \| v^m \|_{L^1(A(k,q)(B_{R(k,\eta)}(x)))} + \| v^{m-1} \|_{W^{k+1,4}(A(k,q)(B_{R(k,\eta)}(x)))} \right)^{ak}.
\]
where \( \tilde{H}_k := H_k \left( 1 + C_k^{1/ak} \| \tilde{q} \|_{A(k,q)(q)+1} \right) > 0 \) is finite and depends on only \( k, \Lambda_k, \lambda, d, q, \eta, \) and \( \rho \).

Now, set \( (k_0, q_0, \eta_0) := (s, 2, 1) \) and define \( \{(k_i, q_i, \eta_i, \theta_i)\}_{i=1}^s \) by
\[
k_i := k_{i-1} - 1, \quad q_i := a_{k_{i-1}} A(k_{i-1}, q_{i-1}) - A(k_{i-1}, q_{i-1}),
\]
\[
\eta_i := R(k_{i-1}, \eta_{i-1}) - 2\eta_{i-1}, \quad \theta_i := a_{k_{i-1}} (Q(k_{i-1}, q_{i-1}) + 1),
\]
for all \( i = 1, \ldots, s \). It then follows from (4.25) and (4.30) that
\[
|D_xv^m(x)| + |D_x^2v^m(x)| \leq C \| v^m \|_{W^{s+2,2}(B_1(x))} \leq C \tilde{H}_s \left( 1 + \| v^n \|_{L^1(B_1(x))} + \| v^{n-1} \|_{W^{s+1,4}(B_1(x))} \right)^{\theta_1}
\]
\[
\leq C \tilde{H}_s \left( 1 + \tilde{H}_s \right)^{1/\theta_2} \left( 1 + \| v^n \|_{L^1(B_1(x))} + \| v^{n-1} \|_{W^{s+1,4}(B_1(x))} + \| v^{n-2} \|_{W^{s+2,4}(B_1(x))} \right)^{\theta_1\theta_2},
\]
where the second inequality follows by taking \( (m, k, q, \eta) = (n, k_0, q_0, \eta_0) \) in (4.30), while the third inequality holds similarly by taking \( (m, k, q, \eta) = (n - i, k_i, q_i, \eta_i) \) in (4.30). By continuing this procedure (i.e., taking \( (m, k, q, \eta) = (n - i, k_i, q_i, \eta_i) \) in (4.30) for all \( i = 2, \ldots, s \), we obtain
\[
|D_xv^m(x)| + |D_x^2v^m(x)| \leq H^* \left( 1 + \sum_{i=0}^{s-1} \| v^{n-i} \|_{L^{q_i+1}(B_{q_i+1}(x))} + \| v^{n-s} \|_{W^{2,s}(B_{q_s}(x))} \right)^{\theta_1\theta_2 \cdots \theta_s},
\]
where \( H^* := C \tilde{H}_s \tilde{H}_{q_s} \left( 1 + \tilde{H}_s \right)^{1/\theta_1} \theta_1 \cdots \theta_{i+1} > 0 \) is finite and depends on only \( s, \Lambda_s, \lambda, d, q, \eta, \) and \( \rho \).

Step 2. We will show that the right-hand-side of (4.31) can be made independent of \( x \in \mathbb{R}^d \). By Lemma 2.1 and Corollary 4.1,
\[
\| v^m \|_{C^0(\mathbb{R}^d)} \leq \max \{ \| v^1 \|_{C^0(\mathbb{R}^d)}, \left( \| r \|_{C^0(\mathbb{R}^d)} + \lambda |\ln(\text{Leb}(U))| \right) / \rho \} \leq: M < \infty, \quad \forall n \in \mathbb{N}.
\]

Then, by taking \( (m, k, q, \eta) = (n - s, 0, q_s, \eta_s) \) in (4.27) and recalling (4.26), we get
\[
\| v^{n-s} \|_{W^{2,s}(B_{q_s}(x))} \leq H_0 \left( 1 + \| b^{n-s} \|_{L^{A(0,q_s)}(B_R(0,q_s)(x))} \right)^{Q(0,q_s)}
\]
\[
\cdot \left( \| v^{n-s} \|_{L^{A(0,q_s)}(B_R(0,q_s)(x))} + \| f^{n-s} \|_{L^{A(0,q_s)}(B_R(0,q_s)(x))} \right)^{Q(0,q_s)}
\]
\[
\leq H_0 \left( 1 + \| b \|_{C^0(\mathbb{R}^d)} (\text{Leb}(B_R(0,q_s)(x))) \right)^{\frac{1}{A(0,q_s)}} \left( 1 + \| r \|_{C^0(\mathbb{R}^d)} + \| \tilde{H} \|_{C^0(\mathbb{R}^d)} + \| \tilde{H} \|_{(\mathbb{R}^d)} \right)^{Q(0,q_s)}
\]
\[
\cdot \left( \text{Leb}(B_R(0,q_s)(x)) \right)^{\frac{1}{A(0,q_s)}} \left( M + \| r \|_{C^0(\mathbb{R}^d)} + \| \tilde{H} \|_{C^0(\mathbb{R}^d)} + \| \tilde{H} \|_{(\mathbb{R}^d)} \right)^{Q(0,q_s)}.
\]
As $\text{Leb}(B_{R(0,n_0)}(x))$ is a constant independent of $x \in \mathbb{R}^d$, we deduce from Corollary 4.2 that the above inequality can be turned into $\|v^{n-s}\|_{W^{2,q}\text{unif}(B_{R_0}(x))} \leq C^*(1 + \ln(1 + \|D_x v^{n-s-1}\|_{C^0(\mathbb{R}^d)}))$, where $C^* > 0$ is finite and depends on only $s$, $\Lambda$, $\lambda$, $d$, $\rho$, $\ell$, $\text{Leb}(U)$, $\Theta$, $\zeta$, and $\alpha$. Thus, (4.31) becomes

\[
|D_x v^n(x)| + |D_x^2 v^n(x)| \\
\leq H^* \left(1 + M \sum_{i=1}^{s} \text{Leb}(B_{R_i}(x))^{1/q_i} + C^* \left(1 + \ln \left(1 + \|D_x v^{n-s-1}\|_{C^0(\mathbb{R}^d)}\right)\right)^{\theta_1 \theta_2 \ldots \theta_s}\right),
\]

(4.34)

where $\bar{C}^* > 0$ is a constant depending on only $s$, $\Lambda$, $\lambda$, $d$, $\rho$, $\ell$, $\text{Leb}(U)$, $\Theta$, $\zeta$, and $\alpha$.

Step 3. We are now ready to prove the desired result (4.4). As we already know from (4.32) that $\sup_{n \in \mathbb{N}} \|v^n\|_{C^0(\mathbb{R}^d)} < \infty$, it suffices to show $\sup_{n \in \mathbb{N}} \|D_x v^n\|_{C^1(\mathbb{R}^d)} < \infty$. By the arbitrariness of $x \in \mathbb{R}^d$ in (4.34) and the fact $\|\cdot\|_{C^0(\mathbb{R}^d)} \leq \|\cdot\|_{C^1(\mathbb{R}^d)}$, we have

\[
\|D_x v^n\|_{C^1(\mathbb{R}^d)} \leq \bar{C}^* \left(1 + \ln \left(1 + \|D_x v^{n-s-1}\|_{C^1(\mathbb{R}^d)}\right)\right)^{\theta_1 \theta_2 \ldots \theta_s}, \quad \forall n \geq s + 2.
\]

(4.35)

Note that $\phi(z) := \bar{C}^* \left(1 + \ln(1 + z)\right)^{\theta_1 \theta_2 \ldots \theta_s}$, for $z \geq 0$, satisfies $\phi(0) = \bar{C}^* > 0$ and is strictly increasing and sublinear (i.e., $\phi(z)/z \to 0$ as $z \to \infty$). Hence, $z_0 := \sup\{z \geq 0 : z \leq \phi(z)\} < \infty$. Consider

$z^* := \max\left\{\max_{1 \leq n \leq s+1} \|D_x v^n\|_{C^1(\mathbb{R}^d)}, z_0\right\} < \infty$. By definition, $\|D_x v^n\|_{C^1(\mathbb{R}^d)} \leq z^*$ for $n = 1, \ldots, s+1$. Moreover, by (4.35),

\[
\max_{s+2 \leq n \leq 2(s+1)} \|D_x v^n\|_{C^1(\mathbb{R}^d)} \leq \bar{C}^* \left(1 + \ln \left(1 + \max_{1 \leq n \leq s+1} \|D_x v^n\|_{C^1(\mathbb{R}^d)}\right)\right)^{\theta_1 \theta_2 \ldots \theta_s}
\]

(4.36)

\[
= \phi\left(\max_{1 \leq n \leq s+1} \|D_x v^n\|_{C^1(\mathbb{R}^d)}\right) \leq \phi(z^*) \leq z^*,
\]

where the last inequality follows from $z^* \geq z_0$ and the definition of $z_0$. That is, $\|D_x v^n\|_{C^1(\mathbb{R}^d)} \leq z^*$ for $n = s+2, \ldots, 2(s+1)$. By repeating the argument in (4.36), we can show $\|D_x v^n\|_{C^1(\mathbb{R}^d)} \leq z^*$ for $n = f(s+1)+1, \ldots, (j+1)(s+1)$, with $j = 2, 3, \ldots$. We then conclude $\sup_{n \in \mathbb{N}} \|D_x v^n\|_{C^1(\mathbb{R}^d)} \leq z^* < \infty$. \hfill \square

As pointed out in Remark 4.2, the “grand plan” in Section 4.1, followed closely by the proof above, seems an overly complicated detour. Such a detour might be necessary as it turns out.

**Remark 4.7.** It is tempting to prove $\sup_{n \in \mathbb{N}} \|D_x v^n\|_{C^1(\mathbb{R}^d)} < \infty$ directly in the space $C^1(\mathbb{R}^d)$. By using a Schauder estimate (precisely for the second inequality below), one gets

\[
\|D_x v^n\|_{C^1(\mathbb{R}^d)} \leq \|D_x v^n\|_{C^1_{\text{unif}}(\mathbb{R}^d)} \leq K_n \left(1 + \|\hat{H}(\cdot, D_x v^{n-1}(\cdot))\|_{C^0_{\text{unif}}(\mathbb{R}^d)}\right) \\
\leq K_n \left(1 + \|\hat{H}(\cdot, D_x v^{n-1}(\cdot))\|_{C^1(\mathbb{R}^d)}\right),
\]

(4.37)

where $K_n > 0$, the usual constant in a Schauder estimate, shall take the form $K_n = g(\|D_x v^{n-1}\|_{C^0_{\text{unif}}(\mathbb{R}^d)})$ for some increasing $g : \mathbb{R}_+ \to \mathbb{R}_+$, in the present context of policy iteration. Ideally, if $\{K_n\}_{n \in \mathbb{N}}$ has a uniform upper bound and $\|\hat{H}(\cdot, D_x v^{n-1}(\cdot))\|_{C^1(\mathbb{R}^d)}$ grows sublinearly in $\|D_x v^{n-1}\|_{C^1(\mathbb{R}^d)}$, the right-hand side of (4.37) will grow sublinearly in $\|D_x v^{n-1}\|_{C^1(\mathbb{R}^d)}$, similarly to (4.35). The estimates below (4.35) can then be applied to get $\sup_{n \in \mathbb{N}} \|D_x v^n\|_{C^1(\mathbb{R}^d)} < \infty$. This argument, however, breaks down because $\|\hat{H}(\cdot, D_x v^{n-1}(\cdot))\|_{C^1(\mathbb{R}^d)}$ grows at least linearly in $\|D_x v^{n-1}\|_{C^1(\mathbb{R}^d)}$ (see Remark 4.6).

That is to say, the true advantage of our “grand plan” is that it allows us to control $\|D_x v^n\|_{C^1(\mathbb{R}^d)}$ by the $C^0(\mathbb{R}^d)$-norm of $\hat{H}(\cdot, D_x v^{n-1}(\cdot))$, instead of its $C^1(\mathbb{R}^d)$-norm (see (4.31) and (4.33)), so that the crucial sublinear growth can be imported from Corollary 4.2.
Now, we are ready to present the main result of this paper: with appropriate regularity on \( r, b, \sigma, \) and the action space \( U, \) the PIA (2.7) yields the optimal value and an optimal relaxed control.

**Theorem 4.1.** Let \( \Lambda_{[d/2]+2} < \infty \) in (2.9) and assume (3.16). Suppose additionally that Assumptions 4.1 and 4.2 hold. Then, the limiting function \( v^* \) in the PIA (2.7), defined as in (4.3), satisfies

(i) \( v^* = V^* \) on \( \mathbb{R}^d \) and \( \pi^*(x, u) := \Gamma(x, D_xv^*(x), u) \in \mathcal{A}_M \) is an optimal relaxed control for (2.3);

(ii) \( v^* \) is the unique solution in \( C^2(\mathbb{R}^d) \) to the HJB equation (2.4);

(iii) \( v^* \in C^{[d/2]+3, \alpha}(\mathbb{R}^d) \) for all \( 0 < \alpha < 1. \)

**Proof.** Consider \( \{v^n\}_{n \in \mathbb{N}} \) in the PIA (2.7) and recall from Corollary 4.1 that its limit \( v^* \) is well-defined. For any compact subset \( E \) of \( \mathbb{R}^d, \) Proposition 4.2 implies that \( \{v^n\}_{n \in \mathbb{N}} \) (resp. \( \{D_xv^n\}_{n \in \mathbb{N}} \)) is uniformly bounded and equicontinuous. By the Arzela–Ascoli theorem, \( v^n \) (resp. \( D_xv^n \)) converges uniformly on \( E, \) up to a subsequence. We then conclude \( v^* \in C^1(E) \) and \( D_xv^* \rightarrow D_xv^* \). Recall from Corollary 3.1 that \( v^n \) satisfies (3.22) for all \( n \in \mathbb{N}. \) By the uniform convergence of \( v^n \) and \( D_xv^n \) on \( E, \) as \( n \rightarrow \infty \) in (3.22), we observe that \( D_x^2v^n \) must also converge uniformly on \( E. \) This in turn implies \( D_xv^* \in C^1(E) \) and \( D_x^2v^* \rightarrow D_x^2v^*. \) Hence, \( v^* \in C^2(E) \) and the limiting equation of (3.22) on \( E \) (as \( n \rightarrow \infty \)) is

\[
\mathcal{L}^\pi v^*(x) + \hat{r}(x, D_xv^*(x)) = 0.
\]  

By the definitions of \( \mathcal{L}^\pi, \hat{r}, \hat{H}, \) and \( \Gamma \) (see (3.15), (2.10), (2.11), and (2.6)), this readily shows that \( v^* \) satisfies the HJB equation (2.4) on \( E. \) As \( E \subset \mathbb{R}^d \) is arbitrary, we conclude that \( v^* \in C^2(\mathbb{R}^d) \) and it satisfies (2.4) on \( \mathbb{R}^d. \) We then have \( \pi^* \in \mathcal{A}_M \) by Proposition 3.1. Also, a modified verification argument (for tackling the entropy term) shows \( v^* = V^* \) on \( \mathbb{R}^d. \) To see this, for any \( x \in \mathbb{R}^d \) and \( \pi \in \mathcal{A}, \) by Itô’s formula (applied to \( v^* (X^\pi) \)) and the boundedness of \( D_xv^* \) and \( \sigma, \) it holds for all \( T > 0 \)

\[
v^*(x) = \mathbb{E} \left[ e^{-\rho T}v^*(X^\pi_T) + \int_0^T e^{-\rho t} \left\{ \rho v^* - \left( \int_U b(\cdot, u)\pi_t(u)du \right) \cdot D_xv^* - \frac{1}{2} \text{tr}(\sigma\sigma^t D_x^2v^*) \right\} (X^\pi_t)dt \right]
\]

\[
= \mathbb{E} \left[ e^{-\rho T}v^*(X^\pi_T) + \int_0^T e^{-\rho t} \left( \int_U r(X^\pi_t, u)\pi_t(u)du \right) dt \right] + \lambda \mathbb{E} \left[ \int_0^T e^{-\rho t} \left( \int_U \pi_t(u) \ln \pi_t(u)du \right) dt \right] - \lambda \mathbb{E} \left[ \int_0^T e^{-\rho t} \left( \int_U \pi_t(u) \ln \pi_t(u)du \right) dt \right] - \lambda \mathbb{E} \left[ \int_0^T e^{-\rho t} \left( \int_U \pi_t(u) \ln \pi_t(u)du \right) dt \right] + \lambda \mathbb{E} \left[ \int_0^T e^{-\rho t} \left( \int_U \pi_t(u) \ln \pi_t(u)du \right) dt \right]
\]

where the inequality holds as \( v^* \) satisfies the HJB equation (2.4). As \( T \rightarrow \infty, \) since \( v^* \) and \( r \) are bounded and \( \int_U \pi_t(u) \ln \pi_t(u)du \) is bounded from below by \( -\ln(\text{Leb}(U)) \) (recall (2.8)), we deduce from the dominated convergence theorem and Fatou’s lemma that

\[
v^*(x) \geq \mathbb{E} \left[ \int_0^T e^{-\rho t} \left( \int_U r(X^\pi_t, u)\pi_t(u)du \right) - \lambda \int_U \pi_t(u) \ln \pi_t(u)du \right] dt = V^\pi(x).
\]

Taking supremum over \( \pi \in \mathcal{A} \) yields \( v^*(x) \geq V^*(x). \) When we take \( \pi = \pi^* \in \mathcal{A}_M \) in (4.39), the inequality becomes an equality as \( \pi^*(x, u) = \Gamma(x, D_xv^*(x), u) \) attains the supremum in (2.4) (with \( v \) taken to be \( v^* \)), thanks to (2.6). This gives \( v^*(x) = \mathbb{E}[e^{-\rho T}v^*(X^\pi_T)] + \mathbb{E}[\int_0^T e^{-\rho t}(\hat{r} - \hat{H})(X^\pi_t, D_xv^*(X^\pi_t))dt]. \) As \( T \rightarrow \infty, \) since \( v^* \) and \( (\hat{r} - \hat{H}) \) are bounded by Lemma 2.2 (i) with \( k = 0, \) the dominated convergence theorem implies \( v^*(x) = V^\pi^*(x) \geq V^*(x). \) We then conclude \( v^* = V^* \) on \( \mathbb{R}^d \) and \( \pi^* \) is an optimal relaxed control for (2.3). For any solution \( w \in C^2(\mathbb{R}^d) \) to (2.4), the same arguments as above (with \( w \) in place of \( v^* \)) entail \( w = V^*. \) Hence, \( v^* = V^* \) is the unique solution in \( C^2(\mathbb{R}^d) \) to (2.4).

As \( v^* \in C^2(\mathbb{R}^d) \subset C^{1,1}(\mathbb{R}^d), \) we take \( v^0 = v^* \) in the PIA (2.7), then \( v^n = v^* \) for all \( n \in \mathbb{N}. \) With \( \Lambda_{[d/2]+2} < \infty, \) Corollary 3.1 readily implies \( v^* \in C^{[d/2]+3, \alpha}(\mathbb{R}^d) \) for all \( 0 < \alpha < 1. \)
Remark 4.8. At the end of [24, Section 6], an open question is raised: under what conditions on \(r, b, \) and \(\sigma\) does the (exploratory) HJB equation (2.4) have a unique solution of the class \(C^3(\mathbb{R}^d)\)? Theorem 4.1 provides an answer to this: Under the conditions that (i) \(r(\cdot, u), b(\cdot, u), \) and \(\sigma(\cdot, u)\) belong to \(C^{d/2+2}(\mathbb{R}^d)\), (ii) \(r(x, \cdot)\) and \(b(x, \cdot)\) are Lipschitz, (iii) \(\sigma\) is uniformly elliptic, and (iv) the action space \(U\) satisfies a uniform cone condition, \(V^*\) in (2.3) is characterized as the unique classical solution to (2.4) with desirable regularity, i.e., \(V^* \in C^{3,\alpha}_{\text{unif}}(\mathbb{R}^d) \subset C^3(\mathbb{R}^d)\).

4.5 Discussion: The Case of a Controlled Diffusion Coefficient

When the diffusion coefficient of (2.1) is also controlled, i.e.,

\[
dX_t^\pi = \left( \int_U b(X_t^\pi, u)\pi_t(u)du \right) dt + \left( \int_U (\sigma\sigma')(X_t^\pi, u)\pi_t(u)du \right)^{1/2} dW_t,
\]

the Gibbs-form function \(\Gamma\) in (2.6) should take the following form: for any \((x, y, z, u) \in \mathbb{R}^d \times \mathbb{R}^d \times S^d \times U\), where \(S^d\) denotes the set of all \(d \times d\) symmetric matrices,

\[
\Gamma(x, y, z, u) := \frac{1}{\int_U \exp \left( \frac{1}{\lambda} [b(x, u) \cdot y + \frac{1}{2} \text{tr}(\sigma\sigma'(x, u)z) + r(x, u)] \right) du}. \tag{4.40}
\]

With sufficient regularity of \(b, r, \) and \(\sigma, \) a policy improvement result similar to Proposition 4.1 still holds; see e.g., [13, Theorem 1]. The convergence of the PIA (i.e., (2.7) with \(\Gamma(x, D_xv^{n-1}(x), u)\) replaced by \(\Gamma(x, D_xv^{n-1}(x), u)\)) is nonetheless much more involved and remains largely open. As explained before, our “grand plan” does not easily apply to the present case.

To obtain a uniform estimate for \(\{v^n\}_{n \in \mathbb{N}}\) in the PIA (2.7), the proof of Proposition 4.2 above relies on the “order reduction” in (4.30): for \(m \in \mathbb{N}\) large, the \(W^{k+2,q}\)-norm of \(v_m^{n-1}\) is controlled by the \(W^{k+1,q}\)-norm of \(v^{m-1}\), with \(q_1 \geq q\). By repeating this procedure, we ultimately control the \(W^{k+2,q}\)-norm of \(v_m^{n-1}\) by the \(W^{2,q_2}\)-norm of \(v^{m-k}\), with \(q_2 \geq q_1\). As shown in (4.33), this \(W^{2,q_2}\)-norm of \(v^{m-k}\) can be expressed in terms of the \(C^0(\mathbb{R}^d)\)-norm of the entropy term \(\tilde{H}\), which is properly controlled in Section 4.3. This then provides a uniform estimate for \(v_m\), with \(m \in \mathbb{N}\) large enough.

With a controlled diffusion coefficient, such “order reduction” no longer holds. As \(\Gamma\) depends additionally on \(z \in S^d, f^m\) and \(b^m\) in (4.26) now also depend on \(D_x^2v^{n-1}(\cdot)\). The same estimation as in Lemma 4.3 then shows that the \(W^{k,q}\)-norms of \(f^m\) and \(b^m\) are controlled by the \(W^{k+q'}\)-norms of not only \(D_xv^{n-1}\) but also \(D_x^2v^{n-1}\), with \(q' \geq q\). Thus, on the right-hand side of (4.29), the \(W^{k+1,q'}\)-norm of \(v^{n-1}\) should now be the \(W^{k+2,q'}\)-norm. When plugging this into (4.27), we obtain an updated version of (4.30), with the \(W^{k+1,q'}\)-norm of \(v^{n-1}\) on the right-hand side replaced by its \(W^{k+2,q'}\)-norm. There is then no order reduction: on both sides of the updated (4.30), we see a Sobolev norm of order \(k + 2\). It is then unclear how Step 2 of the “grand plan” (see (4.5)) can be achieved.

5 An Example of Optimal Consumption

To illustrate how Theorem 4.1 can be applied, let us present an optimal consumption problem from mathematical finance. Take \(d = d = \ell = 1\). Consider a risky asset \(S\) given by \(dS_t/S_t = (r + \mu)dt + \sigma dW_t\), where \(r > 0, \mu \geq 0, \) and \(\sigma > 0\) are the riskfree rate, the risk premium, and the volatility constant, respectively. An agent first invests a fixed proportion \(0 < \eta < 1\) of his wealth in \(S\) and then consumes a proportion \(u \in U := \mathbb{N}, 1 - \eta\) of his wealth following a relaxed control \(\pi \in A\), where \(\zeta > 0\) denotes a subsistence level of consumption. The resulting wealth process is given by

\[
dZ_t^\pi = \left( \int_U (r + \eta \mu - u)Z_t^n\pi_t(u)du \right) dt + a\eta Z_t^\pi dW_t, \quad Z_0 = z \in (0, \infty).
\]
The agent measures utility obtained from consumption by the exponential utility function \( u(y) := -e^{-\alpha y}, y \geq 0 \), for some fixed \( \alpha > 0 \). The agent’s goal is then to choose \( \pi^* \in A \) that achieves the optimal value

\[
\mathcal{W}^*(z) := \sup_{\pi \in A} \mathbb{E}_z \left[ \int_0^\infty e^{-pt} \left( \int_U u(uZ^\pi_t)\pi_t(u)du - \lambda \int_U \ln(\pi_t(u))\pi_t(u)du \right) dt \right]. \tag{5.1}
\]

By Itô’s formula, the process \( X^\pi_t := \ln(Z^\pi_t) \), with \( X_0 = x := \ln z \in \mathbb{R} \), satisfies (2.1) with \( b(y,u) = b(u) := \tau + \mu y - \frac{1}{2}a^2y^2 - u \) and \( \sigma(y) \equiv a \eta \). Then, \( \mathcal{W}^*(z) \) can be expressed as \( V^*(x) \) in (2.3), with \( r(y,u) := -\exp(-\alpha u e^y) \).

We now check the conditions of Theorem 4.1. As \( \sigma > 0 \) and \( U = [\zeta, 1 - \eta], (3.16) \) and Assumption 4.2 hold trivially. A direct calculation shows \( r_y(y,u) = \alpha e^{-\alpha u e^y}y \) and \( r_{yy}(y,u) = \alpha e^{-\alpha u e^y}(-\alpha u^2 e^y + u e^y) \), which implies \( |r(y,u)| \leq 1 \), \( |r_y(y,u)| \leq e^{-1} \) and \( |r_{yy}(y,u)| \leq 4e^{-2} + e^{-1} \).

Hence, \( \sup_{u \in U} \|r(\cdot, u)\|_{C^2(\mathbb{R})} < \infty \). Moreover, \( \sup_{u \in U} \|b(\cdot, u)\|_{C^2(\mathbb{R})} = \sup_{u \in [0,1-\eta]} \|b(\cdot, u)\|_{C^0(\mathbb{R})} = \tau + \mu + \alpha^2 \eta^2/2 + 1 - \eta < \infty \) and \( \sup_{u \in U} \|\sigma(\cdot)\|_{C^2(\mathbb{R})} = a \eta < \infty \). We then conclude \( A_2 = A_1 + d/2 + 2 < \infty \) for \( d = 1 \). Finally, to verify Assumption 4.1, by the definition of \( b \), it suffices to show that \( r \) is Lipschitz in \( u \) uniformly in \( y \). With \( r_u(y,u) = \alpha e^{-\alpha u e^y}y = r_y(y,u)/u \), the estimate above for \( r_{yy}(y,u) \) gives \( |r_u(y,u)| \leq e^{-1}/u \leq e^{-1}/\zeta \), as desired. As all conditions of Theorem 4.1 are fulfilled, the PIA can be used to find the optimal value function \( V^*(x) \) and a corresponding optimal relaxed consumption strategy \( \pi^* \in A \).

Below we numerically implement the PIA (2.7), with parameters \( \zeta = \eta = 0.1, \alpha = 10, \tau + \mu \eta - \frac{1}{2}a^2 \eta^2 = 1 = \alpha = \lambda \) and \( \rho = 0.1 \). Then \( U = [0.1, 0.9] \) and the HJB equation (2.4) reads

\[
-V^*(x) + \frac{1}{2}D_x^2V^*(x) + \sup_{\varpi \in P(U)} \int_{0.1}^{0.9} \left( (1 - u)D_xV^* - \exp(-ue^x) - \ln(\varpi(u)) \right) \varpi(u)du = 0, \tag{5.2}
\]

where the first equality follows from choosing \( \varpi(u) = \Gamma(x, D_xV^*(x), u) = \frac{\exp((1-u)D_xV^*-\exp(-ue^x))}{\int_{0.1}^{0.9} \exp((1-u)D_xV^*-\exp(-ue^x))du} \).

The PIA (2.7) now becomes the following concrete algorithm:

1. Take an initial input \( v^0(x) \) and set \( \pi^1(x,u) := \frac{\exp((1-u)D_xv^0-\exp(-ue^x))}{\int_{0.1}^{0.9} \exp((1-u)D_xv^0-\exp(-ue^x))du} \).

2. For \( n = 1, 2, \ldots \), perform

- **Policy evaluation**: Compute \( v^n = V^{\pi^n} \) by solving the PDE

\[
-0.1v^n + \frac{1}{2}D_x^2v^n + \int_{0.1}^{0.9} \left( (1 - u)D_xv^n(x) - \exp(-ue^x) - \ln(\pi^n(x,u)) \right) \pi^n(x,u)du = 0. \tag{5.3}
\]

- **Policy improvement**: Compute \( \pi^{n+1}(x,u) = \frac{\exp((1-u)D_xv^n-\exp(-ue^x))}{\int_{0.1}^{0.9} \exp((1-u)D_xv^n-\exp(-ue^x))du} \).

In carrying out the above algorithm, we take two different initial guesses \( v^0(x) = \sin(x) \) and \( v^0(x) = \frac{1}{1+x^2} \), solve PDE (5.3) for \( v^n \) by finite difference methods, and compute \( \pi^{n+1} \) using the newly-computed \( v^n \) function. To show that as \( n \) is large enough, we can indeed approximate \( V^* \) and the corresponding optimal relaxed control \( \pi^* \), we solve the HJB equation (5.2) directly for \( V^* \) by finite difference methods, and compute \( \pi^*(x,u) = \frac{\exp((1-u)D_xV^*-\exp(-ue^x))}{\int_{0.1}^{0.9} \exp((1-u)D_xV^*-\exp(-ue^x))du} \). Then, we compute the errors \( \|V^* - v^n\|_{L^\infty(D)} \) and \( \|\pi^* - \pi^n\|_{L^\infty(D \times U)} \), alongside their logarithmic values, for \( D = [-50, 50] \). The numerical results are summarized in Figures 1, 2, 3, 4, and 5, demonstrating the desired convergence of the PIA.\(^4\)

\(^4\)The errors can be computed on an even bigger domain, but it does not change the numerical results presented below.
A Derivation of Lemma 2.2

Recall (2.9)–(2.11). We will prove (i) and (ii) in Lemma 2.2 separately.

Proof for Lemma 2.2 (i). It is sufficient to prove only (2.12), as the rest of the statement in part (i) directly follows. Also, we will only prove (2.12) for

$$f(x, y, u) = r(x, u) - H(x, y, u),$$

as the case of $f = b(x, u)$ is similar and simpler. Let us first define

$$g = g(x, y, u) := \frac{(b(x, u) \cdot y + r(x, u))}{\lambda}, \quad \forall x, y \in \mathbb{R}^d \text{ and } u \in U. \quad (A.2)$$

For any $k, m \in \mathbb{N}_0$, let $y \mapsto p_k^m(y)$ denote a generic polynomial of degree $k$ or smaller, with coefficients being the products of the derivatives of $x \mapsto b(x, u), r(x, u)$ of order $m$ or smaller. It can be checked
Figure 3: Difference between $V^*$ and $v^n$ for $n = 1, 2, \cdots, 10$ with the initial guess $v^0 = \frac{1}{1+x^2}$. The $y$-axis represents $\|V^* - v^n\|_{L^\infty([-50,50])}$ (left panel) and $\ln(\|V^* - v^n\|_{L^\infty([-50,50])})$ (right panel). Notice that the algorithm is set to stop once it reaches the tolerance of the finite difference solver.

Figure 4: Difference between $\pi^*$ and $\pi^n$ for $n = 1, 2, \cdots, 10$ with the initial guess $v^0 = \frac{1}{1+x^2}$. The $y$-axis represents $\|\pi^* - \pi^n\|_{L^\infty([-50,50] \times [0,1.0.9])}$ (left panel) and $\ln(\|\pi^* - \pi^n\|_{L^\infty([-50,50] \times [0,1.0.9])})$ (right panel).

Figure 5: The left panel consists of $V^*$ and $v^{10}$ on $[-10, 10]$ with both initial guesses $v^0 = \sin(x)$ and $v^0 = \frac{1}{1+x^2}$. The right panel displays the graph of $\pi^*(x, u)$ on $[-10, 10] \times [0.1, 0.9]$. 
directly that for any \( i, j = 1, \ldots, d \) and \( k, \bar{k}, m, \bar{m} \in \mathbb{N}_0 \),
\[
\begin{align*}
\{ p^{k,m}(y) \cdot p^{k,\bar{m}}(y) = p^{k+k,m+\bar{m}}(y), \\
D_{y_i}p = p^{0,0}(y), \quad D_{x_j}p = p^{1,1}(y) \\
D_{y_i}p^{0,m}(y) = 0, \quad D_{y_i}p^{k+1,m}(y) = p^{k,m}(y), \quad D_{x_j}p^{k,m}(y) = p^{k,m+1}(y). \quad \text{(A.3)}
\end{align*}
\]

**Step 1.** We will provide a generic formula for the derivatives of \( (x,y) \mapsto \Gamma(x,y,u) \) in (2.6). Given \( k, m \in \mathbb{N}_0 \), for any multi-indices \( a, c \) with \( |a|_{l_1} = k \) and \( |c|_{l_1} = m \), we claim that
\[
D_{x}^{a}(D_{y}^{c}\Gamma) = \frac{\sum p^{n,m}(y) \exp(g) \left( \prod_{n+1}^{n} \int_{U} p^{m,m}(y) \exp(g)du \right)}{(\int_{U} \exp(g)du)^{2k+m}}. \quad \text{(A.4)}
\]

Here, we do not write out the number of terms in the summation “\( \sum \)” explicitly (but note that it is determined by only \( k, m \) and \( d \)), while the superscript of \( “\prod” \) denotes the number of terms in the product. We will use \( \sum \) and \( \prod \) in the same way throughout the proof.

To begin with, let us prove by induction on \( k \in \mathbb{N}_0 \) that for any multi-index \( a \) with \( |a|_{l_1} = k \geq 0 \),
\[
D_{y}^{a}\Gamma = \frac{\sum p^{0,0}(y) \exp(g) \left( \prod_{n+1}^{n} \int_{U} p^{0,0}(y) \exp(g)du \right)}{(\int_{U} \exp(g)du)^{2k}}. \quad \text{(A.5)}
\]

For \( k = 0 \), (A.5) holds as \( \Gamma = \exp(g)/\int_{U} \exp(g)du \) by definition. Suppose that (A.5) holds for \( k = n \in \mathbb{N}_0 \) and we aim to establish (A.5) for \( k = n + 1 \). For any multi-index \( a = (a_1, \ldots, a_d) \) with \( |a|_{l_1} = n + 1 \), assume \( a_1 \geq 1 \) without loss of generality (w.l.o.g.) and set \( a' := (a_1 - 1, \ldots, a_d) \). As \( |a'|_{l_1} = n \), the induction hypothesis implies \( D_{y}^{a'} \Gamma = \text{(I)}/(\text{II}) \), with \( \text{(I)} := \sum p^{0,0}(y) \exp(g) \left( \prod_{n+1}^{n} \int_{U} p^{0,0}(y) \exp(g)du \right) \) and \( \text{(II)} := (\int_{U} \exp(g)du)^{2n} \). Hence,
\[
D_{y}^{a}\Gamma = D_{y}^{a'} \Gamma = \left( \frac{\partial}{\partial y_{1}}(\text{I}) \right) \cdot (\text{II}) - \left( \frac{\partial}{\partial y_{1}}(\text{II}) \right) \cdot (\text{I}) \right) / (\text{II})^2. \quad \text{(A.6)}
\]

By a direct calculation and using (A.3), we have
\[
\left\{ \frac{\partial}{\partial y_{1}}(\text{I}) \right\} \cdot (\text{II}) = \left\{ \sum_{n+1}^{n} \int_{U} p^{0,0}(y) \exp(g)du \right\} \cdot (\text{II}) + \left\{ \sum_{n+1}^{n} \int_{U} p^{0,0}(y) \exp(g)du \right\} \cdot (\text{II}) = \sum_{n+1}^{n} \int_{U} p^{0,0}(y) \exp(g)du \cdot (\text{II}),
\]

where the last line stems from writing \( (\text{II}) = (\int_{U} \exp(g)du)^{2n} \) as \( (\int_{U} p^{0,0}(y) \exp(g)du)^{2n} \). Similarly,
\[
\left\{ \frac{\partial}{\partial y_{1}}(\text{II}) \right\} \cdot (\text{I}) = \left\{ 2^n \left( \int_{U} \exp(g)du \right)^{2n-1} \left( \int_{U} \exp(g)DU \right) \right\} \cdot (\text{I}) = \sum_{n+1}^{n} \int_{U} p^{0,0}(y) \exp(g)du \cdot (\text{II}).
\]

Plugging (A.7) and (A.8) into (A.6) yields (A.5) with \( k = n + 1 \). Thus, (A.5) holds for all \( k \in \mathbb{N}_0 \).
We now prove (A.4). For $m = 0$ and $k \in \mathbb{N}_0$, (A.4) holds as it reduces to (A.5). Suppose that (A.4) holds for $m = n \in \mathbb{N}_0$ and $k \in \mathbb{N}_0$, and we aim to prove (A.4) for $m = n + 1$ and $k \in \mathbb{N}_0$. For any multi-indices $a = (a_1, \ldots, a_d)$ and $c = (c_1, \ldots, c_d)$ with $|a|_{l_1} = k$ and $|c|_{l_1} = n + 1$, assume $c_1 \geq 1$ w.l.o.g. and set $c' = (c_1 - 1, \ldots, c_d)$. As $|c'|_{l_1} = n$, the induction hypothesis gives $D_x^c(D_y^m \Gamma) = (I)/(II)$ with

$$D_x^c(D_y^m \Gamma) = D_{x_1}(D_x^c D_y^m \Gamma) = \left( \frac{\partial}{\partial x_1} (I) \cdot (II) - \frac{\partial}{\partial x_1} (II) \cdot (I) \right) / (II)^2. \quad (A.9)$$

A direct calculation shows that $\frac{\partial}{\partial x_1} (I)$ is equal to

$$\sum \left[ (D_x p^{n,n}(y) + p^{n,n}(y)(D_x g')) \exp(g) \left( \prod_{i=0}^{2k+n-1} \int_U p^{n,n}(y) \exp(g) du \right) \right] \cdot \left( \prod_{i=0}^{2k+n-1} \int_U p^{n,n}(y) \exp(g) du \right).$$

By (A.3), $D_x p^{n,n}(y) = p^{n,n+1}(y)$ and $p^{n,n}(y)(D_x g') = p^{n,n+1}(y)$. Plugging this into the above gives

$$\sum \left( p^{n,n+1}(y) \exp(g) \right) \cdot \left( \prod_{i=0}^{2k+n-1} \int_U p^{n,n+1}(y) \exp(g) du \right). \quad (A.10)$$

By using (A.3) in a similar fashion, we get

$$\sum \left( p^{n,n+1}(y) \exp(g) \right) \cdot \left( \prod_{i=0}^{2k+n-1} \int_U p^{n,n+1}(y) \exp(g) du \right). \quad (A.11)$$

Plugging (A.10) and (A.11) into (A.9) yields (A.4) with $m = n + 1$ and $k \in \mathbb{N}_0$. We then conclude that (A.4) holds for any $m, k \in \mathbb{N}_0$.

**Step 2.** We will provide a generic formula for the derivatives of $(x, y) \mapsto f(x, y, u)$ in (A.1). Given $k, m \in \mathbb{N}_0$ with $k + m \geq 1$, for any multi-indices $a, c$ with $|a|_{l_1} = k$ and $|c|_{l_1} = m$, we claim that

$$D_y^m(D_x^c f) = p^{1,m}(y) + \sum (\prod_{i=0}^{2k+m-1} \int_U \exp(g) p^{m,m}(y) du) \cdot (\prod_{i=0}^{2k+m-1} \int_U \exp(g) du)^{2k+m-1}. \quad (A.12)$$

First, let us prove (A.12) for $k = 0$, i.e., for any multi-index $c = (c_1, \ldots, c_d)$ with $|c|_{l_1} = m \in \mathbb{N},$

$$D_x^c f = p^{1,m}(y) + \sum (\prod_{i=0}^{m-1} \int_U \exp(g) p^{m,m}(y) du) \cdot (\prod_{i=0}^{m-1} \int_U \exp(g) du)^{2m-1}. \quad (A.13)$$

If $m = 1$, as $\mathcal{H} = \lambda \ln \Gamma = \lambda(g - \ln(\int_U \exp(g) du))$ by (2.11) and (A.2), we have

$$D_x^1 f = D_x r + \lambda D_x g - \lambda \frac{\int_U (D_x g) \exp(g) du}{\int_U \exp(g) du} = p^{1,1}(y) + \frac{\int_U p^{1,1}(y) \exp(g) du}{\int_U \exp(g) du} \quad \forall 1 \leq j \leq d. \quad (A.14)$$

That is, (A.13) holds for $m = 1$. Suppose that (A.13) holds for $m = n \in \mathbb{N}$ and we aim to prove (A.13) for $m = n + 1$. For any multi-index $c = (c_1, \ldots, c_d)$ with $|c|_{l_1} = n + 1$, assume $c_1 \geq 1$ w.l.o.g. and set $c' = (c_1 - 1, \ldots, c_d)$. As $|c'|_{l_1} = n$, $D_x^c f$ fulfills (A.13) (by the induction hypothesis), so that

$$D_x^1(D_x^c f) = D_x(D_x^c p^{1,n}(y)) + \left( \frac{\partial}{\partial x_1} (I) \cdot (II) - \frac{\partial}{\partial x_1} (II) \cdot (I) \right) / (II)^2. \quad (A.15)$$
with (I) := \(\sum (\prod_{i=1}^{m-1} \int_U \exp(g) p^m, n(y) \, du)\) and (II) := \((\int_U \exp(g) \, du)^2\). Using (A.3) in the same way as in the derivation of (A.10) and (A.11), we see that both \((\partial_{\nu_1}^m (I)) \cdot (II)\) and \((\partial_{\nu_2}^m (II)) \cdot (I)\) can be expressed as \(\sum (\prod_{i=1}^{m} \int_U \exp(g) p^{n+1, m+1}(y) \, du)\). This, together with (A.15), shows that (A.13) holds for \(m = n + 1\). We then conclude that (A.13) holds for any \(m \in \mathbb{N}\).

We proceed to prove (A.12) for \(k \in \mathbb{N}\). If \(k = 1\) and \(m = 0\), by a calculation similar to (A.14),

\[
D_y f = \lambda D_y g - \frac{\int_U (D_y g) \exp(g)(y) \, du}{\int_U \exp(g)(y) \, du} = p^{0,0}(y) + \frac{\int_U p^{0,0}(y) \exp(g)(y) \, du}{\int_U \exp(g)(y) \, du} \quad \forall 1 \leq i \leq d,
\]

i.e., (A.12) holds in this case. Based on this and (A.13), we suppose that (A.12) holds for \(n \in \mathbb{N}_0\) and \(m \in \mathbb{N}_0\), with \(n + m \geq 1\), and aim to prove (A.12) for \(k = n + 1\) and \(m \in \mathbb{N}_0\). For any multi-indices \(a = (a_1, \ldots, a_d)\) and \(c = (c_1, \ldots, c_d)\) with \(|a|_{l_1} = n + 1\) and \(|c|_{l_1} = m\), assume \(a_1 \geq 1\) w.l.o.g. and set \(a' := (a_1 - 1, \ldots, a_d)\). As \(|a'|_{l_1} = n, D_y^c D_x^c f\) fulfills (A.12) (by the induction hypothesis), so that

\[
D_y (D_y^c D_x^c f) = D_y (p^{1,m}(y)) + \frac{(\partial_{\nu_1}^m (I)) \cdot (II) - (\partial_{\nu_2}^m (II)) \cdot (I)}{(II)^2},
\]

with (I) := \(\sum (\prod_{i=1}^{m+n-1} \int_U \exp(g) p^{m, m}(y) \, du)\) and (II) := \((\int_U \exp(g) \, du)^{2m+n-1}\). Using (A.3) in the same way as in (A.10)-(A.11), we see that both \((\partial_{\nu_1}^m (I)) \cdot (II)\) and \((\partial_{\nu_2}^m (II)) \cdot (I)\) can be expressed as \(\sum (\prod_{i=1}^{m+n} \int_U \exp(g) p^{n+1, m+1}(y) \, du)\). This, together with (A.16), shows that (A.12) holds for \(k = n + 1\) and \(m \in \mathbb{N}_0\). We then conclude that (A.12) holds for any \(k, m \in \mathbb{N}_0\) with \(k + m \geq 1\).

**Step 3.** We will provide an estimate for the derivatives of \(f(x, y) := \int_U f(x, y, u) \Gamma(x, y, u) \, du\), with \(f\) given by (A.1). For two multi-indices \(b^{(1)} = (b_1^{(1)}, \ldots, b_d^{(1)})\), \(j = 1, 2\), we will write \(b^{(1)} \leq b^{(2)}\) if \(b_i^{(1)} \leq b_i^{(2)}\) for all \(i = 1, 2, \ldots, d\). In the sequel, \(C_m > 0\) (resp. \(C_{k,m} > 0\)) is a generic constant that depends on only \(m, \Lambda_m, \lambda, d\) and \(\text{Leb}(U)\) (resp. \(k, m, \Lambda_m, \lambda, d, \text{and } \text{Leb}(U)\)) and may vary from line to line. Given \(k, m \in \mathbb{N}_0\), for any multi-indices \(a, c\) with \(|a|_{l_1} = k\) and \(|c|_{l_1} = m\), we claim that

\[
\left| D_y^a D_x^c f \right| = \left| D_y^a D_x^c \int_U f \, du \right| \leq C_{k+m}(1 + |y|)^{(m+1)2^{k+m}},
\]

For the case \(k = m = 0\), a calculation similar to (3.5) shows

\[
|f| \leq 3A_0(1 + |y|) + \lambda |\ln(\text{Leb}(U))|,
\]

which implies \(\int_U f \Gamma \, du \leq C_{0,0}(1 + |y|)\). That is, (A.17) holds for \(k = m = 0\).

In (A.4) and (A.12), every \(p^{m, m}(y)\) is a polynomial in \(y\) of degree \(m\) or smaller, with coefficients being products of derivatives of \(x \mapsto b(x, u), r(x, u)\) of order \(m\) or smaller; moreover, the number of terms contained in each product depends on how many times differentiation has been carried out, i.e., \(m\). Hence, in (A.4) and (A.12), we have \(|p^{m, m}(y)| \leq C_m(1 + |y|^m)\). Now, for any \(k, m \in \mathbb{N}_0\) and multi-indices \(a, c\) with \(|a|_{l_1} = k\) and \(|c|_{l_1} = m\), applying the above estimate to (A.4) yields

\[
|D_y^c (D_x^a \Gamma)| \leq \sum (C_m(1 + |y|^m))^{2^{k+m}} \exp(g)(\int_U \exp(g) \, du)^{2^{k+m} - 1}
\]

\[
\left(\int_U \exp(g) \, du\right)^{2^{k+m}} \leq C_{k,m}(1 + |y|)^{m2^{k+m}} \Gamma.
\]

Similarly, when \(k + m \geq 1\), applying the same estimate to (A.12) gives

\[
|D_y^a (D_x^c f)| \leq |p^{1,m}(y)| + C_{k,m}(1 + |y|)^{m2^{k+m} - 1} \sum (\int_U \exp(g) \, du)^{2^{k+m} - 1} \left(\int_U \exp(g) \, du\right)^{2^{k+m} - 1} \leq C_{k,m}(1 + |y|)^{(m+1)2^{k+m} - 1}.
\]

28
Combining this and (A.18), which covers the case $k + m = 0$, leads to
\[ |D^a_y(D^c_x f)| \leq C_{k,m}(1 + |y|)^{(m+1)2^{k+m}}. \]  
(A.20)

Finally, given $k, m \in \mathbb{N}_0$ with $k + m \geq 1$, for any multi-indices $a, c$ with $|a|_{l_1} = k$ and $|c|_{l_1} = m$,
\[
|D^a_y D^c_x \int_U f \Gamma du| = \left| \int_U D^a_y D^c_x (f \Gamma) du \right| \leq \sum_{0 \leq p \leq q, 0 \leq q \leq c} \int_U |(D^p_y D^q_x f)(D^{a-p}D^{c-q}_x \Gamma)| du
\]
\[
\leq \sum_{0 \leq p \leq q, 0 \leq q \leq c} C_{|p|_{l_1}, |q|_{l_1}, |a-p|_{l_1}, |c-q|_{l_1}} \cdot (1 + |y|)^{(|q|_{l_1}+1)^2(|p|_{l_1}+1) + (m-|q|_{l_1})2^{k-|p|_{l_1}+m-|c|_{l_1})}
\]
\[
\leq C_{k,m}(1 + |y|)^{(m+1)2^{k+m}} \leq C_{k+m,k+m}(1 + |y|)^{(m+1)2^{k+m}} = C_{k+m}(1 + |y|)^{(m+1)2^{k+m}},
\]
where the second inequality follows from (A.19), (A.20), and $\int_U \Gamma du = 1$. That is, (A.17) holds for $k, m \in \mathbb{N}_0$ with $k + m \geq 1$.

**Step 4.** We now prove (2.12). Fix $k \in \mathbb{N}_0$ and $p = (p_i)_{i=1}^d : \mathbb{R}^d \to \mathbb{R}^d$ in $C^{k,\alpha}(\mathbb{R}^d)$. For $k = 0$, we recover (2.12) from (A.17) by taking $k = m = 0$ and $y = p(x)$ therein. For $k \in \mathbb{N}$, we deduce from (A.17) that for any multi-index $a = (a_1, \ldots, a_d)$ with $|a|_{l_1} = k$,
\[ |D^a_y f(z)| \leq C_k(1 + |y|)^{(k+1)2^k}, \quad \text{with } z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d. \]  
(A.22)

For any $s \in \mathbb{N}$, we say that the multi-index $a$ is decomposed into distinct $q_1, q_2, \ldots, q_s \in \mathbb{N}_0^d$ with multiplicities $m_1, m_2, \ldots, m_s \in \mathbb{N}^{2d}$ if $a = \sum_{j=1}^s m_j q_j$. Note that $q_j$ (resp. $m_j$) is a multi-index of dimension $d$ (resp. $2d$) for all $1 \leq j \leq s$. We denote by $\mathcal{D}$ the set of all $(s, q = (q_j)_{j=1}^s, m = (m_j)_{j=1}^s)$ that form a decomposition of $a$. Now, for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ define $\tilde{3} : \mathbb{R}^d \to \mathbb{R}^{2d}$ by
\[ \tilde{3}(x) = (\tilde{3}_1(x), \ldots, \tilde{3}_d(x), \tilde{3}_{d+1}(x), \ldots, \tilde{3}_{2d}(x)) := (x_1, \ldots, x_d, p_1(x), \ldots, p_d(x)). \]

Then, by the multivariate chain rule (i.e., the Faà di Bruno formula; see [19, Section 6]),
\[ D^a_x f(x, p(x)) = a! \sum_{(s, q, m) \in \mathcal{D}} (D^a_x f)(\tilde{3}(x)) \prod_{j=1}^s \prod_{i=1}^{2d} \frac{1}{m_{j_i}!} \left[ \frac{1}{q_{j_i}!} \cdots \frac{1}{q_{d_i}!} D^q_{x_i} \tilde{3}_i(x) \right]^{m_{j_i}}, \]  
(A.23)

with $a! := \prod_{j=1}^d a_j$! and $\tilde{m} := \sum_{j=1}^s m_j$. By the definition of $\tilde{3}_i$, for any multi-index $c = (c_1, \ldots, c_d)$,
\[ D^c_{\tilde{3}_i}(x) = 1 \quad \text{or} \quad 0 \quad \text{for } 1 \leq i \leq d; \quad D^c_{\tilde{3}_i}(x) = D^c_{x_i} p_i(x) \quad \text{for } d + 1 \leq i \leq 2d. \]  
(A.24)

It follows that
\[
\left| \prod_{j=1}^s \prod_{i=1}^{2d} \left[ D^q_{x_i} \tilde{3}_i \right]^{m_{j_i}} \right| \leq \max \left\{ 1, \prod_{j=1}^s \prod_{i=d+1}^{2d} \left[ D^q_{x_i} \tilde{3}_i \right]^{m_{j_i}} \right\} \leq 1 + \prod_{j=1}^s \prod_{i=d+1}^{2d} \left( \sum_{|c|_{l_1} = 0, \ldots, k} |D^c_{x_i} p_i| \right)^{m_{j_i}}
\]
\[
= 1 + \left( \sum_{|c|_{l_1} = 0, \ldots, k} |D^c_{x_i} p_i| \right)^{\sum_{j=1}^s m_{j_i}} \leq 1 + \left( \sum_{|c|_{l_1} = 0, \ldots, k} |D^c_{x_i} p_i| \right)^{k},
\]
where the second and last inequalities stem from $(\max_{1 \leq j \leq s} |q_j|_{l_1}) \lor (\sum_{j=1}^s m_{j_i}) \leq |a|_{l_1} = k$, a consequence of $(s, q, m) \in \mathcal{D}$. Also, collecting all the constant coefficients in (A.23) results in $a! \sum_{(s, q, m) \in \mathcal{D}} \prod_{j=1}^s \prod_{i=1}^{2d} \left( \frac{1}{q_{j_i}!} \cdots \frac{1}{q_{d_i}!} \right)^{m_{j_i}}$, which is a finite constant depending on only $k$ and $d$. Hence, by applying the estimates (A.22) and (A.25) to (A.23), we get the desired result (2.12). \[\square\]
Now we move to the proof for part (ii) of Lemma 2.2.

Proof of Lemma 2.2 (ii). We will again prove the result only for \((\hat{r} - H)(\cdot, p(\cdot))\), as the case of \(b(\cdot, p(\cdot))\) is similar and simpler. Fix \(k \in \mathbb{N}_0\) and \(0 < \alpha \leq 1\). Consider \(f\) as in (A.1). To prove the desired result \(\hat{f}(\cdot, p(\cdot)) \in C^{k,\alpha}(\mathbb{R}^d)\), it suffices to show \(D^a\hat{f}(\cdot, p(\cdot)) \in C^{0,\alpha}(\mathbb{R}^d)\) for all multi-indices \(a = (a_1, \ldots, a_d)\) with \(|a|_1 \leq k\). First, it can be checked directly that the following implication holds:

\[
h_1, h_2 \in C^{0,\alpha}_{\text{unif}}(\mathbb{R}^d) \implies h_1 + h_2, \ h_1h_2 \in C^{0,\alpha}_{\text{unif}}(\mathbb{R}^d). \tag{A.26}
\]

By (A.23), the above implication indicates that “\(D^a\hat{f}(\cdot, p(\cdot)) \in C^{0,\alpha}_{\text{unif}}(\mathbb{R}^d)\)” holds if \((D^m\hat{f})(\cdot(\cdot))\) and \(D^a\hat{f}\hat{f}\) in (A.23) belong to \(C^{0,\alpha}_{\text{unif}}(\mathbb{R}^d)\). By (A.24) and \(p \in C^{k,\alpha}_{\text{unif}}(\mathbb{R}^d)\), we know \(D^m\hat{f}\hat{f}\in C^{0,\alpha}_{\text{unif}}(\mathbb{R}^d)\). Hence, it remains to show \((D^m\hat{f})(\cdot(\cdot))\) \(\in C^{0,\alpha}_{\text{unif}}(\mathbb{R}^d)\), or equivalently, \((D^a\hat{f})(\cdot, p(\cdot)) \in C^{0,\alpha}(\mathbb{R}^d)\) for any multi-index \(a = (a_1, \ldots, a_{2d})\) with \(|a|_1 \leq k\). It follows from (2.10) that for all \(a \in \mathbb{N}_{0}^{2d}\) with \(|a|_1 \leq k\),

\[
\| (D^a\hat{f})(\cdot, p(\cdot)) \|_{C^{0,\alpha}(B_1(x))} \leq \int_U \| (D^a\hat{f})(\cdot, p(\cdot), u) \|_{C^{0,\alpha}(B_1(x))} \, du.
\]

As Leb\((U) < \infty\), for “\((D^n\hat{f})(\cdot, p(\cdot)) \in C^{0,\alpha}_{\text{unif}}(\mathbb{R}^d)\)” for all \(a \in \mathbb{N}_{0}^{2d}\) with \(|a|_1 \leq k\) to hold, it suffices to show \(\sup_{x \in \mathbb{R}^d, u \in U} \| D^n\hat{f}(\cdot, p(\cdot), u) \|_{C^{0,\alpha}(B_1(x))} < \infty\) for all \(a \in \mathbb{N}_{0}^{2d}\) with \(|a|_1 \leq k\), which is equivalent to:

\[
\sup_{x \in \mathbb{R}^d, u \in U} \| (D^a\hat{f}) (D^b\hat{f}) (D^c\hat{f}) \|_{C^{0,\alpha}(B_1(x))} \|_{C^{0,\alpha}(B_1(x))} < \infty. \tag{A.27}
\]

To establish this, we first observe similarly to (A.26) that, for any \(h_1, h_2 : \mathbb{R}^d \times U \rightarrow \mathbb{R}\),

\[
h_1, h_2 \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d) \implies h_1 + h_2, \ h_1h_2 \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d), \tag{A.28}
\]

where we denote by \(C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\) the set of \(h : \mathbb{R}^d \times U \rightarrow \mathbb{R}\) with \(\sup_{x \in \mathbb{R}^d, u \in U} \| h(\cdot, u) \|_{C^{0,\alpha}(B_1(x))} < \infty\). Thanks to (A.28), we deduce from (A.4) and (A.12) that to establish (A.27), it suffices to show \(p_{m,m}(\cdot, \exp(g), 1/\exp(g)) \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\) in (A.4) and (A.12), with \(y = p(x)\) and \(0 \leq m \leq k\) therein.

With \(y = p(x)\) and \(0 \leq m \leq k\), note that we can express every \(p_{m,m}(y)\) in (A.4) and (A.12) more specifically as a function on \(\mathbb{R}^d \times U\), i.e.,

\[
p_{m,m}(x, u) = p_{m,m}(x, p(x), u) = \sum_i c_i(x, u)(p_1(x))^{a_{i1}} \cdots (p_d(x))^{a_{id}}, \tag{A.29}
\]

where each \(c_i(x, u)\) is a product of the derivatives of \(x \mapsto b(x, u), r(x, u)\) of order \(m\) or smaller, and each \(a_i = (a_{i1}, \ldots, a_{id})\) is a multi-index with \(|a_i|_1 \leq m\). Observe from (2.9) that \(\sup_{x \in U} \| b(\cdot, u) \|_{C^{k,\alpha}(B_1(x)} + \| r(\cdot, u) \|_{C^{k,\alpha}(B_1(x))} \leq 2^{1-N} \Lambda^{k+1} < \infty\) for all \(x \in \mathbb{R}^d\), which implies \(b, r \in C^{k,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\). By (A.28), we get \(c_i \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\). As \(p \in C^{k,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\), each \(p_i\), when viewed as a function on \(\mathbb{R}^d \times U\) (constant on \(U\)), readily lies in \(C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\). Hence, in view of (A.29), we deduce from (A.28) again that \(p_{m,m} \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\).

Set \(\bar{g}(x, u) := g(x, p(x)), u\). It remains to show \(\exp(\bar{g}), 1/\exp(\bar{g}) \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\). As \(b, r, p \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\), by the definition of \(g\) in (A.2) and using (A.28), we have \(\bar{g} \in C^{0,\alpha}_{\text{unif}, U}(\mathbb{R}^d)\). This readily gives the boundedness of \(\exp(\bar{g})\) and \(1/\exp(\bar{g})\). Set \(M := \| \exp(\bar{g}) \|_{L^\infty(\mathbb{R}^d \times U)} < \infty\). By writing \(\bar{g}_i = \bar{g}(x, u), i = 1, 2,\) for any distinct \(x_1, x_2 \in \mathbb{R}^d\) and \(u \in U\), it holds for any \(x \in \mathbb{R}^d\) that

\[
\sup_{x_1, x_2 \in B_1(x)} \frac{|\exp(\bar{g}_1) - \exp(\bar{g}_2)|}{|x_1 - x_2|} = \sup_{x_1, x_2 \in B_1(x)} \frac{|\exp(\bar{g}_1) - \exp(\bar{g}_2)|}{|\bar{g}_1 - \bar{g}_2|} \leq \left( \sup_{x \in B_1(x)} \exp(\bar{g}(x, u)) \right) \| \bar{g}(\cdot, u) \|_{C^{0,\alpha}(B_1(x))} \leq M \| \bar{g}(\cdot, u) \|_{C^{0,\alpha}(B_1(x))}. \tag{A.29}
\]
As \( \bar{g} \in C_{\text{unif},U}(\mathbb{R}^d) \), taking supremum over \( x \in \mathbb{R}^d \) and \( u \in U \) yields a finite upper bound, which implies \( \exp(\bar{g}) \in C_{\text{unif},U}(\mathbb{R}^d) \). Similarly, it holds for any \( x \in \mathbb{R}^d \) and \( u \in U \) that

\[
\sup_{x_1,x_2 \in B_1(x)} \frac{1/\exp(\bar{g}_1) - 1/\exp(\bar{g}_2)}{|x_1 - x_2|^{\alpha}} = \sup_{x_1,x_2 \in B_1(x)} \frac{1}{\exp(\bar{g}_1) \exp(\bar{g}_2)} \frac{|\exp(\bar{g}_1) - \exp(\bar{g}_2)|}{|\bar{g}_1 - \bar{g}_2|} \frac{|\bar{g}_1 - \bar{g}_2|}{|x_1 - x_2|^{\alpha}} \leq N^2 M \|\bar{g}(\cdot, u)\|_{C^{0,\alpha}(B_1(x))},
\]

with \( N := \|1/\exp(\bar{g})\|_{L^\infty(\mathbb{R}^d \times U)} < \infty \). Taking supremum over \( x \in \mathbb{R}^d \) and \( u \in U \) yields a finite upper bound, thanks again to \( \bar{g} \in C_{\text{unif},U}(\mathbb{R}^d) \). This implies \( 1/\exp(\bar{g}) \in C_{\text{unif},U}(\mathbb{R}^d) \). \( \square \)

References

[1] Ya-Zhe Chen and Lan-Cheng Wu. Second order elliptic equations and elliptic systems, volume 174 of Translations of Mathematical Monographs. American Mathematical Society, Providence, RI, 1998. Translated from the 1991 Chinese original by Bei Hu.

[2] Min Dai, Yuchao Dong, and Yanwei Jia. Learning equilibrium mean-variance strategy. Mathematical Finance, 33(4):1166–1212, 2023.

[3] Lawrence C. Evans. Partial differential equations, volume 19 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 1998.

[4] Dena Firoozi and Sebastian Jaimungal. Exploratory LQG mean field games with entropy regularization. Automatica J. IFAC, 139:Paper No. 110177, 12, 2022.

[5] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA.

[6] David Gilbarg and Neil S. Trudinger. Elliptic partial differential equations of second order. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[7] Xin Guo, Renyuan Xu, and Thaleia Zariphopoulou. Entropy regularization for mean field games with learning. Mathematics of Operations research, 47(4):3239–3260, 2022.

[8] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70, pages 1352–1361.

[9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80, pages 1856–1865.

[10] Saul D. Jacka and Aleksandar Mijatović. On the policy improvement algorithm in continuous time. Stochastics, 89(1):348–359, 2017.

[11] E. T. Jaynes. Information theory and statistical mechanics. Phys. Rev. (2), 106:620–630, 1957.

[12] E. T. Jaynes. Information theory and statistical mechanics. II. Phys. Rev. (2), 108:171–190, 1957.

[13] Yanwei Jia and Xun Yu Zhou. q-learning in continuous time. Journal of Machine Learning Research, 24(161):1–61, 2023.

[14] Ioannis Karatzas and Steven E. Shreve. Brownian motion and stochastic calculus, volume 113 of Graduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1991.

[15] B. Kerimkulov, D. Šiška, and L. Szpruch. A modified MSA for stochastic control problems. Appl. Math. Optim., 84(3):3417–3436, 2021.
[16] Bekzhan Kerimkulov, David Šiška, and Lukasz Szpruch. Exponential convergence and stability of Howard’s policy improvement algorithm for controlled diffusions. *SIAM J. Control Optim.*, 58(3):1314–1340, 2020.

[17] Jaeyoung Lee and Richard S Sutton. Policy iterations for reinforcement learning problems in continuous time and space—fundamental theory and methods. *Automatica*, 126:109421, 2021.

[18] Michael Lutter, Shie Mannor, Jan Peters, Dieter Fox, and Animesh Garg. Value iteration in continuous actions, states and time. *arXiv preprint arXiv:2105.04682*, 2021.

[19] Tsoy-Wo Ma. Higher chain formula proved by combinatorics. *Electron. J. Combin.*, 16(1):Note 21, 7, 2009.

[20] M. L. Puterman. On the convergence of policy iteration for controlled diffusions. *J. Optim. Theory Appl.*, 33(1):137–144, 1981.

[21] Christoph Reisinger and Yufei Zhang. Regularity and stability of feedback relaxed controls. *SIAM J. Control Optim.*, 59(5):3118–3151, 2021.

[22] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[23] Wenpin Tang, Hung Vinh Tran, and Yuming Paul Zhang. Policy iteration for the deterministic control problems—a viscosity approach. *arXiv preprint arXiv:2301.00419*, 2023.

[24] Wenpin Tang, Yuming Paul Zhang, and Xun Yu Zhou. Exploratory hjb equations and their convergence. *SIAM Journal on Control and Optimization*, 60(6):3191–3216, 2022.

[25] Brent A Wallace and Jennie Si. Continuous-time reinforcement learning control: A review of theoretical results, insights on performance, and needs for new designs. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[26] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: a stochastic control approach. *J. Mach. Learn. Res.*, 21:Paper No. 198, 34, 2020.

[27] Haoran Wang and Xun Yu Zhou. Continuous-time mean-variance portfolio selection: a reinforcement learning framework. *Math. Finance*, 30(4):1273–1308, 2020.

[28] Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 1255–1262.

[29] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 1433–1438.