



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Iterative methods for solving stochastic optimal control problems

Bekzhan Kerimkulov

Doctor of Philosophy
University of Edinburgh
2021

Declaration

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

The work presented in Chapter 2 was previously published in SIAM Journal on Control and Optimization as “Exponential convergence and stability of Howard’s policy improvement algorithm for controlled diffusions” by myself and my supervisors Dr. David Šiška and Dr. Łukasz Szpruch. This study was conceived by all of the authors.

The work presented in Chapter 3 was previously published in Applied Mathematics and Optimization as “A modified MSA for stochastic control problems” by myself and my supervisors Dr. David Šiška and Dr. Łukasz Szpruch. This study was conceived by all of the authors.

The ideas, development and writing up of the thesis, and the papers included in it, were the principal responsibility of myself, Bekzhan Kerimkulov, working under the supervision of Dr. David Šiška and Dr. Łukasz Szpruch.

(Bekzhan Kerimkulov)

To my loving family and wife

Acknowledgements

First of all, I would like to thank my supervisors, Dr. David Šiška and Dr. Łukasz Szpruch. Their indispensable guidance, constructive critique, and consistent encouragement helped me throughout my doctoral studies. Without their patience and support, this thesis would not have been possible.

I would like acknowledge the support of the Maxwell Institute Graduate School in Analysis and its Applications (MIGSAA). I am also grateful for Dr. Michela Ottobre and Ms. Isabelle Hanlon for their constant support in administrative works.

My time in Edinburgh has been so enjoyable because of the friends I have made along the way. I would like to thank them for all the good times we have shared and the support they have given.

Last but not least, I would like to thank my family – my parents, Kerimkhulle Seyit and Duambekova Bissenkul, my siblings, Ainur, Zhandos and Zhanar, and my wife, Akmaral, for continuously supporting me through all these years and for everything they have done for me.

Abstract

Optimal control problems are inherently hard to solve as the optimization must be performed simultaneously with updating the underlying system. Therefore, most of the time, they have to be solved numerically. In this thesis, we consider two iterative methods for solving stochastic optimal control problems: Howard's policy improvement algorithm and the method of successive approximations (MSA). Starting from an initial guess, Howard's policy improvement algorithm separates the step of updating the trajectory of the dynamical system from the optimization and iterations of this should converge to the optimal control. In the discrete space-time setting this is often the case and even rates of convergence are known. In the continuous space-time setting of controlled diffusion the algorithm consists of solving a linear PDE followed by a maximization problem. This has been shown to converge; in some situations, however no global rate is known. The first main contribution is to establish global rate of convergence for the policy improvement algorithm and a variant, called here the gradient iteration algorithm. The second main contribution is the proof of stability of the algorithms under perturbations to both the accuracy of the linear PDE solution and the accuracy of the maximization step. The proof technique is new in this context as it uses the theory of backward stochastic differential equations.

The classical MSA is an iterative method for solving stochastic control problems and is derived from Pontryagin's optimality principle. It is known that the MSA may fail to converge. Using estimates for the backward stochastic differential equation we propose a modification to the MSA algorithm. This modified MSA is shown to converge for general stochastic control problems with control in both the drift and diffusion coefficients. Under some additional assumptions the rate of convergence is shown. The results are valid without restrictions on the time horizon of the control problem, in contrast to iterative methods based on the theory of forward-backward stochastic differential equations. In addition, we study the MSA for solving stochastic control problems with entropy regularization, where the action space is the space of measures. We modify the classical MSA by the relative entropy of two consecutive controls coming from the algorithm. We establish convergence of the algorithm and show how it can be applied for relaxed stochastic control problems.

Lay summary

Stochastic control problems arise naturally in a range of applications in engineering, economics, and finance. Apart from very specific cases such as linear-quadratic control in engineering or the Merton portfolio optimization task in finance, stochastic control problems typically have no closed form solutions and have to be solved numerically. In this thesis we consider two iterative algorithms to get numerical solutions of stochastic control problems. We study their convergence and under some conditions we obtain convergence rates of the algorithms.

Contents

Abstract	7
1 Introduction	13
1.1 Stochastic control problem	13
1.2 Dynamic programming principle and Hamilton-Jacobi-Bellman equation	15
1.3 Pontryagin's optimality principle	17
1.4 Outline and contribution	21
2 Exponential Convergence and stability of Howard's Policy Im- provement Algorithm for Controlled Diffusions	23
2.1 Introduction	23
2.2 Assumptions and Notation	25
2.3 Convergence of gradient iteration algorithm	28
2.4 Convergence of policy improvement	31
2.5 Policy improvement	34
2.6 Stability under Perturbations to Solution of the Linear PDE	35
2.7 Stability under Perturbation of the Maximization	38
2.8 Example	41
3 A modified MSA for stochastic control problems	45
3.1 Introduction	45
3.2 Main results	47
3.3 Proofs	50
4 A regularized MSA for stochastic control problems	59
4.1 Introduction	59
4.2 Main results	62
4.2.1 Relaxed control	66
4.3 Proofs	68
5 Extensions and Future Work	81
A Supplementary materials for Chapter 2	83
A.1 Some results from theory of BSDEs	83
A.1.1 BSDE with drivers of quadratic growth.	89

B	Supplementary materials for Chapter 3	91
B.1	Auxiliary Lemma	91
C	Supplementary materials for Chapter 4	93
C.1	Flat derivative on measure spaces	93
C.2	Standard SDE estimate	94

Chapter 1

Introduction

The main aim of the thesis is to study iterative algorithms for solving stochastic control problems. In this chapter, we briefly describe stochastic control problems and their connection to the Hamilton-Jacobi-Bellman equation and to the Pontryagin's optimality principle. We also give an outline of the thesis and state the main contributions to the field.

Stochastic control problems arise naturally in a range of applications in engineering, economics, and finance. Apart from very specific cases such as linear-quadratic control in engineering or the Merton portfolio optimization task in finance, stochastic control problems typically have no closed form solutions and have to be solved numerically. There are mainly two approaches to tackle the problem: dynamic programming principle and Pontryagin's optimality principle. Both approaches are used to establish existence (and possibly uniqueness) of solutions to the control problem (either the value function or the optimal control). These approaches can form basis for the derivation of approximation methods.

1.1 Stochastic control problem

We will consider the continuous space, continuous time problem where the controlled system is modelled by an $\mathbb{R}^d$ -valued diffusion process. Let us fix a finite time $T \in (0, \infty)$. Let W be a d' -dimensional Wiener martingale on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0, T]}, \mathbb{P})$. We are given an action set A (in general separable complete metric space). We define the controlled SDE as follows. Let

$$b : [0, T] \times \mathbb{R}^d \times A \rightarrow \mathbb{R}^d \text{ and } \sigma : [0, T] \times \mathbb{R}^d \times A \rightarrow \mathbb{R}^{d \times d'}$$

be measurable functions.

Assumption 1.1.1. Assume that for each $t \in [0, T]$ that $(x, a) \mapsto b(t, x, a)$ and $(x, a) \mapsto \sigma(t, x, a)$ are continuous, assume that for each $t \in [0, T]$ we have $x \mapsto b(t, x, a)$ and $x \mapsto \sigma(t, x, a)$ continuous in x uniformly in $a \in A$ and there is a constant K such that for any $t \in [0, T]$, $x, y \in \mathbb{R}^d$, $a \in A$ we have

$$|b(t, x, a) - b(t, y, a)| + |\sigma(t, x, a) - \sigma(t, y, a)| \leq K|x - y|.$$

Moreover, for all $t \in [0, T]$, $x \in \mathbb{R}^d$, $a \in A$ it holds that

$$|b(t, x, a)| + |\sigma(t, x, a)| \leq K(1 + |x| + |a|).$$

We call the set of admissible controls the set

$$\mathcal{A} := \{\alpha \in \mathbb{H}_T^2 : \forall \omega \in \Omega, t \in [0, T] \alpha_t(\omega) \in A \text{ and } \alpha \text{ is progressively measurable}\},$$

where we write $\mathbb{H}_T^2$ for the set of progressively measurable processes φ such that

$$\|\varphi\|_{\mathbb{H}_T^2} := \left(\mathbb{E} \int_0^T |\varphi_t|^2 dt \right)^{1/2} < \infty.$$

Given a fixed control $\alpha \in \mathcal{A}$, we consider the SDE for $t \in [0, T]$, $s \in [t, T]$

$$dX_s = b(s, X_s, \alpha_s) ds + \sigma(s, X_s, \alpha_s) dW_s, \quad X_t = \xi, \quad (1.1)$$

where ξ is $\mathcal{F}_t$ -measurable and square integrable. Notice that the SDE (1.1) is a special case of an SDE with random coefficients. Indeed, if we fix $\alpha \in \mathcal{A}$ and denote $\tilde{b}(t, x) := b(t, x, \alpha_t)$ and $\tilde{\sigma}(t, x) := \sigma(t, x, \alpha_t)$, one can show the progressive measurability of $\tilde{b}$ and $\tilde{\sigma}$. Moreover, by Assumption 1.1.1

$$\mathbb{E} \int_0^T |b(t, 0, \alpha_t)|^2 dt \leq \mathbb{E} \int_0^T K^2(1 + |\alpha_t|^2) dt \leq 2K^2T + 2K^2\|\alpha\|_{\mathbb{H}_T^2}^2 < \infty,$$

and, similarly, $\mathbb{E} \int_0^T |\sigma(t, 0, \alpha_t)|^2 dt < \infty$. Hence, due to [1, Theorem 1.3.15] we have the following result.

Theorem 1.1.2 (Existence and uniqueness of a strong solution to the controlled SDE). *Let $t \in [0, T]$ and $\alpha \in \mathcal{A}$. For any $\mathcal{F}_t$ -measurable random variable ξ valued in $\mathbb{R}^d$, such that $\mathbb{E}[|\xi|^2] < \infty$, there is a unique strong solution $X = X^{t, \xi, \alpha}$ to the SDE (1.1) on the interval $[t, T]$ such that there exists a constant C_T such that*

$$\sup_{\alpha \in \mathcal{A}} \mathbb{E} \sup_{s \in [t, T]} |X_s|^2 \leq C_T(1 + \mathbb{E}|\xi|^2).$$

Let $f : [0, T] \times \mathbb{R}^d \times A \rightarrow \mathbb{R}$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be two measurable functions. We then define the objective functional:

$$J(t, \xi, \alpha) := \mathbb{E} \left[\int_t^T f(s, X_s^{t, \xi, \alpha}, \alpha_s) ds + g(X^{t, \xi, \alpha}) \right].$$

We refer to f as the running gain and to g as the terminal gain. Under some assumptions on f and g , e.g. f and g are locally Lipschitz continuous in space uniformly in time and action and have quadratic growth in space and action, one can show that $J(t, \xi, \alpha) < \infty$. Therefore, the functional J is well-defined. We wish to solve the optimization problem, i.e., to maximize over control processes from the set of all admissible controls the objective functional J , and we introduce

the associated value function:

$$v(t, x) = \sup_{\alpha \in \mathcal{A}} J(t, x, \alpha). \quad (1.2)$$

One of the mathematical difficulties in stochastic control theory is that we do not even know at this point whether v is measurable or not. However, under the above assumptions on f and g one can show that J and v are locally Lipschitz, see [2, Lemma 3.13]. Hence, as an important consequence of this we have that if we fix t then $x \mapsto v(t, x)$ is measurable (as continuous functions are measurable). In many cases, there is no optimal control process α^* for which we would have $v(t, x) = J(t, x, \alpha^*)$. However, there is always an ε -optimal control. We provide the definition of ε -optimal controls.

Definition 1.1.3 (ε -optimal controls). *Let $t \in [0, T]$, $x \in \mathbb{R}^d$ and $\varepsilon \geq 0$. A control $\alpha^\varepsilon \in \mathcal{A}$ is said to be ε -optimal if*

$$v(t, x) \leq \varepsilon + J(t, x, \alpha^\varepsilon).$$

1.2 Dynamic programming principle and Hamilton-Jacobi-Bellman equation

After setting up the stochastic control problem, we now ready to discuss the dynamic programming principle and the Hamilton-Jacobi-Bellman equation. Dynamic programming is one of the most popular approaches for solving the stochastic control problem. The main idea was originated from the so-called Bellman's principle, which states:

“An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.”

We state the Bellman's principle / dynamic programming in the following theorem, see [1, Theorem 3.3.1] for more details.

Theorem 1.2.1 (Dynamic programming principle). *For any $0 \leq t \leq s \leq T$, for any $x \in \mathbb{R}^d$, we have*

$$v(t, x) = \sup_{\alpha \in \mathcal{A}[t, s]} \mathbb{E} \left[\int_t^s f(u, X_u^{t, x, \alpha}, \alpha_u) du + v(s, X_s^{t, x, \alpha}) \middle| X_t^{t, x, \alpha} = x \right], \quad (1.3)$$

where $\mathcal{A}[t, s]$ denotes the set of all admissible controls on the time interval $[t, s]$ and $X^{t, x, \alpha}$ denotes the controlled SDE starting from x at time t and controlled by $\alpha \in \mathcal{A}$.

Assume that the value function $v = v(t, x)$ is $C^{1,2}([0, T] \times \mathbb{R}^d)$. To get the Hamilton-Jacobi-Bellman equation, we can apply Itô's formula to v and X . For compactness of writing we will denote $\sigma^a(t, x) := \sigma(t, x, a)$, $b^a(t, x) := b(t, x, a)$ and $f^a(t, x) := f(t, x, a)$. We then define

$$\mathcal{L}^a v := b^a D_x v + \frac{1}{2} \text{tr}[\sigma^a (\sigma^a)^\top D_x^2 v].$$

Having this, we now state the Hamilton-Jacobi-Bellman equation, see [2, Theorem 4.7] for more details.

Theorem 1.2.2 (Hamilton-Jacobi-Bellman). *Assume that the Bellman principle holds. Assume that b and σ are bounded and right continuous in t uniformly in x variable. If the value function v is $C^{1,2}([0, T] \times \mathbb{R}^d)$, then it satisfies*

$$\begin{aligned} \partial_t v + \sup_{a \in A} (\mathcal{L}^a v + f^a) &= 0 \quad \text{on } [0, T] \times \mathbb{R}^d \\ v(T, x) &= g(x) \quad \forall x \in \mathbb{R}^d. \end{aligned} \tag{1.4}$$

If we have a classical solution to (1.4), then we give the following verification theorem, see [2, Theorem 4.8] for more details.

Theorem 1.2.3 (HJB verification). *If some $u \in C^{1,2}([0, T] \times \mathbb{R}^d)$ satisfies (1.4) and we have that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ there is some measurable function $a : [0, T] \times \mathbb{R}^d \rightarrow A$ such that*

$$a(t, x) \in \arg \max_{a \in A} ((\mathcal{L}^a u)(t, x) + f^a(t, x)) , \tag{1.5}$$

and if

$$dX_s^* = b(s, X_s^*, a(s, X_s^*)) ds + \sigma(s, X_s^*, a(s, X_s^*)) dW_s, \quad X_t^* = x$$

admits a unique solution, and if the process

$$t' \mapsto \int_t^{t'} D_x u(s, X_s^*) \sigma(s, X_s^*, a(s, X_s^*)) dW_s$$

is a martingale in $t' \in [t, T]$, then

$$\alpha_s^* := a(s, X_s^*) \quad s \in [t, T]$$

is optimal for problem and $v(t, x) = u(t, x)$.

We refer to Theorem 1.2.3 as the verification theorem. This theorem is important for solving the control problem, because if we know the value function v , then the dynamic optimisation problem turns into a static optimisation problems at each point $(t, x) \in [0, T] \times \mathbb{R}^d$. Notice that (1.5) is calculated pointwise over $(t, x) \in [0, T] \times \mathbb{R}^d$.

It is rarely possible to find a closed form solution to (1.4) and so various approximations have to be employed. One may, for example, choose to use a finite difference method to discretize (1.4) and indeed this has been widely studied; see, e.g., [3] or [4] and references therein. This results in a high dimensional nonlinear system of equations that still retains the structure of (1.4). To solve this nonlinear system one may apply the Howard's policy improvement algorithm. The convergence and the rate of convergence would then follow from results available on discrete space-time control problems. However, to check that the assumptions required for convergence are satisfied is not straightforward and moreover it is dependent on the discretization scheme used.

An alternative approach is to linearize (1.4) and to iterate. The classical approach is the Bellman–Howard policy improvement/iteration algorithm. The algorithm is initialized with a “guess” of the Markovian control. Given a Markovian control strategy at step n one solves a linear PDE with the given control fixed and then one uses the solution to the linear PDE to update the Markovian control. In Chapter 2, we will show that this policy improvement algorithm (see Algorithm 1) and a variant which we call the gradient iteration algorithm (see Algorithm 2) converge, under appropriate assumptions, exponentially fast.

Iterative algorithms for the solution of optimal control problems go back to the work of Bellman [5, 6] where the value iteration algorithms for finite space-time problems are developed and their convergences are shown. Howard [7] proposed the policy improvement algorithm in the context of the discrete space-time Markovian decision process. Puterman and Brumelle’s paper [8] was one of the first results on the convergence properties for the policy iteration for MDP problems. The abstract function space setting employed in the paper applies to both discrete and continuous settings. Their main observation is that the policy iteration can be viewed as a type of Newton’s method. Hence similar convergence results to those known for Newton’s method follow: in particular, if the initial guess is in a neighborhood of the true solution, then the convergence will be quadratic. Puterman [9] applied this in a setting very similar to that of this paper to prove quadratic convergence in the neighborhood of the limit. Santos and Rust [10] consider the discrete time but continuous space and controls setting. They extend the results of Puterman and Brumelle [8] to show global convergence, but without global rate, and quadratic local convergence rate of policy iteration and superlinear local convergence under more general conditions. In the case of stochastic control problems with jump-diffusion processes, Bäuerle and Rieder [11] have proved a convergence result of the Howard’s policy improvement algorithm with the help of martingale techniques. In the fully discrete space and time setting Bokanowski, Maroso, and Zidani [12] have shown global superlinear convergence, under a monotonicity assumption on the matrices defining the control problem. Convergence of policy iteration has been recently proved by Jacka and Mijatović [13] and Jacka, Mijatović, and Siraj [14]. Further, Maeda and Jacka [15] have shown quadratic local convergence of the policy iteration algorithm for the time-independent control problem. The local quadratic convergence is similar to the result of Puterman [9] but the specific control problem is different and moreover they employ a completely different technique based on Schauder estimates for linear PDEs.

1.3 Pontryagin’s optimality principle

In this section, we consider another approach for solving stochastic optimal control problems based on backward stochastic differential equations (BSDEs), Pontryagin’s optimality principle, see [16, Chapter 4] for more details. The theory of BSDEs was pioneered by Pardoux and Peng [17]. It became an important field of research due to its connections with stochastic control, mathematical finance, and partial differential equations. BSDEs provide a probabilistic representation

of nonlinear PDEs, which extends the famous Feynman-Kac formula for linear PDEs.

Recall that given running gain f and terminal gain g our aim is to optimally control the dynamic

$$dX_t^\alpha = b(t, X_t^\alpha, \alpha_t) dt + \sigma(t, X_t^\alpha, \alpha_t) dW_t, \quad t \in [0, T], \quad X_0^\alpha = x, \quad (1.6)$$

where $\alpha \in \mathcal{A}$ and we assume that Assumption 1.1.1 holds. This means that we want to find a control which will maximize

$$J(\alpha) := \mathbb{E} \left[\int_0^T f(t, X_t^\alpha, \alpha_t) dt + g(X_T^\alpha) \right] \quad (1.7)$$

over $\alpha \in \mathcal{A}$. In this section, we assume that the process starts from state x at time 0 and we will assume that A is a convex subset of $\mathbb{R}^p$ and that the conditions for a control process α to be admissible are stable under convex combinations. We assume that the coefficients b, σ, f are continuously differentiable with respect to the variables $(x, a) \in \mathbb{R}^d \times A$ with Lipschitz continuous and bounded partial derivatives uniformly in $t \in [0, T]$. Similarly, we assume that the function g is continuous differentiable with respect to $x \in \mathbb{R}^d$, with Lipschitz continuous and bounded partial derivatives uniformly in $t \in [0, T]$. For $\alpha, \beta \in \mathcal{A}$ and for each $\varepsilon > 0$ small enough, define the admissible control $\alpha^\varepsilon = \alpha + \varepsilon\beta$. We consider the variation process V as the solution of the following linear SDE

$$\begin{aligned} dV_t = & [D_x b(t, X_t^\alpha, \alpha_t) V_t + D_a b(t, X_t^\alpha, \alpha_t) \beta_t] dt \\ & + [D_x \sigma(t, X_t^\alpha, \alpha_t) V_t + D_a \sigma(t, X_t^\alpha, \alpha_t) \beta_t] dW_t \end{aligned} \quad (1.8)$$

with the initial condition $V_0 = 0$. The existence and uniqueness of V comes from assumption on the boundedness of the partial derivatives of the coefficients. After defining V , one can compute the Gateaux derivative of the functional J .

Lemma 1.3.1 (Lemma 4.8 in [16]). *The function $\alpha \mapsto J(\alpha)$ is Gateaux differentiable and*

$$\begin{aligned} & \frac{d}{d\varepsilon} J(\alpha + \varepsilon\beta) \Big|_{\varepsilon=0} \\ &= \mathbb{E} \left[\int_0^T [D_x f(t, X_t, \alpha_t) V_t + D_a f(t, X_t, \alpha_t) \beta_t] dt + D_x g(X_T) V_T \right]. \end{aligned} \quad (1.9)$$

Let us now define the Hamiltonian $H : [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d \times d'} \times A \rightarrow \mathbb{R}$ of the dynamical system as

$$H(t, x, y, z, a) := b(t, x, a)y + \text{tr} [\sigma^\top(t, x, a)z] + f(t, x, a).$$

Assumption 1.3.2. Assume that $x \mapsto H(t, x, y, z, a)$ is differentiable for all $a \in A, t \in [0, T], y \in \mathbb{R}^d, z \in \mathbb{R}^{d \times d'}$ with derivative bounded uniformly in a, t, y, z . Assume that g is differentiable in x with the derivative having at most linear growth in x .

For each $\alpha \in \mathcal{A}$ consider the adjoint BSDE:

$$dY_t^\alpha = -D_x H(t, X_t^\alpha, Y_t^\alpha, Z_t^\alpha, \alpha_t) dt + Z_t^\alpha dW_t, \quad Y_T^\alpha = D_x g(X_T^\alpha). \quad (1.10)$$

Notice that under Assumption 1.3.2 and Assumption 1.1.1 one can show that the BSDE (1.10) has the unique solution (Y^α, Z^α) . Moreover, (1.10) is a linear BSDE and we can derive an explicit formula for Y^α . Applying Itô's product rule we can derive an expression of the Gâteaux derivative of the cost functional in terms of the Hamiltonian of the system.

Lemma 1.3.3 (Corollary 4.11 in [16]).

$$\frac{d}{dt} J(\alpha + \varepsilon \beta) \Big|_{\varepsilon=0} = \mathbb{E} \int_0^T D_a H(t, X_t, Y_t, Z_t, \alpha_t) \beta_t dt.$$

Hence, we can state a necessary condition of optimality.

Theorem 1.3.4 (Theorem 4.12 in [16]). *Under the above assumptions, if the admissible control $\alpha \in \mathcal{A}$ is optimal, X is the associated controlled state, and (Y, Z) are associated adjoint processes solving the adjoint equation (1.10), then*

$$\forall a \in A, D_a H(t, X_t, Y_t, Z_t, \alpha_t)(a - \alpha_t) \leq 0 \quad \text{a.e. in } t \in [0, T], \text{ a.s.} \quad (1.11)$$

Under mild regularity condition from (1.11) we can conclude that

$$\forall a \in A, H(t, X_t, Y_t, Z_t, \alpha_t) \geq H(t, X_t, Y_t, Z_t, a) \quad \text{a.e. in } t \in [0, T], \text{ a.s.}$$

If we assume additionally convexity of coefficients, then the necessary condition can be turned into a sufficient condition for optimality.

Theorem 1.3.5 (Theorem 4.14 in [16]). *Let Assumption 1.1.1 and 1.3.2 hold. Let $A \subset \mathbb{R}^p$. Let g be concave. Let $\beta \in \mathcal{A}$ and X^β be the associated controlled diffusion and (Y^β, Z^β) the solution of the adjoint BSDE. If $\beta \in \mathcal{A}$ is such that*

$$H(t, X_t^\beta, Y_t^\beta, Z_t^\beta, \beta_t) = \max_{a \in A} H(t, X_t^\beta, Y_t^\beta, Z_t^\beta, a) < \infty \quad \text{a.s. for almost all } t \in [0, T]$$

holds and if

$$(x, a) \mapsto H(t, x, Y_t^\beta, Z_t^\beta, a)$$

is differentiable and concave, then $J(\beta) = \sup_{\alpha \in \mathcal{A}} J(\alpha)$ i.e. β is an optimal control.

As we mentioned previously, the Pontryagin's optimality principle can be base to derive numerical algorithms to get numerical solutions of stochastic control problems. The method of successive approximations (MSA), see Algorithm 3, is essentially a way of applying the Pontryagin's optimality principle to derive a numerical algorithm. The classical MSA (i.e. case $\rho = 0$ in Algorithm 3) for numerical solution of deterministic control problems was proposed already in [18]. Recent application of the modified MSA to a deep learning problem has been studied in [19], where they formulated the training of deep neural networks

as an optimal control problem and introduced the modified method of successive approximations as an alternative training algorithm for deep learning. In Chapter 3, we will study the modified MSA in more details and show convergence result.

It is known also that the solution of the stochastic optimal control problem can be obtained from a corresponding forward backward stochastic differential equation (FBSDE) via the stochastic optimality principle, see [20, Chapter 8.1]. Indeed, let us consider (1.6) and (1.10), and recall from the stochastic optimality principle that for the optimal control $\alpha^* = (\alpha_s^*)_{s \in [0, T]}$ we have that (1.11) holds. Assume that under some conditions on b, σ and f we have that the first order condition uniquely determines α^* for $s \in [0, T]$ by

$$\alpha_s^* = \varphi(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}), \quad (1.12)$$

for some function φ . Therefore, after plugging (1.12) into (1.6) and (1.10), we obtain the following coupled FBSDE:

$$\begin{aligned} dX_s^{\alpha^*} &= \bar{b}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}) ds + \bar{\sigma}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}) dW_s, \quad s \in [0, T], \quad X_0^{\alpha^*} = x. \\ dY_s^{\alpha^*} &= -D_x \bar{\mathcal{H}}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}) ds + Z_s^{\alpha^*} dW_s, \quad Y_T = D_x g(X_T^{\alpha^*}), \quad s \in [0, T], \end{aligned} \quad (1.13)$$

where $(\bar{b}, \bar{\sigma})(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}) = (b, \sigma)(s, X_s^{\alpha^*}, \varphi(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}))$ and $\bar{\mathcal{H}}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}) = \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \varphi(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}))$. It is worth mentioning that when σ does not depend on the control $\bar{\sigma}$ will depend on the forward process and time only. This means that $\bar{\sigma}$ does not have Y and Z components. Therefore, numerical methods for solving (1.13) can be form for numerical methods for the stochastic control problem.

The theory of FBSDE has been studied widely and there are several methods to show the existence and uniqueness result, and a number of numerical algorithms have been proposed based on those methods. First is the method of contraction mapping. It was first studied by Antonelli [21] and later by Pardoux and Tang [22]. The main idea there is to show that a certain map is a contraction, and then to apply a fixed point argument. However, it turns out that this method works only for small enough time horizon T . In the case when $\bar{\sigma}$ does not depend on Y and Z , having small T is sufficient to get contraction. Otherwise, one needs to assume additionally that the Lipschitz constants of $\bar{\sigma}$ in z and that of g in x satisfy a certain inequality, see [20, Theorem 8.2.1]. Using the method of contraction mapping one can then implement a Picard-iteration-type numerical algorithm and show exponential convergence for small T . The second method is the Four Step Scheme. It was introduced by Ma, Protter and Yong, see [23], and was later studied by Delarue [24]. The idea is to use a decoupling function and then study an associated quasi-linear PDE. We note that in [23, 24] the forward diffusion coefficient $\bar{\sigma}$ does not depend on Z . This corresponds to stochastic control problems with the uncontrolled diffusion coefficient. The numerical algorithms based on this method exploits the numerical solution of the associated quasi-linear PDE and therefore faces some limitations for high dimensional problems, see Douglas, Ma and Protter [25], Milstein and Tretyakov [26], Ma, Shen and Zhao [27] and Delarue and Menozzi [28]. Guo, Zhang and Zhuo [29]

proposed a numerical scheme for high-dimensional quasi-linear PDE associated with the coupled FBSDE when $\bar{\sigma}$ does not depend on Z , which is based on a monotone scheme and on probabilistic approach. Finally, there is the method of continuation. This method was developed by Hu and Peng [30], Peng and Wu [31] and by Yong [32]. It allows them to show the existence and uniqueness result for arbitrary T under monotonicity conditions on the coefficients, which one would not expect to apply to FBSDEs arising from a control problem as described by (1.12), (1.13). Recently, deep learning methods have been applied to solving FBSDEs. In [33], three algorithms for solving fully coupled FBSDEs which have good accuracy and performance for high-dimensional problems are provided. One of the algorithms is based on the Picard iteration and it converges, but only for small enough T . Such method for solving high-dimensional FBSDEs has also been proposed in [34].

1.4 Outline and contribution

In this thesis, we consider iterative methods for solving stochastic optimal control problems (1.2), which are based on two discussed approaches. We study their convergence and, in some cases, derive rates of convergence.

In Chapter 2, we apply Howard’s policy iteration by guessing an initial Markovian control, by solving a linear PDE at each iteration and by updating the control with the solution to the linear PDE. The main contributions are to establish a global rate of convergence and stability for the policy iteration algorithm and a variant, which we call the gradient iteration algorithm. The analysis is carried out using backward stochastic differential equations (BSDEs) and to the ours best knowledge this is the first time BSDEs have been used to study convergence of the policy iteration algorithm. The assumptions required for this are effectively Lipschitz dependence in the drift, diffusion, instantaneous payoff, and terminal payoff functions and independence of the diffusion matrix on the control. The stability results show that the policy iteration remains stable even if the linear PDE is solved only approximately and even if the maximization is step performed approximately. Moreover, they allow one to devise computationally efficient algorithms as they show that in the initial steps it is sufficient to solve the linear PDE with very low accuracy, and a highly accurate PDE solver is only required for the final few iterations of the algorithms.

In Chapter 3, we consider a modification to the method of successive approximations, see Algorithm 3. The modified MSA consists of successive integrations of the state and adjoint systems and updates to the control. One of the motivations to explore the modified MSA for stochastic control problems is to obtain convergence, ideally with rate, of an iterative algorithm, applicable to problems with the control in the diffusion part of the controlled dynamics. This is in contrast to the result in Chapter 2 where convergence rate of an the Bellman–Howard policy iteration is shown but only for control problems with no control in the diffusion part of the controlled dynamics. The main contributions are the probabilistic proof of convergence of the modified method of successive approximations and establishing convergence rate for a specific type of optimal control problems.

In Chapter 4, we continue exploring MSA for entropy regularized stochastic control problems. In particular, we modify the Hamiltonian using the relative entropy of two consecutive controls appearing in Algorithm 4. From empirical studies, we know that the relative entropy penalization is important in learning and makes the learning more stable, since at each iteration we penalize for being far from the control of the previous iteration. As in Chapter 3, the main contribution is the probabilistic proof of convergence of Algorithm 4 for solving entropy regularized optimal control problems.

Finally, in Chapter 5, a short discourse on potential extensions of the work contained in this thesis is provided.

Chapter 2

Exponential Convergence and stability of Howard's Policy Improvement Algorithm for Controlled Diffusions

2.1 Introduction

In this chapter, we consider the policy iteration algorithm and gradient iteration algorithm; see Algorithms 1 and 2. These are effectively a linearization method for the inherently nonlinear problem and play an essential role in numerical solutions of stochastic control problems.

We will consider the continuous space, continuous time problem where the controlled system is modelled by an $\mathbb{R}^d$ -valued diffusion process. Let W be a d' -dimensional Wiener martingale on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$. Let us fix a finite time $T \in (0, \infty)$ and consider the controlled SDE

$$dX_s = b^\alpha(s, X_s) ds + \sigma(s, X_s) dW_s, \quad s \in [t, T], \quad X_t = x. \quad (2.1)$$

Here $\alpha = (\alpha_s)$ is a control belonging to the space of admissible controls $\mathcal{A}$, valued in $A \subseteq \mathbb{R}^m$, and we will write $X^{t,x,\alpha}$ to denote the solution of (2.1) which starts from x at time t while being controlled by α . We shall consider the gain functional in the form

$$J(t, x, \alpha) := \mathbb{E} \left[\int_t^T f^\alpha(s, X_s^{t,x,\alpha}) ds + g(X_T^{t,x,\alpha}) \right] \quad (2.2)$$

for all $(t, x) \in [0, T] \times \mathbb{R}^d$ and $\alpha \in \mathcal{A}$. The value function $v = v(t, x)$ is given for all $t \in [0, T]$ and $x \in \mathbb{R}^d$ by

$$v(t, x) = \sup_{\alpha \in \mathcal{A}} J(t, x, \alpha). \quad (2.3)$$

We wish to solve the optimization problem, i.e., to find either the value function v or the optimal control α^* which achieves the maximum (or, if the supremum cannot be reached by $\alpha \in \mathcal{A}$, then an ε -optimal control $\alpha^\varepsilon \in \mathcal{A}$ such that $v(t, x) \leq$

$J(t, x, \alpha^\varepsilon) + \varepsilon$). It is well known that (see, e.g., Krylov [35]) that under reasonable assumptions the value function satisfies the Bellman PDE:

$$\begin{aligned} \partial_t v + \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v) + \sup_{a \in A} (b^a D_x v + f^a) &= 0 \text{ on } [0, T) \times \mathbb{R}^d, \\ v(T, x) &= g(x) \text{ on } x \in \mathbb{R}^d. \end{aligned} \quad (2.4)$$

Moreover (again see Krylov [35]), it is sufficient to consider Markovian controls, i.e., processes $\alpha_s = a(s, X_s^{t,x,\alpha})$ for some measurable function $a : [0, T] \times \mathbb{R}^d \rightarrow A$. Thus if we have obtained the value function, then we can find the optimal control (if it exists) as

$$a^*(t, x) = \arg \max_{a \in A} (b^a(t, x)(D_x v)(t, x) + f^a(t, x)).$$

The main contributions of this chapter are to establish a global rate of convergence and stability for the policy iteration algorithm and a variant, which we call the gradient iteration algorithm. The analysis is carried out using backward stochastic differential equations (BSDEs). The assumptions required for this are effectively Lipschitz dependence in the drift, diffusion, instantaneous payoff, and terminal payoff functions and independence of the diffusion matrix on the control; see (2.1). The stability results show that the policy iteration remains stable even if the linear PDE is solved only approximately and even if the maximization is step performed approximately. Moreover they allow one to devise computationally efficient algorithms as they show that in the initial steps it is sufficient to solve the linear PDE with very low accuracy, and a highly accurate PDE solver is only required for the final few iterations of the algorithms.

The chapter is organized as follows. In Section 2.2 we introduce all the assumptions and notation used throughout the chapter. In Sections 2.3 and 2.4 we state and prove the results concerning convergence of the gradient iteration algorithm and the policy improvement algorithm, respectively. Section 2.5 justifies the name “policy improvement algorithm” in that it shows that the value functions increase monotonically with iterations and it also shows that the algorithm converges under weaker assumptions than those required for obtaining the rate. Sections 2.6 and 2.7 prove the stability of the algorithms. In Section 2.8 we present an example that fits the setting of this chapter. In Appendix A.1, we collect several known results from the theory of BSDEs that are essential for the proofs.

We would like to emphasize that Algorithm 1 and Algorithm 2 are different, although they look rather similar. In Algorithm 1, v^n is the value function for the Markov control a^n , since it solves the PDE (2.5). In Algorithm 2, v^n is not the value function for the Markov control a^n . This is due to the term $b^{a^n} D_x v^{n-1}$ in the linear PDE (2.8).

Algorithm 1 Policy improvement algorithm.

Initialization: make a guess of the control $a^0 = a^0(t, x)$.

while difference between v^{n+1} and v^n is large **do**

 Given a control $a^n = a^n(t, x)$ solve the *linear* PDE

$$\begin{aligned} \partial_t v^n + \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v^n) + b^{a^n} D_x v^n + f^{a^n} &= 0 \text{ on } [0, T) \times \mathbb{R}^d, \\ v^n(T, \cdot) &= g \text{ on } x \in \mathbb{R}^d. \end{aligned} \quad (2.5)$$

Update the control

$$a^{n+1}(t, x) = \arg \max_{a \in A} [(b^a D_x v^n + f^a)(t, x)]. \quad (2.6)$$

end while

return v^n, a^{n+1} .

Algorithm 2 Gradient iteration algorithm.

Initialization: make a guess of the value function $v^0 = v^0(t, x)$.

while difference between v^n and v^{n-1} is large **do**

 Given value function $v^{n-1} = v^{n-1}(t, x)$ update the control

$$a^n(t, x) = \arg \max_{a \in A} [(b^a D_x v^{n-1} + f^a)(t, x)]. \quad (2.7)$$

Solve the *linear* PDE

$$\begin{aligned} \partial_t v^n + \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v^n) + b^{a^n} D_x v^{n-1} + f^{a^n} &= 0 \text{ on } [0, T) \times \mathbb{R}^d, \\ v^n(T, \cdot) &= g \text{ on } x \in \mathbb{R}^d. \end{aligned} \quad (2.8)$$

end while

return v^n, a^n .

2.2 Assumptions and Notation

We fix a finite horizon $T \in (0, \infty)$. We assume that for some $m \in \mathbb{N}$ we have $A \subseteq \mathbb{R}^m$ such that $0 \in A$. This is the space where the control processes α take values. We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$. Let $W = (W_t)_{t \in [0, T]}$ be a d' -dimensional Wiener martingale on this space. Moreover, we have the following:

- (i) For $\gamma > 0$ and a predictable process ϕ let us define

$$\|\phi\|_{\mathbb{H}_\gamma^2} := \left(\mathbb{E} \int_0^T e^{\gamma s} |\phi_s|^2 ds \right)^{\frac{1}{2}}.$$

For $\gamma = 0$ we will write $\|\cdot\|_{\mathbb{H}^2}$. We will use $\mathbb{H}^2$ to denote the set of all predictable processes ϕ such that $\|\phi\|_{\mathbb{H}^2} < \infty$. Note that the norm $\|\cdot\|_{\mathbb{H}^2}$ is

equivalent to the norm $\|\cdot\|_{\mathbb{H}_\gamma^2}$ for any $\gamma \geq 0$.

- (ii) Let $\mathcal{S}^2$ be the set of real valued $\mathbb{F}$ -adapted continuous processes ϕ on $[0, T]$ such that

$$\|\phi\|_{\mathcal{S}^2} := \mathbb{E} \left[\sup_{0 \leq r \leq T} |\phi_r|^2 \right] < \infty.$$

- (iii) For adapted processes ϕ such that $\int_0^t |\phi_s|^2 ds < \infty$ almost surely we will define

$$(\phi \bullet W)_t := \int_0^t \phi_s dW_s.$$

- (iv) For any continuous local martingale M let with $(\langle M \rangle_t)_{t \in [0, T]}$ denote the quadratic variation process and moreover let

$$\mathcal{E}(M)_t := \exp \left(M_t - \frac{1}{2} \langle M \rangle_t \right).$$

We are given measurable functions

$$b : A \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d \text{ and } \sigma : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d'}.$$

The state of the system is governed by the controlled SDE (2.1).

Assumption 2.2.1. The functions b and σ are continuous in t . There exists $K \geq 0$ and such that $\forall x, y \in \mathbb{R}^d, \forall a \in A, \forall t \in [0, T]$,

$$|b^a(t, x) - b^a(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq K|x - y| \quad (2.9)$$

and

$$|\sigma(t, x)| \leq K(1 + |x|), \quad |b^a(t, x)| \leq K(1 + |x| + |a|). \quad (2.10)$$

Under Assumption 2.2.1 we know that for any $(t, x) \in [0, T] \times \mathbb{R}^d$ and for any progressively measurable A -valued control process $\alpha = (\alpha_s)$ there is a unique strong solution to (2.1) which we denote $(X_s^{t, x, \alpha})_{s \in [t, T]}$. Let

$$f : A \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R} \text{ and } g : \mathbb{R}^d \rightarrow \mathbb{R}$$

be two given measurable functions. Let us assume the following for the running gain function f and the terminal gain function g appearing in (2.2).

Assumption 2.2.2. There is a constant $K \geq 0$ such that $\forall x, y \in \mathbb{R}^d, \forall a \in A, \forall t \in [0, T]$

$$|g(x) - g(y)| + |f^a(t, x) - f^a(t, y)| \leq K|x - y| \quad (2.11)$$

and

$$|f^a(t, x)| \leq K(1 + |x| + |a|), \quad |g(x)| \leq K. \quad (2.12)$$

Under Assumption 2.2.2 the gain functional J given by (2.2) and the value function v given by (2.3) are well defined. Moreover, the value function v satisfies the

Bellman equation (with derivatives existing almost everywhere, see Krylov [35, Chapter 4], or in the sense of viscosity solutions, see, e.g., Pham [1] or Fleming and Soner [36])

$$\begin{aligned} \partial_t v + \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v) + \sup_{a \in A} (b^a D_x v + f^a) &= 0 \text{ on } [0, T) \times \mathbb{R}^n, \\ v(T, x) &= g(x) \text{ on } x \in \mathbb{R}^d. \end{aligned} \quad (2.13)$$

Let us now state the additional assumptions required for our convergence result.

Assumption 2.2.3. Assume that σ is non-degenerate and invertible. Let us define for each fixed $(t, x, z) \in [0, T] \times \mathbb{R}^d \times \mathbb{R}^d$ the function

$$a(t, x, z) := \arg \max_{a \in A} (b^a(t, x) \sigma^{-1}(t, x) z + f^a(t, x)). \quad (2.14)$$

We assume that the function $a(t, x, z)$ is measurable.

If the function $a \mapsto (b^a(t, x) \sigma^{-1}(t, x) z + f^a(t, x))$ is convex for each fixed (t, x, z) , which is in $[0, T] \times \mathbb{R}^d \times \mathbb{R}^d$, one can immediately see that Assumption 2.2.3 holds. More generally, this assumption can be verified using an appropriate measurable selection theorem. For example, if A is compact, then [37, Proposition D.5] shows that an appropriate measurable selection exists. If A is not compact but f is bounded, then [37, Proposition D.6] gives the same conclusion (using also that $z = \sigma(t, x) D_x v(t, x)$ and Remark 2.2.8).

Assumption 2.2.4. There are constants $K, \theta \geq 0$ such that the following hold:

1. (On the drift) For all $t \in [0, T]$, $x \in \mathbb{R}^d$, $a, a' \in A$,

$$|b^a(t, x) - b^{a'}(t, x)| \leq \sqrt{\theta} |a - a'| \quad (2.15)$$

and σ is non-degenerate, invertible and for all $t \in [0, T]$, $x \in \mathbb{R}^d$, $a \in A$ we have

$$|(b^a \sigma^{-1})(t, x)| < K. \quad (2.16)$$

2. (On the control function) For all $t \in [0, T]$, $x, x', z, z' \in \mathbb{R}^d$, $a, a' \in A$ we have that

$$\begin{aligned} |a(t, x, z) - a(t, x, z')| &\leq \sqrt{\theta} |z - z'|, \\ |a(t, x, z) - a(t, x', z)| &\leq K |x - x'| \text{ and } |a(t, 0, 0)| \leq K. \end{aligned}$$

3. (On the running reward)

$$|f^a(t, x) - f^{a'}(t, x)| \leq \sqrt{\theta} |a - a'| \quad \forall t \in [0, T], \forall x \in \mathbb{R}^d, \forall a, a' \in A.$$

Remark 2.2.5. Under Assumptions 2.2.2 and 2.2.4 we have that for all $t \in [0, T]$, $x, z, z' \in \mathbb{R}^d$ the following hold:

$$|f^{a(t, x, z)}(t, x) - f^{a(t, x, z')}(t, x)| \leq \theta |z - z'|$$

and

$$|f^{a(t,x,0)}(t,x)| \leq (K + K^2)(1 + |x|).$$

Under Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 there is an optimal control process and this fact will be used to prove the main results.

Remark 2.2.6. Due to results of Krylov [35] we know that (2.4) has a unique solution and moreover the map $[0, T] \times \mathbb{R}^d \ni (t, x) \mapsto D_x v(t, x) \in \mathbb{R}^d$ is bounded; see [35, Chapter 4, section 1, Theorem 1]. Hence, by Assumptions 2.2.3 and 2.2.4 we know that $(t, x) \mapsto a(t, x, \sigma(t, x)D_x v(t, x))$ is jointly measurable and Lipschitz in x . Thus, for each $(t, x) \in [0, T] \times \mathbb{R}^d$, the SDE

$$dX_s = b^{a(s, X_s, \sigma(s, X_s)D_x v(s, X_s))}(s, X_s) ds + \sigma(s, X_s) dW_s, \quad s \in [t, T], X_t = x$$

has a unique solution $X^{t,x}$. Then by the verification theorem, the process $\alpha_s^* := a(s, X_s, \sigma(s, X_s)D_x v(s, X_s))$ is the optimal control process for (2.3).

All the proofs will be completed in a new measure $\hat{\mathbb{P}}$ on $(\Omega, \mathcal{F})$ given in the following lemma. We will use $\hat{\mathbb{E}}$ to denote the expectation under the measure $\hat{\mathbb{P}}$.

Lemma 2.2.7. *Let Assumptions 2.2.1 and 2.2.2 together with (2.16) hold. Let $(t, x) \in [0, T] \times \mathbb{R}^d$. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (2.1) started from (t, x) and controlled by the optimal control process α^* . Then*

$$d\hat{\mathbb{P}} := \mathcal{E}((b^{\alpha^*} \sigma^{-1})(\cdot, X) \bullet W)_T d\mathbb{P}$$

is a probability measure equivalent to $\mathbb{P}$ and the process

$$\widehat{W}_s := W_s + \int_0^s b^{\alpha_r^*}(r, X_r) \sigma^{-1}(r, X_r) dr$$

is a $\hat{\mathbb{P}}$ -Wiener process.

Proof. This is an immediate consequence of (2.16) and Girsanov's theorem. $\square$

Remark 2.2.8. From Krylov [35, Chapter 4, section 1, Theorem 1] we get that there is a constant $C > 0$ such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ we have that $|D_x v(t, x)| \leq C$.

2.3 Convergence of gradient iteration algorithm

The following theorem gives the convergence result for Algorithm 2.

Theorem 2.3.1. *Let Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 hold. Let v be the solution to (2.4) and let $(v^n)_{n \in \mathbb{N}}$ be the approximation sequence given by Algorithm 2. Then there is $q \in (0, 1)$ depending only on K, θ, T and the initial guess $v^0 = v^0(t, x)$ such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ there exists $C = C(t, x)$ such that*

$$|v(t, x) - v^n(t, x)| \leq C(t, x)q^{n/2}.$$

The main idea of the proof consists of noticing that Algorithm 2 can be seen as an iteration on the level of BSDEs. Using Lemma A.1.2 we see that on the level of BSDEs this iteration is contractive. Finally we need to use known results on the connection between BSDEs and solutions to the HJB equation.

Proof of Theorem 2.3.1. We prove the main result in several steps. First, we show how to rewrite the gradient iteration algorithm as an iteration on the level of BSDEs. On the n -th step of the algorithm we need to solve the linear PDE with Lipschitz continuous coefficients (2.8). Let v^n be the solution to (2.8) and recall that

$$a^n(t, x) = \arg \max_{a \in A} \left((b^a D_x v^{n-1})(t, x) + f^a(t, x) \right) = a(t, x, \sigma(t, x) D_x v^{n-1}(t, x)).$$

Since we are working with the linear PDE with Lipschitz continuous coefficients, we have v^n in $C^{1,2}([0, T] \times \mathbb{R}^d)$. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (2.1) started from (t, x) and controlled by the optimal control process α^* ; see Remark 2.2.6. From Itô's formula we then get that

$$\begin{aligned} dv^n(s, X_s) &= \left[\partial_t v^n(s, X_s) + \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v^n)(s, X_s) + (b^{\alpha_s^*} D_x v^n)(s, X_s) \right] ds \\ &\quad + (D_x v^n \sigma)(s, X_s) dW_s \\ &= \left[(b^{\alpha_s^*} D_x v^n)(s, X_s) - (b^{a^n} D_x v^{n-1})(s, X_s) - f^{a^n}(s, X_s) \right] ds \\ &\quad + (D_x v^n \sigma)(s, X_s) dW_s. \end{aligned}$$

Let

$$F_s(z) := b^{a(s, X_s, z)}(s, X_s) \sigma^{-1}(s, X_s) z + f^{a(s, X_s, z)}(s, X_s)$$

and

$$Y_t^n := v^n(t, X_t), \quad Z_t^n := \sigma(t, X_t) D_x v^n(t, X_t), \quad \xi := g(X_T). \quad (2.17)$$

Then we may write

$$Y_t^n = \xi - \int_t^T \left[(b^{\alpha_s^*} \sigma^{-1})(s, X_s) Z_s^n - F_s(Z_s^{n-1}) \right] ds - \int_t^T Z_s^n dW_s. \quad (2.18)$$

Let $\hat{\mathbb{P}}$ and $\widehat{W}$ be given by Lemma 2.2.7. Hence (2.18) becomes

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1}) ds - \int_t^T Z_s^n d\widehat{W}_s. \quad (2.19)$$

Consider now the following BSDE:

$$\begin{aligned} Y_{t'}^{t,x} &= \xi + \int_{t'}^T b^{a(s, X_s^{t,x}, Z_s^{t,x})}(s, X_s^{t,x}) \sigma^{-1}(s, X_s^{t,x}) Z_s^{t,x} + f^{a(s, X_s^{t,x}, Z_s^{t,x})}(s, X_s^{t,x}) ds \\ &\quad - \int_{t'}^T Z_s^{t,x} d\widehat{W}_s, \quad t' \in [t, T], \end{aligned} \quad (2.20)$$

where the superscript means that the forward process started from (t, x) . Hence, we can define

$$w(t, x) := Y_t^{t,x} \quad \text{and} \quad \sigma(t, x)D_x w(t, x) := Z_t^{t,x}. \quad (2.21)$$

Therefore by (2.14) we have

$$b^{a(t,x,z)}(t, x)\sigma^{-1}(t, x)z + f^{a(t,x,z)}(t, x) = \max_{a \in A} (b^a(t, x)\sigma^{-1}(t, x)z + f^a(t, x)).$$

Thus, by Pham [1, Theorem 6.3.3], the function $w = w(t, x)$ solves the HJB equation (2.4). Notice that here is the crucial point where the fact that we use the optimal control α^* plays a role. Indeed with other control processes we couldn't claim that w solves the HJB equation. By uniqueness of the viscosity solution to the HJB equation (see the strong comparison principle from [1, Theorem 4.4.5]), we can conclude that $w = v$ and therefore w is the value function of our stochastic control problem. Therefore, the BSDE (2.20) is the BSDE corresponding to the value function. Notice that (2.20) is a quadratic BSDE, since in the generator we have a product of two Lipschitz functions which depend on Z . The existence of the solution to (2.20) under our assumptions can be obtained by applying Theorem A.1.9 in the case when the terminal cost is bounded for our stochastic control problem.

Using Remark 2.2.8, the fact that $\sigma^{-1}(s, X_s)Z_s = D_x v(s, X_s)$, and Assumption 2.2.4, for all $s \in [t, T]$ we get that

$$\begin{aligned} & |F_s(Z_s) - F_s(Z_s^{n-1})| \\ & \leq |\sigma^{-1}(s, X_s)Z_s| |b^{a(s, X_s, Z_s)}(s, X_s) - b^{a(s, X_s, Z_s^{n-1})}(s, X_s)| \\ & \quad + |b^{a(s, X_s, Z_s^{n-1})}(s, X_s)\sigma^{-1}(s, X_s)| |Z_s - Z_s^{n-1}| + \theta |Z_s - Z_s^{n-1}| \\ & \leq (C\theta + K + \theta) |Z_s - Z_s^{n-1}|. \end{aligned} \quad (2.22)$$

Moreover, recalling $\xi = g(X_T)$, we get that $\xi \in L^2(\Omega, \mathcal{F}_T, \hat{\mathbb{P}})$ from the higher moment estimates for the solution of the SDE and from the Lipschitz property of g . Similarly $F_s(0) \in \hat{\mathbb{H}}^2$ by Assumption 2.2.4. We may thus apply Lemma A.1.2 and hence, due to (2.19) and (2.22), we have $q \in (0, 1)$ and $\gamma \geq 0$ such that for all $t \in [0, T]$

$$e^{\gamma t} \hat{\mathbb{E}} |Y_t - Y_t^n|^2 + \|Z - Z^n\|_{\hat{\mathbb{H}}_\gamma^2}^2 \leq q \|Z - Z^{n-1}\|_{\hat{\mathbb{H}}_\gamma^2}^2. \quad (2.23)$$

Therefore, from (2.21) and (2.17), we have $v(t, x) = Y_t^{t,x}$ and $v^n(t, x) = Y_t^{n,t,x}$ and by (2.23) we obtain

$$\begin{aligned} e^{\gamma t} |v(t, x) - v^n(t, x)|^2 & \leq \hat{\mathbb{E}} [e^{\gamma t} |Y_t^{t,x} - Y_t^{n,t,x}|^2] + \|Z^{t,x} - Z^{n,t,x}\|_{\hat{\mathbb{H}}_\gamma^2}^2 \\ & \leq q^n \|Z^{t,x} - Z^{0,t,x}\|_{\hat{\mathbb{H}}_\gamma^2}^2. \end{aligned}$$

Hence

$$\begin{aligned} & |v(t, x) - v^n(t, x)|^2 \\ & \leq q^n \hat{\mathbb{E}} \int_t^T e^{\gamma(T-t)} |\sigma(s, X_s^{t,x,\alpha^*})|^2 |D_x v(s, X_s^{t,x,\alpha^*}) - D_x v^0(s, X_s^{t,x,\alpha^*})|^2 ds. \end{aligned}$$

This finishes the proof. $\square$

2.4 Convergence of policy improvement

Theorem 2.4.1. *Let Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 hold. Let v be the solution to (2.4) and let $(v^n)_{n \in \mathbb{N}}$ be the approximation sequence given by Algorithm 1. Then there is $q \in (0, 1)$ depending only on K, θ, T and the initial guess $v^0 = v^0(t, x)$ such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ there exists $C = C(t, x)$ such that*

$$|v(t, x) - v^n(t, x)| \leq C(t, x) q^{n/2}.$$

The proof of Theorem 2.4.1 is similar to that of Theorem 2.3.1 except that the iteration on the level of BSDEs is nonstandard.

Proof of Theorem 2.4.1. Let v^n be the solution to (2.5) and recall that

$$a^n(t, x) = \arg \max_{a \in A} ((b^a D_x v^{n-1})(t, x) + f^a(t, x)) = a(t, x, \sigma(t, x) D_x v^{n-1}(t, x)).$$

As before, let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (2.1) started from (t, x) and controlled by the optimal control process α^* ; see Remark 2.2.6. By Itô's formula

$$\begin{aligned} dv^n(s, X_s) &= \left[\partial_t v^n(s, X_s) + \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v^n)(s, X_s) + (b^{\alpha_s^*} D_x v^n)(s, X_s) \right] ds \\ &\quad + (D_x v^n \sigma)(s, X_s) dW_s \\ &= [(b^{\alpha_s^*} D_x v^n)(s, X_s) - (b^{a^n} D_x v^n)(s, X_s) - f^{a^n}(s, X_s)] ds \\ &\quad + (D_x v^n \sigma)(s, X_s) dW_s. \end{aligned}$$

Let

$$F_s(z, Z) := b^{a(s, X_s, z)}(s, X_s) \sigma^{-1}(s, X_s) Z + f^{a(s, X_s, z)}(s, X_s).$$

Recalling that the control α^* and the associated diffusion X are fixed we can write

$$Y_t^n = \xi - \int_t^T [(b^{\alpha_s^*} \sigma^{-1})(s, X_s) Z_s^n - F_s(Z_s^{n-1}, Z_s^n)] ds - \int_t^T Z_s^n dW_s. \quad (2.24)$$

Let $\hat{\mathbb{P}}$ and $\widehat{W}$ be given by Lemma 2.2.7. Then (2.24) becomes

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1}, Z_s^n) ds - \int_t^T Z_s^n d\widehat{W}_s. \quad (2.25)$$

Similarly as in Theorem 2.3.1, consider the BSDE

$$\begin{aligned} Y_{t'}^{t,x} &= \xi + \int_{t'}^T b^{a(s, X_s^{t,x}, Z_s^{t,x})}(s, X_s^{t,x}) \sigma^{-1}(s, X_s^{t,x}) Z_s^{t,x} + f^{a(s, X_s^{t,x}, Z_s^{t,x})}(s, X_s^{t,x}) ds \\ &\quad - \int_{t'}^T Z_s^{t,x} d\widehat{W}_s, \quad t' \in [t, T]. \end{aligned} \quad (2.26)$$

In same way we can show that $v(t, x) = Y_t^{t,x}$ is the value function of our stochastic control problem. As before, from Krylov [35, Chapter 4, section 1, Theorem 1], we get that there is a constant $C > 0$ such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ we have that $|D_x v(t, x)| \leq C$. Moreover, as before, using Remark 2.2.8, the fact that $\sigma^{-1}(s, X_s) Z_s = D_x v(s, X_s)$, and Assumption 2.2.4, for all $s \in [t, T]$ we get that

$$\begin{aligned} |F_s(Z_s, Z_s) - F_s(Z_s^{n-1}, Z_s^n)| &\leq \theta |\sigma^{-1}(s, X_s) Z_s| |Z_s - Z_s^{n-1}| \\ &\quad + |b^{a(s, X_s, Z_s^{n-1})}(s, X_s) \sigma^{-1}(s, X_s)| |Z_s - Z_s^n| + \theta |Z_s - Z_s^{n-1}| \\ &\leq \theta C |Z_s - Z_s^{n-1}| + K |Z_s - Z_s^n| + \theta |Z_s - Z_s^{n-1}|. \end{aligned} \quad (2.27)$$

Finally we note that $\xi \in L^2(\Omega, \mathcal{F}_T, \hat{\mathbb{P}})$ and $F_s(0, 0) \in \hat{\mathbb{H}}^2$, so by Lemma A.1.5, together with (2.25) and (2.26), we have $q \in (0, 1)$ and $\gamma \geq 0$ such that for all $t \in [0, T]$

$$e^{\gamma t} \hat{\mathbb{E}} |Y_t - Y_t^n|^2 + \|Z - Z^n\|_{\hat{\mathbb{H}}_\gamma^2}^2 \leq q \|Z - Z^{n-1}\|_{\hat{\mathbb{H}}_\gamma^2}^2. \quad (2.28)$$

Similarly as before, using (2.28), we conclude that

$$\begin{aligned} |v(t, x) - v^n(t, x)|^2 &\leq q^n \hat{\mathbb{E}} \int_t^T e^{\gamma(T-s)} |\sigma(s, X_s^{t,x,\alpha^*})|^2 |D_x v(s, X_s^{t,x,\alpha^*}) - D_x v^0(s, X_s^{t,x,\alpha^*})|^2 ds. \end{aligned}$$

This concludes the proof of the theorem. $\square$

Remark 2.4.2. Consider briefly the situation where the diffusion coefficient also depends on the control, i.e., $\sigma = \sigma^a(t, x)$. After applying Itô's formula to $v^n(s, X_s)$ and substituting the solution to the linear PDE for v^n we get

$$\begin{aligned} dv^n(s, X_s) &= [(b^{\alpha_s^*} D_x v^n)(s, X_s) + \frac{1}{2} \text{tr}(\sigma^{\alpha_s^*} (\sigma^{\alpha_s^*})^\top D_x^2 v^n)(s, X_s) \\ &\quad - (b^{a^n} D_x v^n)(s, X_s) - \frac{1}{2} \text{tr}(\sigma^{a^n} (\sigma^{a^n})^\top D_x^2 v^n)(s, X_s) - f^{a^n}(s, X_s)] ds \\ &\quad + (D_x v^n \sigma)(s, X_s) dW_s. \end{aligned}$$

The resulting object can be seen as a second order BSDE (2BSDE). Analysis of 2BSDEs goes beyond the scope of this thesis.

Remark 2.4.3. Let us briefly consider the infinite-time-horizon control problem.

In this case we consider a constant $\lambda > 0$ and the gain functional:

$$J(x, \alpha) = \mathbb{E} \left[\int_0^\infty e^{-\lambda s} f^\alpha(X_s^x) ds \right].$$

It is known that the Bellman PDE for the value function is

$$\lambda v - \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v) - \sup_{a \in A} (b^a D_x v + f^a) = 0 \text{ on } \mathbb{R}^d.$$

The linear PDE from the iteration of the policy improvement algorithm then is

$$\lambda v^n - \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v^n) - b^{a^n} D_x v^n - f^{a^n} = 0 \text{ on } \mathbb{R}^d,$$

where

$$a^n(x) = \arg \max_{a \in A} (b^a(x) D_x v^{n-1}(x) + f^a(x)).$$

After applying Itô's formula we get

$$\begin{aligned} dv^n(X_s) &= \left[\frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v^n)(X_s) + (b^{\alpha_s^*} D_x v^n)(X_s) \right] ds + (D_x v^n \sigma)(X_s) dW_s \\ &= [(b^{\alpha_s^*} D_x v^n)(X_s) - (b^{a^n} D_x v^n)(X_s) - f^{a^n}(X_s) + \lambda v^n(X_s)] ds \\ &\quad + (D_x v^n \sigma)(X_s) dW_s. \end{aligned}$$

Let $Y_t^n := v^n(X_t)$ and $Z_t^n := \sigma(X_t) D_x v^n(X_t)$. Then after change of measure we may write

$$dY_s^n = -[(b^{a(X_s, Z_s^{n-1})} \sigma^{-1})(X_s) Z_s^n + f^{a(X_s, Z_s^{n-1})}(X_s) - \lambda Y_s^n] ds + Z_s^n d\widehat{W}_s. \quad (2.29)$$

Let

$$F_s(z, Z) := (b^{a(X_s, z)} \sigma^{-1})(X_s) Z + f^{a(X_s, z)}(X_s).$$

Hence (2.29) becomes

$$dY_s^n = [-F_s(Z_s^{n-1}, Z_s^n) + \lambda Y_s^n] ds + Z_s^n d\widehat{W}_s, \quad s \in [0, \infty). \quad (2.30)$$

To proceed, we need a suitable contraction-type inequality for this infinite time horizon BSDE. Buckdahn and Peng [38] studied infinite time horizon BSDEs and have proved existence and uniqueness of their solutions for sufficiently large values of λ . To get the required contraction-type inequality we can use similar calculations as in Fuhrman and Tessitore [39, Theorems 3.2 and 3.7], where they use Banach's fixed point theorem to show existence and uniqueness of solutions to infinite time horizon BSDEs. Hence, for sufficiently large $\lambda > 0$ we would obtain results analogous to Theorem 2.4.1 as well as the other theorems in the chapter.

2.5 Policy improvement

We want to show that the policy obtained at each step of Algorithm 1 is an improvement on the one from the previous step. This is formulated as Theorem 2.5.1 below. Note that we do not require Assumption 2.2.4 here.

Theorem 2.5.1. *Let Assumptions 2.2.1, 2.2.2, and 2.2.3 hold. Assume that there exists $K \geq 0$ such that $\forall t \in [0, T]$, $\forall x \in \mathbb{R}^d$, and $\forall a \in A$*

$$|b^a \sigma^{-1}(t, x)| < K.$$

Fix $n \in \mathbb{N}$. Let v^n and v^{n+1} be the solutions of (2.5) at steps n and $n+1$ of the algorithm. Then for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that

$$v^{n+1}(t, x) \geq v^n(t, x).$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (2.1) started from (t, x) and controlled by the optimal control process α^* ; see Remark 2.2.6. Then, as in the proof of Theorem 2.4.1, we get that for $k = n, n+1$ with $Y^k = Y^{k,t,x} = v^k(\cdot, X^{t,x,\alpha^*})$ and with $Z^k = Z^{k,t,x} = (\sigma D_x v^k)(\cdot, X^{t,x,\alpha^*})$ we have the BSDE representation

$$Y_t^k = \xi + \int_t^T F_s(Z_s^{k-1}, Z_s^k) ds - \int_t^T Z_s^k d\widehat{W}_s, \quad k = n, n+1,$$

where

$$F_s(z, Z) := b^{a(s, X_s, z)}(s, X_s) \sigma^{-1}(s, X_s) Z + f^{a(s, X_s, z)}(s, X_s).$$

Let us denote for $s \in [t, T]$ and $z \in \mathbb{R}^d$

$$\phi_s^2(z) := F_s(Z_s^n, z) \quad \text{and} \quad \phi_s^1(z) := F_s(Z_s^{n-1}, z).$$

Hence, notice that by the definition of the a^{n+1} (see (2.6)), we have for all $s \in [t, T]$ that

$$\begin{aligned} \phi_s^2(Z_s^n) &= F_s(Z_s^n, Z_s^n) = b^{a^{n+1}}(s, X_s) \sigma^{-1}(s, X_s) Z_s^n + f^{a^{n+1}}(s, X_s) \\ &= \max_{a \in A} ((b^a D_x v^n)(s, X_s) + f^a(s, X_s)) \geq (b^{a^n} D_x v^n)(s, X_s) + f^{a^n}(s, X_s) \\ &= F_s(Z_s^{n-1}, Z_s^n) = \phi_s^1(Z_s^n). \end{aligned}$$

Therefore by the comparison principle for BSDEs (see Lemma A.1.6), we get

$$Y_{t'}^{n+1} \geq Y_{t'}^n \quad \forall t' \in [t, T].$$

Hence, we have

$$v^{n+1}(t, x) = Y_t^{n+1,t,x} \geq Y_t^{n,t,x} = v^n(t, x).$$

□

Remark 2.5.2. It is perhaps interesting to note that the comparison principle for BSDEs cannot be used to deduce that in the gradient iteration algorithm we have

an “improvement” at each step. Indeed, let us write the BSDE representation of the two steps of gradient iteration for $n, n+1 \in \mathbb{N}$,

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1})ds - \int_t^T Z_s^n d\widehat{W}_s$$

and

$$Y_t^{n+1} = \xi + \int_t^T F_s(Z_s^n)ds - \int_t^T Z_s^{n+1} d\widehat{W}_s,$$

where

$$F_s(z) := b^{a(s, X_s, z)}(s, X_s)\sigma^{-1}(s, X_s)z + f^{a(s, X_s, z)}(s, X_s).$$

In order to apply a comparison principle for BSDEs (see Lemma A.1.6), we would need to have $F_s(Z_s^{n-1}) \leq F_s(Z_s^n)$. Nevertheless we observe that

$$\begin{aligned} F_s(Z_s^{n-1}) &= b^{a(s, X_s, Z_s^{n-1})}\sigma^{-1}(s, X_s)Z_s^{n-1} + f^{a(s, X_s, Z_s^{n-1})} \\ &= b^{a^n}(s, X_s)\sigma^{-1}(s, X_s)Z_s^{n-1} + f^{a^n}(s, X_s) \\ &= \max_{a \in A}(b^a(s, X_s)D_x v^{n-1}(s, X_s) + f^a(s, X_s)). \end{aligned}$$

Similarly,

$$\begin{aligned} F_s(Z_s^n) &= b^{a(s, X_s, Z_s^n)}\sigma^{-1}(s, X_s)Z_s^n + f^{a(s, X_s, Z_s^n)} \\ &= b^{a^{n+1}}(s, X_s)\sigma^{-1}(s, X_s)Z_s^n + f^{a^{n+1}}(s, X_s) \\ &= \max_{a \in A}(b^a(s, X_s)D_x v^n(s, X_s) + f^a(s, X_s)). \end{aligned}$$

From the above calculations we have no way to conclude that $F_s(Z_s^{n-1}) \leq F_s(Z_s^n)$. Thus the gradient iteration algorithm is not guaranteed to be improving the policy with each step.

2.6 Stability under Perturbations to Solution of the Linear PDE

In this section we study a stability property of the policy improvement algorithm under perturbations to solutions of the linear PDE (2.5) since in practical applications one will only solve this equation approximately. Of course the maximization step (2.6) of Algorithm 1 can now be performed only with this approximate solution, thus feeding the errors into further iterations.

Let ε be a parameter (or a set of parameters), which determines the accuracy of our approximation to the solution of the linear PDE (2.5). Let π_ε^n be the policy at iteration n obtained from an approximate solution to the linear PDE. Let v_ε^n denote the solution to

$$\begin{aligned} \partial_t v_\varepsilon^n + \frac{1}{2}\text{tr}(\sigma\sigma^\top D_x^2 v_\varepsilon^n) + b^{\pi_\varepsilon^n} D_x v_\varepsilon^n + f^{\pi_\varepsilon^n} &= 0 \text{ on } [0, T) \times \mathbb{R}^d, \\ v_\varepsilon^n(T, \cdot) &= g \text{ on } x \in \mathbb{R}^d. \end{aligned} \tag{2.31}$$

At step n of Algorithm 1 we approximate the solution to the equation above (this is PDE (2.5) but with π_ε^n replacing a^n everywhere). We will denote such approximation by $\tilde{v}_\varepsilon^n$. The policy function for the next iteration step is then given by

$$\pi_\varepsilon^{n+1}(t, x) = a(t, x, (\sigma D_x \tilde{v}_\varepsilon^n)(t, x)) = \arg \max_{a \in A} [(b^a D_x \tilde{v}_\varepsilon^n)(t, x) + f^a(t, x)] ,$$

recalling that the function $a = a(t, x, z)$ was defined in (2.14). We need to assume that $(t, x) \mapsto D_x \tilde{v}_\varepsilon^n$ is bounded so that π_ε^{n+1} is Lipschitz in x so that the solution to (2.31) is $C^{1,2}([0, T] \times \mathbb{R}^d)$. This assumption is not really a restriction as we know that the gradient of the value function is bounded under our assumptions; see Krylov [35, Chapter 4, section 1, Theorem 1] and also Remark 2.2.6. Any reasonable approximation should retain this property.

Theorem 2.6.1. *Let Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 hold. Let $(v^n)_{n \in \mathbb{N}}$ be the approximation sequence given by Algorithm 1. Let $(v_\varepsilon^n)_{n \in \mathbb{N}}$ be the approximation sequence given by (2.31). Let α^* and X^{t,x,α^*} be the optimal control process for (2.3) and the associated diffusion started from $(t, x) \in [0, T] \times \mathbb{R}^d$. Assume that $D_x \tilde{v}_\varepsilon^n$ is uniformly bounded. Define*

$$E_{t,x}^k := \left\| \left[(\sigma(D_x v_\varepsilon^k - D_x \tilde{v}_\varepsilon^k))(\cdot, X^{t,x,\alpha^*}) \right] \mathcal{E}^{-1/2}((b^{\alpha^*} \sigma^{-1})(\cdot, X^{t,x,\alpha^*}) \bullet W)_T \right\|_{\mathbb{H}^2} .$$

Then there is $q \in (0, 1)$ and $\gamma > 0$, depending only on K, θ, T , such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ there exists $C = C(t, x)$ such that

$$|v^n(t, x) - v_\varepsilon^n(t, x)|^2 \leq C(t, x) q^n + 2e^{\gamma(T-t)} \sum_{k=1}^n q^k E^{n-k} .$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (2.1) started from (t, x) and controlled by the optimal control process α^* ; see Remark 2.2.6. By applying Itô's formula to v_ε^n we get

$$\begin{aligned} dv_\varepsilon^n(s, X_s) &= \left[\partial_t v_\varepsilon^n(s, X_s) + \frac{1}{2} \text{tr}(\sigma \sigma^\top D_x^2 v_\varepsilon^n)(s, X_s) + (b^{\alpha_s^*} D_x v_\varepsilon^n)(s, X_s) \right] ds \\ &\quad + (D_x v_\varepsilon^n \sigma)(s, X_s) dW_s \\ &= \left[(b^{\alpha_s^*} D_x v_\varepsilon^n)(s, X_s) - (b^{\pi_\varepsilon^n} D_x v_\varepsilon^n + f^{\pi_\varepsilon^n})(s, X_s) \right] ds \\ &\quad + (D_x v_\varepsilon^n \sigma)(s, X_s) dW_s . \end{aligned}$$

Let us denote

$$Y_{t,\varepsilon}^n := v_\varepsilon^n(t, X_t) , \quad Z_{t,\varepsilon}^n := \sigma(t, X_t) D_x v_\varepsilon^n(t, X_t) , \quad \xi := g(X_T) ,$$

$$F_s(z, Z) := b^{a(s, X_s, z)}(s, X_s) \sigma^{-1}(s, X_s) Z + f^{a(s, X_s, z)}(s, X_s)$$

and

$$\tilde{Z}_{t,\varepsilon}^{n-1} := \sigma(t, X_t) D_x \tilde{v}_\varepsilon^{n-1}(t, X_t) ,$$

where $\tilde{v}_\varepsilon^{n-1}$ is an approximate solution to corresponding PDE. Then using this notation, we may write

$$Y_{t,\varepsilon}^n = \xi - \int_t^T \left[(b^{\alpha_s^*} \sigma^{-1})(s, X_s) Z_{s,\varepsilon}^n - F_s(\tilde{Z}_{s,\varepsilon}^{n-1}, Z_{s,\varepsilon}^n) \right] ds - \int_t^T Z_{s,\varepsilon}^n dW_s.$$

Let $\hat{\mathbb{P}}$ and $\widehat{W}$ be given by Lemma 2.2.7. Then the above equation becomes

$$Y_{t,\varepsilon}^n = \xi + \int_t^T F_s(\tilde{Z}_{s,\varepsilon}^{n-1}, Z_{s,\varepsilon}^n) ds - \int_t^T Z_{s,\varepsilon}^n d\widehat{W}_s.$$

We want to study the difference of $(Y_\varepsilon^n, Z_\varepsilon^n)$ with (Y^n, Z^n) , where (Y^n, Z^n) solves the BSDE (2.25).

$$\begin{aligned} \hat{\mathbb{E}}[e^{\gamma t} |Y_t^n - Y_{t,\varepsilon}^n|^2] + \|Z^n - Z_\varepsilon^n\|_{\mathbb{H}_\gamma^2}^2 &\leq 2\hat{\mathbb{E}}[e^{\gamma t} |Y_t - Y_t^n|^2] + 2\|Z - Z^n\|_{\mathbb{H}_\gamma^2}^2 \\ &\quad + 2\hat{\mathbb{E}}[e^{\gamma t} |Y_t - Y_{t,\varepsilon}^n|^2] + 2\|Z - Z_\varepsilon^n\|_{\mathbb{H}_\gamma^2}^2, \end{aligned} \quad (2.32)$$

where (Y, Z) solves the BSDE (2.26). Due to (2.27) and

$$|F_s(Z_s, Z_s) - F_s(\tilde{Z}_{s,\varepsilon}^{n-1}, Z_{s,\varepsilon}^n)| \leq (C\theta + \theta)|Z_s - \tilde{Z}_{s,\varepsilon}^{n-1}| + K|Z_s - Z_{s,\varepsilon}^n|,$$

we can apply Lemma A.1.5. Hence, there is $\tilde{q} \in (0, 1/2)$ and $\gamma > 0$ such that

$$\hat{\mathbb{E}}[e^{\gamma t} |Y_t - Y_t^n|^2] + \|Z - Z^n\|_{\mathbb{H}_\gamma^2}^2 \leq \tilde{q} \|Z - Z^{n-1}\|_{\mathbb{H}_\gamma^2}^2$$

and

$$\hat{\mathbb{E}}[e^{\gamma t} |Y_t - Y_{t,\varepsilon}^n|^2] + \|Z - Z_\varepsilon^n\|_{\mathbb{H}_\gamma^2}^2 \leq \tilde{q} \|Z - \tilde{Z}_\varepsilon^{n-1}\|_{\mathbb{H}_\gamma^2}^2.$$

Therefore we continue the estimate (2.32)

$$\begin{aligned} &\hat{\mathbb{E}}[e^{\gamma t} |Y_t^n - Y_{t,\varepsilon}^n|^2] + \|Z^n - Z_\varepsilon^n\|_{\mathbb{H}_\gamma^2}^2 \\ &\leq 2\hat{\mathbb{E}}[e^{\gamma t} |Y_t - Y_t^n|^2] + 2\|Z - Z^n\|_{\mathbb{H}_\gamma^2}^2 + 2\hat{\mathbb{E}}[e^{\gamma t} |Y_t - Y_{t,\varepsilon}^n|^2] + 2\|Z - Z_\varepsilon^n\|_{\mathbb{H}_\gamma^2}^2 \\ &\leq 2\tilde{q} \|Z - Z^{n-1}\|_{\mathbb{H}_\gamma^2}^2 + 2\tilde{q} \|Z - \tilde{Z}_\varepsilon^{n-1}\|_{\mathbb{H}_\gamma^2}^2 \\ &\leq 2\tilde{q} \|Z - Z^{n-1}\|_{\mathbb{H}_\gamma^2}^2 + 4\tilde{q} \|Z - Z_\varepsilon^{n-1}\|_{\mathbb{H}_\gamma^2}^2 + 4\tilde{q} \|Z_\varepsilon^{n-1} - \tilde{Z}_\varepsilon^{n-1}\|_{\mathbb{H}_\gamma^2}^2 \\ &\leq 2\tilde{q}^2 \|Z - Z^{n-2}\|_{\mathbb{H}_\gamma^2}^2 + 4\tilde{q}^2 \|Z - \tilde{Z}_\varepsilon^{n-2}\|_{\mathbb{H}_\gamma^2}^2 + 4\tilde{q} \|Z_\varepsilon^{n-1} - \tilde{Z}_\varepsilon^{n-1}\|_{\mathbb{H}_\gamma^2}^2 \\ &\leq 2\tilde{q}^2 \|Z - Z^{n-2}\|_{\mathbb{H}_\gamma^2}^2 + 8\tilde{q}^2 \|Z - Z_\varepsilon^{n-2}\|_{\mathbb{H}_\gamma^2}^2 + 8\tilde{q}^2 \|Z_\varepsilon^{n-2} - \tilde{Z}_\varepsilon^{n-2}\|_{\mathbb{H}_\gamma^2}^2 \\ &\quad + 4\tilde{q} \|Z_\varepsilon^{n-1} - \tilde{Z}_\varepsilon^{n-1}\|_{\mathbb{H}_\gamma^2}^2 \leq \dots \\ &\leq 2\tilde{q}^n \|Z - Z^0\|_{\mathbb{H}_\gamma^2}^2 + 2^{n+1}\tilde{q}^n \|Z - Z_\varepsilon^0\|_{\mathbb{H}_\gamma^2}^2 + \sum_{k=1}^n 2^{k+1}\tilde{q}^k \|Z_\varepsilon^{n-k} - \tilde{Z}_\varepsilon^{n-k}\|_{\mathbb{H}_\gamma^2}^2. \end{aligned}$$

This concludes the proof of the theorem. $\square$

2.7 Stability under Perturbation of the Maximization

In this section we study a stability property of the gradient iteration algorithm under perturbations to maximization procedure (2.7). Let $\bar{v}^n$ be the solution to corresponding PDE at iteration n of the gradient iteration algorithm, where instead of obtaining the control function corresponding to the exact maximum

$$a^n(t, x) = a(t, x, (\sigma D_x \bar{v}^n)(t, x)) = \arg \max_{a \in A} ((b^a D_x \bar{v}^n + f^a)(t, x))$$

we only solve this maximization problem approximately and so we are dealing with a control function of the form

$$\bar{a}(t, x, (\sigma D_x \bar{v}^n)(t, x)) := a(t, x, (\sigma D_x \bar{v}^n)(t, x)) + \varepsilon(t, x, (\sigma D_x \bar{v}^n)(t, x)),$$

where the function $\varepsilon = \varepsilon(t, x, z)$ determines the accuracy of our approximation.

Theorem 2.7.1. *Let Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 hold. Let $(v^n)_{n \in \mathbb{N}}$ be the approximation sequence given by Algorithm 2. Let $(\bar{v}^n)_{n \in \mathbb{N}}$ be the approximation sequence given by the perturbations to the maximization procedure and assume that $v^0 = \bar{v}^0$. Let α^* and X^{t,x,α^*} be the optimal control process for (2.3) and the associated diffusion started from $(t, x) \in [0, T] \times \mathbb{R}^d$. Define*

$$E_{t,x}^{k+1} := \left\| \left[1 + |D_x \bar{v}^k(\cdot, X^{t,x,\alpha^*})| \right] \mathcal{E}^{-1/2}((b^{\alpha^*} \sigma^{-1})(\cdot, X^{t,x,\alpha^*}) \bullet W)_T \right\|_{\mathbb{H}^2}^2,$$

$$\varepsilon^{k+1} = \sup_{(s,y) \in [t,T] \times \mathbb{R}^d} |\varepsilon(s, y, (\sigma D_x \bar{v}^k)(s, y))|^2.$$

Then there is $q \in (0, 1)$ and $\gamma > 0$, depending only on K, θ, T , such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ there exists $C = C(t, x)$ such that

$$|\bar{v}^n(t, x) - v^n(t, x)|^2 \leq C(t, x) q^n + e^{\gamma(T-t)} \frac{2\theta}{C\theta + K + \theta} \sum_{k=1}^n q^{n-k+1} \varepsilon^k E_{t,x}^k.$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (2.1) started from (t, x) and controlled by the optimal control process α^* ; see Remark 2.2.6. As in the proof of Theorem 2.3.1 we can write two BSDEs we get after the change of measure given by Lemma 2.2.7. The first BSDE arises from the perturbations of the maximization:

$$\bar{Y}_t^n = \xi + \int_t^T \bar{F}_s(\bar{Z}_s^{n-1}) ds - \int_t^T \bar{Z}_s^n d\widehat{W}_s,$$

where

$$\bar{F}_s(z) = b^{\bar{a}(s, X_s, z)} \sigma^{-1}(s, X_s) z + f^{\bar{a}(s, X_s, z)}(s, X_s).$$

The second BSDE arises from the gradient iteration algorithm with the maximization performed exactly:

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1}) ds - \int_t^T Z_s^n d\widehat{W}_s,$$

where

$$F_s(z) = b^{a(s, X_s, z)} \sigma^{-1}(s, X_s) z + f^{a(s, X_s, z)}(s, X_s).$$

We want to study the difference of $(\bar{Y}^n, \bar{Z}^n)$ with (Y^n, Z^n) . Hence, notice that

$$\begin{aligned} e^{\gamma t} \hat{\mathbb{E}} |\bar{Y}_t^n - Y_t^n|^2 + \|\bar{Z}^n - Z^n\|_{\mathbb{H}_\gamma^2}^2 &\leq 2e^{\gamma t} \hat{\mathbb{E}} |Y_t - Y_t^n|^2 + 2\|Z - Z^n\|_{\mathbb{H}_\gamma^2}^2 \\ &\quad + 2e^{\gamma t} \hat{\mathbb{E}} |Y_t - \bar{Y}_t^n|^2 + 2\|Z - \bar{Z}^n\|_{\mathbb{H}_\gamma^2}^2, \end{aligned} \quad (2.33)$$

where (Y, Z) solves (2.20). Therefore, since

$$|F_s(Z_s) - F_s(\bar{Z}_s^{n-1})| \leq (C\theta + K + \theta) |Z_s - \bar{Z}_s^{n-1}|,$$

we can apply Lemma A.1.7 so that there is $q \in (0, 1)$ and $\gamma > 0$ such that

$$\begin{aligned} e^{\gamma t} \hat{\mathbb{E}} |Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|_{\mathbb{H}_\gamma^2}^2 &\leq q \|Z - \bar{Z}^{n-1}\|_{\mathbb{H}_\gamma^2}^2 \\ &\quad + \frac{q}{C\theta + K + \theta} \|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|_{\mathbb{H}_\gamma^2}^2. \end{aligned} \quad (2.34)$$

Now we need to estimate the second term of the right-hand side (RHS). Notice that by Assumption 2.2.4 the following holds:

$$\begin{aligned} &|\bar{F}_s(\bar{Z}_s^{n-1}) - F_s(\bar{Z}_s^{n-1})| \\ &\leq |\sigma^{-1}(s, X_s) \bar{Z}_s^{n-1}| |b^{\bar{a}(s, X_s, \bar{Z}_s^{n-1})}(s, X_s) - b^{a(s, X_s, \bar{Z}_s^{n-1})}(s, X_s)| \\ &\quad + |f^{\bar{a}(s, X_s, \bar{Z}_s^{n-1})}(s, X_s) - f^{a(s, X_s, \bar{Z}_s^{n-1})}(s, X_s)| \\ &\leq \sqrt{\theta} |\sigma^{-1}(s, X_s) \bar{Z}_s^{n-1}| |\varepsilon(s, X_s, \bar{Z}_s^{n-1})| + \sqrt{\theta} |\varepsilon(s, X_s, \bar{Z}_s^{n-1})|. \end{aligned} \quad (2.35)$$

Hence by (2.35) we have

$$\|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|_{\mathbb{H}_\gamma^2}^2 \leq \theta \|(1 + |\sigma^{-1}(\cdot, X) \bar{Z}^{n-1}|) \varepsilon(\cdot, X, \bar{Z}^{n-1})\|_{\mathbb{H}_\gamma^2}^2. \quad (2.36)$$

By inequalities (2.33), (2.34), (2.36), and the result of Theorem 2.3.1 and since $\bar{Y}_t^{t,x,n} = \bar{v}^n(t, x)$, $Y_t^{t,x,n} = v^n(t, x)$ as well as $Z^0 = \bar{Z}^0$, we conclude that

$$\begin{aligned} &e^{\gamma t} |\bar{v}^n(t, x) - v^n(t, x)|^2 \\ &\leq 2e^{\gamma t} \hat{\mathbb{E}} |Y_t - Y_t^n|^2 + \|Z - Z^n\|_{\mathbb{H}_\gamma^2}^2 + 2e^{\gamma t} \hat{\mathbb{E}} |Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|_{\mathbb{H}_\gamma^2}^2 \\ &\leq 2q \|Z - Z^{n-1}\|_{\mathbb{H}_\gamma^2}^2 + 2q \|Z - \bar{Z}^{n-1}\|_{\mathbb{H}_\gamma^2}^2 + \frac{2q}{C\theta + K + \theta} \|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|_{\mathbb{H}_\gamma^2}^2 \\ &\leq 4q^n \|Z - Z^0\|_{\mathbb{H}_\gamma^2}^2 + \sum_{k=1}^n q^k \frac{2\theta}{C\theta + K + \theta} \|(1 + |\sigma^{-1}(\cdot, X) \bar{Z}^{n-k}|) \varepsilon(\cdot, X, \bar{Z}^{n-k})\|_{\mathbb{H}_\gamma^2}^2. \end{aligned}$$

□

We obtain the same result for the policy improvement algorithm.

Theorem 2.7.2. *Let Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 hold. Let $(v^n)_{n \in \mathbb{N}}$ be the approximation sequence given by Algorithm 1. Let $(\bar{v}^n)_{n \in \mathbb{N}}$ be the approximation sequence given by the perturbations to the maximization procedure. Let α^**

and X^{t,x,α^*} be the optimal control process for (2.3) and the associated diffusion started from $(t, x) \in [0, T] \times \mathbb{R}^d$. Define

$$E_{t,x}^k := \left\| [1 + D_x \bar{v}^k(\cdot, X^{t,x,\alpha^*})] \mathcal{E}^{-1/2}((b^{\alpha^*} \sigma^{-1})(\cdot, X^{t,x,\alpha^*}) \bullet W)_T \right\|_{\mathbb{H}^2}^2,$$

$$\varepsilon^{k+1} = \sup_{(s,y) \in [t,T] \times \mathbb{R}^d} |\varepsilon(s, y, (\sigma D_x \bar{v}^k)(s, y))|^2.$$

Then there is $q \in (0, 1)$ and $\gamma > 0$, depending only on K, θ, T , such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ there is $C = C(t, x)$ such that

$$|\bar{v}^n(t, x) - v^n(t, x)|^2 \leq C(t, x) q^n + \frac{2\theta}{\max(C\theta + \theta, K)} e^{\gamma(T-t)} \sum_{k=1}^n q^k \varepsilon^{n-k+1} E_{t,x}^{n-k+1}.$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (2.1) started from (t, x) and controlled by the optimal control process α^* ; see Remark 2.2.6. Due to Theorem 2.4.1 we can write two BSDEs we get after the change of measure: first from the perturbation and second from the gradient iteration

$$\bar{Y}_t^n = \xi + \int_t^T \bar{F}_s(\bar{Z}_s^{n-1}, \bar{Z}_s^n) ds - \int_t^T \bar{Z}_s^n d\widehat{W}_s,$$

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1}, Z_s^n) ds - \int_t^T Z_s^n d\widehat{W}_s,$$

where

$$\bar{F}_s(z, Z) := b^{\bar{a}(s, X_s, z)}(s, X_s) \sigma^{-1}(s, X_s) Z + f^{\bar{a}(s, X_s, z)}(s, X_s),$$

$$F_s(z, Z) := b^{a(s, X_s, z)}(s, X_s) \sigma^{-1}(s, X_s) Z + f^{a(s, X_s, z)}(s, X_s).$$

Similarly, we want to study the difference of $(\bar{Y}^n, \bar{Z}^n)$ with (Y^n, Z^n) . Hence, notice that

$$e^{\gamma t} \mathbb{E} |\bar{Y}_t^n - Y_t^n|^2 + \|\bar{Z}^n - Z^n\|_{\hat{\mathbb{H}}_\gamma^2}^2 \leq 2e^{\gamma t} \mathbb{E} |Y_t - Y_t^n|^2 + \|Z - Z^n\|_{\hat{\mathbb{H}}_\gamma^2}^2$$

$$+ 2e^{\gamma t} \mathbb{E} |Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|_{\hat{\mathbb{H}}_\gamma^2}^2, \quad (2.37)$$

where (Y, Z) solves (2.4.1). Therefore, since

$$|F_s(Z_s, Z_s) - F_s(\bar{Z}_s^{n-1}, \bar{Z}_s^n)| \leq \theta C |Z_s - \bar{Z}_s^{n-1}| + K |Z_s - \bar{Z}_s^n| + \theta |Z_s - \bar{Z}_s^{n-1}|,$$

we can apply Lemma A.1.8 so that there is $q \in (0, 1)$ and $\gamma > 0$ such that

$$e^{\gamma t} \mathbb{E} |Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|_{\hat{\mathbb{H}}_\gamma^2}^2 \leq q \|Z - \bar{Z}^{n-1}\|_{\hat{\mathbb{H}}_\gamma^2}^2$$

$$+ \frac{q}{\max(C\theta + \theta, K)} \|\bar{F}(\bar{Z}^{n-1}, \bar{Z}^n) - F(\bar{Z}^{n-1}, \bar{Z}^n)\|_{\hat{\mathbb{H}}_\gamma^2}^2. \quad (2.38)$$

Now we need to estimate the second term of the RHS. Notice that by Assumption

2.2.4 we have that

$$\begin{aligned}
& |\bar{F}_s(\bar{Z}_s^{n-1}, \bar{Z}_s^n) - F_s(\bar{Z}_s^{n-1}, \bar{Z}_s^n)| \\
& \leq \left| b^{\bar{a}(s, X_s, \bar{Z}_s^{n-1})}(s, X_s) \sigma^{-1}(s, X_s) \bar{Z}_s^n - b^{a(s, X_s, \bar{Z}_s^{n-1})}(s, X_s) \sigma^{-1}(s, X_s) \bar{Z}_s^n \right| \\
& \quad + \left| f^{\bar{a}(s, X_s, \bar{Z}_s^{n-1})}(s, X_s) - f^{a(s, X_s, \bar{Z}_s^{n-1})}(s, X_s) \right| \\
& \leq \sqrt{\theta} |\sigma^{-1}(s, X_s) \bar{Z}_s^n| |\varepsilon(s, X_s, \bar{Z}_s^{n-1})| + \sqrt{\theta} |\varepsilon(s, X_s, \bar{Z}_s^{n-1})|.
\end{aligned} \tag{2.39}$$

Hence by (2.39) we have

$$\|\bar{F}(\bar{Z}^{n-1}, \bar{Z}^n) - F(\bar{Z}^{n-1}, \bar{Z}^n)\|_{\mathbb{H}_\gamma^2}^2 \leq \theta \|(1 + |\sigma^{-1}(s, X) \bar{Z}^n|) \varepsilon(\cdot, X, \bar{Z}^{n-1})\|_{\mathbb{H}_\gamma^2}^2. \tag{2.40}$$

By inequalities (2.37), (2.38), (2.40), by the result of Theorem 2.4.1, and by $\bar{Y}_t^{t,x,n} = \bar{v}^n(t, x)$, $Y_t^{t,x,n} = v^n(t, x)$ we conclude that

$$\begin{aligned}
& e^{\gamma t} |\bar{v}^n(t, x) - v^n(t, x)|^2 \\
& \leq 2e^{\gamma t} \hat{\mathbb{E}} |Y_t - Y_t^n|^2 + \|Z - Z^n\|_{\mathbb{H}_\gamma^2}^2 + 2e^{\gamma t} \hat{\mathbb{E}} |Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|_{\mathbb{H}_\gamma^2}^2 \\
& \leq 2q \|Z - Z^{n-1}\|_{\mathbb{H}_\gamma^2}^2 + 2q \|Z - \bar{Z}^{n-1}\|_{\mathbb{H}_\gamma^2}^2 \\
& \quad + \frac{2q}{\max(C\theta + \theta, K)} \|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|_{\mathbb{H}_\gamma^2}^2 \\
& \leq 2q^n \|Z - Z^0\|_{\mathbb{H}_\gamma^2}^2 + 2q^n \|Z - \bar{Z}^0\|_{\mathbb{H}_\gamma^2}^2 \\
& \quad + \sum_{k=1}^n q^k \frac{2\theta}{\max(C\theta + \theta, K)} \|(1 + |\sigma^{-1}(\cdot, X) \bar{Z}^{n-k+1}|) \varepsilon(\cdot, X, \bar{Z}^{n-k})\|_{\mathbb{H}_\gamma^2}^2.
\end{aligned}$$

□

2.8 Example

In this section we would like to consider an example when Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 hold. Let $t \mapsto s(t)$ and $t \mapsto k(t)$ be continuous functions for $t \in [0, T]$. Consider the state which is governed by the controlled SDE

$$dX_t = s(t) \sin \alpha_t dt + \sqrt{2} dW_t, \quad t \in [0, T],$$

and consider the cost functional

$$J(t, x, \alpha) = \mathbb{E} \left[\int_t^T k(s) \cos \alpha_s ds + g(X_T) \right].$$

The aim is to maximize J over admissible controls $\alpha \in \mathcal{A}$. The value function $v = \sup_{\alpha \in \mathcal{A}} J(t, x, \alpha)$ satisfies the Bellman PDE

$$\partial_t v + D_x^2 v + \sup_{a \in A} [s(t) \sin a D_x v + k(t) \cos a] = 0, \quad \text{on } [0, T] \times \mathbb{R},$$

with the terminal condition $v(T, x) = g(x) := \arctan(x)$. Hence, the optimal control is

$$a(t, x) = \arctan\left(\frac{s(t)D_x v}{k(t)}\right).$$

It is easy to check that Assumptions 2.2.1, 2.2.2, 2.2.3, and 2.2.4 hold for this problem. Therefore, the Bellman PDE becomes

$$\partial_t v + D_x^2 v + \frac{\frac{(s(t)D_x v)^2}{k(t)}}{\sqrt{1 + \left(\frac{s(t)D_x v}{k(t)}\right)^2}} + \frac{k(t)}{\sqrt{1 + \left(\frac{s(t)D_x v}{k(t)}\right)^2}} = 0. \quad (2.41)$$

We can solve this problem using the policy improvement algorithm by approximating the Bellman PDE with a sequence of linear PDEs:

Step 1. Make an initial choice of control $a^0(t, x)$.

Step 2. For $n = 0, 1, \dots$:

- Evaluation step: Find a solution $v^n = v^n(t, x)$ to the linear PDE

$$\partial_t v^n + D_x^2 v^n + s(t) \sin a^n D_x v^n + \cos a^n = 0. \quad (2.42)$$

- Improvement step: Find a new policy $a^{n+1} = a^{n+1}(t, x)$ such that

$$a^{n+1}(t, x) = \arctan\left(\frac{s(t)D_x v^n}{k(t)}\right).$$

Step 3. Iterate the process until no changes occur in the controls updates.

One can do similar calculations in the case of the gradient iteration algorithm.

We will solve (2.41) and (2.42) by the finite difference method. For simplicity, let us choose $s(t) = 1$ and $k(t) = 1$ for all $t \in [0, T]$. In Figure 2.8.1, one can see the logarithm of the error between the value function obtained by the iterative methods, by the policy improvement algorithm, and by the gradient iteration algorithm at every step and the value function obtained by the solution of the Bellman PDE. This shows the fast convergence of the policy improvement method for our example in one dimension. In Figure 2.8.2, we can see that after only a few steps the policies obtained from the policy improvement algorithm are close to the exact one. Finally, in Figure 2.8.3, we plot the value function and the policy from the solution of the Bellman PDE.

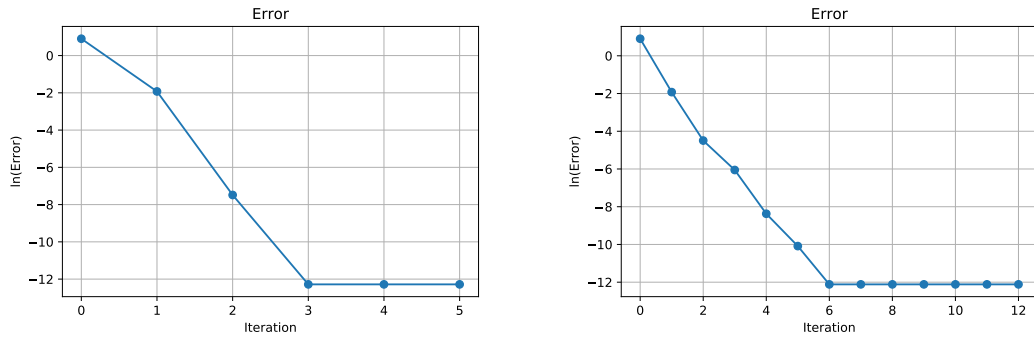


Figure 2.8.1: Plot for the logarithm of the error between the value function obtained by the iterative methods, by the policy improvement, and by the gradient iteration algorithms (from left to right) at every step and the value function obtained by the solution of the Bellman PDE. Note that the convergence stops once we have reached the accuracy of the finite-difference solver.

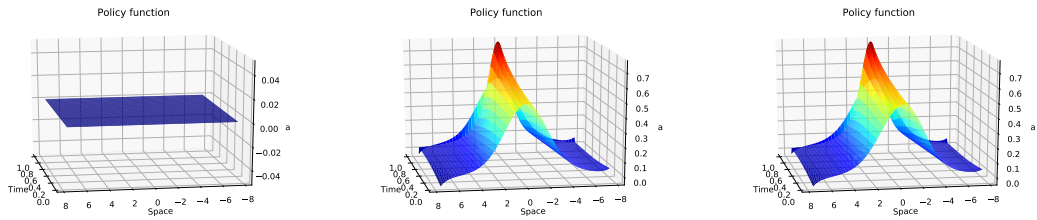


Figure 2.8.2: Plot of the initial policy and policies obtained by the policy improvement algorithm at Steps 1 and 5

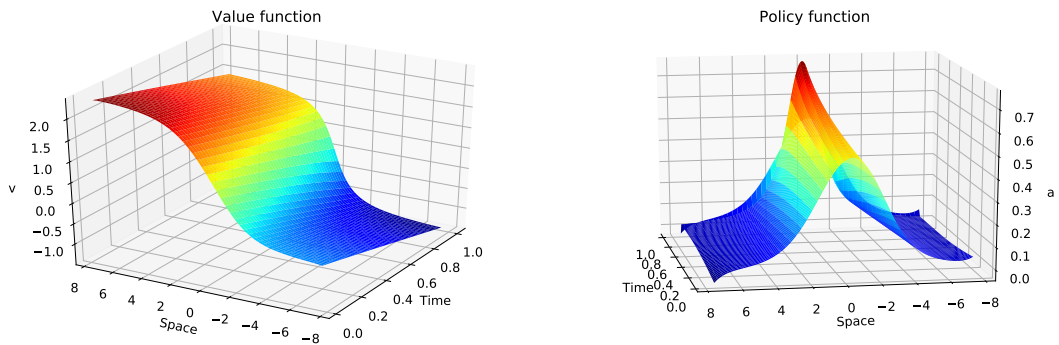


Figure 2.8.3: Plot of the true value function and the policy

Chapter 3

A modified MSA for stochastic control problems

3.1 Introduction

In this chapter, we consider a modification to the method of successive approximations (MSA), see Algorithm 3. The MSA is essentially a way of applying the Pontryagin's optimality principle to get numerical solutions of stochastic control problems.

We will consider the continuous space, continuous time problem where the controlled system is modelled by an $\mathbb{R}^d$ -valued diffusion process. Let W be a d' -dimensional Wiener martingale on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$. We will provide exact assumptions we need in Section 3.2. For now, let us fix a finite time $T \in (0, \infty)$ and consider the controlled stochastic differential equation (SDE) for given measurable functions $b : [0, T] \times \mathbb{R}^d \times A \rightarrow \mathbb{R}^d$ and $\sigma : [0, T] \times \mathbb{R}^d \times A \rightarrow \mathbb{R}^{d \times d'}$

$$dX_s = b(s, X_s, \alpha_s) ds + \sigma(s, X_s, \alpha_s) dW_s, \quad s \in [0, T], \quad X_0 = x. \quad (3.1)$$

Here $\alpha = (\alpha_s)_{s \in [0, T]}$ is a control process belonging to the space of admissible controls $\mathcal{A}$, valued in a separable metric space A and we will write X^α to denote the unique solution of (3.1) which starts from x at time 0 whilst being controlled by α . Furthermore let $f : [0, T] \times \mathbb{R}^d \times A \rightarrow \mathbb{R}$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be given measurable functions and consider the gain functional

$$J(x, \alpha) := \mathbb{E} \left[\int_0^T f(s, X_s^\alpha, \alpha_s) ds + g(X_T^\alpha) \right] \quad (3.2)$$

for all $x \in \mathbb{R}^d$ and $\alpha \in \mathcal{A}$. We want to solve the optimisation problem i.e. to find the optimal control α^* which achieves the minimum of (3.2) (or, if the infimum cannot be reached by $\alpha \in \mathcal{A}$ then an ε -optimal control $\alpha^\varepsilon \in \mathcal{A}$ such that $\inf_{\alpha \in \mathcal{A}} J(x, \alpha) \leq J(x, \alpha^\varepsilon) + \varepsilon$).

In the present chapter, we study an approach based on Pontryagin's optimality principle, see e.g. [40], [41] or [16]. The main idea is to consider optimality conditions for controls of the problem (3.2). Given b, σ and f we define the

Hamiltonian $\mathcal{H} : [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d \times d'} \times A \rightarrow \mathbb{R}$ as

$$\mathcal{H}(t, x, y, z, a) = b(t, x, a) \cdot y + \text{tr}(\sigma^\top(t, x, a)z) + f(t, x, a). \quad (3.3)$$

Consider for each $\alpha \in \mathcal{A}$, the BSDE, called the adjoint equation

$$dY_s^\alpha = -D_x \mathcal{H}(s, X_s^\alpha, Y_s^\alpha, Z_s^\alpha, \alpha_s) ds + Z_s^\alpha dW_s, \quad Y_T^\alpha = D_x g(X_T^\alpha), \quad s \in [0, T]. \quad (3.4)$$

It is well known from Pontryagin's optimality principle that, if an admissible control $\alpha^* \in \mathcal{A}$ is optimal, X^{α^*} is the corresponding optimally controlled dynamic (3.1) and $(Y^{\alpha^*}, Z^{\alpha^*})$ is the solution to the associated adjoint equation (3.4), then $\forall a \in A$ and $\forall s \in [0, T]$ the following holds

$$\mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*) \leq \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, a) \quad \text{a.s.} \quad (3.5)$$

We now define the augmented Hamiltonian $\tilde{\mathcal{H}} : [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d \times d'} \times A \times A \rightarrow \mathbb{R}$ for some $\rho \geq 0$ by

$$\begin{aligned} \tilde{\mathcal{H}}(t, x, y, z, a', a) &:= \mathcal{H}(t, x, y, z, a) + \frac{1}{2}\rho |b(t, x, a) - b(t, x, a')|^2 \\ &\quad + \frac{1}{2}\rho |\sigma(t, x, a) - \sigma(t, x, a')|^2 + \frac{1}{2}\rho |D_x \mathcal{H}(t, x, y, z, a) - D_x \mathcal{H}(t, x, y, z, a')|^2. \end{aligned} \quad (3.6)$$

Notice that when $\rho = 0$ we have exactly the definition of Hamiltonian (3.3). Given the augmented Hamiltonian, let us introduce the modified MSA in Algorithm 3 which consists of successive integrations of the state and adjoint systems and updates to the control. Notice that the backward SDE depends on the Hamiltonian $\mathcal{H}$, while the control update step comes from minimizing the augmented Hamiltonian $\tilde{\mathcal{H}}$.

Algorithm 3 Modified Method of Successive Approximations:

Initialisation: make a guess of the control $\alpha^0 = (\alpha_s^0)_{s \in [0, T]}$.

while difference between $J(x, \alpha^n)$ and $J(x, \alpha^{n-1})$ is large **do**

 Given a control $\alpha^{n-1} = (\alpha_s^{n-1})_{s \in [0, T]}$ solve the following forward SDE, then solve backward SDE:

$$\begin{aligned} dX_s^n &= b(s, X_s^n, \alpha_s^{n-1}) ds + \sigma(s, X_s^n, \alpha_s^{n-1}) dW_s, \quad X_0^n = x, \\ dY_s^n &= -D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}) ds + Z_s^n dW_s, \quad Y_T^n = D_x g(X_T^n). \end{aligned} \quad (3.7)$$

 Update the control

$$\alpha_s^n \in \arg \min_{a \in A} \tilde{\mathcal{H}}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}, a), \quad \forall s \in [0, T]. \quad (3.8)$$

end while

return α^n .

The main motivation to explore the modified MSA for stochastic control prob-

lems is to obtain convergence, ideally with rate, of an iterative algorithm, applicable to problems with the control in the diffusion part of the controlled dynamics. This is in contrast to [42] where convergence rate of an the Bellman–Howard policy iteration is shown but only for control problems with no control in the diffusion part of the controlled dynamics.

In Lemma 3.2.3, which can be established using careful BSDE estimates, we can see the estimate on the change of J when we do a minimization step of Hamiltonian as in (3.8). If the sum of the last three terms of (3.12) is bigger than the first term, then for classical MSA algorithm (i.e. case $\rho = 0$) we cannot guarantee that we do an update of the control in optimal descent direction of J . That means that the method of successive approximations may diverge. To overcome this, we need to modify the algorithm in such way so that we ensure convergence. With this in mind the desirability of the the augmented Hamiltonian (3.6) for updating the control becomes clear, as long as it still characterises optimal controls like $\mathcal{H}$ does. Theorem 3.2.4 answers this question affirmatively which opens the way to the modified MSA. In Theorem 3.2.5 we show that the modified method of successive approximations, converges for arbitrary T , and in Corollary 3.2.6, we show logarithmic convergence rate for certain stochastic control problems.

We observe that the forward and backward dynamics in (3.7) are decoupled, due to the iteration used. Therefore, it can be efficiently approximated, even in high dimension, using deep learning methods, see [43] and [44]. However, the minimization step (3.8) might be computationally expensive for some problems. A possible approach circumventing this is to replace the full minimization of (3.8) by gradient descent. A continuous version of this gradient flow is analyzed in [45].

The main contributions are the probabilistic proof of convergence of the modified method of successive approximations and establishing convergence rate for a specific type of optimal control problems.

This chapter is organised as follow: In Section 3.2 we state the assumptions and main results. In Section 3.3 we collect all proofs. Finally, in Appendix B.1 we recall an auxiliary lemma which is needed in the proof of Corollary 3.2.6.

3.2 Main results

We fix a finite horizon $T \in (0, \infty)$. Let A be a separable metric space. This is the space where the control processes α take values. We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P} = (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$. Let $W = (W_t)_{t \in [0, T]}$ be a d' -dimensional Wiener martingale on this space. By $\mathbb{E}_t$ we denote the conditional expectation with respect to $\mathcal{F}_t$. Let $|\cdot|$ denote any norm in a finite dimensional Euclidean space. By $\|\cdot\|_{L^\infty}$ we denote the norm in $L^\infty(\Omega)$. Let $\|Z\|_{\mathbb{H}^\infty} := \text{ess sup}_{(t, \omega)} |Z_t(\omega)|$ for any predictable process Z . We understand the following as $D_x \sigma = D_{x_l} \sigma^{ij}$, $D_x^2 b = D_{x_l x_n}^2 b^i$ and $D_x^2 \sigma = D_{x_l x_n}^2 \sigma^{ij}$, where $i, l, n = 1, 2, \dots, d$ and $j = 1, 2, \dots, d'$. By $Z^\top$ we denote the transpose of Z . The state of the system is governed by the controlled SDE (3.1). The corresponding adjoint equation satisfies (3.4).

Assumption 3.2.1. The functions b and σ are jointly continuous in t and twice

differentiable in x . There exists $K \geq 0$ such that $\forall x \in \mathbb{R}^d, \forall a \in A, \forall t \in [0, T]$,

$$|D_x b(t, x, a)| + |D_x \sigma(t, x, a)| + |D_x^2 b(t, x, a)| \leq K. \quad (3.9)$$

Moreover, assume that $D_x^2 \sigma(t, x, a) = 0 \forall x \in \mathbb{R}^d, \forall a \in A, \forall t \in [0, T]$.

Clearly the assumption (3.10) implies that $\forall x, x' \in \mathbb{R}^d, \forall a \in A, \forall t \in [0, T]$ we have

$$|b(t, x, a) - b(t, x', a)| + |\sigma(t, x, a) - \sigma(t, x', a)| \leq K|x - x'|. \quad (3.10)$$

The assumption that $D_x^2 \sigma(t, x, a) = 0 \forall x \in \mathbb{R}^d, \forall a \in A, \forall t \in [0, T]$ is needed so that (3.19), in the proof of Lemma 3.3.1, holds. Without this assumption (3.19) would only hold if we could show that $\|Z^\alpha\|_{\mathbb{H}^\infty} < \infty$. Without additional regularization of the control problem this is impossible. Indeed, with [46, Proposition 5.3] we see that Z_t^α is a version of $D_t Y_t^\alpha$ (the Malliavin derivative of Y_t^α) and $D_t Y_t^\alpha$ itself satisfies an a linear BSDE. However, to obtain the estimates using this representation, one term that arises is $D_t \alpha_s$ where $t \in [0, T]$ and $s \in [t, T]$. So we would need $\text{ess sup}_{\omega \in \Omega, t \in (0, T), s \in (t, T)} |D_t \alpha_s(\omega)| < \infty$. This is not necessarily the case here.

Assumption 3.2.2. The functions f is jointly continuous in t , and f and σ are twice differentiable in x . There is a constant $K \geq 0$ such that $\forall x, \forall a \in A, \forall t \in [0, T]$

$$|D_x g(x)| + |D_x f(t, x, a)| + |D_x^2 g(x)| + |D_x^2 f(t, x, a)| \leq K. \quad (3.11)$$

Under these assumptions, we can obtain the following estimate.

Lemma 3.2.3. *Let Assumption 3.2.1 and 3.2.2 hold. Then for any admissible controls φ and θ there exists a constant $C > 0$ such that*

$$\begin{aligned} J(x, \varphi) - J(x, \theta) &\leq \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)] ds \\ &\quad + C \mathbb{E} \int_0^T |b(s, X_s^\theta, \varphi_s) - b(s, X_s^\theta, \theta_s)|^2 ds \\ &\quad + C \mathbb{E} \int_0^T |\sigma(s, X_s^\theta, \varphi_s) - \sigma(s, X_s^\theta, \theta_s)|^2 ds \\ &\quad + C \mathbb{E} \int_0^T |D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)|^2 ds. \end{aligned} \quad (3.12)$$

The proof will be given in Section 3.3. We now state a necessary condition for optimality for the augmented Hamiltonian.

Theorem 3.2.4 (Extended Pontryagin's optimality principle). *Let α^* be the (locally) optimal control, X^{α^*} be the associated controlled state solving (3.1), and $(Y^{\alpha^*}, Z^{\alpha^*})$ be the associated adjoint processes solving (3.4). Then for any $a \in A$ we have*

$$\tilde{\mathcal{H}}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*, \alpha_s^*) \leq \tilde{\mathcal{H}}(t, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*, a), \quad \forall s \in [0, T]. \quad (3.13)$$

The proof of Theorem 3.2.4 will come in Section 3.3. We are now ready to present the main result of the chapter.

Theorem 3.2.5. *Let Assumptions 3.2.1 and 3.2.2 hold. Let $\{\alpha^n\}_{n \geq 0}$ be the sequence given by Algorithm 3. For $n \geq 1$, define*

$$\mu^{n-1} := \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})] ds.$$

Then, for sufficiently large $\rho > 0$, the sequence $\{\mu^n\}_{n \geq 1}$ converges to 0 as $n \rightarrow \infty$.

Theorem 3.2.5 will be proved in Section 3.3. It can be seen from the proof that ρ needs to be two times larger than the constant appearing in Lemma 3.2.3, which itself depends increases with T, d and constants from Assumption 3.2.1 and 3.2.2.

We cannot guarantee that the Algorithm 3 converges to the optimal control which minimizes (3.2), since the extended Pontryagin's optimality principle, see Theorem 3.2.4, is the necessary condition for optimality. The sufficient condition for optimality tells us that to get the optimal control we need to assume convexity of the Hamiltonian in state and control variables, and need to assume convexity of the terminal cost function. To that end, we need to assume convexity of b, σ, f and g in x and a .

In the following corollary, we show that under a particular setting of the problem we have logarithmic convergence of the modified method of successive approximations to the true solution of the problem.

Corollary 3.2.6. *Let Assumptions 3.2.1 and 3.2.2 hold. Moreover, if b, σ and f are in the form of*

$$\begin{aligned} b(t, x, a) &= b_1(t)x + b_2(t, a), \\ \sigma(t, x, a) &= \sigma_1(t)x + \sigma_2(t, a), \\ f(t, x, a) &= f_1(t, x) + f_2(t, a) \end{aligned}$$

for $\forall t \in [0, T], \forall x \in \mathbb{R}^d, \forall a \in A$. In addition, assume that f and g are convex in x , f_2, b_2, σ_2 are convex in a . Then we have the following estimate for the sequence $(\alpha^n)_{n \in \mathbb{N}}$ from Algorithm 3:

$$0 \leq J(x, \alpha^n) - J(x, \alpha^*) \leq \frac{C}{n},$$

where α^* is the optimal control for (3.2) and C is a positive constant.

The proof of Corollary 3.2.6 will be given in Section 3.3. Theorem 3.2.5 and Corollary 3.2.6 are extensions of the result in [18] to the stochastic case.

3.3 Proofs

We start working towards the proof of Theorem 3.2.5. Recall the adjoint equation for an admissible control α :

$$dY_s^\alpha = -D_x \mathcal{H}(s, X_s^\alpha, Y_s^\alpha, Z_s^\alpha, \alpha_s) ds + Z_s^\alpha dW_s, \quad s \in [0, T], \quad Y_T = D_x g(X_T^\alpha). \quad (3.14)$$

From now on, we shall use Einstein notation, so that repeated indices in a single term imply summation over all the values of that index.

Lemma 3.3.1. *Assume that there exists $K \geq 0$ such that $\forall x \in \mathbb{R}^d, \forall a \in A, \forall t \in [0, T]$ we have*

$$|D_x b(t, x, a)| + |D_x \sigma(t, x, a)| \leq K,$$

and

$$|D_x g(x)| + |D_x f(t, x, a)| \leq K.$$

Then $\|Y^\alpha\|_{\mathbb{H}^\infty}$ is bounded.

Proof. From the definition of the Hamiltonian (3.3) we have

$$\begin{aligned} D_{x_i} \mathcal{H}(s, X_s^\alpha, Y_s^\alpha, Z_s^\alpha, \alpha_s) &= D_{x_i} b^j(s, X_s^\alpha, \alpha_s) (Y_s^\alpha)^j + D_{x_i} \sigma^{jp}(s, X_s^\alpha, \alpha_s) (Z_s^\alpha)^{jp} \\ &\quad + D_{x_i} f(s, X_s^\alpha, \alpha_s), \quad \forall s \in [0, T], \quad i = 1, 2, \dots, d. \end{aligned}$$

Hence, one can observe that (3.14) is a linear BSDE. Therefore, from [47, Proposition 3.2] we can write the formula for the solution of (3.14):

$$Y_t^\alpha = \mathbb{E}_t \left[S_t^{-1} S_T D_x g(X_T^\alpha) + \int_t^T S_t^{-1} S_s D_x f(s, X_s^\alpha, \alpha_s) ds \right],$$

where the process S is the unique strong solution of

$$dS_t^{ij} = S_t^{il} D_{x_l} b^j(t, X_t^\alpha, \alpha_t) dt + S_t^{il} D_{x_l} \sigma^{jp}(t, X_t^\alpha, \alpha_t) dW_t^p, \quad i, j = 1, 2, \dots, d, \quad S_0 = I_d,$$

and S^{-1} is the inverse process of S . Thus, due to [47, Corollary 3.7] and assumptions of lemma we have the following bound:

$$\|Y^\alpha\|_{\mathbb{H}^\infty} \leq C \|D_x g(X_T^\alpha)\|_{L^\infty} + CT \|D_x f(\cdot, X^\alpha, \alpha)\|_{\mathbb{H}^\infty}.$$

Hence, due to assumptions of lemma we conclude that $\|Y^\alpha\|_{\mathbb{H}^\infty}$ is bounded. $\square$

Proof of Lemma 3.2.3. Let φ and θ be some generic admissible controls. We will write $(X_s^\varphi)_{s \in [0, T]}$ for the solution of (3.1) controlled by φ and $(X_s^\theta)_{s \in [0, T]}$ for the solution of (3.1) controlled by θ . We denote solutions of corresponding adjoint equations by $(Y_s^\varphi, Z_s^\varphi)_{s \in [0, T]}$ and $(Y_s^\theta, Z_s^\theta)_{s \in [0, T]}$. Due to Taylor's theorem, we

note that for some $R^1(\omega) \in [0, 1]$, we have $\forall \omega \in \Omega$ that

$$\begin{aligned}
g(X_T^\varphi) - g(X_T^\theta) &= (D_x g(X_T^\theta))^\top (X_T^\varphi - X_T^\theta) \\
&\quad + \frac{1}{2} (X_T^\varphi - X_T^\theta)^\top D_x^2 g(X_T^\theta + R^1(X_T^\varphi - X_T^\theta)) (X_T^\varphi - X_T^\theta) \\
&\leq (D_x g(X_T^\theta))^\top (X_T^\varphi - X_T^\theta) \\
&\quad + \frac{1}{2} (X_T^\varphi - X_T^\theta)^\top |D_x^2 g(X_T^\theta + R^1(X_T^\varphi - X_T^\theta))| (X_T^\varphi - X_T^\theta) \\
&\leq (D_x g(X_T^\theta))^\top (X_T^\varphi - X_T^\theta) + \frac{K}{2} |X_T^\varphi - X_T^\theta|^2.
\end{aligned}$$

The last inequality holds due to Assumption 3.2.2. Recall that $Y_T^\theta = D_x g(X_T^\theta)$. Hence, using Itô's product rule, we get

$$\begin{aligned}
\mathbb{E}[g(X_T^\varphi) - g(X_T^\theta)] &\leq \mathbb{E} \left[(Y_T^\theta)^\top (X_T^\varphi - X_T^\theta) + \frac{K}{2} |X_T^\varphi - X_T^\theta|^2 \right] \\
&\leq \mathbb{E} \int_0^T (X_s^\varphi - X_s^\theta)^\top dY_s^\theta + \mathbb{E} \int_0^T (Y_s^\theta)^\top [dX_s^\varphi - dX_s^\theta] \\
&\quad + \mathbb{E} \int_0^T \text{tr}[(\sigma(s, X_s^\varphi, \varphi_s) - \sigma(s, X_s^\theta, \theta_s))^\top Z_s^\theta] ds + \frac{K}{2} \mathbb{E} [|X_T^\varphi - X_T^\theta|^2].
\end{aligned}$$

From this, the forward SDE (3.1) and the adjoint equation (3.4) we thus get

$$\begin{aligned}
&\mathbb{E}[g(X_T^\varphi) - g(X_T^\theta)] \\
&\leq -\mathbb{E} \int_0^T (X_s^\varphi - X_s^\theta)^\top D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s) ds \\
&\quad + \mathbb{E} \int_0^T (Y_s^\theta)^\top [b(s, X_s^\varphi, \varphi_s) - b(s, X_s^\theta, \theta_s)] ds \\
&\quad + \mathbb{E} \int_0^T \text{tr}[(\sigma(s, X_s^\varphi, \varphi_s) - \sigma(s, X_s^\theta, \theta_s))^\top Z_s^\theta] ds + \frac{K}{2} \mathbb{E} [|X_T^\varphi - X_T^\theta|^2].
\end{aligned} \tag{3.15}$$

On the other hand, by definition of the Hamiltonian we have

$$\begin{aligned}
&\mathbb{E} \int_0^T [f(s, X_s^\varphi, \varphi_s) - f(s, X_s^\theta, \theta_s)] ds \\
&= \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^\varphi, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)] ds \\
&\quad - \mathbb{E} \int_0^T (Y_s^\theta)^\top [b(s, X_s^\varphi, \varphi_s) - b(s, X_s^\theta, \theta_s)] ds \\
&\quad - \mathbb{E} \int_0^T \text{tr}[(\sigma(s, X_s^\varphi, \varphi_s) - \sigma(s, X_s^\theta, \theta_s))^\top Z_s^\theta] ds.
\end{aligned} \tag{3.16}$$

Summing up (3.15) and (3.16) we get

$$\begin{aligned}
& J(x, \varphi) - J(x, \theta) \\
&= \mathbb{E}[g(X_T^\varphi) - g(X_T^\theta)] + \mathbb{E} \int_0^T [f(s, X_s^\varphi, \varphi_s) - f(s, X_s^\theta, \theta_s)] ds \\
&\leq \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^\varphi, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s) \\
&\quad - (X_s^\varphi - X_s^\theta)^\top D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)] ds + \frac{K}{2} \mathbb{E} [|X_T^\varphi - X_T^\theta|^2] .
\end{aligned} \tag{3.17}$$

Due to Taylor's theorem, there exists $(R_s^2(\omega))_{s \in [0, T]} \in [0, 1]$ such that $\forall \omega \in \Omega$ we have

$$\begin{aligned}
& \mathcal{H}(s, X_s^\varphi, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s) \\
&= \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s) \\
&+ (X_s^\varphi - X_s^\theta)^\top D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) \\
&+ \frac{1}{2} (X_s^\varphi - X_s^\theta)^\top D_x^2 \mathcal{H}(s, X_s^\theta + R_s^2(X_s^\varphi - X_s^\theta), Y_s^\theta, Z_s^\theta, \varphi_s) (X_s^\varphi - X_s^\theta) .
\end{aligned} \tag{3.18}$$

Since $D_x^2 \sigma(s, X_s^\theta + R_s^2(X_s^\varphi - X_s^\theta), \varphi_s) = 0$ by Assumption 3.2.1, we have that

$$\begin{aligned}
& \left| D_{x_i x_j}^2 \mathcal{H}(s, X_s^\theta + R_s^2(X_s^\varphi - X_s^\theta), Y_s^\theta, Z_s^\theta, \varphi_s) \right| \\
&= \left| D_{x_i x_j}^2 b^l(s, X_s^\theta + R_s^2(X_s^\varphi - X_s^\theta), \varphi_s) (Y_s^\theta)^l \right. \\
&\quad \left. + D_{x_i x_j}^2 f(s, X_s^\theta + R_s^2(X_s^\varphi - X_s^\theta), \varphi_s) \right| , \quad i, j = 1, 2, \dots, d .
\end{aligned}$$

From Lemma 3.3.1 we know that $|Y_s^\theta|$ is bounded a.s. for all $s \in [0, T]$. Hence by Assumption 3.2.1 and 3.2.2 we have

$$|D_x^2 \mathcal{H}(s, X_s^\theta + R_s^2(X_s^\varphi - X_s^\theta), Y_s^\theta, Z_s^\theta, \varphi_s)| < \infty . \tag{3.19}$$

Therefore, after substituting (3.18) into (3.17), and by 3.19 we get

$$\begin{aligned}
J(x, \varphi) - J(x, \theta) &= \mathbb{E} \left[\int_0^T [\mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s) \right. \\
&\quad \left. + (X_s^\varphi - X_s^\theta)^\top (D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)) \right. \\
&\quad \left. + \frac{K}{2} |X_s^\varphi - X_s^\theta|^2 ds \right] + \frac{K}{2} \mathbb{E} [|X_T^\varphi - X_T^\theta|^2] .
\end{aligned}$$

Let us now get a standard SDE estimate for the difference of X^φ and X^θ . From $(a + b)^2 \leq 2a^2 + 2b^2$, from taking the expectation, from Hölder's inequality, from Assumption 3.2.1, from the Burkholder-Davis-Gundy inequality and from

Gronwall's inequality we obtain

$$\begin{aligned} \mathbb{E} \sup_{0 \leq t \leq T} |X_t^\varphi - X_t^\theta|^2 &\leq C \mathbb{E} \int_0^T |b(s, X_s^\theta, \varphi_s) - b(s, X_s^\theta, \theta_s)|^2 ds \\ &\quad + C \mathbb{E} \int_0^T |\sigma(s, X_s^\theta, \varphi_s) - \sigma(s, X_s^\theta, \theta_s)|^2 ds. \end{aligned} \quad (3.20)$$

Young's inequality allows us to get the estimate

$$\begin{aligned} J(x, \varphi) - J(x, \theta) &\leq \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)] ds + \frac{1}{2} \mathbb{E} \int_0^T |X_s^\varphi - X_s^\theta|^2 ds \\ &\quad + \frac{1}{2} \mathbb{E} \left[\int_0^T |D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)|^2 \right. \\ &\quad \left. + \frac{K}{2} |X_s^\varphi - X_s^\theta|^2 ds \right] + \frac{K}{2} \mathbb{E} [|X_T^\varphi - X_T^\theta|^2]. \end{aligned}$$

Hence, from (3.20) we have that

$$\begin{aligned} J(x, \varphi) - J(x, \theta) &\leq \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)] ds \\ &\quad + C \mathbb{E} \int_0^T |b(s, X_s^\theta, \varphi_s) - b(s, X_s^\theta, \theta_s)|^2 ds \\ &\quad + C \mathbb{E} \int_0^T |\sigma(s, X_s^\theta, \varphi_s) - \sigma(s, X_s^\theta, \theta_s)|^2 ds \\ &\quad + C \mathbb{E} \int_0^T |D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \varphi_s) - D_x \mathcal{H}(s, X_s^\theta, Y_s^\theta, Z_s^\theta, \theta_s)|^2 ds, \end{aligned}$$

for some constant $C > 0$, which depends on K, T , and d . $\square$

Proof of Theorem 3.2.4. Since α^* is the (locally) optimal control for the problem (3.2), the Pontryagin's optimality principle holds, see e.g. [1]. Hence for any $a \in A$ we have

$$\mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*) \leq \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, a), \quad \forall s \in [0, T]. \quad (3.21)$$

By definition of the augmented Hamiltonian (3.6) for all $s \in [0, T]$ we have

$$\begin{aligned} \tilde{\mathcal{H}}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*, a) &= \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, a) \\ &\quad + \frac{1}{2} \rho |b(s, X_s^{\alpha^*}, a) - b(s, X_s^{\alpha^*}, \alpha_s^*)|^2 + \frac{1}{2} \rho |\sigma(s, X_s^{\alpha^*}, a) - \sigma(s, X_s^{\alpha^*}, \alpha_s^*)|^2 \\ &\quad + \frac{1}{2} \rho |D_x \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, a) - D_x \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*)|^2. \end{aligned} \quad (3.22)$$

Therefore, due to (3.21) and (3.22) we have

$$\begin{aligned}
\tilde{\mathcal{H}}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*, \alpha_s^*) &= \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*) \\
&\leq \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, a) + \frac{1}{2}\rho|b(s, X_s^{\alpha^*}, a) - b(s, X_s^{\alpha^*}, \alpha_s^*)|^2 \\
&\quad + \frac{1}{2}\rho|\sigma(s, X_s^{\alpha^*}, a) - \sigma(s, X_s^{\alpha^*}, \alpha_s^*)|^2 \\
&\quad + \frac{1}{2}\rho|D_x \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, a) - D_x \mathcal{H}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*)|^2 \\
&= \tilde{\mathcal{H}}(s, X_s^{\alpha^*}, Y_s^{\alpha^*}, Z_s^{\alpha^*}, \alpha_s^*, a).
\end{aligned}$$

This concludes the proof. $\square$

Proof of Theorem 3.2.5. Let us apply Lemma 3.2.3 for $\varphi = \alpha^n$ and $\theta = \alpha^{n-1}$. Hence, for some $C > 0$ we have

$$\begin{aligned}
J(x, \alpha^n) - J(x, \alpha^{n-1}) &\leq \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})] ds \\
&\quad + C\mathbb{E} \int_0^T |b(s, X_s^n, \alpha_s^n) - b(s, X_s^n, \alpha_s^{n-1})|^2 ds \\
&\quad + C\mathbb{E} \int_0^T |\sigma(s, X_s^n, \alpha_s^n) - \sigma(s, X_s^n, \alpha_s^{n-1})|^2 ds \\
&\quad + C\mathbb{E} \int_0^T |D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})|^2 ds.
\end{aligned} \tag{3.23}$$

Let

$$\mu(\alpha^{n-1}) = \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})] ds.$$

Due to the definition of α^n (3.8) and (3.13) we have for all $s \in [0, T]$

$$\begin{aligned}
&\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) + \frac{1}{2}\rho|b(s, X_s^n, \alpha_s^n) - b(s, X_s^n, \alpha_s^{n-1})|^2 \\
&\quad + \frac{1}{2}\rho|\sigma(s, X_s^n, \alpha_s^n) - \sigma(s, X_s^n, \alpha_s^{n-1})|^2 \\
&\quad + \frac{1}{2}\rho|D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})|^2 \\
&\leq \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}).
\end{aligned}$$

Therefore, we can observe that $\mu(\alpha^{n-1}) \leq 0$. Hence we can rewrite the inequality (3.23) as

$$J(x, \alpha^n) - J(x, \alpha^{n-1}) \leq \mu(\alpha^{n-1}) - \frac{2C}{\rho}\mu(\alpha^{n-1}) = D\mu(\alpha^{n-1}), \tag{3.24}$$

where $D := 1 - \frac{2C}{\rho}$. By choosing $\rho > 2C$ we have that $D > 0$. Notice that for any integer $M > 1$ we have

$$\begin{aligned} \sum_{n=1}^M (-\mu(\alpha^{n-1})) &\leq D^{-1} \sum_{n=1}^M (J(x, \alpha^{n-1}) - J(x, \alpha^n)) \\ &= D^{-1} (J(x, \alpha^0) - J(x, \alpha^M)) \leq D^{-1} (J(x, \alpha^0) - \inf_{\alpha \in \mathcal{A}} J(x, \alpha)) < \infty. \end{aligned}$$

Since $(-\mu(\alpha^{n-1})) \geq 0$ and $\sum_{n=1}^{\infty} (-\mu(\alpha^{n-1})) < +\infty$ we have that $\mu(\alpha^{n-1}) \rightarrow 0$ as $n \rightarrow \infty$. This concludes the proof. $\square$

We need to introduce new notation, which will be used in the proof of Corollary 3.2.6. Denote the set

$$I_{\tau,h} := [\tau - h, \tau + h] \cap [0, T], \quad \tau \in [0, T], \quad h \in [0, +\infty). \quad (3.25)$$

Let us define for all $s \in [0, T]$

$$\Delta_{\alpha^{n-1}} \mathcal{H}(s) := \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}),$$

and

$$\mu(\alpha^{n-1}) := \mathbb{E} \int_0^T \Delta_{\alpha^{n-1}} \mathcal{H}(s) ds.$$

By definition of α^n notice that $\Delta_{\alpha^{n-1}} \mathcal{H}(t) \leq 0$ for all $t \in [0, T]$. Let us show an auxiliary lemma.

Lemma 3.3.2. *For any $h > 0$ there exists τ , which depends on h and α^{n-1} , such that*

$$\mathbb{E} \int_{I_{\tau,h}} \Delta_{\alpha^{n-1}} \mathcal{H}(t) dt \leq \frac{h\mu(\alpha^{n-1})}{T}.$$

Proof. We will prove by contradiction. Assume that there exists $h^* > 0$ such that $\forall \tau \in [0, T]$ we have

$$\mathbb{E} \int_{I_{\tau,h^*}} \Delta_{\alpha^{n-1}} \mathcal{H}(t) dt > \frac{h^*\mu(\alpha^{n-1})}{T}. \quad (3.26)$$

Denote $\tau_i = ih^*$, $i = 0, 1, \dots, N(h^*)$, where $N(h^*) = \lceil T/h^* \rceil$ - integer part. Since $\Delta_{\alpha^{n-1}} \mathcal{H}(t) \leq 0$ for all $t \in [0, T]$ by definition of α^n and $\cup_{i=0}^{N(h^*)} I_{\tau_i, h^*}$ is a superset of $[0, T]$ we have

$$\mu(\alpha^{n-1}) = \mathbb{E} \int_0^T \Delta_{\alpha^{n-1}} \mathcal{H}(t) dt \geq \sum_{i=0}^{N(h^*)} \mathbb{E} \int_{I_{\tau_i, h^*}} \Delta_{\alpha^{n-1}} \mathcal{H}(t) dt. \quad (3.27)$$

Hence, by (3.26) we get

$$\mu(\alpha^{n-1}) > \frac{h^* N(h^*)}{T} \mathbb{E} \int_0^T \Delta_{\alpha^{n-1}} \mathcal{H}(t) dt > \mu(\alpha^{n-1}).$$

Hence we get the contradiction. $\square$

Now we are ready to prove Corollary 3.2.6.

Proof of Corollary 3.2.6. First, observe that

$$\begin{aligned} b(s, X_s^n, \alpha_s^n) - b(s, X_s^n, \alpha_s^{n-1}) &= b_2(s, \alpha_s^n) - b_2(s, \alpha_s^{n-1}), \\ \sigma(s, X_s^n, \alpha_s^n) - \sigma(s, X_s^n, \alpha_s^{n-1}) &= \sigma_2(s, \alpha_s^n) - \sigma_2(s, \alpha_s^{n-1}), \\ D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}) &= 0. \end{aligned}$$

Let us consider the set $I_{\tau,h}$ given by (3.25). We will specify the choice of τ and h later. Hence, after applying Lemma 3.2.3 for α^n and α^{n-1} we have for some $C > 0$

$$\begin{aligned} &J(x, \alpha^n) - J(x, \alpha^{n-1}) \\ &\leq \mathbb{E} \int_{I_{\tau,h}} [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})] ds \\ &\quad + C \mathbb{E} \int_{I_{\tau,h}} |b_2(s, \alpha_s^n) - b_2(s, \alpha_s^{n-1})|^2 + |\sigma_2(s, \alpha_s^n) - \sigma_2(s, \alpha_s^{n-1})|^2 ds \\ &\quad + \mathbb{E} \int_{[0,T] \setminus I_{\tau,h}} [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})] ds \\ &\quad + C \mathbb{E} \int_{[0,T] \setminus I_{\tau,h}} |b_2(s, \alpha_s^n) - b_2(s, \alpha_s^{n-1})|^2 + |\sigma_2(s, \alpha_s^n) - \sigma_2(s, \alpha_s^{n-1})|^2 ds. \end{aligned}$$

Since the following holds for all $s \in [0, T]$ and $\rho \geq 0$:

$$\begin{aligned} &\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}) \\ &\quad + \frac{1}{2} \rho |b_2(s, \alpha_s^n) - b_2(s, \alpha_s^{n-1})|^2 + \frac{1}{2} \rho |\sigma_2(s, \alpha_s^n) - \sigma_2(s, \alpha_s^{n-1})|^2 \leq 0, \end{aligned}$$

we have for $\rho \geq 2C$

$$\begin{aligned} &J(x, \alpha^n) - J(x, \alpha^{n-1}) \\ &\leq \mathbb{E} \int_{I_{\tau,h}} [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})] ds \\ &\quad + C \mathbb{E} \int_{I_{\tau,h}} |b_2(s, \alpha_s^n) - b_2(s, \alpha_s^{n-1})|^2 + |\sigma_2(s, \alpha_s^n) - \sigma_2(s, \alpha_s^{n-1})|^2 ds. \end{aligned}$$

Therefore, from Lemma 3.3.2 and from similar calculations as in (3.24), there exists τ such that

$$\begin{aligned} &J(x, \alpha^n) - J(x, \alpha^{n-1}) \\ &\leq \left(1 - \frac{2C}{\rho}\right) \mathbb{E} \int_{I_{\tau,h}} [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1})] ds \\ &\leq \left(1 - \frac{2C}{\rho}\right) \frac{h\mu(\alpha^{n-1})}{T}. \end{aligned}$$

Let us choose $h = -(\rho - 2C)\mu(\alpha^{n-1})/(\rho T)$. Hence

$$J(x, \alpha^n) - J(x, \alpha^{n-1}) \leq -(\rho - 2C)^2(\mu(\alpha^{n-1}))^2/(\rho^2 T^2). \quad (3.28)$$

Let α^* be the optimal control. Indeed, by the sufficient condition for optimality, see e.g. [1], and by assumptions of corollary, we have the existence of the optimal control. Therefore, by convexity of g , and by Itô's product rule we have

$$\begin{aligned} 0 &\leq J(x, \alpha^{n-1}) - J(x, \alpha^*) \\ &= \mathbb{E} \left[\int_0^T (f(s, X_s^n, \alpha_s^{n-1}) - f(s, X_s, \alpha_s^*)) ds + g(X_T^n) - g(X_T) \right] \\ &\leq \mathbb{E} \left[\int_0^T (f(s, X_s^n, \alpha_s^{n-1}) - f(s, X_s, \alpha_s^*)) ds \right] + \mathbb{E}[(D_x g(X^n))^\top (X_T^n - X_T)] \\ &\leq \mathbb{E} \left[\int_0^T (f(s, X_s^n, \alpha_s^{n-1}) - f(s, X_s, \alpha_s^*)) ds \right] \\ &\quad + \mathbb{E} \left[\int_0^T (Y_s^n)^\top d(X_s^n - X_s) + \int_0^T (X_s^n - X_s)^\top dY_s^n \right] \\ &\quad + \mathbb{E} \left[\int_0^T \text{tr}((\sigma(s, X_s^n, \alpha_s^{n-1}) - \sigma(s, X_s, \alpha_s^*))^\top Z_s^n) ds \right]. \end{aligned}$$

Hence, we have that

$$\begin{aligned} 0 &\leq J(x, \alpha^{n-1}) - J(x, \alpha^*) \leq \mathbb{E} \left[\int_0^T f(s, X_s^n, \alpha_s^{n-1}) - f(s, X_s, \alpha_s^*) ds \right] \\ &\quad + \mathbb{E} \left[\int_0^T (Y_s^n)^\top (b(s, X_s^n, \alpha_s^{n-1}) - b(s, X_s, \alpha_s^*)) ds \right] \\ &\quad - \mathbb{E} \left[\int_0^T (X_s^n - X_s)^\top D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}) ds \right] \\ &\quad + \mathbb{E} \left[\int_0^T \text{tr}((\sigma(s, X_s^n, \alpha_s^{n-1}) - \sigma(s, X_s, \alpha_s^*))^\top Z_s^n) ds \right]. \end{aligned}$$

Recalling the form of b, σ and observing that

$$D_x \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}) = b_1(s)Y_s^n + \sigma_1(s)Z_s^n + D_x f(s, X_s^n, \alpha_s^{n-1}),$$

we have

$$\begin{aligned} 0 &\leq J(x, \alpha^{n-1}) - J(x, \alpha^*) \leq \mathbb{E} \left[\int_0^T f(s, X_s^n, \alpha_s^{n-1}) - f(s, X_s, \alpha_s^*) ds \right] \\ &\quad + \mathbb{E} \left[\int_0^T \text{tr}((\sigma_2(s, \alpha_s^{n-1}) - \sigma_2(s, \alpha_s^*))^\top Z_s^n) ds \right] \\ &\quad + \mathbb{E} \left[\int_0^T (Y_s^n)^\top (b_2(s, \alpha_s^{n-1}) - b_2(s, \alpha_s^*)) ds - \int_0^T (X_s^n - X_s)^\top D_x f(s, X_s^n, \alpha_s^{n-1}) ds \right]. \end{aligned}$$

Since f is convex in x we have for all $s \in [0, T]$ that

$$f(s, X_s, \alpha_s^{n-1}) \geq f(s, X_s^n, \alpha_s^{n-1}) + (X_s - X_s^n)^\top D_x f(s, X_s^n, \alpha_s^{n-1}).$$

Therefore, we obtain

$$\begin{aligned} J(x, \alpha^{n-1}) - J(x, \alpha^*) &\leq \mathbb{E} \int_0^T [\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^{n-1}) - \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^*)] ds \\ &\leq -\mu(\alpha^{n-1}), \end{aligned} \tag{3.29}$$

where the second inequality holds due to

$$\mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^n) \leq \mathcal{H}(s, X_s^n, Y_s^n, Z_s^n, \alpha_s^*).$$

Let $b^n := J(x, \alpha^n) - J(x, \alpha)$, then due to (3.28) and (3.29) we have that

$$b^n - b^{n-1} \leq \frac{-(\rho - 2C)^2 \mu(\alpha^{n-1})^2}{(\rho^2 T^2)} \leq \frac{-(\rho - 2C)^2 (b^{n-1})^2}{\rho^2 T^2}.$$

Therefore, due to Lemma B.1.1 we have

$$J(x, \alpha^n) - J(x, \alpha^*) \leq \frac{C_1}{n}.$$

for some constant $C_1 > 0$. This concludes the proof. $\square$

Chapter 4

A regularized MSA for stochastic control problems

4.1 Introduction

In this chapter, we study the method of successive approximations (MSA) for entropy regularized stochastic control problems, where the action space is the space of measures. We fix a finite horizon $T \in (0, \infty)$. We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$. Let $W = (W_t)_{t \in [0, T]}$ be a d' -dimensional Wiener martingale on this space. Given some metric space E , let $\mathcal{P}_2(E)$ denote the set of probability measures defined on E with finite 2-nd moment, endowed with the 2-Wasserstein distance. Let $\mathcal{P}_0(E) = \mathcal{P}(E)$ denote the set of probability measures and let $\mathcal{M}(E)$ denote the set of measures on E . Let $E = [0, T] \times \mathbb{R}^p$ and let

$$\mathcal{M}_2 := \left\{ \pi \in \mathcal{M}(E) : \pi_t \in \mathcal{P}(\mathbb{R}^p), \pi(da, dt) = \pi_t(a) da dt \text{ for a.a. } t \in [0, T], \right. \\ \left. \int_0^T \int |a|^2 \pi_t(a) da dt < \infty \right\},$$

where here and elsewhere any integral without an explicitly stated domain of integration is over $\mathbb{R}^p$. We define a set of admissible controls as

$$\mathcal{A} := \left\{ \pi : \Omega \rightarrow \mathcal{M}_2 : \mathbb{E} \int_0^T \int |a|^2 \pi_t(a) da dt < \infty \right. \\ \left. \text{and } \pi_t \text{ is } \mathcal{F}_t\text{-measurable } \forall t \in [0, T] \right\}.$$

We consider the controlled stochastic differential equation for given measurable functions $b : [0, T] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^p) \rightarrow \mathbb{R}^d$, $\sigma : [0, T] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^p) \rightarrow \mathbb{R}^{d \times d'}$, and for the initial state $x \in \mathbb{R}^d$, the admissible control $\pi \in \mathcal{A}$

$$dX_s(\pi) = b_s(X_s(\pi), \pi_s) ds + \sigma_s(X_s(\pi), \pi_s) dW_s, \quad s \in [0, T], \quad X_0(\pi) = x. \quad (4.1)$$

For now let us assume sufficient regularity of b and σ such that (4.1) admits an unique square integrable solution. See Assumptions 4.2.1 and 4.2.2. Denote the relative entropy of $m, m' \in \mathcal{P}(\mathbb{R}^p)$ as

$$\text{Ent}(m|m') := \int [\log m(a) - \log m'(a)] m(a) da$$

for m, m' that are absolutely continuous with respect to the Lebesgue measure, and for the not absolutely continuous ones let $\text{Ent}(m|m') = \infty$. Let $f : [0, T] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^p) \rightarrow \mathbb{R}$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be given measurable functions and consider the cost functional

$$J^\tau(\pi) := \mathbb{E} \left[\int_0^T [f_s(X_s(\pi), \pi_s) + \tau \text{Ent}(\pi_s | e^{-U})] ds + g(X_T(\pi)) \right] \quad (4.2)$$

for all $\pi \in \mathcal{A}$ and $\tau \geq 0$, where U is such that $\int e^{-U(a)} da = 1$. We want to solve the optimisation problem i.e. to find the optimal control $\pi^* \in \mathcal{A}$ which achieves the minimum of (4.2) for some fixed $\tau \geq 0$ and the initial state $x \in \mathbb{R}^d$.

Given b, σ, f, τ and $\lambda \geq 0$ we define the augmented Hamiltonian $H^{\tau, \lambda} : [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d \times d'} \times \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p) \rightarrow \mathbb{R}$ as

$$\begin{aligned} H_t^{\tau, \lambda}(x, y, z, m, m') \\ := b_t(x, m) \cdot y + \text{tr}(\sigma_t^\top(x, m)z) + f_t(x, m) + \tau \text{Ent}(m | e^{-U}) + \lambda \text{Ent}(m|m'). \end{aligned} \quad (4.3)$$

Note that $H^{\tau, 0}$ corresponds to usual Hamiltonian for (4.1)-(4.2). When $\lambda = 0$ and $\tau = 0$, we use $H_t^0(x, y, z, m) := H_t^{0, 0}(x, y, z, m, m')$. Consider for each $\pi \in \mathcal{A}$, the BSDE, called the adjoint equation

$$\begin{aligned} dY_s(\pi) &= -D_x H_s^0(X_s(\pi), Y_s(\pi), Z_s(\pi), \pi_s) ds + Z_s(\pi) dW_s, \\ Y_T(\pi) &= D_x g(X_T(\pi)), \quad s \in [0, T]. \end{aligned} \quad (4.4)$$

To ease our notation, for each $\pi \in \mathcal{A}$ we define $(\Theta_s(\pi))_{s \in [0, T]}$ as $\Theta_s(\pi) := (X_s(\pi), Y_s(\pi), Z_s(\pi))$ for $s \in [0, T]$. Note that trivially $D_x H^0 = D_x H^{\tau, \lambda}$. Now we are ready to introduce the regularized MSA in Algorithm 4.

Recently, there have been some progress on application of MSA to stochastic control problems. In [48], the authors studied the convergence of a modified MSA algorithm for solving stochastic control problems. They showed the convergence of the algorithm and obtained a rate of convergence for a specific problem setting. An application of MSA for decoupled forward-backward stochastic control systems is established in [49]. They extended the modified MSA convergence result from [48] to the more general stochastic recursive optimal control problem on a non-convex control domain and imposed weaker assumptions on coefficients enabled by appropriate BMO estimates.

In this chapter entropic regularization appears in two places. It may be intrinsic to the control problem we are solving i.e. we penalize the running cost by entropy ($\tau > 0$). This has proved extremely useful in Reinforcement learning (RL) problems, e.g. see [50], where the authors include the entropy directly in a

Algorithm 4 Regularized Method of Successive Approximations:

Initialization: make a guess of the control $\pi^0 = (\pi_s^0)_{s \in [0, T]}$.

while difference between $J^\tau(\pi^{n+1})$ and $J^\tau(\pi^n)$ is large **do**

 Given a control $\pi^n = (\pi_s^n)_{s \in [0, T]}$ solve the following forward SDE, then solve backward SDE:

$$\begin{aligned} dX_s(\pi^n) &= b_s(X_s(\pi^n), \pi_s^n) ds + \sigma_s(X_s(\pi^n), \pi_s^n) dW_s, \quad X_0(\pi^n) = x, \\ dY_s(\pi^n) &= -D_x H_s^0(\Theta_s(\pi^n), \pi_s^n) ds + Z_s(\pi^n) dW_s, \quad Y_T(\pi^n) = D_x g(X_T(\pi^n)). \end{aligned} \quad (4.5)$$

Update the control

$$\pi_s^{n+1} \in \arg \min_{\pi \in \mathcal{P}(\mathbb{R}^p)} H_s^{\tau, \lambda}(\Theta_s(\pi^n), \pi, \pi_s^n), \quad \forall s \in [0, T]. \quad (4.6)$$

end while

return π^{n+1} .

reward function to enforce the exploration of the algorithm. In fact to balance exploitation and exploration one only needs to adjust τ . Empirical studies (e.g. [51]) show that such a regularization helps obtain more robust decision making. In stochastic control, this has been studied in [52, 45, 53, 54]. Of particular note is [52], where it is shown that entropic regularization implies that the optimal feedback (i.e. Markovian) controls are Lipschitz continuous in the space variable. Thus the optimal control is shown to be robust under uncertainty. We also introduce entropic regularization on the level of the algorithm (see Algorithm 4). Here it plays an important role ensuring stability of the algorithm as we explain below.

Our contributions

In this chapter, we study the application of MSA to the problem (4.1)-(4.2).

- We modify the Hamiltonian $H^{\tau, 0}$ using the relative entropy of two consecutive controls appearing in Algorithm 4. Having done this we are able to solve the optimization problem (4.6) at each iteration of the algorithm and get a minimizer in explicit form (4.7) which is not too far from the previous control. In Lemma 4.2.5, which can be established using BSDE estimates, we estimate the change of the cost functional when we do a minimization step of Hamiltonian as in (4.6). If the last term of (4.9) is bigger than the first term, then for classical MSA algorithm (i.e $\lambda = 0$ in Algorithm 4) we cannot guarantee that we do an update of the control in optimal descent direction of the cost functional. Therefore, it could happen that the classical MSA may diverge. To overcome this, we need to modify the algorithm using the relative entropy of two consecutive controls from the iteration. The relative entropy penalization is also important in learning and makes the learning more stable, since at each iteration we penalize for being far from the control of the previous iteration.

- The Kullback-Leibler regularization is well studied in RL and it has been shown that regularization helps to get a better convergence and achieves averaging of the errors made over iteration, see [50]. In Lemma 4.2.7, we show that the combination of entropic and the Kullback-Leibler regularization leads to averaging effect of the estimation errors. This is due to the fact that the control updates can be written in the Polyak-Ruppert type update rule.
- We observe that the forward and backward dynamics in (4.5) are decoupled, due to the iteration used. Therefore, it can be efficiently approximated, even in high dimension, using deep learning methods, see [43] and [44]. Notice that in Algorithm 4 one needs to solve nonlinear minimization problem (4.6) at each iteration of the algorithm. Therefore, the minimization step might be computationally expensive for some problems. However, there are various tools to deal with this. Indeed, recalling the notion of flat (functional) derivative, see C.1.1, we observe that from (4.6) the first order condition, see e.g. [55, Proposition 2.5], we have $\forall a \in \mathbb{R}^p$

$$\begin{aligned} \pi_s^{n+1}(a) &= (\mathcal{Z}_s^{n+1})^{-1} (\pi_s^n(a))^{\frac{\lambda}{\lambda+\tau}} \exp \left(-\frac{\tau}{\lambda+\tau} U(a) \right) \\ &\quad \times \exp \left(-\frac{1}{\lambda+\tau} \frac{\delta H_s^0}{\delta m} (\Theta_s(\pi^n), \pi_s^{n+1}, a) \right), \end{aligned} \quad (4.7)$$

where

$$\begin{aligned} \mathcal{Z}_s^{n+1} &:= \int (\pi_s^n(\tilde{a}))^{\frac{\lambda}{\lambda+\tau}} \exp \left(-\frac{\tau}{\lambda+\tau} U(\tilde{a}) \right) \\ &\quad \times \exp \left(-\frac{1}{\lambda+\tau} \frac{\delta H_s^0}{\delta m} (\Theta_s(\pi^n), \pi_s^{n+1}, \tilde{a}) \right) d\tilde{a}. \end{aligned}$$

Observe that (4.7) is nonlinear, since we have π^{n+1} in the both sides of the equation. Numerical methods from mean-field Langevin equations or from statistical sampling theory can be used to approximate π^{n+1} .

- The main contribution of this chapter is the probabilistic proof of convergence of the modified method of successive approximations for an entropy regularized optimal control problems with the action space being on the space of measures.

This chapter is organised as follow: In Section 4.2 we state the assumptions and main results. In Section 4.3 we collect all proofs. Finally, in Appendix C we recall a linear functional derivative and the standard SDE estimate.

4.2 Main results

In this section, we will present the main results of the chapter. We start by introducing necessary notations. Let $|\cdot|$ denote any norm in a finite dimensional

Euclidean space and let $\|\cdot\|_{\text{TV}}$ denote the total variation distance of probability measures.

Assumption 4.2.1. The functions b and σ are jointly continuous in t and once differentiable in x , and functional derivatives $\frac{\delta b}{\delta m}$, $\frac{\delta \sigma}{\delta m}$ exist. There exists $K \geq 0$ such that $\forall x \in \mathbb{R}^d, \forall m \in \mathcal{P}_2(\mathbb{R}^p), \forall t \in [0, T]$,

$$|D_x b_t(x, m)| + |D_x \sigma_t(x, m)| \leq K,$$

and such that $\forall x \in \mathbb{R}^d, \forall m, m' \in \mathcal{P}_2(\mathbb{R}^p), \forall t \in [0, T]$,

$$|b_t(x, m) - b_t(x, m')| + |\sigma_t(x, m) - \sigma_t(x, m')| \leq K \|m - m'\|_{\text{TV}}.$$

Moreover, $\forall x, x' \in \mathbb{R}^d, \forall m, m' \in \mathcal{P}_2(\mathbb{R}^p), \forall t \in [0, T]$

$$\begin{aligned} |D_x b_t(x, m) - D_x b_t(x', m')| &\leq K|x - x'| + K\|m - m'\|_{\text{TV}}, \\ |D_x \sigma_t(x, m) - D_x \sigma_t(x', m')| &\leq K|x - x'| + K\|m - m'\|_{\text{TV}}, \end{aligned}$$

and $\forall x, x' \in \mathbb{R}^d, \forall m \in \mathcal{P}_2(\mathbb{R}^p), \forall a \in \mathbb{R}^p, \forall t \in [0, T]$

$$\left| \frac{\delta b_t}{\delta m}(x, m, a) \right| + \left| \frac{\delta \sigma_t}{\delta m}(x, m, a) \right| \leq K,$$

$$\left| \frac{\delta b_t}{\delta m}(x, m, a) - \frac{\delta b_t}{\delta m}(x', m, a) \right| + \left| \frac{\delta \sigma_t}{\delta m}(x, m, a) - \frac{\delta \sigma_t}{\delta m}(x', m, a) \right| \leq K|x - x'|.$$

Assumption 4.2.2. The function f is jointly continuous in t , f and g are once differentiable in x , and the linear functional $\frac{\delta f}{\delta m}$ exists. There is a constant $K \geq 0$ such that $\forall x \in \mathbb{R}^d, \forall m \in \mathcal{P}_2(\mathbb{R}^p), \forall t \in [0, T]$

$$|D_x g(x)| + |D_x f_t(x, m)| \leq K.$$

Moreover, $\forall x, x' \in \mathbb{R}^d, \forall m, m' \in \mathcal{P}_2(\mathbb{R}^p), \forall a \in \mathbb{R}^p, \forall t \in [0, T]$,

$$\begin{aligned} |D_x f_t(x, m) - D_x f_t(x', m')| &\leq K|x - x'| + K\|m - m'\|_{\text{TV}}, \\ |D_x g(x) - D_x g(x')| &\leq K|x - x'|, \\ \left| \frac{\delta f_t}{\delta m}(x, m, a) \right| &\leq K, \\ |D_x g(x) - D_x g(x')| &\leq K|x - x'|, \\ \left| \frac{\delta f_t}{\delta m}(x, m, a) - \frac{\delta f_t}{\delta m}(x', m, a) \right| &\leq K|x - x'|. \end{aligned}$$

In subsection 4.2.1, we provide an example when Assumptions 4.2.1 and 4.2.2 hold.

Lemma 4.2.3. *Let Assumptions 4.2.1 and 4.2.2 hold and fix $\pi \in \mathcal{A}$. Then (3.1) and (3.4) have a unique solution.*

We will not prove Lemma 4.2.3 since the existence of the unique solution to (3.1) is classical under the assumptions above and can be found e.g. in

Krylov [35, Chapter 3, Section 1]. The adjoint equation (3.4) is affine, with $(D_x b, D_x \sigma, D_x f, D_x g)$ being bounded due to Assumption 4.2.1 and 4.2.2, and can be solved directly utilising the moment bound for $X(\pi)$. Uniqueness follows from e.g. [20, Th 4.3.1].

We now state a necessary condition for optimality with regards to the augmented Hamiltonian $H^{\tau, \lambda}$, which is important to characterize the optimal solution of the problem (3.1)-(3.2).

Theorem 4.2.4 (Extended Pontryagin's optimality principle). *Let $\pi^* \in \mathcal{A}$ be the (locally) optimal control, $X(\pi^*)$ be the associated controlled state solving (3.1), and $(Y(\pi^*), Z(\pi^*))$ be the associated adjoint processes solving (3.4). Then*

$$H_s^{\tau, \lambda}(\Theta_s(\pi^*), \pi_s^*, \pi_s^*) = \min_{m \in \mathcal{P}(\mathbb{R}^p)} H_s^{\tau, \lambda}(\Theta_s(\pi^*), m, \pi_s^*), \quad \forall s \in [0, T]. \quad (4.8)$$

The proof of Theorem 4.2.4 will come in Section 4.3. We call $(X(\pi^*), Y(\pi^*), Z(\pi^*), \pi^*)$ a solution to (4.8). It is worth mentioning that not all solutions to (4.8) are solutions to the problem (3.1)-(3.2), since Theorem 4.2.4 is only a necessary condition for optimality. Some convexity assumption of the Hamiltonian is needed to conclude the sufficient condition for optimality.

Under Assumptions 4.2.1 and 4.2.2, we obtain the following key estimate. It says that the difference of the cost functional evaluated at two controls can be estimated by the difference of Hamiltonians and by the relative entropy of two controls. If we fix the control π in (4.9) and wish to choose π' so that the cost functional is decreasing, it is not enough to point-wise minimize the difference of Hamiltonians. Since the relative entropy is nonnegative, we might get that RHS of (4.9) is nonnegative as well. Hence, the classical MSA (i.e. $\lambda = 0$ in Algorithm 4) may diverge in this case. By updating the control as in (4.6) in Algorithm 4, we ensure that the updated control makes RHS of (4.9) nonpositive for sufficiently large $\lambda > 0$.

Lemma 4.2.5. *Let Assumptions 4.2.1 and 4.2.2 hold. Then for any admissible controls $\pi, \pi' \in \mathcal{A}$ there exists a constant $C > 0$ such that*

$$\begin{aligned} J^\tau(\pi') - J^\tau(\pi) &\leq \mathbb{E} \int_0^T [H_s^{\tau, 0}(\Theta_s(\pi), \pi'_s) - H_s^{\tau, 0}(\Theta_s(\pi), \pi_s)] ds \\ &\quad + C \mathbb{E} \int_0^T \text{Ent}(\pi'_s | \pi_s) ds. \end{aligned} \quad (4.9)$$

The proof will be given in Section 4.3. We are now ready to present the main result of the chapter.

Theorem 4.2.6. *Let Assumptions 4.2.1 and 4.2.2 hold. Let $\{\pi^n\}_{n \geq 0}$ be the sequence given by Algorithm 4. For $n \geq 0$, define*

$$\mu^n := \mathbb{E} \int_0^T [H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^{n+1}) - H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^n)] ds.$$

Then, for sufficiently large $\lambda > 0$, the sequence $\{\mu^n\}_{n \geq 0}$ converges to 0 as $n \rightarrow \infty$.

Theorem 4.2.6 will be proved in Section 4.3. It can be seen from the proof that λ needs to be larger than the constant appearing in Lemma 4.2.5, which itself depends on T, d and constants from Assumption 4.2.1 and 4.2.2.

In the lemma below, we discuss the benefit of using entropy regularization. In particular, it helps to average out the errors introduced by the numerical approximation of the forward and the backward SDEs in Algorithm 4.

Lemma 4.2.7. *Let $\tau, \lambda > 0$. Let $\{\pi^n\}_{n \geq 0}$ be the sequence given by Algorithm 4 and let $\{\tilde{\pi}^n\}_{n \geq 0}$ be the sequence of controls given by introducing numerical errors on the approximation of the forward and the backward SDEs in Algorithm 4. Denote by $\varepsilon_s^i(a) = \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^i), \pi_s^{i+1}, a) - \frac{\delta H_s^0}{\delta m}(\tilde{\Theta}_s(\tilde{\pi}^i), \tilde{\pi}_s^{i+1}, a)$ the corresponding numerical error on the (i) -th iteration, where $\tilde{\Theta}(\tilde{\pi}^i)$ is the numerical approximations of forward and backward processes. Then*

$$\tilde{\pi}_s^{n+1}(a) \propto \pi_s^{n+1}(a) e^{-\frac{1}{\lambda+\tau} \sum_{i=0}^n \left(\frac{\lambda}{\lambda+\tau}\right)^{n-i} \varepsilon_s^i(a)}.$$

Proof. We define $\beta = \frac{\lambda}{\lambda+\tau}$ (and thus $1 - \beta = \frac{\tau}{\lambda+\tau}$ and $\frac{\beta}{\lambda} = \frac{1}{\lambda+\tau}$). By induction from (4.7) we have

$$\begin{aligned} \pi_s^{n+1}(a) &\propto (\pi_s^n(a))^\beta e^{-(1-\beta)U(a)} e^{-\frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a)} \\ &\propto (\pi_s^{n-1}(a))^{\beta^2} e^{-(1-\beta^2)U(a)} e^{-\frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a) - \frac{\beta^2}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^{n-1}), \pi_s^n, a)} \\ &\propto \dots \\ &\propto (\pi_s^0(a))^{\beta^{n+1}} e^{-(1-\beta^{n+1})U(a)} e^{-\frac{\beta}{\lambda} \sum_{i=0}^n \beta^{n-i} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^i), \pi_s^{i+1}, a)}. \end{aligned}$$

Define w_s^n as the moving average of past values, with $w_s^0 := (1-\beta) \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^0), \pi_s^1, a)$,

$$\begin{aligned} w_s^n &:= \beta w_s^{n-1} + (1-\beta) \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a) \\ &= (1-\beta) \sum_{i=0}^n \beta^{n-i} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^i), \pi_s^{i+1}, a). \end{aligned}$$

This might remind the reader the Polyak-Ruppert type update. Noticing that $\frac{\beta}{\lambda(1-\beta)} = \frac{1}{\tau}$ we get

$$\pi_s^{n+1}(a) \propto (\pi_s^0(a))^{\beta^{n+1}} e^{-(1-\beta^{n+1})U(a)} e^{-\frac{1}{\tau} w_s^n}.$$

We now redo these calculations with numerical errors involving on the level of solving the forward and backward SDE to show that errors average out in w_s^n . We assume that optimization step (4.6) can be evaluated exactly. Indeed, let $(\tilde{\Theta}_s(\pi))_{s \in [0, T]} = (\tilde{X}_s(\pi), \tilde{Y}_s(\pi), \tilde{Z}_s(\pi))_{s \in [0, T]}$ be numerical approximations of forward and backward processes $(\Theta_s(\pi))_{s \in [0, T]} = (X_s(\pi), Y_s(\pi), Z_s(\pi))_{s \in [0, T]}$, and let ε_s^i be the corresponding numerical error on the (i) -th iteration such that

$$\varepsilon_s^i(a) = \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^i), \pi_s^{i+1}, a) - \frac{\delta H_s^0}{\delta m}(\tilde{\Theta}_s(\tilde{\pi}^i), \tilde{\pi}_s^{i+1}, a).$$

Notice that ε_s^i involves two type of errors, since

$$\begin{aligned}\varepsilon_s^i(a) &= \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^i), \pi_s^{i+1}, a) - \frac{\delta H_s^0}{\delta m}(\Theta_s(\tilde{\pi}^i), \tilde{\pi}_s^{i+1}, a) \\ &\quad + \frac{\delta H_s^0}{\delta m}(\Theta_s(\tilde{\pi}^i), \tilde{\pi}_s^{i+1}, a) - \frac{\delta H_s^0}{\delta m}(\tilde{\Theta}_s(\tilde{\pi}^i), \tilde{\pi}_s^{i+1}, a) := \varepsilon_s^{i,1}(a) + \varepsilon_s^{i,2}(a).\end{aligned}$$

The error $\varepsilon_s^{i,2}$ is the local error at each iteration and the error $\varepsilon_s^{i,1}$ can be estimated by the stability result from the BSDE theory, see Lemma 4.3.3. Let us find the representation of $\tilde{\pi}_s^{n+1}$:

$$\begin{aligned}\tilde{\pi}_s^{n+1}(a) &\propto (\tilde{\pi}_s^n(a))^\beta e^{-(1-\beta)U(a)} e^{-\frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\tilde{\Theta}_s(\tilde{\pi}^n), \tilde{\pi}_s^{n+1}, a)} \\ &\propto (\tilde{\pi}_s^n(a))^\beta e^{-(1-\beta)U(a)} e^{-\frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a)} e^{\frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a) - \frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\tilde{\Theta}_s(\tilde{\pi}^n), \tilde{\pi}_s^{n+1}, a)} \\ &\propto (\tilde{\pi}_s^{n-1}(a))^{\beta^2} e^{-(1-\beta^2)U(a)} e^{-\frac{\beta^2}{\lambda} \frac{\delta H_s^0}{\delta m}(\tilde{\Theta}_s(\tilde{\pi}^{n-1}), \tilde{\pi}_s^n, a)} e^{-\frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a)} e^{\frac{\beta}{\lambda} \varepsilon_s^n(a)} \\ &\propto (\tilde{\pi}_s^{n-1}(a))^{\beta^2} e^{-(1-\beta^2)U(a)} e^{-\frac{\beta^2}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^{n-1}), \pi_s^n, a) - \frac{\beta}{\lambda} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a)} \\ &\quad \times e^{\frac{\beta}{\lambda} \varepsilon_s^n(a) + \frac{\beta^2}{\lambda} \varepsilon_s^{n-1}(a)} \propto \dots \\ &\propto (\pi_s^0(a))^{\beta^{n+1}} e^{-(1-\beta^{n+1})U(a)} e^{-\frac{\beta}{\lambda} \sum_{i=0}^n \beta^{n-i} \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^i), \pi_s^{i+1}, a)} e^{\frac{\beta}{\lambda} \sum_{i=0}^n \beta^{n-i} \varepsilon_s^i(a)} \\ &\propto \pi_s^n(a) e^{\frac{\beta}{\lambda} \sum_{i=0}^n \beta^{n-i} \varepsilon_s^i(a)}.\end{aligned}$$

Let us now define $\tilde{w}_s^0(a) = \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^0), \pi_s^1, a) - \varepsilon_s^0$ and $\tilde{w}_s^n$ for $n \geq 1$ as the average of past $n+1$ values

$$\begin{aligned}\tilde{w}_s^n(a) &= \beta \sum_{i=0}^n \beta^{n-i} \left[\frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^i), \pi_s^{i+1}, a) - \varepsilon_s^i \right] \\ &= \beta \tilde{w}_s^{n-1}(a) + (1-\beta) \left[\frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^n), \pi_s^{n+1}, a) - \varepsilon_s^n \right].\end{aligned}$$

Therefore, we get that

$$\tilde{\pi}_s^{n+1}(a) \propto (\tilde{\pi}_s^0)^{\beta^{n+1}} e^{-(1-\beta^{n+1})U(a)} \exp\left(-\frac{1}{\tau} \tilde{w}_s^n(a)\right).$$

This concludes the proof. $\square$

4.2.1 Relaxed control

In this subsection, we apply Algorithm 4 to relaxed control problems. Assume that $d = d' = 1$. Consider the controlled diffusion process $X(\pi)$ for each $\pi \in \mathcal{A}$ satisfying the following SDE:

$$dX_s(\pi) = b_s(X_s(\pi), \pi_s) ds + \sigma_s(X_s(\pi), \pi_s) dW_s, \quad s \in [0, T], \quad X_0(\pi) = x \in \mathbb{R},$$

where

$$b_t(x, m) := \int \tilde{b}_t(x, a)m(a) da, \quad \sigma_t(x, m) := \sqrt{\int \tilde{\sigma}_t^2(x, a)m(a) da},$$

for some functions $\tilde{b}, \tilde{\sigma}$ for which we specify in Assumption 4.2.8. Consider the gain functional

$$J^\tau(\pi) := \mathbb{E} \left[\int_0^T [f_s(X_s(\pi), \pi_s) + \tau \text{Ent}(\pi_s | e^{-U})] ds + g(X_T(\pi)) \right],$$

where

$$f_t(x, m) := \int \tilde{f}_t(x, a)m(a) da,$$

for some function $\tilde{f}$ which will be specified in Assumption 4.2.8 as well. This setting is similar to one in [53], but in finite horizon case. In order to apply Algorithm 4, we need to check Assumptions 4.2.1 and 4.2.2 are satisfied.

Assumption 4.2.8. There exists $K \geq 0$ such that $\forall x \in \mathbb{R}^d, \forall a \in \mathbb{R}^p, \forall t \in [0, T]$

$$|\tilde{b}_t(x, a)| + |D_x \tilde{b}_t(x, a)| + |\tilde{\sigma}_t^2(x, a)| + |D_x \tilde{\sigma}_t^2(x, a)| + |\tilde{f}_t(x, a)| + |D_x \tilde{f}_t(x, a)| \leq K$$

and

$$|\tilde{\sigma}_t^2(x, a)| \geq K^{-1}.$$

Moreover, assume that $\tilde{b}, \tilde{\sigma}^2, \tilde{f}, D_x \tilde{b}, D_x \tilde{\sigma}$ and $D_x \tilde{f}$ are Lipschitz continuous with respect to x uniformly in $(t, a) \in [0, T] \times \mathbb{R}^p$.

Then, by Assumption 4.2.8 we have that $\forall x, x' \in \mathbb{R}, m, m' \in \mathcal{P}_2(\mathbb{R}^p)$

$$\begin{aligned} |b_t(x, m) - b_t(x', m')| &\leq |b_t(x, m) - b_t(x, m')| + |b_t(x, m') - b_t(x', m')| \\ &= \left| \int \tilde{b}_t(x, a)(m - m')(a) da \right| + \left| \int (\tilde{b}_t(x, a) - \tilde{b}_t(x', a))m'(a) da \right| \\ &\leq K\|m - m'\|_{\text{TV}} + K|x - x'|, \end{aligned}$$

and

$$\begin{aligned} |\sigma_t(x, m) - \sigma_t(x', m')| &\leq |\sigma_t(x, m) - \sigma_t(x, m')| + |\sigma_t(x, m') - \sigma_t(x', m')| \\ &= \frac{|\int \tilde{\sigma}_t^2(x, a)(m - m')(a) da|}{\sigma_t(x, m) + \sigma_t(x, m')} + \frac{|\int (\tilde{\sigma}_t^2(x, a) - \tilde{\sigma}_t^2(x', a))m'(a) da|}{\sigma_t(x, m') + \sigma_t(x', m')} \\ &\leq K\|m - m'\|_{\text{TV}} + K|x - x'|. \end{aligned}$$

Moreover,

$$\begin{aligned} |D_x b_t(x, m) - D_x b_t(x', m')| &\leq \left| \int D_x \tilde{b}_t(x, a)(m - m')(a) da \right| + \left| \int (D_x \tilde{b}_t(x, a) - D_x \tilde{b}_t(x', a))m'(a) da \right| \\ &\leq K\|m - m'\|_{\text{TV}} + K|x - x'|, \end{aligned}$$

and

$$\begin{aligned}
& |D_x \sigma_t(x, m) - D_x \sigma_t(x', m')| \\
& \leq \left| \frac{\int D_x \tilde{\sigma}_t(x, a) m(a) da}{\sigma(x, m)} - \frac{\int D_x \tilde{\sigma}_t(x', a) m(a) da}{\sigma(x', m)} \right| \\
& \quad + \left| \frac{\int D_x \tilde{\sigma}_t(x', a) m(a) da}{\sigma(x', m)} - \frac{\int D_x \tilde{\sigma}_t(x', a) m'(a) da}{\sigma(x', m')} \right| \\
& \leq K \|m - m'\|_{\text{TV}} + K |x - x'|.
\end{aligned}$$

One can also observe that the linear functional derivative of b and σ are

$$\frac{\delta b_t}{\delta m}(x, m, a) = \tilde{b}_t(x, a), \forall t \in [0, T], x \in \mathbb{R}, a \in \mathbb{R}^p, m \in \mathcal{P}_2(\mathbb{R}^p),$$

$$\frac{\delta \sigma_t}{\delta m}(x, m, a) = \frac{\tilde{\sigma}_t^2(x, a)}{2\sigma_t(x, m)}, \forall t \in [0, T], x \in \mathbb{R}, a \in \mathbb{R}^p, m \in \mathcal{P}_2(\mathbb{R}^p),$$

and under Assumption 4.2.8 $\frac{\delta b}{\delta m}$ and $\frac{\delta \sigma}{\delta m}$ satisfy Assumption 4.2.1. Similarly, we can show that f satisfy Assumption 4.2.2 under Assumption 4.2.8 and we assume that g satisfies Assumption 4.2.2. Hence, we observe that

$$\frac{\delta H_t^0}{\delta m}(x, y, z, m, a) = \tilde{b}_t(x, a) y + \tilde{f}_t(x, a) + \frac{\tilde{\sigma}_t^2(x, a)}{2\sigma_t(x, m)}.$$

Therefore, Algorithm 4 can be applied in the described setting to relaxed control problems.

In the case when the diffusion coefficient σ does not depend on the control variable, we have that the Hamiltonian is linear in control variable. Thus, we have that

$$\frac{\delta H_t^0}{\delta m}(x, y, z, m, a) = \tilde{b}_t(x, a) y + \tilde{f}_t(x, a).$$

Hence, the nonlinear minimization problem (4.6) becomes linear, and the algorithm can be rewritten as in Algorithm 5 below.

4.3 Proofs

In this section, we provide proofs of the results above. We start by defining necessary notations. We then show some auxiliary results which are needed to prove Lemma 4.2.5. Finally, we prove Theorem 4.2.6.

By $\mathbb{E}_t$ we denote the conditional expectation with respect to $\mathcal{F}_t$. By $\|\cdot\|_{L^\infty}$ we denote the norm in $L^\infty(\Omega)$. Let $\|Z\|_{\mathbb{H}^\infty} := \text{ess sup}_{(t, \omega)} |Z_t(\omega)|$ for any progressively measurable Z . For a uniformly integrable martingale M with $M_0 = 0$ we set

$$\|M\|_{\text{BMO}} = \sup_{\tau} \left\| (\mathbb{E}_\tau [\langle M \rangle_T - \langle M \rangle_\tau])^{\frac{1}{2}} \right\|_{L^\infty},$$

where $\langle M \rangle$ is quadratic variation, and the supremum is taken over all stopping times $\tau \in [0, T]$. Let Z be an $\mathbb{F}$ -adapted process and denote by $Z \cdot W$ the stochastic

Algorithm 5 Regularized Method of Successive Approximations for relaxed controls:

Initialization: make a guess of the control $\pi^0 = (\pi_s^0)_{s \in [0, T]}$.

while difference between $J^\tau(\pi^{n+1})$ and $J^\tau(\pi^n)$ is large **do**

 Given a control $\pi^n = (\pi_s^n)_{s \in [0, T]}$ solve the following forward SDE, then solve backward SDE:

$$\begin{aligned} dX_s(\pi^n) &= b_s(X_s(\pi^n), \pi_s^n) ds + \sigma_s(X_s(\pi^n)) dW_s, \quad X_0(\pi^n) = x, \\ dY_s(\pi^n) &= -D_x H_s^0(\Theta_s(\pi^n), \pi_s^n) ds + Z_s(\pi^n) dW_s, \\ Y_T(\pi^n) &= D_x g(X_T(\pi^n)). \end{aligned}$$

Update the control $\forall s \in [0, T]$

$$\pi_s^{n+1}(a) = (\mathcal{Z}_s^{n+1})^{-1}(\pi_s^n(a)) \frac{\lambda}{\lambda + \tau} e^{-\frac{\tau}{\lambda + \tau} U(a)} e^{-\frac{1}{\lambda + \tau} [\tilde{b}_s(X_s(\pi^n), a) Y_s(\pi^n) + \tilde{f}_s(X_s(\pi^n), a)]}, \quad (4.10)$$

where

$$\mathcal{Z}_s^{n+1} := \int (\pi_s^n(\tilde{a}))^{\frac{\lambda}{\lambda + \tau}} e^{-\frac{\tau}{\lambda + \tau} U(\tilde{a})} e^{-\frac{1}{\lambda + \tau} [\tilde{b}_s(X_s(\pi^n), \tilde{a}) Y_s(\pi^n) + \tilde{f}_s(X_s(\pi^n), \tilde{a})]} d\tilde{a}.$$

end while

return π^{n+1} .

integral of Z with respect to the Brownian motion W , that is, $(Z \cdot W)_t := \int_0^t Z_s dW_s$ for $t \in [0, T]$. We understand the following as $D_x \sigma = D_{x_l} \sigma^{ij}$, where $i, l = 1, 2, \dots, d$ and $j = 1, 2, \dots, d'$. By $Z^\top$ we denote the transpose of Z .

We start working towards the proof of Theorem 4.2.6 by recalling the adjoint equation for an admissible control $\pi \in \mathcal{A}$:

$$\begin{aligned} dY_s(\pi) &= -D_x H_s^0(\Theta_s(\pi), \pi_s) ds + Z_s(\pi) dW_s, \quad s \in [0, T], \\ Y_T(\pi) &= D_x g(X_T(\pi)). \end{aligned} \quad (4.11)$$

From now on, we shall use Einstein notation, so that repeated indices in a single term imply summation over all the values of that index.

Lemma 4.3.1. *Assume that there exists $K \geq 0$ such that $\forall x \in \mathbb{R}^d, \forall m \in \mathcal{P}_2(\mathbb{R}^p), \forall t \in [0, T]$ we have*

$$|D_x b_t(x, m)| + |D_x \sigma_t(x, m)| \leq K,$$

and

$$|D_x g(x)| + |D_x f_t(x, m)| \leq K.$$

Then $\sup_{\pi \in \mathcal{A}} \|Y(\pi)\|_{\mathbb{H}^\infty}$ and $\sup_{\pi \in \mathcal{A}} \|Z(\pi) \cdot W\|_{\text{BMO}}$ are bounded.

Proof. Step 1. Bound for $Y(\pi)$. From the definition of the Hamiltonian (3.3) we

have

$$\begin{aligned} D_{x_i} H_s^0(\Theta_s(\pi), \pi_s) &= (D_{x_i} b_s^j)(X_s(\pi), \pi_s)(Y_s(\pi))^j + (D_{x_i} \sigma_s^{jp})(X_s(\pi), \pi_s)(Z_s(\pi))^{jp} \\ &\quad + (D_{x_i} f_s)(X_s(\pi), \pi_s), \quad \forall s \in [0, T], \quad i = 1, 2, \dots, d. \end{aligned}$$

Hence, one can observe that (4.11) is a linear BSDE. Therefore, from [47, Proposition 3.2] we can write the explicit formula for the solution of (4.11):

$$Y_t(\pi) = \mathbb{E}_t \left[S_t^{-1} S_T (D_x g)(X_T(\pi)) + \int_t^T S_t^{-1} S_s (D_x f_s)(X_s(\pi), \pi_s) ds \right],$$

where the process S is the unique strong solution of

$$\begin{aligned} dS_t^{ij} &= S_t^{il} (D_{x_l} b_t^j)(X_t(\pi), \pi_t) dt + S_t^{il} (D_{x_l} \sigma_t^{jp})(X_t(\pi), \pi_t) dW_t^p, \quad i, j = 1, 2, \dots, d, \\ S_0 &= I_d, \end{aligned}$$

and S^{-1} is the inverse process of S . Thus, due to [47, Corollary 3.7], the fact that $\mathbb{E}_t |S_t^{-1} S_T| < \infty, \forall t \in [0, T]$, and assumptions of the lemma we have the following bound: $\exists C > 0$ s.t. $\forall \pi \in \mathcal{A}$

$$\|Y(\pi)\|_{\mathbb{H}^\infty} \leq C \| (D_x g)(X_T(\pi)) \|_{L^\infty} + CT \| (D_x f)(X(\pi), \pi) \|_{\mathbb{H}^\infty}.$$

Hence, due to assumptions of the lemma we conclude that $\sup_{\pi \in \mathcal{A}} \|Y(\pi)\|_{\mathbb{H}^\infty} < \infty$.

Step 2. Bound on $Z(\pi)$. First we show that $\exists C > 0$ s.t. $\forall \pi \in \mathcal{A}$ and $\forall s \in [0, T]$ $\mathbb{E}_s \int_s^T |Z_t(\pi)|^2 dt \leq C$. By Itô's formula we obtain

$$\begin{aligned} \int_s^T |Z_t(\pi)|^2 dt &\leq | (D_x g)(X_T(\pi)) |^2 - 2 \int_s^T Z_t(\pi)^\top Y_t(\pi) dW_t \\ &\quad + \int_s^T 2Y_t(\pi) (D_x H_t^0)(\Theta_t(\pi), \pi_t) dt. \end{aligned}$$

Then, by taking the conditional expectation we get that

$$\begin{aligned} \mathbb{E}_s \int_s^T |Z_t(\pi)|^2 dt &\leq \mathbb{E}_s | (D_x g)(X_T(\pi)) |^2 + \mathbb{E}_s \int_s^T \left[2Y_t(\pi) \right. \\ &\quad \times \left. \left((D_x b_t)(X_t(\pi), \pi_t) Y_t(\pi) + (D_x \sigma_t)(X_t(\pi), \pi_t) Z_t(\pi) + (D_x f_t)(X_t(\pi), \pi_t) \right) \right] dt. \end{aligned}$$

From assumptions of the lemma, since $\sup_{\pi \in \mathcal{A}} \|Y(\pi)\|_{\mathbb{H}^\infty} < C$ and by Young's inequality we have

$$\begin{aligned} \mathbb{E}_s \int_s^T |Z_t(\pi)|^2 dt &\leq \| (D_x g)(X_T(\pi)) \|_{L^\infty}^2 + \mathbb{E}_s \int_s^T \left[2\|Y(\pi)\|_{\mathbb{H}^\infty}^2 \| (D_x b)(X(\pi), \pi) \|_{\mathbb{H}^\infty} \right. \\ &\quad + \| (D_x \sigma)(X(\pi), \pi) \|_{\mathbb{H}^\infty} (\gamma^{-1} \|Y(\pi)\|_{\mathbb{H}^\infty}^2 + \gamma |Z_t(\pi)|^2) \\ &\quad \left. + 2\|Y(\pi)\|_{\mathbb{H}^\infty} \| (D_x f)(X(\pi), \pi) \|_{\mathbb{H}^\infty} \right] dt. \end{aligned}$$

By choosing γ such that $0 < 1 - K\gamma < 1$, we conclude that there is $C > 0$ such that $\sup_{\pi \in \mathcal{A}} \mathbb{E}_s \int_s^T |Z_t(\pi)|^2 dt \leq C$ for all $s \in [0, T]$. Notice that the constant C

does not depend on π and s . Hence, by the definition of BMO norm we have

$$\sup_{\pi \in \mathcal{A}} \|Z(\pi) \cdot W\|_{\text{BMO}} = \sup_{\pi \in \mathcal{A}} \sup_{\tau} \left\| \left(\mathbb{E}_{\tau} \left[\int_{\tau}^T |Z_t(\pi)|^2 dt \right] \right)^{1/2} \right\| \leq C$$

where the supremum is taken over all stopping times $\tau \in [0, T]$ $\square$

Let us recall the well-known result that the space BMO can be identified with the dual space, a space of all martingales N such that $\sup_{0 \leq t \leq T} |N_t| \in L^1(\Omega)$.

Theorem 4.3.2 (Fefferman's inequality). *If $M \in \text{BMO}$ and N is a martingale such that $\sup_{0 \leq t \leq T} |N_t| \in L^1(\Omega)$, then*

$$\mathbb{E} \left[\int_0^T |d\langle M, N \rangle_t| \right] \leq \sqrt{2} \|M\|_{\text{BMO}} \mathbb{E} \left[\langle N \rangle_T^{\frac{1}{2}} \right].$$

Proof. See [56, Th 2.5]. $\square$

We will use Fefferman's inequality in the proof of Lemma 4.3.3 and later.

Lemma 4.3.3. *Let Assumption 4.2.1 and 4.2.2 hold. Then, there exists a constant C such that for any $\pi, \pi' \in \mathcal{A}$*

$$\mathbb{E} \sup_{0 \leq s \leq T} |Y_s(\pi) - Y_s(\pi')|^2 + \mathbb{E} \int_0^T |Z_t(\pi) - Z_t(\pi')|^2 dt \leq C \mathbb{E} \int_0^T \|\pi_t - \pi'_t\|_{\text{TV}}^2 dt.$$

Proof. Given a constant β to be fixed later, we compute by Itô's formula that:

$$\begin{aligned} e^{\beta s} |Y_s(\pi) - Y_s(\pi')|^2 &= e^{\beta T} |(D_x g)(X_T(\pi)) - (D_x g)(X_T(\pi'))|^2 \\ &+ 2 \int_s^T e^{\beta t} (Z_t(\pi) - Z_t(\pi'))^\top (Y_t(\pi) - Y_t(\pi')) dW_t + \int_s^T e^{\beta t} \left(2(Y_t(\pi) - Y_t(\pi')) \right. \\ &\quad \times \left(- (D_x H_t^0)(\Theta_t(\pi), \pi_t) + (D_x H_t^0)(\Theta_t(\pi'), \pi'_t) \right) \\ &\quad \left. - |Z_t(\pi) - Z_t(\pi')|^2 - \beta |Y_t(\pi) - Y_t(\pi')|^2 \right) dt. \end{aligned}$$

By taking the conditional expectation we get

$$\begin{aligned} e^{\beta s} |Y_s(\pi) - Y_s(\pi')|^2 &+ \mathbb{E}_s \int_s^T e^{\beta t} |Z_t(\pi) - Z_t(\pi')|^2 dt \\ &= \mathbb{E}_s e^{\beta T} |(D_x g)(X_T(\pi)) - (D_x g)(X_T(\pi'))|^2 + \mathbb{E}_s \int_s^T e^{\beta t} \left(2(Y_t(\pi) - Y_t(\pi')) \right. \\ &\quad \times \left(- (D_x H_t^0)(\Theta_t(\pi), \pi_t) + (D_x H_t^0)(\Theta_t(\pi'), \pi'_t) \right) - \beta |Y_t(\pi) - Y_t(\pi')|^2 \Big) dt. \end{aligned} \tag{4.12}$$

Note that

$$\begin{aligned}
& - (D_x H_t^0) (\Theta_t(\pi), \pi_t) + (D_x H_t^0) (\Theta_t(\pi'), \pi'_t) \\
& = - (D_x b_t) (X_t(\pi), \pi_t) (Y_t(\pi) - Y_t(\pi')) \\
& \quad + Y_t(\pi') \left((D_x b_t) (X_t(\pi'), \pi'_t) - (D_x b_t) (X_t(\pi), \pi_t) \right) \\
& \quad - (D_x \sigma_t) (X_t(\pi), \pi_t) (Z_t(\pi) - Z_t(\pi')) \\
& \quad + Z_t(\pi') \left((D_x \sigma_t) (X_t(\pi'), \pi'_t) - (D_x \sigma_t) (X_t(\pi), \pi_t) \right) \\
& \quad - (D_x f_t) (X_t(\pi), \pi_t) + (D_x f_t) (X_t(\pi'), \pi'_t).
\end{aligned}$$

Therefore, by Assumption 4.2.1 and 4.2.2 and by Young's inequality we get that

$$\begin{aligned}
& \mathbb{E}_s \int_s^T 2e^{\beta t} (Y_t(\pi) - Y_t(\pi')) (- (D_x H_t^0) (\Theta_t(\pi), \pi_t) + (D_x H_t^0) (\Theta_t(\pi'), \pi'_t)) dt \\
& \leq \mathbb{E}_s \int_s^T 2e^{\beta t} \| (D_x b) (X(\pi), \pi) \|_{\mathbb{H}^\infty} |Y_t(\pi) - Y_t(\pi')|^2 dt \\
& \quad + \mathbb{E}_s \int_s^T e^{\beta t} \| (D_x \sigma) (X(\pi), \pi) \|_{\mathbb{H}^\infty} (\gamma |Z_t(\pi) - Z_t(\pi')|^2 + \gamma^{-1} |Y_t(\pi) - Y_t(\pi')|^2) dt \\
& \quad + \mathbb{E}_s \int_s^T e^{\beta t} \left(\gamma^{-1} |Y_t(\pi) - Y_t(\pi')|^2 + \gamma \left((D_x f_t) (X_t(\pi'), \pi'_t) - (D_x f_t) (X_t(\pi), \pi_t) \right)^2 \right) dt \\
& \quad + \mathbb{E}_s \int_s^T 2e^{\beta t} (Y_t(\pi) - Y_t(\pi')) Z_t(\pi') \left((D_x \sigma_t) (X_t(\pi'), \pi'_t) - (D_x \sigma_t) (X_t(\pi), \pi_t) \right) dt \\
& \quad + \mathbb{E}_s \int_s^T 2e^{\beta t} (Y_t(\pi) - Y_t(\pi')) Y_t(\pi') \left((D_x b_t) (X_t(\pi'), \pi'_t) - (D_x b_t) (X_t(\pi), \pi_t) \right) dt \\
& := I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned} \tag{4.13}$$

From Lemma 4.3.1 and Fefferman's inequality, see Theorem 4.3.2, we have that

$$\begin{aligned}
I_4 & = \mathbb{E}_s \int_s^T e^{\beta t} 2(Y_t(\pi) - Y_t(\pi')) Z_t(\pi') \left((D_x \sigma_t) (X_t(\pi'), \pi'_t) - (D_x \sigma_t) (X_t(\pi), \pi_t) \right) dt \\
& \leq 2\mathbb{E} \int_0^T e^{\beta t} |Z_t(\pi')| |Y_t(\pi) - Y_t(\pi')| \left| (D_x \sigma_t) (X_t(\pi'), \pi'_t) - (D_x \sigma_t) (X_t(\pi), \pi_t) \right| dt \\
& = 2\mathbb{E} \int_0^T \left| d \left\langle Z(\pi') \cdot W, e^{\beta \cdot} |Y(\pi) - Y(\pi')| \left| (D_x \sigma) (X(\pi'), \pi') - (D_x \sigma) (X(\pi), \pi) \right| \cdot W \right\rangle_t \right| \\
& \leq 2\sqrt{2} \|Z(\pi') \cdot W\|_{\text{BMO}} \mathbb{E} \left[\left\langle e^{\beta \cdot} |Y(\pi) - Y(\pi')| \left| (D_x \sigma) (X(\pi'), \pi') - (D_x \sigma) (X(\pi), \pi) \right| \cdot W \right\rangle_T^{\frac{1}{2}} \right] \\
& \leq 2\sqrt{2} C \mathbb{E} \left[\left(\int_0^T e^{2\beta t} |Y_t(\pi) - Y_t(\pi')|^2 | (D_x \sigma_t) (X_t(\pi'), \pi'_t) - (D_x \sigma_t) (X_t(\pi), \pi_t) |^2 dt \right)^{\frac{1}{2}} \right].
\end{aligned}$$

Applying Young's inequality we obtain

$$\begin{aligned}
I_4 &\leq \sqrt{2}C\mathbb{E}\left[\gamma \sup_{0\leq t\leq T} e^{\beta t}|Y_t(\pi) - Y_t(\pi')|^2 \right. \\
&\quad \left. + \gamma^{-1} \int_0^T e^{\beta t} |(D_x \sigma_t)(X_t(\pi'), \pi'_t) - (D_x \sigma_t)(X_t(\pi), \pi_t)|^2 dt \right] \\
&\leq \sqrt{2}C\mathbb{E}\left[\gamma \sup_{0\leq t\leq T} e^{\beta t}|Y_t(\pi) - Y_t(\pi')|^2 \right. \\
&\quad \left. + 2\gamma^{-1}K^2 \int_0^T e^{\beta t} (|X_t(\pi) - X_t(\pi')|^2 + \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt \right]. \tag{4.14}
\end{aligned}$$

Again, from Lemma 4.3.1 and Young's inequality we get that

$$\begin{aligned}
I_5 &= \mathbb{E}_s \int_s^T e^{\beta t} 2(Y_t(\pi) - Y_t(\pi'))Y_t(\pi') \left((D_x b_t)(X_t(\pi'), \pi'_t) - (D_x b_t)(X_t(\pi), \pi_t) \right) dt \\
&\leq \|Y(\pi')\|_{\mathbb{H}^\infty} \mathbb{E}_s \int_s^T e^{\beta t} 2|Y_t(\pi) - Y_t(\pi')| \left| (D_x b_t)(X_t(\pi'), \pi'_t) - (D_x b_t)(X_t(\pi), \pi_t) \right| dt \\
&\leq C\mathbb{E}_s \int_s^T e^{\beta t} (\gamma^{-1}|Y_t(\pi) - Y_t(\pi')|^2 + 2\gamma K^2 |X_t(\pi) - X_t(\pi')|^2 + 2\gamma K^2 \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt. \tag{4.15}
\end{aligned}$$

Therefore, combining (4.12), (4.13), (4.15) and (4.14), and by Assumption 4.2.1 we obtain

$$\begin{aligned}
&e^{\beta s}|Y_s(\pi) - Y_s(\pi')|^2 + \mathbb{E}_s \int_s^T e^{\beta t}|Z_t(\pi) - Z_t(\pi')|^2 dt \\
&\leq K^2 e^{\beta T} \mathbb{E}_s |X_T(\pi) - X_T(\pi')|^2 + \mathbb{E}_s \int_s^T e^{\beta t} (2K - \beta) |Y_t(\pi) - Y_t(\pi')|^2 dt \\
&\quad + \mathbb{E}_s \int_s^T e^{\beta t} K (\gamma |Z_t(\pi) - Z_t(\pi')|^2 + \gamma^{-1} |Y_t(\pi) - Y_t(\pi')|^2) dt \\
&\quad + \mathbb{E}_s \int_s^T e^{\beta t} \left(\gamma^{-1} |Y_t(\pi) - Y_t(\pi')|^2 + 2\gamma K^2 (|X_t(\pi) - X_t(\pi')|^2 + \|\pi_t - \pi'_t\|_{\text{TV}}^2) \right) dt \\
&\quad + \sqrt{2}C\mathbb{E}\left[\gamma \sup_{0\leq t\leq T} e^{\beta t}|Y_t(\pi) - Y_t(\pi')|^2 \right. \\
&\quad \left. + 2\gamma^{-1}K^2 \int_0^T e^{\beta t} (|X_t(\pi) - X_t(\pi')|^2 + \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt \right] \\
&\quad + C\mathbb{E}_s \int_s^T e^{\beta t} (\gamma^{-1}|Y_t(\pi) - Y_t(\pi')|^2 + 2\gamma K^2 |X_t(\pi) - X_t(\pi')|^2 + 2\gamma K^2 \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt.
\end{aligned}$$

By choosing $\beta = 2K + (K + C + 1)\gamma^{-1}$ we get

$$\begin{aligned}
& e^{\beta s} |Y_s(\pi) - Y_s(\pi')|^2 + \mathbb{E}_s \int_s^T e^{\beta t} |Z_t(\pi) - Z_t(\pi')|^2 dt \\
& \leq K^2 e^{\beta T} \mathbb{E}_s |X_T(\pi) - X_T(\pi')|^2 + K\gamma \mathbb{E}_s \int_s^T e^{\beta t} |Z_t(\pi) - Z_t(\pi')|^2 dt \\
& + 2\gamma K^2 (1 + C) \mathbb{E}_s \int_s^T e^{\beta t} (|X_t(\pi) - X_t(\pi')|^2 + \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt \\
& + \sqrt{2}C \mathbb{E} \left[\gamma \sup_{0 \leq t \leq T} e^{\beta t} |Y_t(\pi) - Y_t(\pi')|^2 \right. \\
& \quad \left. + 2\gamma^{-1} K^2 \int_0^T e^{\beta t} (|X_t(\pi) - X_t(\pi')|^2 + \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt \right]. \tag{4.16}
\end{aligned}$$

By letting $s = 0$ and by choosing γ such that $0 < 1 - K\gamma < 1$ we have

$$\begin{aligned}
& (1 - K\gamma) \mathbb{E} \int_0^T e^{\beta t} |Z_t(\pi) - Z_t(\pi')|^2 dt \\
& \leq K^2 e^{\beta T} \mathbb{E} |X_T(\pi) - X_T(\pi')|^2 + \sqrt{2}C \mathbb{E} \left[\gamma \sup_{0 \leq t \leq T} e^{\beta t} |Y_t(\pi) - Y_t(\pi')|^2 \right] \\
& + 2K^2 (2\gamma(1 + C) + \sqrt{2}C\gamma^{-1}) \mathbb{E} \int_0^T e^{\beta t} (|X_t(\pi) - X_t(\pi')|^2 + \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt. \tag{4.17}
\end{aligned}$$

By taking the supremum over $s \in [0, T]$ in (4.16), and, additionally, choosing γ such that $0 < 1 - \sqrt{2}\gamma C < 1$ hold we get that there is $C_1 > 0$ such that

$$\begin{aligned}
& \mathbb{E} \sup_{0 \leq s \leq T} |Y_s(\pi) - Y_s(\pi')|^2 \leq C_1 \mathbb{E} |X_T(\pi) - X_T(\pi')|^2 \\
& + C_1 \mathbb{E} \int_0^T (|X_t(\pi) - X_t(\pi')|^2 + \|\pi_t - \pi'_t\|_{\text{TV}}^2) dt. \tag{4.18}
\end{aligned}$$

By Lemma C.2.1 we have that

$$\mathbb{E} \sup_{0 \leq t \leq T} |X_t(\pi) - X_t(\pi')|^2 \leq C_2 \mathbb{E} \int_0^T \|\pi_t - \pi'_t\|_{\text{TV}}^2 dt. \tag{4.19}$$

Hence, combining this and (4.18) we have that there exists $C_3 > 0$ such that

$$\mathbb{E} \sup_{0 \leq s \leq T} |Y_s(\pi) - Y_s(\pi')|^2 \leq C_3 \mathbb{E} \int_0^T \|\pi_t - \pi'_t\|_{\text{TV}}^2 dt.$$

Substituting this, 4.19 and 4.18 into 4.17 we conclude the proof. $\square$

Let us introduce the following result, where we derive an expression of the Gâteaux derivative of J^0 in terms of the functional derivative of the Hamiltonian.

Lemma 4.3.4. *Under Assumption 4.2.1 and 4.2.2, for $\pi, \pi' \in \mathcal{A}$ we have that*

$$\frac{d}{d\varepsilon} J^0(\pi + \varepsilon(\pi' - \pi))|_{\varepsilon=0} = \mathbb{E} \int_0^T \left[\int \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s, a)(\pi'_s - \pi_s)(a) da \right] ds.$$

Proof. See [45, Lemma 3.5]. □

Proof of Lemma 4.2.5. Step 1. Let $\pi, \pi' \in \mathcal{A}$ be admissible controls. We first obtain the expression for the difference $J^\tau(\pi') - J^\tau(\pi)$. To begin with, observe that by the fundamental theorem of calculus we have

$$J^0(\pi') - J^0(\pi) = \int_0^1 \lim_{\delta \rightarrow 0+} \frac{J^0(\pi^{\varepsilon+\delta}) - J^0(\pi^\varepsilon)}{\delta} d\varepsilon,$$

where $\pi^\varepsilon = \pi + \varepsilon(\pi' - \pi)$ for $\varepsilon \in (0, 1)$ and $\pi^{\varepsilon+\delta} = \pi + (\varepsilon + \delta)(\pi' - \pi)$ for $\delta \in (0, 1 - \varepsilon)$. Therefore, from Lemma 4.3.4 we have

$$\lim_{\delta \rightarrow 0+} \frac{J^0(\pi^{\varepsilon+\delta}) - J^0(\pi^\varepsilon)}{\delta} = \mathbb{E} \int_0^T \left[\int \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^\varepsilon), \pi_s^\varepsilon, a)(\pi'_s - \pi_s)(a) da \right] ds.$$

Hence,

$$\begin{aligned} J^\tau(\pi') - J^\tau(\pi) &= \int_0^1 \mathbb{E} \int_0^T \left[\int \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^\varepsilon), \pi_s^\varepsilon, a)(\pi'_s - \pi_s)(a) da \right] ds d\varepsilon \\ &\quad + \tau \mathbb{E} \int_0^T [\text{Ent}(\pi'_s | e^{-U}) - \text{Ent}(\pi_s | e^{-U})] ds \\ &= \int_0^1 \mathbb{E} \int_0^T \left[\int \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s^\varepsilon, a)(\pi'_s - \pi_s)(a) da \right] ds d\varepsilon \\ &\quad + \int_0^1 \mathbb{E} \int_0^T \left[\int \left(\frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s^\varepsilon, a) \right) (\pi'_s - \pi_s)(a) da \right] ds d\varepsilon \\ &\quad + \tau \mathbb{E} \int_0^T [\text{Ent}(\pi'_s | e^{-U}) - \text{Ent}(\pi_s | e^{-U})] ds. \end{aligned} \tag{4.20}$$

Notice that the definition of the flat (functional) derivative, see C.1.1, we have that

$$\begin{aligned} &\int_0^1 \mathbb{E} \int_0^T \left[\int \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s^\varepsilon, a)(\pi'_s - \pi_s)(a) da \right] ds d\varepsilon \\ &= \mathbb{E} \int_0^T [H_s^0(\Theta_s(\pi), \pi'_s) - H_s^0(\Theta_s(\pi), \pi_s)] ds. \end{aligned}$$

Therefore, (4.20) becomes

$$\begin{aligned}
J^\tau(\pi') - J^\tau(\pi) &= \mathbb{E} \int_0^T \left[H_s^0(\Theta_s(\pi), \pi'_s) - H_s^0(\Theta_s(\pi), \pi_s) \right] ds \\
&+ \int_0^1 \mathbb{E} \int_0^T \left[\int \left(\frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s^\varepsilon, a) \right) (\pi'_s - \pi_s)(a) da \right] ds d\varepsilon \\
&+ \tau \mathbb{E} \int_0^T [\text{Ent}(\pi'_s | e^{-U}) - \text{Ent}(\pi_s | e^{-U})] ds \\
&= \mathbb{E} \int_0^T \left[H_s^{\tau,0}(\Theta_s(\pi), \pi'_s) - H_s^{\tau,0}(\Theta_s(\pi), \pi_s) \right] ds \\
&+ \int_0^1 \mathbb{E} \int_0^T \left[\int \left(\frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s^\varepsilon, a) \right) (\pi'_s - \pi_s)(a) da \right] ds d\varepsilon.
\end{aligned} \tag{4.21}$$

Step 2. The next step is to estimate the second term in (4.21). By Assumption 4.2.1 and 4.2.2 and Young's inequality we get

$$\begin{aligned}
&\int \left[\frac{\delta H_s^0}{\delta \pi}(\Theta_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s^\varepsilon, a) \right] (\pi'_s - \pi_s)(a) da \\
&= \int \left[\left(\frac{\delta b_s}{\delta m}(X_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta b_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) \right) Y_s(\pi^\varepsilon) \right. \\
&\quad + \frac{\delta b_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) (Y_s(\pi^\varepsilon) - Y_s(\pi)) \\
&\quad + \left(\frac{\delta \sigma_s}{\delta m}(X_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta \sigma_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) \right) Z_s(\pi^\varepsilon) \\
&\quad + \frac{\delta \sigma_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) (Z_s(\pi^\varepsilon) - Z_s(\pi)) \\
&\quad \left. + \frac{\delta f_s}{\delta m}(X_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta f_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) \right] (\pi'_s - \pi_s)(a) da \\
&\leq \int \left[\frac{\delta b_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) (Y_s(\pi^\varepsilon) - Y_s(\pi)) + \frac{\delta \sigma_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) (Z_s(\pi^\varepsilon) - Z_s(\pi)) \right] \\
&\quad \times (\pi'_s - \pi_s)(a) da \\
&\quad + \sup_a \left| \frac{\delta b_s}{\delta m}(X_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta b_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) \right| \cdot |Y_s(\pi^\varepsilon)| \cdot \|\pi'_s - \pi_s\|_{\text{TV}} \\
&\quad + \sup_a \left| \frac{\delta \sigma_s}{\delta m}(X_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta \sigma_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) \right| \cdot |Z_s(\pi^\varepsilon)| \cdot \|\pi'_s - \pi_s\|_{\text{TV}} \\
&\quad + \sup_a \left| \frac{\delta f_s}{\delta m}(X_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta f_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) \right| \cdot \|\pi'_s - \pi_s\|_{\text{TV}} \\
&\leq \int \left[\frac{\delta b_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) (Y_s(\pi^\varepsilon) - Y_s(\pi)) + \frac{\delta \sigma_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a) (Z_s(\pi^\varepsilon) - Z_s(\pi)) \right] \\
&\quad \times (\pi'_s - \pi_s)(a) da \\
&\quad + K |X_s(\pi^\varepsilon) - X_s(\pi)| \cdot \|\pi'_s - \pi_s\|_{\text{TV}} (|Y_s(\pi^\varepsilon)| + |Z_s(\pi^\varepsilon)| + 1).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
& \int_0^1 \mathbb{E} \int_0^T \left[\int \left(\frac{\delta H_s^0}{\delta m}(\Theta_s(\pi^\varepsilon), \pi_s^\varepsilon, a) - \frac{\delta H_s^0}{\delta m}(\Theta_s(\pi), \pi_s, a) \right) (\pi'_s - \pi_s)(a) da \right] ds d\varepsilon \\
& \leq \int_0^1 \mathbb{E} \int_0^T \int \left(\frac{\delta b_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a)(Y_s(\pi^\varepsilon) - Y_s(\pi)) \right. \\
& \quad \left. + \frac{\delta \sigma_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a)(Z_s(\pi^\varepsilon) - Z_s(\pi)) \right) (\pi'_s - \pi_s)(a) da ds d\varepsilon \\
& + \int_0^1 \mathbb{E} \int_0^T |Y_s(\pi^\varepsilon)| \left(\frac{K}{2} |X_s(\pi^\varepsilon) - X_s(\pi)|^2 + \frac{K}{2} \|\pi'_s - \pi_s\|_{\text{TV}}^2 \right) ds d\varepsilon \\
& + K \int_0^1 \mathbb{E} \int_0^T |Z_s(\pi^\varepsilon)| |X_s(\pi^\varepsilon) - X_s(\pi)| \cdot \|\pi'_s - \pi_s\|_{\text{TV}} ds d\varepsilon \\
& + \int_0^1 \mathbb{E} \int_0^T \left(\frac{K}{2} |X_s(\pi^\varepsilon) - X_s(\pi)|^2 + \frac{K}{2} \|\pi'_s - \pi_s\|_{\text{TV}}^2 \right) ds d\varepsilon =: I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Notice that by the Cauchy–Schwarz inequality we obtain

$$\begin{aligned}
I_1 & \leq \int_0^1 \left(\mathbb{E} \int_0^T |Y_s(\pi^\varepsilon) - Y_s(\pi)|^2 ds \right)^{1/2} \\
& \quad \times \left(\mathbb{E} \int_0^T \left| \int \frac{\delta b_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a)(\pi'_s - \pi_s)(a) da \right|^2 ds \right)^{1/2} d\varepsilon \\
& + \int_0^1 \left(\mathbb{E} \int_0^T |Z_s(\pi^\varepsilon) - Z_s(\pi)|^2 ds \right)^{1/2} \\
& \quad \times \left(\mathbb{E} \int_0^T \left| \int \frac{\delta \sigma_s}{\delta m}(X_s(\pi), \pi_s^\varepsilon, a)(\pi'_s - \pi_s)(a) da \right|^2 ds \right)^{1/2} d\varepsilon.
\end{aligned} \tag{4.22}$$

By Lemma 4.3.3 and by Assumption 4.2.1, (4.22) becomes

$$\begin{aligned}
I_1 & \leq 2 \int_0^1 \left(C \mathbb{E} \int_0^T \|\pi_s^\varepsilon - \pi_s\|_{\text{TV}}^2 ds \right)^{1/2} \left(C \mathbb{E} \int_0^T \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \right)^{1/2} d\varepsilon \\
& \leq 2 \int_0^1 \varepsilon \left(C \mathbb{E} \int_0^T \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \right) d\varepsilon,
\end{aligned}$$

since $\pi^\varepsilon = \pi + \varepsilon(\pi' - \pi)$. Also, recall Csiszár-Kullback-Pinsker inequality, see e.g. [57, 58, 59], for π' and π :

$$\mathbb{E} \int_0^T \|\pi'_t - \pi_t\|_{\text{TV}}^2 dt \leq 2 \mathbb{E} \int_0^T \text{Ent}(\pi'_t | \pi_t) dt.$$

Hence,

$$I_1 \leq C \mathbb{E} \int_0^T \text{Ent}(\pi'_t | \pi_t) dt.$$

Since $\sup_{\pi \in \mathcal{A}} \|Y_s(\pi)\|_{\mathbb{H}^\infty} < \infty$ due to Lemma 4.3.1 and by Lemma C.2.1 we have

$$\begin{aligned} I_2 &= \int_0^1 \mathbb{E} \int_0^T |Y_s(\pi^\varepsilon)| \left(\frac{K}{2} |X_s(\pi^\varepsilon) - X_s(\pi)|^2 + \frac{K}{2} \|\pi'_s - \pi_s\|_{\text{TV}}^2 \right) ds d\varepsilon \\ &\leq \frac{KC}{2} \int_0^1 \mathbb{E} \left[\sup_{0 \leq s \leq T} |X_s(\pi^\varepsilon) - X_s(\pi)|^2 + \int_0^T \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \right] d\varepsilon \\ &\leq C \mathbb{E} \int_0^T \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \leq C \mathbb{E} \int_0^T \text{Ent}(\pi'_s | \pi_s) ds. \end{aligned}$$

Similarly,

$$\begin{aligned} I_4 &= \int_0^1 \mathbb{E} \int_0^T \left(\frac{K}{2} |X_s(\pi^\varepsilon) - X_s(\pi)|^2 + \frac{K}{2} \|\pi'_s - \pi_s\|_{\text{TV}}^2 \right) ds d\varepsilon \\ &\leq C \mathbb{E} \int_0^T \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \leq C \mathbb{E} \int_0^T \text{Ent}(\pi'_s | \pi_s) ds. \end{aligned}$$

To estimate I_3 , we apply Fefferman's inequality, see Theorem 4.3.2, Lemma 4.3.1, Young's inequality and Lemma C.2.1:

$$\begin{aligned} I_3 &\leq K \int_0^1 \mathbb{E} \int_0^T |Z_s(\pi^\varepsilon)| |X_s(\pi^\varepsilon) - X_s(\pi)| \cdot \|\pi'_s - \pi_s\|_{\text{TV}} ds d\varepsilon \\ &= K \int_0^1 \mathbb{E} \int_0^T d \langle Z(\pi^\varepsilon) \cdot W, |X(\pi^\varepsilon) - X(\pi)| \cdot \|\pi' - \pi\|_{\text{TV}} \cdot W \rangle_s d\varepsilon \\ &\leq \sqrt{2}K \int_0^1 \|Z(\pi^\varepsilon) \cdot W\|_{\text{BMO}} \mathbb{E} \left(\int_0^T |X_s(\pi^\varepsilon) - X_s(\pi)|^2 \cdot \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \right)^{\frac{1}{2}} d\varepsilon \\ &\leq C \int_0^1 \mathbb{E} \sup_{0 \leq s \leq T} |X_s(\pi^\varepsilon) - X_s(\pi)|^2 d\varepsilon + C \mathbb{E} \int_0^T \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \\ &\leq C \mathbb{E} \int_0^T \|\pi'_s - \pi_s\|_{\text{TV}}^2 ds \leq C \mathbb{E} \int_0^T \text{Ent}(\pi'_s | \pi_s) ds. \end{aligned}$$

Step 3. Combining estimates for I_1, I_2, I_3 and I_4 , and substituting them into (4.21) we obtain that

$$\begin{aligned} J^\tau(\pi') - J^\tau(\pi) &\leq \mathbb{E} \int_0^T \left[H_s^{\tau,0}(\Theta_s(\pi), \pi'_s) - H_s^{\tau,0}(\Theta_s(\pi), \pi_s) \right] ds \\ &\quad + C \mathbb{E} \int_0^T \text{Ent}(\pi'_t | \pi_t) dt. \end{aligned}$$

Here the constant $C > 0$ depends on K, T and d . This concludes the proof. $\square$

Proof of Theorem 4.2.4. Since $\pi^* \in \mathcal{A}$ is the (locally) optimal control for the problem (3.2), then Pontryagin's optimality principle holds, see [45, Th 2.12]. Hence, for any $m \in \mathcal{P}(\mathbb{R}^p)$ we have $\forall s \in [0, T]$

$$H_s^{\tau,0}(\Theta_s(\pi^*), \pi_s^*) \leq H_s^{\tau,0}(\Theta_s(\pi^*), m), \quad \text{a.s.} \quad (4.23)$$

By definition of augmented Hamiltonian for all $s \in [0, T]$ we have

$$H_s^{\tau, \lambda}(\Theta_s(\pi^*), m, \pi_s^*) = H_s^{\tau, 0}(\Theta_s(\pi^*), m) + \lambda \text{Ent}(m | \pi_s^*). \quad (4.24)$$

Therefore, due to (4.23) and (4.24), and the fact that $\text{Ent}(m | \pi_s^*) \geq 0$ we have

$$\begin{aligned} H_s^{\tau, \lambda}(\Theta_s(\pi^*), \pi_s^*, \pi_s^*) &= H_s^{\tau, 0}(\Theta_s(\pi^*), \pi_s^*) \leq H_s^{\tau, 0}(\Theta_s(\pi^*), m) \\ &\leq H_s^{\tau, 0}(\Theta_s(\pi^*), m) + \lambda \text{Ent}(m | \pi_s^*) = H_s^{\tau, \lambda}(\Theta_s(\pi^*), m, \pi_s^*). \end{aligned}$$

This concludes the proof. $\square$

Proof of Theorem 4.2.6. Let us apply Lemma 4.2.5 for $\pi' = \pi^{n+1}$ and $\pi = \pi^n$. Hence, for some $C > 0$ we have

$$\begin{aligned} J^\tau(\pi^{n+1}) - J^\tau(\pi^n) &\leq \mathbb{E} \int_0^T \left[H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^{n+1}) - H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^n) \right] ds \\ &\quad + C \mathbb{E} \int_0^T \text{Ent}(\pi_s^{n+1} | \pi_s^n) ds. \end{aligned} \quad (4.25)$$

Let

$$\mu(\pi^n) := \mathbb{E} \int_0^T \left[H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^{n+1}) - H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^n) \right] ds.$$

Due to the definition of π^n (4.6) we have that for all $s \in [0, T]$

$$\begin{aligned} H_s^{\tau, \lambda}(\Theta_s(\pi^n), \pi_s^{n+1}, \pi_s^n) &= H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^{n+1}) + \lambda \text{Ent}(\pi_s^{n+1} | \pi_s^n) \\ &\leq H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^n) = H_s^{\tau, \lambda}(\Theta_s(\pi^n), \pi_s^n, \pi_s^n). \end{aligned}$$

Therefore, we can observe that

$$0 \leq \lambda \text{Ent}(\pi_s^{n+1} | \pi_s^n) \leq H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^n) - H_s^{\tau, 0}(\Theta_s(\pi^n), \pi_s^{n+1}).$$

Hence, we can rewrite the inequality (4.25) as

$$J^\tau(\pi^{n+1}) - J^\tau(\pi^n) \leq \mu(\pi^n) - \frac{C}{\lambda} \mu(\pi^n) = D \mu(\pi^n), \quad (4.26)$$

where $D := 1 - \frac{C}{\lambda}$. By choosing $\lambda > C$ we get that $D > 0$. Notice that for any integer $M > 1$ we have

$$\begin{aligned} 0 &\leq \lambda \sum_{n=0}^M \mathbb{E} \int_0^T \text{Ent}(\pi_s^{n+1} | \pi_s^n) ds \leq \sum_{n=0}^M (-\mu(\pi^n)) \leq D^{-1} \sum_{n=0}^M (J^\tau(\pi^n) - J^\tau(\pi^{n+1})) \\ &= D^{-1} (J^\tau(\pi^0) - J^\tau(\pi^{M+1})) \leq D^{-1} (J^\tau(\pi^0) - \inf_{\pi \in \mathcal{A}} J^\tau(\pi)) < \infty. \end{aligned}$$

Hence, we obtain that $\mu(\pi^n) \rightarrow 0$ as $n \rightarrow \infty$. Let $\mu(\pi^n) = 0$ for some n . This leads to the fact that $\mathbb{E} \int_0^T \text{Ent}(\pi_s^{n+1} | \pi_s^n) ds = 0$, where

$$\pi_s^{n+1} \in \arg \min_{\pi \in \mathcal{P}(\mathbb{R}^p)} H_s^{\tau, \lambda}(\Theta_s(\pi_s^n), \pi, \pi_s^n).$$

Hence,

$$H_s^{\tau,\lambda}(\Theta_s(\pi_s^n), \pi_s^{n+1}, \pi_s^n) = H_s^{\tau,\lambda}(\Theta_s(\pi_s^n), \pi_s^n, \pi_s^n), \forall s \in [0, T] \quad \text{a.s.}$$

Thus, by the definition of π^{n+1} we have

$$H_s^{\tau,\lambda}(\Theta_s(\pi_s^n), \pi_s^n, \pi_s^n) = \min_{m \in \mathcal{P}(\mathbb{R}^p)} H_s^{\tau,\lambda}(\Theta_s(\pi_s^n), m, \pi_s^n), \forall s \in [0, T] \quad \text{a.s.}$$

Therefore, π^n satisfy the extended Pontryagin's optimality principle stated in Theorem [4.2.4](#). This concludes the proof. $\square$

Chapter 5

Extensions and Future Work

There are a few directions in which the work contained in this thesis could be extended:

1. In Chapter 2, we considered the case when the diffusion coefficient in (2.1) does not depend on the control. However, it would be interesting to get analogous result as in Chapter 2 when the diffusion coefficient does depend on the control. In Remark 2.4.2, we briefly mentioned that in this case one need to deal with a second order BSDE.
2. Also in Chapter 2, we assumed in Assumption 2.2.2 that the running gain f has a linear growth in space and action variables. This assumption does not hold, for example, for the linear-quadratic optimal control problems. If we want to consider the quadratic growth in space and action variables of the running gain, it will require studying the application of a quadratic BSDE theory to our problem. Another approach would be to apply clipping to the running gain and estimate an error it would bring.
3. In Chapter 3, we could get the convergence rate of the Algorithm 3 only for certain case. The extension on this would be to get the convergence in more general case or to study the rate of convergence empirically through the experiments for different problem settings.
4. In Chapter 4, the extension is the same as in previous item, to study the convergence rate. It would also be interesting to consider the case when the coefficient λ is big enough, so that the policy in (4.6) has a unique solution. However, it is known in RL theory that having a big penalization coefficient in front of the KL term leads to slower convergence of the algorithm, see [50].

Appendix A

Supplementary materials for Chapter 2

A.1 Some results from theory of BSDEs

We fix a finite horizon $T \in (0, \infty)$. We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$. Let there be a d' -dimensional Wiener martingale on this space.

Lemma A.1.1. *Let $F : \Omega \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function that satisfies the following conditions: The process $(F_t(0))_{t \in [0, T]}$ is in $\mathbb{H}^2$. Moreover there is a constant $\theta > 0$ such that a.s.*

$$|F_t(z)| \leq |F_t(0)| + \theta|z| \quad \forall z \in \mathbb{R}^d, t \in [0, T].$$

Then, for every $\xi \in L^2(\Omega, \mathcal{F}_T)$ and $z \in \mathbb{H}^2$, there is a unique solution $(Y, Z) \in \mathcal{S}^2 \times \mathbb{H}^2$ to

$$Y_t = \xi + \int_t^T F_s(z_s) ds - \int_t^T Z_s dW_s, \quad t \in [0, T], \quad \mathbb{P}\text{-a.s.} \quad (\text{A.1})$$

Proof. This follows immediately from, e.g., Pham [1, Theorem 6.2.1]. $\square$

Lemma A.1.2. *Let $F : \Omega \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy the hypothesis of Lemma A.1.1. Fix $\xi \in L^2(\Omega, \mathcal{F}_T)$. Let $\Phi : \mathbb{H}^2 \ni z \mapsto (Y, Z) \in \mathcal{S}^2 \times \mathbb{H}^2$, where (Y, Z) is the unique solution to (A.1). Moreover assume that for all $z^1, z^2 \in \mathbb{R}^d$ and $t \in [0, T]$ the following condition satisfies that there is a constant $\theta > 0$ such that a.s.*

$$|F_t(z^1) - F_t(z^2)| \leq \theta|z^1 - z^2|. \quad (\text{A.2})$$

Then there is $\gamma > 0$ and $q \in (0, 1)$ such that for $(Y^i, Z^i) := \Phi(z^i)$, $i = 1, 2$, and any $t \in [0, T]$ we have

$$\mathbb{E} [e^{\gamma t} |Y_t^1 - Y_t^2|^2] + \|Z^1 - Z^2\|_{\mathbb{H}_\gamma^2}^2 \leq q \|z^1 - z^2\|_{\mathbb{H}_\gamma^2}^2.$$

The proof is well known and is included, e.g., as part of Pham [1, Proof of Theorem 6.2.1]. We provide it here for the convenience of the reader and before we proceed we need to make the following observation.

Remark A.1.3. Assume that $Y \in \mathcal{S}^2$ and $Z \in \mathbb{H}^2$ and let

$$M_t := \int_0^t e^{\gamma s} Z_s Y_s dW_s.$$

Then $\sup_{t \leq T} |M_t| \in L^1(\Omega, \mathcal{F}_T)$ and hence M_t is a uniformly integrable martingale. Indeed, from the Burkholder–Davis–Gundy inequality and the Young inequality we get

$$\begin{aligned} \mathbb{E} \left[\sup_{t \leq T} |M_t| \right] &\leq C_1 \mathbb{E} \left[\left(\int_0^T e^{2\gamma s} |Y_s|^2 |Z_s|^2 ds \right)^{1/2} \right] \\ &\leq e^{\gamma T} C_1 \mathbb{E} \left[\left(\sup_{t \leq T} |Y_t|^2 \int_0^T |Z_s|^2 ds \right)^{1/2} \right] \\ &\leq \frac{e^{\gamma T}}{2} C_1 \mathbb{E} \left[\sup_{t \leq T} |Y_t|^2 + \int_0^T |Z_s|^2 ds \right] < \infty. \end{aligned}$$

Proof of Lemma A.1.2. Consider $\gamma > 0$ which we will fix later. We denote $\delta z := z^1 - z^2$, $\delta Z := Z^1 - Z^2$, $\delta Y := Y^1 - Y^2$, and $\delta F := F(z^1) - F(z^2)$. We then apply Itô's formula to $e^{\gamma t} |\delta Y_t|^2$:

$$\begin{aligned} e^{\gamma t} |\delta Y_t|^2 + \int_t^T e^{\gamma s} |\delta Z_s|^2 ds &= \int_t^T e^{\gamma s} (2\delta Y_s \delta F_s - \gamma |\delta Y_s|^2) ds \\ &\quad - 2 \int_t^T e^{\gamma s} \delta Z_s \delta Y_s dW_s. \end{aligned}$$

Due to Remark A.1.3, the stochastic integral vanishes by taking expectation. Hence

$$\mathbb{E} \left[e^{\gamma t} |\delta Y_t|^2 + \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \right] = \mathbb{E} \left[\int_t^T e^{\gamma s} (2\delta Y_s \delta F_s - \gamma |\delta Y_s|^2) ds \right].$$

By the Lipschitz property of the generator and by the Young inequality we continue our estimate, noting that for any $\varepsilon > 0$, we have

$$\begin{aligned} \mathbb{E} \left[e^{\gamma t} |\delta Y_t|^2 + \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \right] &\leq \mathbb{E} \left[\int_t^T e^{\gamma s} (2\theta |\delta Y_s| |\delta z_s| - \gamma |\delta Y_s|^2) ds \right] \\ &\leq \mathbb{E} \left[\int_t^T e^{\gamma s} (\theta(\varepsilon |\delta Y_s|^2 + \varepsilon^{-1} |\delta z_s|^2) - \gamma |\delta Y_s|^2) ds \right]. \end{aligned}$$

Choose ε such that $\gamma = \varepsilon\theta$. Thus

$$\mathbb{E} \left[e^{\gamma t} |\delta Y_t|^2 + \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \right] \leq \mathbb{E} \left[\int_t^T e^{\gamma s} (\theta \varepsilon^{-1} |\delta z_s|^2) ds \right] \leq \frac{\theta^2}{\gamma} \|\delta z\|_{\mathbb{H}_\gamma^2}^2. \quad (\text{A.3})$$

Hence, from (A.3) we have that for $\gamma > \theta^2$ and any $t \in [0, T]$

$$\mathbb{E} [e^{\gamma t} |Y_t^1 - Y_t^2|^2] + \|Z^1 - Z^2\|_{\mathbb{H}_\gamma^2}^2 \leq q \|z^1 - z^2\|_{\mathbb{H}_\gamma^2}^2,$$

where $q = \frac{\theta^2}{\gamma} \in (0, 1)$. This concludes the proof of the lemma. $\square$

Lemma A.1.4. *Let $F : \Omega \times [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function such that the process $(F_t(0, 0))_{t \in [0, T]}$ is in $\mathbb{H}^2$ and such that there are $\theta, K > 0$ so that for all $t \in [0, T]$, $z, Z \in \mathbb{R}^d$ we have*

$$|F_t(z, Z)| \leq |F_t(0, 0)| + \theta|z| + K|Z| \quad a.s.$$

If $\xi \in L^2(\Omega, \mathcal{F}_T)$ and $z \in \mathbb{H}^2$, then there is a unique solution $(Y, Z) \in \mathcal{S}^2 \times \mathbb{H}^2$ to

$$Y_t = \xi + \int_t^T F_s(z_s, Z_s) ds - \int_t^T Z_s dW_s, \quad t \in [0, T], \quad \mathbb{P}\text{-a.s.} \quad (\text{A.4})$$

Proof. This follows immediately from, e.g., Pham [1, Theorem 6.2.1]. $\square$

Lemma A.1.5. *Let $F : \Omega \times [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy the hypothesis of Lemma A.1.4. Fix $\xi \in L^2(\Omega, \mathcal{F}_T)$. Let $\Phi : \mathbb{H}^2 \ni z \mapsto (Y, Z) \in \mathcal{S}^2 \times \mathbb{H}^2$, where (Y, Z) is the unique solution to (A.4). Moreover assume that for $z^1, z^2 \in \mathbb{H}^2$ the following condition satisfies that there are constants $\theta, K > 0$ such that a.s.*

$$|F_t(z_t^1, Z_t^1) - F_t(z_t^2, Z_t^2)| \leq \theta|z_t^1 - z_t^2| + K|Z_t^1 - Z_t^2|, \quad t \in [0, T],$$

where $(Y^i, Z^i) := \Phi(z^i)$, $i = 1, 2$. Then there is $\gamma > 0$ and $q \in (0, 1)$ such that for any $t \in [0, T]$ we have

$$\mathbb{E} [e^{\gamma t} |Y_t^1 - Y_t^2|^2] + \|Z^1 - Z^2\|_{\mathbb{H}_\gamma^2}^2 \leq q \|z^1 - z^2\|_{\mathbb{H}_\gamma^2}^2.$$

Moreover, there is $\gamma > 0$ such that $q \in (0, 1/2)$.

Proof. Consider $\gamma > 0$ which we will fix later. We denote $\delta z := z^1 - z^2$, $\delta Z := Z^1 - Z^2$, and $\delta Y := Y^1 - Y^2$. We then apply Itô's formula to $e^{\gamma t} |\delta Y_t|^2$:

$$\begin{aligned} e^{\gamma t} |\delta Y_t|^2 &+ \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \\ &= \int_t^T e^{\gamma s} (2\delta Y_s (F_s(z_s^1, Z_s^1) - F_s(z_s^2, Z_s^2)) - \gamma |\delta Y_s|^2) ds \\ &\quad - 2 \int_t^T e^{\gamma s} \delta Z_s \delta Y_s dW_s. \end{aligned}$$

The expectation of the stochastic integral is 0 due to Remark A.1.3. Hence, by taking expectation we derive from the equality above that

$$\begin{aligned} \mathbb{E} \left[e^{\gamma t} |\delta Y_t|^2 + \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \right] \\ = \mathbb{E} \int_t^T e^{\gamma s} (2\delta Y_s (F_s(z_s^1, Z_s^1) - F_s(z_s^2, Z_s^2)) - \gamma |\delta Y_s|^2) ds. \end{aligned}$$

By the Lipschitz property of the generator and by the Young inequality we observe

that, for any $\varepsilon > 0$,

$$\begin{aligned} \mathbb{E} \left[e^{\gamma t} |\delta Y_t|^2 + \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \right] &\leq \mathbb{E} \int_t^T e^{\gamma s} (2\delta Y_s (\theta |\delta z_s| + K |\delta Z_s|) - \gamma |\delta Y_s|^2) ds \\ &\leq \mathbb{E} \int_t^T e^{\gamma s} ((\theta + K)\varepsilon |\delta Y_s|^2 + \theta \varepsilon^{-1} |\delta z_s|^2 + K \varepsilon^{-1} |\delta Z_s|^2 - \gamma |\delta Y_s|^2) ds. \end{aligned}$$

Take $\gamma > 0$ sufficiently large so that $\tilde{q} := \max(\frac{(\theta+K)K}{\gamma}, \frac{(\theta+K)\theta}{\gamma}) \in (0, 1/2)$. Choose ε such that $\gamma = (\theta + K)\varepsilon$. Thus

$$\mathbb{E} \left[e^{\gamma t} |\delta Y_t|^2 + (1 - \tilde{q}) \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \right] \leq \mathbb{E} \left[\int_t^T e^{\gamma s} \tilde{q} |\delta z_s|^2 ds \right]. \quad (\text{A.5})$$

Dividing by $1 - \tilde{q} \in (1/2, 1)$ we obtain

$$\mathbb{E} \left[e^{\gamma t} |\delta Y_t|^2 + \int_t^T e^{\gamma s} |\delta Z_s|^2 ds \right] \leq q \mathbb{E} \left[\int_t^T e^{\gamma s} |\delta z_s|^2 ds \right],$$

where $q := \frac{\tilde{q}}{1-\tilde{q}}$. Since $0 < \tilde{q} < 1/2$ we have that $q \in (0, 1)$. Therefore from (A.5) we have for any $t \in [0, T]$

$$\mathbb{E} [e^{\gamma t} |Y_t^1 - Y_t^2|^2] + \|Z^1 - Z^2\|_{\mathbb{H}_\gamma^2}^2 \leq q \|z^1 - z^2\|_{\mathbb{H}_\gamma^2}^2.$$

By choosing γ such that $\tilde{q} \in (0, 1/3)$ we get that $q \in (0, 1/2)$. This concludes the proof of the lemma. $\square$

We now state a comparison principle for BSDEs.

Lemma A.1.6. *Consider the following BSDEs:*

$$Y_t^i = \xi^i + \int_t^T \phi^i(s, Z_s^i) ds - \int_t^T Z_s^i dW_s, \quad t \in [0, T], \quad i = 1, 2.$$

Assume that $\xi^i \in L^2(\Omega, \mathcal{F}_T)$, $i = 1, 2$, and $\xi^1 \leq \xi^2$ a.s. Let $\phi^i : \Omega \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$, $i = 1, 2$, be such that for all $z \in \mathbb{R}^d$ the processes $(\phi^i(t, z))_{t \in [0, T]}$ are progressively measurable, $\phi^i(t, 0) \in \mathbb{H}^2$, and such that there is $\theta > 0$ so that for all $t \in [0, T]$, $z, z' \in \mathbb{R}^d$ we have

$$|\phi^i(t, z) - \phi^i(t, z')| \leq \theta |z - z'| \text{ a.s.}$$

Moreover, suppose that for $Z^1, Z^2 \in \mathbb{H}^2$ it holds that

$$\phi^1(t, Z_t^1) \leq \phi^2(t, Z_t^1) \quad \forall t \in [0, T], \text{ a.s.}$$

Then $Y_t^1 \leq Y_t^2$ for all $0 \leq t \leq T$ a.s.

Proof. This follows from, e.g., Pham [1, Theorem 6.2.2]. $\square$

The following two lemmas are auxiliary results we need in Section 2.7.

Lemma A.1.7. Let $F, \bar{F} : \Omega \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ be measurable functions and let F satisfy the hypotheses of Lemmas A.1.1 and A.1.2. Fix $\xi \in L^2(\Omega, \mathcal{F}_T)$. Let $\bar{z}, z, Z, \bar{Z} \in \mathbb{H}^2$ and $Y, \bar{Y} \in \mathcal{S}^2$ be such that

$$\bar{Y}_t = \xi + \int_t^T \bar{F}_s(\bar{z}_s) ds - \int_t^T \bar{Z}_s dW_s, \quad t \in [0, T],$$

and

$$Y_t = \xi + \int_t^T F_s(z_s) ds - \int_t^T Z_s dW_s, \quad t \in [0, T].$$

Then there is $\gamma > 0$ and $q \in (0, 1)$ such that for $t \in [0, T]$ we have

$$e^{\gamma t} \mathbb{E} |\bar{Y}_t - Y_t|^2 + \|\bar{Z} - Z\|_{\mathbb{H}_\gamma^2}^2 \leq q \|\bar{z} - z\|_{\mathbb{H}_\gamma^2}^2 + \frac{q}{\theta} \|\bar{F}(\bar{z}) - F(\bar{z})\|_{\mathbb{H}_\gamma^2}^2.$$

Proof. Consider $\gamma > 0$ which we will fix later. We denote $\tilde{Y} := \bar{Y} - Y$, $\tilde{Z} := \bar{Z} - Z$, and $\tilde{z} := \bar{z} - z$. We then apply Itô's formula to $e^{\gamma t} |\tilde{Y}_t|^2$:

$$\begin{aligned} e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds &= \int_t^T e^{\gamma s} (2\tilde{Y}_s (\bar{F}_s(\bar{z}_s) - F_s(z_s)) - \gamma |\tilde{Y}_s|^2) ds \\ &\quad - 2 \int_t^T e^{\gamma s} \tilde{Z}_s \tilde{Y}_s dW_s. \end{aligned}$$

Due to Remark A.1.3, the stochastic integral vanishes by taking expectation. Hence

$$\mathbb{E} \left[e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds \right] = \mathbb{E} \left[\int_t^T e^{\gamma s} (2\tilde{Y}_s (\bar{F}_s(\bar{z}_s) - F_s(z_s)) - \gamma |\tilde{Y}_s|^2) ds \right]. \quad (\text{A.6})$$

Notice that due to (A.2) for all $s \in [t, T]$ it holds that

$$|\bar{F}_s(\bar{z}_s) - F_s(z_s)| \leq |\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)| + |F_s(\bar{z}_s) - F_s(z_s)| \leq |\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)| + \theta |\bar{z}_s - z_s|.$$

Then by the Young inequality we continue our estimate (A.6), noting that for any $\delta > 0$, we have

$$\begin{aligned} &\mathbb{E} \left[e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds \right] \\ &\leq \mathbb{E} \left[\int_t^T e^{\gamma s} (2|\tilde{Y}_s| (|\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)| + \theta |\bar{z}_s - z_s|) - \gamma |\tilde{Y}_s|^2) ds \right] \\ &\leq \mathbb{E} \left[\int_t^T e^{\gamma s} \left((1 + \theta) \delta |\tilde{Y}_s|^2 + \delta^{-1} (\theta |\bar{z}_s - z_s|^2 + |\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)|^2) - \gamma |\tilde{Y}_s|^2 \right) ds \right]. \end{aligned}$$

Fix $\gamma > (1 + \theta)\theta$ and $q = (1 + \theta)\theta/\gamma$. Let $\delta = \gamma/(1 + \theta)$. Then

$$\begin{aligned} \mathbb{E} \left[e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds \right] &\leq \mathbb{E} \left[\int_t^T e^{\gamma s} q \left(|\tilde{z}_s|^2 + \frac{1}{\theta} |\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)|^2 \right) ds \right] \\ &\leq q \|\tilde{z}\|_{\mathbb{H}_\gamma^2}^2 + \frac{q}{\theta} \|\bar{F}(\bar{z}) - F(\bar{z})\|_{\mathbb{H}_\gamma^2}^2. \end{aligned}$$

This concludes the proof of the lemma. $\square$

Lemma A.1.8. *Let $\bar{F} : \Omega \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function and let F satisfies the hypotheses of Lemmas A.1.4 and A.1.5. Fix $\xi \in L^2(\Omega, \mathcal{F}_T)$. Let $\bar{z}, z, \bar{Z}, Z \in \mathbb{H}^2$ and $\bar{Y}, Y \in \mathcal{S}^2$ be such that*

$$\bar{Y}_t = \xi + \int_t^T \bar{F}_s(\bar{z}_s, \bar{Z}_s) ds - \int_t^T \bar{Z}_s dW_s, \quad t \in [0, T],$$

and

$$Y_t = \xi + \int_t^T F_s(z_s, Z_s) ds - \int_t^T Z_s dW_s, \quad t \in [0, T].$$

Then there is $\gamma > 0$ and $q \in (0, 1)$ such that for $t \in [0, T]$ we have

$$e^{\gamma t} \mathbb{E} |\bar{Y}_t - Y_t|^2 + \|\bar{Z} - Z\|_{\mathbb{H}_\gamma^2}^2 \leq q \|\bar{z} - z\|_{\mathbb{H}_\gamma^2}^2 + \frac{q}{\max(K, \theta)} \|\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)\|_{\mathbb{H}_\gamma^2}^2.$$

Proof. Consider $\gamma > 0$ which we will fix later. We denote $\tilde{Y} := \bar{Y} - Y$, $\tilde{Z} := \bar{Z} - Z$, and $\tilde{z} := \bar{z} - z$. We then apply Itô's formula to $e^{\gamma t} |\tilde{Y}_t|^2$:

$$\begin{aligned} e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds &= \int_t^T e^{\gamma s} (2\tilde{Y}_s (\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)) - \gamma |\tilde{Y}_s|^2) ds \\ &\quad - 2 \int_t^T e^{\gamma s} \tilde{Z}_s \tilde{Y}_s dW_s. \end{aligned}$$

Due to Remark A.1.3, the stochastic integral vanishes by taking expectation. Hence

$$\begin{aligned} \mathbb{E} \left[e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds \right] &= \mathbb{E} \left[\int_t^T e^{\gamma s} (2\tilde{Y}_s (\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)) - \gamma |\tilde{Y}_s|^2) ds \right]. \end{aligned}$$

Notice that by assumptions of the lemma for all $s \in [t, T]$ it holds that

$$\begin{aligned} |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)| &\leq |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)| + |F_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)| \\ &\leq |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)| + \theta |\bar{z}_s - z_s| + K |\bar{Z}_s - Z_s|. \end{aligned}$$

Then by the Young inequality for any $\delta > 0$, we have

$$\begin{aligned}
& \mathbb{E} \left[e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds \right] \\
& \leq \mathbb{E} \left[\int_t^T e^{\gamma s} (2|\tilde{Y}_s|(|\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)| + \theta|\bar{z}_s - z_s| + K|\bar{Z}_s - Z_s|) - \gamma|\tilde{Y}_s|^2) ds \right] \\
& \leq \mathbb{E} \left[\int_t^T e^{\gamma s} \left((\theta + K + 1)\delta|\tilde{Y}_s|^2 + \delta^{-1}(K|\tilde{Z}_s|^2 + \theta|\bar{z}_s|^2 + |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)|^2) \right. \right. \\
& \quad \left. \left. - \gamma|\tilde{Y}_s|^2 \right) ds \right].
\end{aligned}$$

Let us take $\gamma > 0$ sufficiently large so that $\tilde{q} := \max(\frac{(1+\theta+K)K}{\gamma}, \frac{(1+\theta+K)\theta}{\gamma}) \in (0, 1/2)$. Let $\delta := \gamma/(1 + \theta + K)$ so that

$$\begin{aligned}
& \mathbb{E} \left[e^{\gamma t} |\tilde{Y}_t|^2 + (1 - \tilde{q}) \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds \right] \\
& \leq \mathbb{E} \left[\int_t^T e^{\gamma s} \tilde{q} \left(|\bar{z}_s|^2 + \frac{1}{\max(K, \theta)} |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)|^2 \right) ds \right] \\
& \leq \tilde{q} \|\bar{z}\|_{\mathbb{H}_\gamma^2}^2 + \frac{\tilde{q}}{\max(K, \theta)} \|\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)\|_{\mathbb{H}_\gamma^2}^2.
\end{aligned}$$

Dividing by $(1 - \tilde{q}) \in (1/2, 1)$ we obtain

$$\mathbb{E} \left[e^{\gamma t} |\tilde{Y}_t|^2 + \int_t^T e^{\gamma s} |\tilde{Z}_s|^2 ds \right] \leq q \|\bar{z}\|_{\mathbb{H}_\gamma^2}^2 + \frac{q}{\max(K, \theta)} \|\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)\|_{\mathbb{H}_\gamma^2}^2,$$

where $q := \frac{\tilde{q}}{1 - \tilde{q}}$. Since $0 < \tilde{q} < 1/2$ we have that $q \in (0, 1)$. $\square$

A.1.1 BSDE with drivers of quadratic growth.

Since we are using BSDE theory in the proof of the main result, we would like to present some results on BSDE with drivers of quadratic growth. We refer to [60].

Consider the following system:

$$\begin{aligned}
X_t &= x + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s, \\
Y_t &= g(X_T) + \int_t^T f(s, X_s, Z_s) ds - \int_t^T Z_s dW_s.
\end{aligned} \tag{A.7}$$

Theorem A.1.9 (Theorem 3.6 in [60]). *Let $b : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ and $\sigma : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d'}$ be Lipschitz continuous with Lipschitz constant C and $|b(t, 0)| \leq C$ and $|\sigma(t, 0)| \leq C$ for all $t \in [0, T]$. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ and $f : [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be measurable functions and let us assume that there exists constant C such that*

for all $r \in \mathbb{R}^+, t \in [0, T], x, x' \in \mathbb{R}^d$, and $z, z' \in \mathbb{R}^d$

$$\begin{aligned} |f(t, x, z)| &\leq C(1 + |z|^2), \\ |g(x)| &\leq C, \\ |f(t, x, z) - f(t, x, z')| &\leq C(1 + |z| + |z'|)|z - z'|, \\ |f(t, x, z) - f(t, x', z)| &\leq C(1 + |x|^r + |x'|^r)|x - x'|, \\ |g(x) - g(x')| &\leq C(1 + |x|^r + |x'|^r)|x - x'|. \end{aligned}$$

There exists a solution (Y, Z) of the Markovian BSDE (A.7) in $\mathcal{S}^2 \times \mathbb{H}^2$ and this solution is unique among solutions $(Y, Z) \in \mathcal{S}^2 \times \mathbb{H}^2$ such that Y is bounded. Moreover, we have

$$|Z_t| \leq C(1 + |X_t|^{1+r}) \quad \forall t \in [0, T],$$

and

$$\|Z \bullet W\|_{BMO} < \infty,$$

where

$$\|M\|_{BMO} := \sup_{\tau \in \mathcal{T}_0^T} \|\mathbb{E}[\langle M \rangle_T - \langle M \rangle_\tau \mid \mathcal{F}_\tau]\|_\infty;$$

here the supremum is taken over all stopping times in $[0, T]$.

Appendix B

Supplementary materials for Chapter 3

B.1 Auxiliary Lemma

Lemma B.1.1. *Let $\{b_k\}_{k \in \mathbb{N}}$ be the sequence of nonnegative numbers such that*

$$b_{k+1} \leq b_k - qb_k^2,$$

where q is a positive constant. Then $b_k = O(1/k)$.

One can find the proof in [61, Lemma 1.4, p. 93]. However, the proof is written in Russian. For convenience of the reader we provide it here.

Proof. Let $b_k = \frac{c_k}{k}$ for some nonnegative sequence $(c_k)_{k \in \mathbb{N}}$. Then it is enough to show that c_k is bounded for all $k \in \mathbb{N}$. By assumption we have

$$b_k - b_{k+1} = \frac{c_k}{k} - \frac{c_{k+1}}{k+1} = \frac{c_k}{k} \left(1 - \frac{c_{k+1}}{c_k} \frac{k}{k+1} \right) \geq q \frac{c_k^2}{k^2}.$$

Therefore,

$$1 - \frac{c_{k+1}}{c_k} \frac{k}{k+1} \geq q \frac{c_k}{k}.$$

After some transformation, we can rewrite the equation above as

$$\left(1 + \frac{1}{k} \right) \left(1 - q \frac{c_k}{k} \right) \geq \frac{c_{k+1}}{c_k}.$$

Thus

$$1 + \frac{1}{k} (1 - qc_k) - q \frac{c_k}{k^2} \geq \frac{c_{k+1}}{c_k}.$$

If $1 - qc_k < 0$ we have

$$1 > 1 + \frac{1}{k} (1 - qc_k) - q \frac{c_k}{k^2} \geq \frac{c_{k+1}}{c_k}.$$

Hence $c_{k+1} < c_k$. On the other hand, if $1 - qc_k \geq 0$, we have $c_k \leq \frac{1}{q}$. Therefore,

we conclude that for all k we have

$$c_k \leq \max \left\{ c_1, \frac{1}{q} \right\} .$$

□

Appendix C

Supplementary materials for Chapter 4

C.1 Flat derivative on measure spaces

We define flat derivative on $\mathcal{P}_2(\mathbb{R}^p)$. See e.g. [62, Section 5.4.1].

Definition C.1.1. A functional $U : \mathcal{P}_2(\mathbb{R}^p) \mapsto \mathbb{R}$ is said to admit a linear derivative if there is a (continuous on $\mathcal{P}_2(\mathbb{R}^p)$) map $\frac{\delta U}{\delta m} : \mathcal{P}_2(\mathbb{R}^p) \times \mathbb{R}^d \mapsto \mathbb{R}$, such that $|\frac{\delta U}{\delta m}(a, \mu)| \leq C(1 + |a|^2)$ and, for all $m, m' \in \mathcal{P}_2(\mathbb{R}^p)$, it holds that

$$U(m') - U(m) = \int_0^1 \int \frac{\delta U}{\delta m}(m + \lambda(m' - m), a)(m' - m)(da) d\lambda. \quad (\text{C.1})$$

Since $\frac{\delta U}{\delta m}$ is only defined up to a constant we make a choice by demanding $\int \frac{\delta U}{\delta m}(m, a)m(da) = 0$. The functional $\frac{\delta U}{\delta m}$ is then called the linear (functional) derivative of U on $\mathcal{P}_2(\mathbb{R}^p)$.

Note that if $\frac{\delta U}{\delta m}$ exists according to Definition C.1.1 then $\forall \nu, \mu \in \mathcal{P}_2(\mathbb{R}^p)$

$$\lim_{\varepsilon \rightarrow 0+} \frac{U(\nu + \varepsilon(\mu - \nu)) - U(\nu)}{\varepsilon} = \int \frac{\delta U}{\delta m}(\nu, a)(\mu - \nu)(da). \quad (\text{C.2})$$

Indeed (C.1) immediately implies (C.2). To see the implication in the other direction take $\nu^\lambda := \nu + \lambda(\mu - \nu)$ and $\mu^\lambda := \mu - \nu + \nu^\lambda$ and notice that (C.2) ensures for all $\lambda \in [0, 1]$ that

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0+} \frac{U(\nu^\lambda + \varepsilon(\mu - \nu)) - U(\nu^\lambda)}{\varepsilon} &= \lim_{\varepsilon \rightarrow 0+} \frac{U(\nu^\lambda + \varepsilon(\mu^\lambda - \nu^\lambda)) - U(\nu^\lambda)}{\varepsilon} \\ &= \int \frac{\delta U}{\delta m}(\nu^\lambda, a)(\mu^\lambda - \nu^\lambda)(da) = \int \frac{\delta U}{\delta m}(\nu^\lambda, a)(\mu - \nu)(da). \end{aligned}$$

By the fundamental theorem of calculus

$$U(\mu) - U(\nu) = \int_0^1 \lim_{\varepsilon \rightarrow 0+} \frac{U(\nu^{\lambda+\varepsilon}) - U(\nu^\lambda)}{\varepsilon} d\lambda = \int_0^1 \int \frac{\delta U}{\delta m}(\nu^\lambda, a)(\mu - \nu)(da) d\lambda.$$

C.2 Standard SDE estimate

Lemma C.2.1. *Assume that there exists $K \geq 0$ such that $\forall x, x' \in \mathbb{R}^d, \forall m, m' \in \mathcal{P}(\mathbb{R}^p), \forall t \in [0, T]$,*

$$|b_t(x, m) - b_t(x, m')| + |\sigma_t(x, m) - \sigma_t(x, m')| \leq K|x - x'| + K\|m - m'\|_{\text{TV}}.$$

Then there exists constant $C > 0$ such that

$$\mathbb{E} \sup_{0 \leq t \leq T} |X_t(\pi) - X_t(\pi')|^2 \leq C \mathbb{E} \int_0^T \|\pi_s - \pi'_s\|_{\text{TV}}^2 ds.$$

Proof. Let us consider the difference

$$\begin{aligned} & X_t(\pi) - X_t(\pi') \\ &= \int_0^t (b_s(X_s(\pi), \pi_s) - b_s(X_s(\pi'), \pi'_s)) ds + \int_0^t (\sigma_s(X_s(\pi), \pi_s) - \sigma_s(X_s(\pi'), \pi'_s)) dW_s. \end{aligned}$$

From the inequality $(a + b)^2 \leq 2a^2 + 2b^2$ and from Hölder's inequality we have that

$$\begin{aligned} & |X_t(\pi) - X_t(\pi')|^2 \\ &= \left| \int_0^t (b_s(X_s(\pi), \pi_s) - b_s(X_s(\pi'), \pi'_s)) ds + \int_0^t (\sigma_s(X_s(\pi), \pi_s) - \sigma_s(X_s(\pi'), \pi'_s)) dW_s \right|^2 \\ &\leq 2 \left| \int_0^t (b_s(X_s(\pi), \pi_s) - b_s(X_s(\pi'), \pi'_s)) ds \right|^2 + 2 \left| \int_0^t (\sigma_s(X_s(\pi), \pi_s) - \sigma_s(X_s(\pi'), \pi'_s)) dW_s \right|^2 \\ &\leq 2t \int_0^t |b_s(X_s(\pi), \pi_s) - b_s(X_s(\pi'), \pi'_s)|^2 ds + 2 \left| \int_0^t (\sigma_s(X_s(\pi), \pi_s) - \sigma_s(X_s(\pi'), \pi'_s)) dW_s \right|^2. \end{aligned}$$

After taking supremum over $t \in [0, T]$ we get that

$$\begin{aligned} \sup_{0 \leq t \leq T} |X_t(\pi) - X_t(\pi')|^2 &\leq 2T \int_0^T |b_s(X_s(\pi), \pi_s) - b_s(X_s(\pi'), \pi'_s)|^2 ds \\ &\quad + 2 \sup_{0 \leq t \leq T} \left| \int_0^t (\sigma_s(X_s(\pi), \pi_s) - \sigma_s(X_s(\pi'), \pi'_s)) dW_s \right|^2. \end{aligned}$$

By taking the expectation and by applying the Burkholder-Davis-Gundy inequality we obtain

$$\begin{aligned} \mathbb{E} \sup_{0 \leq t \leq T} |X_t(\pi) - X_t(\pi')|^2 &\leq 2T \mathbb{E} \int_0^T |b_s(X_s(\pi), \pi_s) - b_s(X_s(\pi'), \pi'_s)|^2 ds \\ &\quad + 2 \mathbb{E} \int_0^T |\sigma_s(X_s(\pi), \pi_s) - \sigma_s(X_s(\pi'), \pi'_s)|^2 ds. \end{aligned}$$

Due to assumption of the lemma we have that

$$\begin{aligned} \mathbb{E} \sup_{0 \leq t \leq T} |X_t(\pi) - X_t(\pi')|^2 &\leq 4TK^2 \mathbb{E} \int_0^T (|X_s(\pi) - X_s(\pi')|^2 + \|\pi_s - \pi'_s\|_{\text{TV}}^2) ds \\ &\quad + 4K^2 \mathbb{E} \int_0^T (|X_s(\pi) - X_s(\pi')|^2 + \|\pi_s - \pi'_s\|_{\text{TV}}^2) ds. \end{aligned}$$

Hence, by Gronwall's inequality we obtain

$$\mathbb{E} \sup_{0 \leq t \leq T} |X_t(\pi) - X_t(\pi')|^2 \leq C \mathbb{E} \int_0^T \|\pi_s - \pi'_s\|_{\text{TV}}^2 ds.$$

□

Bibliography

- [1] H. Pham, *Continuous-time stochastic control and optimization with financial applications*, vol. 61. Springer Science & Business Media, 2009.
- [2] G. dos Reis and D. Šiška, *Stochastic Control and Dynamic Asset Allocation*. Lecture notes, <https://www.maths.ed.ac.uk/~dsiska/LecNotesSCDAA.pdf>, 2020.
- [3] H. Dong and N. V. Krylov, “The rate of convergence of finite-difference approximations for parabolic Bellman equations with Lipschitz coefficients in cylindrical domains,” *Applied Mathematics and Optimization*, vol. 56, no. 1, pp. 37–66, 2007.
- [4] I. Gyöngy and D. Šiška, “On finite-difference approximations for normalized Bellman equations,” *Applied Mathematics and Optimization*, vol. 60, no. 3, pp. 297–339, 2009.
- [5] R. Bellman, “Functional equations in the theory of dynamic programming. v. positivity and quasi-linearity,” *Proceedings of the National Academy of Sciences*, vol. 41, no. 10, pp. 743–746, 1955.
- [6] R. Bellman, *Dynamic programming*. Princeton University Press, New Jersey, USA, 1957.
- [7] R. A. Howard, *Dynamic programming and markov processes*. MIT Press, Cambridge, USA, 1960.
- [8] M. L. Puterman and S. L. Brumelle, “On the convergence of policy iteration in stationary dynamic programming,” *Mathematics of Operations Research*, vol. 4, no. 1, pp. 60–69, 1979.
- [9] M. Puterman, “On the convergence of policy iteration for controlled diffusions,” *Journal of Optimization Theory and Applications*, vol. 33, no. 1, pp. 137–144, 1981.
- [10] M. S. Santos and J. Rust, “Convergence properties of policy iteration,” *SIAM Journal on Control and Optimization*, vol. 42, no. 6, pp. 2094–2115, 2004.
- [11] N. Bäuerle and U. Rieder, “Control improvement for jump-diffusion processes with applications to finance,” *Applied Mathematics and Optimization*, vol. 65, no. 1, pp. 1–14, 2012.

- [12] O. Bokanowski, S. Maroso, and H. Zidani, “Some convergence results for Howard’s algorithm,” *SIAM Journal on Numerical Analysis*, vol. 47, no. 4, pp. 3001–3026, 2009.
- [13] S. D. Jacka and A. Mijatović, “On the policy improvement algorithm in continuous time,” *Stochastics*, vol. 89, no. 1, pp. 348–359, 2017.
- [14] S. D. Jacka, A. Mijatovic, and D. Siraj, “Coupling and a generalised policy iteration algorithm in continuous time,” *arXiv preprint arXiv:1707.07834*, 2017.
- [15] J. Maeda and S. D. Jacka, “Evaluation of the rate of convergence in the PIA,” *arXiv preprint arXiv:1709.06466*, 2017.
- [16] R. Carmona, *Lectures on BSDEs, stochastic control, and stochastic differential games with financial applications*. SIAM, 2016.
- [17] E. Pardoux and S. Peng, “Backward stochastic differential equations and quasilinear parabolic partial differential equations,” in *Stochastic partial differential equations and their applications*, pp. 200–217, Springer, 1992.
- [18] F. Chernousko and A. Lyubushin, “Method of successive approximations for solution of optimal control problems,” *Optimal Control Applications and Methods*, vol. 3, no. 2, pp. 101–114, 1982.
- [19] Q. Li, L. Chen, C. Tai, and W. E, “Maximum principle based algorithms for deep learning,” *Journal of Machine Learning Research*, vol. 18, pp. 1–29, 2018.
- [20] J. Zhang, *Backward Stochastic Differential Equations: From Linear to Fully Nonlinear Theory*, vol. 86. Springer, 2017.
- [21] F. Antonelli, *Backward forward stochastic differential equations*. PhD thesis, Purdue University, 1993.
- [22] E. Pardoux and S. Tang, “Forward-backward stochastic differential equations and quasilinear parabolic PDEs,” *Probability Theory and Related Fields*, vol. 114, no. 2, pp. 123–150, 1999.
- [23] J. Ma, P. Protter, and J. Yong, “Solving forward-backward stochastic differential equations explicitly—a four step scheme,” *Probability theory and related fields*, vol. 98, no. 3, pp. 339–359, 1994.
- [24] F. Delarue, “On the existence and uniqueness of solutions to FBSDEs in a non-degenerate case,” *Stochastic processes and their applications*, vol. 99, no. 2, pp. 209–286, 2002.
- [25] J. Douglas Jr, J. Ma, and P. Protter, “Numerical methods for forward-backward stochastic differential equations,” *The Annals of Applied Probability*, vol. 6, no. 3, pp. 940–968, 1996.

- [26] G. Milstein and M. Tretyakov, “Discretization of forward–backward stochastic differential equations and related quasi-linear parabolic equations,” *IMA journal of numerical analysis*, vol. 27, no. 1, pp. 24–44, 2007.
- [27] J. Ma, J. Shen, and Y. Zhao, “On numerical approximations of forward-backward stochastic differential equations,” *SIAM Journal on Numerical Analysis*, vol. 46, no. 5, pp. 2636–2661, 2008.
- [28] F. Delarue and S. Menozzi, “A forward–backward stochastic algorithm for quasi-linear PDEs,” *The Annals of Applied Probability*, vol. 16, no. 1, pp. 140–184, 2006.
- [29] W. Guo, J. Zhang, and J. Zhuo, “A monotone scheme for high-dimensional fully nonlinear PDEs,” *The Annals of Applied Probability*, vol. 25, no. 3, pp. 1540–1580, 2015.
- [30] Y. Hu and S. Peng, “Solution of forward-backward stochastic differential equations,” *Probability Theory and Related Fields*, vol. 103, no. 2, pp. 273–283, 1995.
- [31] S. Peng and Z. Wu, “Fully coupled forward-backward stochastic differential equations and applications to optimal control,” *SIAM Journal on Control and Optimization*, vol. 37, no. 3, pp. 825–843, 1999.
- [32] J. Yong, “Finding adapted solutions of forward–backward stochastic differential equations: method of continuation,” *Probability Theory and Related Fields*, vol. 107, no. 4, pp. 537–572, 1997.
- [33] S. Ji, S. Peng, Y. Peng, and X. Zhang, “Three algorithms for solving high-dimensional fully coupled FBSDEs through deep learning,” *IEEE Intelligent Systems*, vol. 35, no. 3, pp. 71–84, 2020.
- [34] J. Han and J. Long, “Convergence of the deep BSDE method for coupled FB-SDEs,” *Probability, Uncertainty and Quantitative Risk*, vol. 5, no. 1, pp. 1–33, 2020.
- [35] N. V. Krylov, *Controlled diffusion processes*, vol. 14. Springer Science & Business Media, 2008.
- [36] W. H. Fleming and H. M. Soner, *Controlled Markov processes and viscosity solutions*, vol. 25. Springer Science & Business Media, 2006.
- [37] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, vol. 30. Springer Science & Business Media, 2012.
- [38] R. Buckdahn and S. Peng, “Ergodic backward SDE and associated PDE,” in *Seminar on Stochastic Analysis, Random Fields and Applications*, pp. 73–85, Springer, 1999.

- [39] M. Fuhrman and G. Tessitore, “Infinite horizon backward stochastic differential equations and elliptic equations in Hilbert spaces,” *The Annals of Probability*, vol. 32, no. 1B, pp. 607–660, 2004.
- [40] V. G. Boltyanskii, R. V. Gamkrelidze, and L. S. Pontryagin, “The theory of optimal processes. i. the maximum principle,” *Izv. Akad. Nauk SSSR Ser. Mat.*, vol. 24, no. 1, pp. 3–42, 1960.
- [41] L. S. Pontryagin, *Mathematical theory of optimal processes*. CRC press, 1987.
- [42] B. Kerimkulov, D. Šiška, and L. Szpruch, “Exponential convergence and stability of Howard’s policy improvement algorithm for controlled diffusions,” *SIAM Journal on Control and Optimization*, vol. 58, no. 3, pp. 1314–1340, 2020.
- [43] J. Han, A. Jentzen, and W. E, “Solving high-dimensional partial differential equations using deep learning,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 34, pp. 8505–8510, 2018.
- [44] M. S. Vidales, D. Šiška, and L. Szpruch, “Unbiased deep solvers for parametric PDEs,” *arXiv preprint arXiv:1810.05094*, 2018.
- [45] D. Šiška and L. Szpruch, “Gradient flows for regularized stochastic control problems,” *arXiv preprint arXiv:2006.05956*, 2020.
- [46] N. El Karoui, S. Peng, and M. C. Quenez, “Backward stochastic differential equations in finance,” *Mathematical finance*, vol. 7, no. 1, pp. 1–71, 1997.
- [47] J. Harter and A. Richou, “A stability approach for solving multidimensional quadratic BSDEs,” *Electronic Journal of Probability*, vol. 24, pp. 1–51, 2019.
- [48] B. Kerimkulov, D. Šiška, and L. Szpruch, “A modified MSA for stochastic control problems,” *Applied Mathematics and Optimization*, vol. 84, p. 3417–3436, 2021.
- [49] S. Ji and R. Xu, “A modified method of successive approximations for forward-backward stochastic control systems,” *arXiv preprint arXiv:2102.01484*, 2021.
- [50] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, and M. Geist, “Leverage the average: an analysis of KL regularization in RL,” *Neural Information Processing Systems*, 2020.
- [51] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, “Reinforcement learning with deep energy-based policies,” in *International Conference on Machine Learning*, pp. 1352–1361, PMLR, 2017.
- [52] C. Reisinger and Y. Zhang, “Regularity and stability of feedback relaxed controls,” *SIAM Journal on Control and Optimization*, vol. 59, no. 5, pp. 3118–3151, 2021.

- [53] H. Wang, T. Zariphopoulou, and X. Y. Zhou, “Reinforcement learning in continuous time and space: A stochastic control approach,” *Journal of Machine Learning Research*, vol. 21, no. 198, pp. 1–34, 2020.
- [54] H. Wang and X. Y. Zhou, “Continuous-time mean–variance portfolio selection: A reinforcement learning framework,” *Mathematical Finance*, vol. 30, no. 4, pp. 1273–1308, 2020.
- [55] K. Hu, Z. Ren, D. Šiška, and L. Szpruch, “Mean-field Langevin dynamics and energy landscape of neural networks,” *arXiv preprint arXiv:1905.07769*, 2019.
- [56] N. Kazamaki, *Continuous exponential martingales and BMO*. Springer, 2006.
- [57] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observation,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.
- [58] S. Kullback, “A lower bound for discrimination information in terms of variation (corresp.),” *IEEE transactions on Information Theory*, vol. 13, no. 1, pp. 126–127, 1967.
- [59] M. S. Pinsker, *Information and information stability of random variables and processes*. Holden-Day, 1964.
- [60] A. Richou, “Markovian quadratic and superquadratic BSDEs with an unbounded terminal condition,” *Stochastic Processes and their Applications*, vol. 122, no. 9, pp. 3173–3208, 2012.
- [61] V. Dem’yanov and A. Rubinov, “Approximate methods for solving extremal problems (in Russian),” *Izdat.Leningrad.Gos.Univ.*, 1968.
- [62] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games with Applications I: Mean Field Games with Common Noise and Master Equations*, vol. 83. Springer, 2018.