Probabilistic Framework of Howard's Policy Iteration: BML Evaluation and Robust Convergence Analysis

Yutian Wang, Yuan-Hua Ni, Zengqiang Chen, and Ji-Feng Zhang, *Fellow, IEEE*

***Abstract*— This paper aims to build a probabilistic framework for Howard's policy iteration algorithm using the language of forward-backward stochastic differential equations (FBSDEs). As opposed to conventional formulations based on partial differential equations, our FBSDE-based formulation can be easily implemented by optimizing criteria over sample data, and is therefore less sensitive to the state dimension. In particular, both on-policy and off-policy evaluation methods are discussed by constructing different FBSDEs. The backward-measurability-loss (BML) criterion is then proposed for solving these equations. By choosing specific weight functions in the proposed criterion, we can recover the popular Deep BSDE method or the martingale approach for BSDEs. The convergence results are established under both ideal and practical conditions, depending on whether the optimization criteria are decreased to zero. In the ideal case, we prove that the policy sequences produced by proposed FBSDE-based algorithms and the standard policy iteration have the same performance, and thus have the same convergence rate. In the practical case, the proposed algorithm is still proved to converge robustly under mild assumptions on optimization errors.**

***Index Terms*— forward-backward stochastic differential equations, policy iteration, stochastic optimal control**

I. Introduction

As an abstract description of policy-based methods, such as policy iteration (PI) [1]–[6] and policy gradient methods [7]–[9], the general policy iteration (GPI) for optimal control problems works as follows.

1) *(Initialization.)* Given an initial policy α^0 and set $n \leftarrow 0$.
2) *(GPI Subroutine.)* Given a policy α^{n-1}, find a new policy α^n in the policy space $\mathcal{A}$.
3) Set $n \leftarrow n+1$ and go back to step 2.

This work was supported in part by National Key R&D Program of China under Grant 2018YFA0703800, and in part by the National Natural Science foundation of China under Grants 62173191 and 61973175. (*Corresponding Author: Yuan-Hua Ni*).

Yutian Wang is with the Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: yutian.wang@connect.polyu.hk). Yuan-Hua Ni and Zengqiang Chen are with the College of Artificial Intelligence, Nankai University, Tianjin, China (e-mail: yhni@nankai.edu.cn; chenzq@nankai.edu.cn). Ji-Feng Zhang is with the Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100149, China (e-mail: jif@iss.ac.cn).

The key element of GPI is step 2, referred to as the GPI subroutine in this paper, which takes the current policy α^{n-1} as inputs, along with some other arguments if needed, and returns a new policy α^n. For example, in policy gradient methods, the new policy is obtained via gradient descent in the policy space. That subroutine is carefully designed such that the generated policy sequence $\{\alpha^n\}$ of GPI converges to, or approaches in some sense, an optimal policy α^*.

Originally developed by Howard for the Markov process model [1], Howard's policy improvement procedure (an instance of GPI subroutines), along with the policy iteration method, has been widely applied to optimal control problems, from discrete to continuous, deterministic to stochastic and linear to nonlinear systems [10]–[14]. A major advantage of Howard's policy iteration (hereafter referred to as the standard PI) is its fast convergence rate. For discrete time and state problems, Puterman and Brumelle [2] pointed out that the standard PI can be regarded as an instance of Newton's method, noting that both are finding zeros of a nonlinear operator. Based on this crucial observation, they successfully established a local quadratic convergence rate, which is also a standard result for Newton's iterative scheme in root finding problems. For linear quadratic regulation (LQR) problems in continuous time and state, the value function sequence generated by the standard policy iteration also converges quadratically [10]. Another interesting property of the standard PI is its robustness against numerical errors. For stochastic nonlinear systems, Kerimkulov *et al*. [15] analyzed the standard PI with perturbation errors. They employed the theory of backward stochastic differential equations (BSDEs) to estimate the performance error bound; see also [16] for perturbation discussion on continuous-time LQR problem.

Howard's policy improvement procedure is usually recognized as two consecutive steps: policy evaluation and policy improvement. The purpose of policy evaluation is to collect quantitative information on the current policy, or more specifically, the value function of the policy. Based on this information, the policy improvement step constructs a new policy that guarantees a monotone increase in performance. In this work, we focus on policy evaluation, and assume that a minimizing function for policy improvement exists and is accessible [4], [15]. Most early methods of policy evaluation obtain value functions by solving the differential Bellman equation, a first or second order linear partial differential equation (PDE)

[10], [17], [18]. Since traditional finite difference methods for PDEs generally suffer from the curse of dimensionality [19], integral PI [11] and temporal difference learning [20], [21] are preferred in practice. In addition to aforementioned works that focus on deterministic case, Jia and Zhou [22] investigated policy evaluation in stochastic settings with a finite planning horizon. They extended temporal difference learning to stochastic systems, and proposed a martingale approach which can be viewed as the stochastic counterpart of integral PI. It is worth noting that their martingale approach utilized a forward-backward stochastic differential equation (FBSDE), which is precisely the stochastic representation of the value function. From this point of view, their work is closely related to early policy evaluation methods utilizing PDEs, as Feynman-Kac's formula relates FBSDEs and PDEs [23]. On the other hand, Han *et al.* [24] proposed Deep BSDE method as a numerical approach for high-dimensional PDEs, where the problem is transformed into an optimization problem subject to FBSDEs by nonlinear Feynman-Kac's formula.

We conclude the literature review with a brief introduction to FBSDE computations, specifically focusing on the connection between FBSDEs and PDEs. Feynman-Kac type formulae established a relationship between FBSDEs and PDEs, enabling the transformation of FBSDE problems into PDE problems. To solve FBSDEs, various numerical methods have been employed, including finite element, finite difference, and sparse grid methods [25], [26]. However, these methods encounter challenges when dealing with high-dimensional problems due to their exponential complexity. In recent years, deep learning approaches, such as the Deep BSDE method, have emerged as promising solutions for handling high-dimensional problems without requiring space discretization [24], [27]. These methods leverage neural networks to provide efficient and accurate solutions to FBSDEs and related PDE problems. Additionally, techniques like multi-layer Picard iteration and neural network optimization have shown potential in solving nonlinear and high-dimensional PDEs [28]. The utilization of deep learning-based methods has proven valuable in practical applications, as they provide general solutions to FBSDEs and related PDE problems. These methods offer a powerful toolset for tackling complex systems in various fields.

Contributions. The main contributions of this paper are as follows. **1)** Motivated by these two parallel applications of Feynman-Kac type formulae [22], [24], we rigorously build the FBSDE-based framework of policy evaluation. In particular, we propose two FBSDE-based GPI subroutines are proposed that, under certain assumptions, are shown to be equivalent to conventional PDE-based subroutine used in Howard's policy iteration. This in turn shows GPI equipped with proposed subroutines converges as fast as the standard PI. **2)** We propose a novel optimization-based formulation of policy evaluation, whereby value function gradients are evaluated rather than the value function itself. In the case of inexact policy evaluation, we present a robust convergence result in terms of the optimization errors. **3)** We propose a versatile criterion for the optimization problem in policy evaluation. As the solution to the FBSDE constraint is not known a priori, we prove that it is equivalent to optimizing the proposed backward-measurability-loss (BML) criterion. By selecting different weight functions in the BML criterion, we are able to recover the Deep BSDE method in [24] as well as the martingale approach in [22]. Combining with the time discretization scheme in [29], our method can also be used to solving FBSDEs and Feynman-Kac type PDEs. See also Figure 1 for an overview of our policy iteration framework.

Organizations. This paper is organized as follows. In Section II, we set up the stochastic optimal control problem and review the concept of value functions. In Section III, we state the standard policy iteration algorithm and present a global linear convergence result. Two FBSDE-based policy iteration algorithms are introduced and analyzed in Section IV. In addition to the ideal convergence results, a robust convergence analysis is offered regarding optimization errors. Section V discusses the optimization problems in proposed algorithms. Numerical examples are present in Section VI [1]

Notations. Notations to be used frequently are summarized as follows. **1)** About probability theory and stochastic analysis. An element $\xi \in L^2_{\mathcal{F}}$ is a $\mathcal{F}$-measurable function with $\mathbb{E}\left\|\xi\right\|^2 < \infty$. $W^{t,T} \equiv \{W^{t,T}_s : t \le s \le T\}$ denotes a d-dimensional Brownian motion starting at $W^{t,T}_t = 0$. $\mathbb{S}^2(t,T)$ denotes the set of adapted process Y satisfying $\mathbb{E}[\sup_{t\le s\le T}|Y_s|^2] < \infty$. $\mathbb{H}^2(t,T)$ denotes the set of adapted process Z satisfying $\mathbb{E}\int_t^T \left\|Z_s\right\|^2 ds < \infty$. When there is no ambiguity, we drop the dependencies on t and T in these notations. **2)** About optimal control and reinforcement learning. We use $x \in \mathbb{R}^n$ and $a \in \mathbb{R}^m$ to denote the state and the action (control). A function α is termed a (feedback-control) policy if it maps time-state pairs to control values. We use F^α to indicate that a quantity F depends on a policy α and F^* to indicate the quantity corresponding to the optimal policy. Moreover, for a quantity $F(t,x,a)$ depending on the time-state-action triple, we write $F^a(\cdot,\cdot) \equiv F(\cdot,\cdot,a)$ and $F^\alpha(\cdot,\cdot) \equiv F(\cdot,\cdot,\alpha(\cdot,\cdot))$ if a is a control value and α is a control policy. **3)** About vector space. For elements in Euclidean space, $\left\|\cdot\right\|$ stands for the L^2 norm and $\langle\cdot,\cdot\rangle$ stands for the standard inner product. **4)** About functional classes. We use $w \in C^{1,2}$ to say that w is continuously differentiable with respect to the first variable and twice continuously differentiable with respect to the second variable. In Section III-A, we also introduce the notation $\phi \in C_b^{\mathrm{UniLip}}$ to say that ϕ is uniformly Lipschitz continuous and uniformly bounded.

II. Preliminaries

In this section, we review some basic concepts and results in general stochastic optimal control theory. For a comprehensive description of this subject, please refer to the monograph [30].

A. Problem settings

We consider an optimal control problem with system dynamics governed by the stochastic differential equation (SDE):

$$X_s = x + \int_t^s b^\alpha(\tau, X_\tau)\, d\tau + \int_t^s \sigma(\tau, X_\tau)\, dW_\tau. \quad (1)$$

[1] The code and additional numerical experiments are available at `https://github.com/Dou-Meishi/TAC-PIBSDE`.

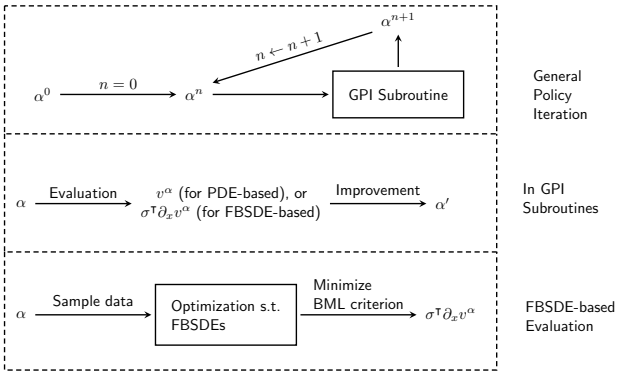


Fig. 1. Hierarchical illustration of the proposed policy iteration framework. At the top level of the hierarchy is GPI, which iterates in the policy space. At the midlevel is the GPI subroutine, and at the bottom is the optimization formulation of policy evaluation.

The solution to this equation, denoted by $X^{\alpha,t,x}$ or simply X^α, is a controlled diffusion process, depending on both the policy α and the starting point (t,x). Let us fix the initial time-state pair (t,x) at first. Eq. (1) is studied on an underlying probability space $(\Omega,\mathcal{F},\mathbb{P})$, which is required to be complete and admits a standard d-dimensional Brownian motion $\{W_s\}_{t\le s\le T}$ with $W_t=0$. Here, $T<\infty$ is the planning horizon. We equip $(\Omega,\mathcal{F},\mathbb{P})$ with the natural filtration $\{\mathcal{F}_s\}_{t\le s\le T}$ generated by $\{W_s\}_{t\le s\le T}$. Note that the definition of $\{W_s,\mathcal{F}_s; t\le s\le T\}$ relies on the choice of $t\in[0,T]$.[2] We develop our theory with fixed (t,x) and the generalization to varying (t,x) is straightforward by substituting specific values.

The (controlled) drift coefficient b^α and diffusion coefficient σ are measurable functions defined on $[0,T]\times\mathbb{R}^n$. In particular, b^α is defined by another measurable functions $b:[0,T]\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^n$ and a policy $\alpha:[0,T]\times\mathbb{R}^n\to\mathbb{R}^m$, i.e., $b^\alpha:(t,x)\mapsto b(t,x,\alpha(t,x))$. Under certain conditions on b^α and σ, there exists an adapted process $X^{\alpha,t,x}$ satisfying Eq. (1) $\mathbb{P}$-a.s. for any $s\in[t,T]$; see, for example, Karatzas and Shreve [23]. Here, by saying a process is adapted, we mean it is progressively measurable [3].

The cost of a policy α starting at (t,x) is measured by the following expectation:

$$v^\alpha(t,x)\coloneqq\mathbb{E}\left[\int_t^T f^\alpha(s,X_s^{\alpha,t,x})\,ds+g(X_T^{\alpha,t,x})\right]. \tag{2}$$

Here, $f:[0,T]\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ and $g:\mathbb{R}^n\to\mathbb{R}$ are measurable functions, and f^α is defined in terms of f and α, in the same way as b^α is defined in terms of b and α. A control policy is said to be admissible if it takes value in $A\subset\mathbb{R}^m$ and the solution to Eq. (1) uniquely exists. We denote by $\mathcal{A}$ the collection of all admissible policies. When the policy α is fixed, the function $v^\alpha:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ is called the value function of α. In addition, the following infimum:

$$v^*(t,x)\coloneqq\inf_{\alpha\in\mathcal{A}} v^\alpha(t,x) \tag{3}$$

is called the optimal value function.

The stochastic optimal control problem, in view of Eq. (1)–(3), is then stated as finding $\alpha^*\in\mathcal{A}$ such that $v^*(t,x)=v^{\alpha^*}(t,x)$ for a given pair (t,x).

B. Characterizing value functions via PDEs

Using dynamic programming, we can link value functions to a family of PDEs. Specifically, the dynamic programming principle states that

$$v^*(t,x)=\inf_{\alpha\in\mathcal{A}}\mathbb{E}\left[\int_t^{t+\epsilon} f^\alpha(s,X_s^{\alpha,t,x})\,ds+v^*(t+\epsilon,X_{t+\epsilon}^{\alpha,t,x})\right]. \tag{4}$$

Recall that for any sufficient smooth v, there is $\mathscr{L}^\alpha v(t,x)=\lim_{\epsilon\to0}\frac{1}{\epsilon}\mathbb{E}\big[v(t+\epsilon,X_{t+\epsilon}^{\alpha,t,x})-v(t,x)\big]$ with $\mathscr{L}^\alpha$ the infinitesimal generator the associate to Eq. (1)

$$\mathscr{L}^\alpha v\coloneqq\partial_t v+\langle b^\alpha,\partial_x v\rangle+\frac{1}{2}\operatorname{tr}\{\sigma\sigma^\intercal\partial_{xx}v\}. \tag{5}$$

Here, we drop the dependency on (t,x) for simplicity. Dividing Eq. (4) by ϵ and taking $\epsilon\to0$ lead to a second order partial differential equation. Setting $t=T$ in the definition (3) yields a boundary condition. Putting these all together and varying (t,x) lead to the following second order nonlinear Cauchy problem for the optimal value function

$$\begin{cases}0=\inf\limits_{a\in A}\{\mathscr{L}^a v^*(t,x)+f^a(t,x)\}, & \forall(t,x)\in[0,T)\times\mathbb{R}^n,\\ v^*(T,x)=g(x), & \forall x\in\mathbb{R}^n,\end{cases} \tag{6}$$

which is exactly the Hamilton-Jacobi-Bellman (HJB) equation.

Following the similar arguments of Eq. (4)–(6) leads to the following linear Cauchy problem for the value function

$$\begin{cases}0=\mathscr{L}^\alpha v^\alpha(t,x)+f^\alpha(t,x), & \forall(t,x)\in[0,T)\times\mathbb{R}^n,\\ v^\alpha(T,x)=g(x), & \forall x\in\mathbb{R}^n,\end{cases} \tag{7}$$

where the infimum is absent because this value function might be not optimal. We refer to this as the PDE characterization of value functions.

C. Characterizing value functions via FBSDEs

As a result of Feynman-Kac's formula, solutions to PDEs (7) admit FBSDEs representation, and therefore it is possible to characterized value functions with FBSDEs. To see this, one may apply Itô's rule to find that

$$dv^\alpha(s,X_s^\alpha)=\mathscr{L}^\alpha v^\alpha(s,X_s^\alpha)\,ds+\langle\sigma^\intercal\partial_x v^\alpha(s,X_s^\alpha),dW_s\rangle.$$

Substituting Eq. (7) into this equality and combining Eq. (1) yield the FBSDE characterization of v^α

$$\begin{cases}X_s=x+\int_t^s b^\alpha(\tau,X_\tau)\,d\tau+\int_t^s\sigma(\tau,X_\tau)\,dW_\tau,\\ Y_s=g(X_T)+\int_s^T f^\alpha(\tau,X_\tau)\,d\tau-\int_s^T\langle Z_\tau,dW_\tau\rangle,\\ Y_s=v^\alpha(s,X_s),\quad\forall s\in[t,T],\quad d\mathbb{P}\text{-a.s.},\\ Z_s=\sigma^\intercal\partial_x v^\alpha(s,X_s),\quad ds\otimes d\mathbb{P}\text{-a.e. on }[t,T]\times\Omega\end{cases} \tag{8}$$

under some conditions ensuring the solution's existence and uniqueness. We shall point out that this FBSDE is not in the

[2]This is known as the weak formulation of stochastic optimal control problems in [30]. The main motivation of this formulation is that we can deal with a family of stochastic optimal control problems by varying (t,x).

[3]Strictly speaking, an adapted process need not be progressively measurable. But, if it is also measurable, then it has a stochastic equivalent process which is indeed progressively measurable [31].

most general form. In Eq. (8), the forward SDE does not contain the backward part Y as well as the control part Z. This means that the FBSDE is decoupled, and we can separately solve the forward SDE and the backward SDE.

The PDE characterization Eq. (7) and FBSDE characterization Eq. (8), along with HJB Eq. (6), are fundamental motivations of this paper. However, in deriving these equations, we implicitly assume that v^* and v^α are sufficiently smooth. This is nontrivial, especially for HJB Eq. (6), which is strongly nonlinear. Though the nonlinear Feynman-Kac formula is still valid in viscosity settings, the difficulty lies in connecting the Z process of the FBSDE to the optimal control u^* when $\partial_x v^*$ does not exist. Nevertheless, we focus on problems such that this assumption holds, as the nonsmooth solution to HJB equation is already a broad topic, in which the concept of viscosity solutions must be introduced [32]. Extensions to the nonsmooth case might be considered in future works.

To conclude this section, we point out that the HJB Eq. (6) characterizing the optimal value function is a nonlinear PDE, while its reduced form Eq. (7), satisfied by the value function of a given policy, is linear. From this point of view, the standard PI manages to approximate the solution to a nonlinear PDE with a sequence of solutions to linear PDEs. This linearization coincides with the idea of Newton's method for finding zeros, regarding some abstract arguments of general derivatives. However, as discussed in the last section, solving PDEs directly generally suffers the curse of dimensionality, and thus prevents applications in large-scale problems. This is the reason why we need the probabilistic formulation Eq. (8).

III. The PDE-based Policy Iteration Algorithm

In this section, we reformulate the system dynamics, state our assumptions, and present a global linear convergence result of the standard policy iteration algorithm. At last, we highlight two key issues with this PDE-based algorithm.

A. Problem reformulation and assumptions

In this paper, we consider a slightly different system description other than the general form Eq. (1). Specifically, we require the drift coefficient can be decomposed in a way such that the control-dependent term is explicitly coupled with the diffusion coefficient: $\forall (t,x,a) \in [0,T] \times \mathbb{R}^n \times A$,

$$b(t,x,a) = \bar{b}(t,x) + \sigma(t,x)\hat{b}(t,x,a). \tag{9}$$

Namely, $b(t,x,a)$ can be split into two parts; one $\bar{b}(t,x)$ is independent of control, and the other one $\sigma(t,x)\hat{b}(t,x,a)$ is control-dependent. It seems too restrictive at the first glance. But, if $\sigma\sigma^\intercal$ is nondegenerate, i.e., $(\sigma\sigma^\intercal)^{-1}$ exists on $[0,T] \times \mathbb{R}^n$, then the desired decomposition exists. Indeed, we can choose $\bar{b} \equiv 0$ and $\hat{b} \equiv \sigma^\intercal(\sigma\sigma^\intercal)^{-1} b$. Also, we require that a measurable minimizing function μ is given such that for any $(t,x,z) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^d$,

$$\mu(t,x,z) \in \operatorname*{arginf}_{a\in A} \big\{ \langle \hat{b}^a(t,x), z\rangle + f^a(t,x) \big\}. \tag{10}$$

This function is useful in canceling the painful infimum operator in HJB equation. To see this, we note that the diffusion coefficient σ is independent of control, and thus, for any (t,x) and smooth function $v(\cdot,\cdot)$,

$$\begin{aligned} &\operatorname*{arginf}_{a\in A} \{\mathscr{L}^a v(t,x) + f^a(t,x)\} \\ =\ &\operatorname*{arginf}_{a\in A} \{\langle \bar{b} + \sigma\hat{b}^a, \partial_x v(t,x)\rangle + f^a(t,x)\} \\ =\ &\mu(t,x,\sigma^\intercal \partial_x v(t,x)). \end{aligned}$$

We should stress that this property holds only for $b = \bar{b} + \sigma\hat{b}$. Without the explicit appearance of $\sigma\hat{b}$, the definition of μ would be problematic. However, for the affine system and quadratic control cost, which is main topic of adaptive dynamic programming [11], [33]–[37], the minimizer of the right-hand side of Eq. (10) uniquely exists and admits a closed analytic form. In particular, suppose that b is linear in a (then so is $\hat{b}$) and that f is quadratic in a, and that A is closed and convex. Then, μ can be obtained by projecting the minimizer of a quadratic function onto a closed convex set. See also [15] for a more general discussion on the existence of μ. Nevertheless, μ is regarded as an abstract representation of the minimizer. For complex systems where Eq. (10) does not admit closed form solutions, an intermediate numerical solver could be embedded. A rigorous analysis in this direction exceeds the scope of this paper. Interested readers may refer to [38] for a possible resolution.

In order to rigorously state our algorithm and establish the desired convergence results, we need to pose some conditions on our problem. At first, we recall the useful uniform Lipschitz continuity and uniform boundness, which are able to ensure the existence and uniqueness of solutions to SDEs and BSDEs.

Definition 1 (Uniform Lipschitz continuity and boundness)**.** *A continuous function $\phi(t,x,y)$ is said to be uniformly Lipschitz continuous in x,y with respect to t if there exists a positive constant L such that for any $t \in E^1,\ x,x' \in E^2,\ y,y' \in E^3$,*

$$\|\phi(t,x,y) - \phi(t,x',y')\| \le L\|x-x'\| + L\|y-y'\|, \tag{11}$$

where E^1, E^2, E^3 are nonempty subsets of Euclidean spaces with proper dimensions.

Further, ϕ is said to be uniformly bounded if there exists a constant L such that (suppose $0 \in E^2, E^3$)

$$\|\phi(t,0,0)\| \le L, \quad \forall t \in E^1. \tag{12}$$

For convenience, let $C^{\mathrm{UniLip}}(E^1 \times E^2 \times E^3)$ denote the collection of functions satisfying Eq. (11)*, and $C_b^{\mathrm{UniLip}}(E^1 \times E^2 \times E^3)$ denote the collection of functions satisfying both Eq.* (11) *and Eq.* (12)*.*

Convention 1. *For continuous function $\phi^1(t,x)$ or $\phi^2(x)$, we mean ϕ^1 or $\phi^2 \in C^{\mathrm{UniLip}}(E^1 \times E^2 \times E^3)$ if the extended function $\tilde{\phi}^1$ or $\tilde{\phi}^2 \in C^{\mathrm{UniLip}}(E^1 \times E^2 \times E^3)$, where*

$$\tilde{\phi}^1(t,x,\cdot) \equiv \phi^1(t,x), \quad \tilde{\phi}^2(\cdot,x,\cdot) \equiv \phi^2(x).$$

We apply this simplification to C_b^{UniLip} too.

Assumption 1. *Let the following assumptions hold.*

1) *The functions $\bar{b}, \hat{b}, \sigma, f, g \in C^{\mathrm{UniLip}}([0,T] \times \mathbb{R}^n \times A)$. Moreover, the given minimizing function $\mu \in C^{\mathrm{UniLip}}([0,T] \times \mathbb{R}^n \times \mathbb{R}^d)$.*

2) *The functions* $\bar{b}, \mu, f$ *are uniformly bounded:* $\forall t \in [0,T]$,

$$\|\bar{b}(t,0,0)\| + \|\mu(t,0,0)\| + |f(t,0,0)| \leq L;$$

and $\hat{b}, \sigma$ *are bounded:* $\forall (t,x,a) \in [0,T] \times \mathbb{R}^n \times A$,

$$\|\hat{b}(t,x,a)\| + \|\sigma(t,x)\| \leq L.$$

3) *For any* $\alpha \in C_b^{\mathrm{UniLip}}([0,T] \times \mathbb{R}^n)$, *the linear Cauchy problem Eq.* (7) *has a smooth solution* $w^\alpha \in C^{1,2}([0,T] \times \mathbb{R}^n)$ *such that* $\partial_x w^\alpha \in C_b^{\mathrm{UniLip}}$. *Moreover, the HJB Eq.* (6) *has such a smooth solution* v^* *too.*

Remark III.1. As a matter of fact, one essential condition on the existence of a smooth solution to HJB equation (6) is the uniform elliptic condition: $\exists \delta > 0$ such that $y^\intercal \sigma \sigma^\intercal y \geq \delta y^\intercal y$ holds for any $(t,x,y) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^n$. Clearly, this condition is also sufficient to ensure the existence of $\bar{b}$ and $\hat{b}$.

Remark III.2. Under Assumption 1.1 and Assumption 1.2, we have $b, \sigma, f, g, \mu \in C_b^{\mathrm{UniLip}}$, and thus, for any policy $\alpha \in C_b^{\mathrm{UniLip}}$ taking values in A, there is $b^\alpha \in C_b^{\mathrm{UniLip}}$. Hence, the solution to Eq. (1) uniquely exists for any (t,x). Moreover, for any $\ell > 1$, $\mathbb{E}[\sup_{t \leq s \leq T} \|X_s^{\alpha,t,x}\|^\ell]$ is finite [39].

Remark III.3. In linear quadratic problems, the assumptions that f and g are bounded and Lipschitz are violated. In practice, however, we can make some minor modifications to the problem in order to satisfy these assumptions. The idea is manually clipping the control and state in these functions below a certain threshold. For example, if $f(t,x,a) = x^\intercal Q x + a^\intercal R a$, then $\tilde{f}(t,x,a) = f(t,\tilde{x},\tilde{a})$ may be used, where $\tilde{x}$ and $\tilde{a}$ are component-wise clipped versions of x and a, respectively. By choosing a sufficiently large threshold, we can still obtain a satisfactory suboptimal control policy of the original problem.

We stress that Assumption 1 might not be the most general condition to make above assertions. But, it is very convenient to illustrate our key ideas without getting too involved into abstract theories of PDEs and SDEs. In particular, we have the following lemma to characterize value functions, which also serves as a starting point for the following subsections.

Lemma 1. *Let Assumption 1 hold. Then, for any policy* $\alpha \in C_b^{\mathrm{UniLip}}([0,T] \times \mathbb{R}^n)$ *valued in* A, *the value function* v^α, *defined by Eq.* (1) *and Eq.* (2) *is a unique solution to PDE* (7) *with* $v^\alpha \in C^{1,2}$. *Moreover,* v^α *admits the stochastic representation Eq.* (8).

Proof: This is a direct consequence of Remark III.2 and [30, Theorem 7.4.1]. ∎

Remark III.4. Under Assumption 1.3, $\partial_x v^\alpha \in C_b^{\mathrm{UniLip}}$, and thus, $\mu(\cdot,\cdot,\sigma^\intercal \partial_x v^\alpha(\cdot,\cdot))$ is a policy valued in A and lies in C^{UniLip}. It is also important to note that in the stochastic representation Eq (8), the term $\sigma^\intercal \partial_x v^\alpha$ is encoded in the Z process. Therefore, obtaining Z is to some extent sufficient to construct the policy $\mu(\cdot,\cdot,\sigma^\intercal \partial_x v^\alpha(\cdot,\cdot))$.

B. The standard policy iteration subroutine

Let us focus on the HJB Eq. (6) and the PDE characterization Eq. (7). Suppose α is an optimal policy, then v^α satisfies both of these equations. Combining Eq. (6) and Eq. (7), for any $(t,x) \in [0,T] \times \mathbb{R}^n$, we have

$$\mathscr{L}^\alpha v^\alpha(t,x) + f^\alpha(t,x) = \inf_{a \in A} \{\mathscr{L}^a v^\alpha(t,x) + f^a(t,x)\}. \quad (13)$$

Conversely, if this equation is satisfied by some policy α, then its value function v^α satisfies the HJB equation. Hence, the central idea of policy iteration is to force Eq. (13) to hold.

The standard policy iteration algorithm works as follows.

1) Given a policy α, find its value function v^α by Eq. (7).
2) Given v^α, find a policy α' such that for any (t,x),

$$\alpha'(t,x) = \operatorname*{arginf}_{a \in A} \{\mathscr{L}^a v^\alpha(t,x) + f^a(t,x)\}.$$

Alternatively repeating these two steps generates a sequence of policies. The first step is also known as policy evaluation, and the second step is policy improvement. According to Eq. (10), the policy improvement step can also be realized by setting

$$\alpha'(t,x) \coloneqq \mu(t,x,z(t,x)), \quad \forall (t,x) \in [0,T] \times \mathbb{R}^n, \quad (14)$$

where $z(\cdot,\cdot) = \sigma^\intercal \partial_x v^\alpha(\cdot,\cdot)$. For simplicity, we combine policy evaluation and policy improvement into a single procedure and refer to it as the standard policy iteration subroutine, or the PDE-based subroutine; see Algorithm 1. The global convergence result of GPI equipped with this subroutine is provided in Proposition 1.

Algorithm 1 A PDE-based subroutine of GPI.

Input: a feedback control policy α.
Output: a feedback control policy α' not worse than α.
1: Obtain the value function v^α by Eq. (7).
2: Construct the output policy by Eq. (14) with $z \leftarrow \sigma^\intercal \partial_x v^\alpha$.

Proposition 1. *Let Assumption 1 hold. Starting at an initial policy* α^0 *valued in* A, *let* $\{\alpha_n\}_{n \in \mathbb{N}}$ *denote the policy sequence generated by GPI equipped with Algorithm 1. If* $\alpha^0 \in C_b^{\mathrm{UniLip}}([0,T] \times \mathbb{R}^n)$ *is valued in* A, *then* α^n *is admissible for any* $n \geq 0$. *For any* $(t,x) \in [0,T] \times \mathbb{R}^n$, *the cost sequence* $\{v^{\alpha^n}(t,x)\}_{n \in \mathbb{N}}$ *is monotonically decreasing to* $v^*(t,x)$. *Moreover, there exists a constant* $C = C(t,x)$ *depending on* (t,x) *and a constant* $q \in (0,1)$ *independent to* (t,x) *such that*

$$|v^{\alpha^n}(t,x) - v^*(t,x)| \leq C(t,x) q^n, \quad \textit{for any } n \geq 0. \quad (15)$$

Proof: The assertion of admissibility is a direct consequence of Remark III.4. The monotonicity is also expected due to the definition of μ [12]. Under our assumptions, Eq. (15) can be demonstrated by following the proof of [15, Theorem 4.1], so we omit this technical proof here. The proof of Eq. (15) can also be viewed as a simplified version of the proof of Theorem 4; see Remark IV.4 for more details. ∎

C. Two key issues

To this end, we have formulated the PDE-based subroutine in Algorithm 1 and developed corresponding convergence results. Sadly, we have to admit that the global linear convergence rate in Proposition 1 generally cannot be achieved

with a practical program. The dilemma arises from the policy evaluation step.

The first issue is the design of numerical methods for policy evaluation. In Algorithm 1, policy evaluation is formulated as solving PDEs, which generally has no closed form solution and has to be solved with numerical methods. Traditional numerical ways for PDEs require discretizing the time-state space, and thus, suffer from the curse of dimensionality. Moreover, extending traditional ways to model-free settings seems to be challenging. Based on these considerations, another two policy iteration subroutines utilizing the FBSDE characterization of value functions are proposed in Section IV. We also develop a numerical method for solving FBSDEs by optimizing a novel criterion; see Section V.

The second issue is more subtle. Since numerical methods cannot be expected to provide the exact solution, especially after time discretization, approximation errors are generally inevitable. Consequently, the improved policy based on this inexact solution is different from the expected output policy. To address this issue, we quantify these approximation errors as ϵ_n and analyze the convergence of the policy iteration with $\epsilon_n > 0$. We discuss this topic at the end of Section IV.

IV. FBSDE-based Policy Iteration Algorithms

In this section, we propose two FBSDE-based policy iteration algorithms. The convergence result is established by showing the equivalence between the PDE-based and FBSDE-based policy iteration subroutines. At last, we present a robust convergence result with respect to approximation errors. In all following sections, the initial pair of time states (t, x) is fixed.

A. The on-policy subroutine

In the PDE-based subroutine, the next trial policy is constructed by μ and $\sigma^\intercal \partial_x v^\alpha$, where the latter is obtained via solving PDE (7). In view of Lemma 1, it is very natural to consider carrying out policy evaluation by solving FBSDE (8). We formulate this idea in Algorithm 2.

The second step of Algorithm 2 is the key of this work. Instead of evaluating v^α via a linear PDE and substituting $\partial_x v^\alpha$ into the policy improvement step, we directly obtain a z^α term via an optimization problem, and then construct the next trial policy based on it. We will discuss in detail how to minimize the objective function Eq. (16) in Section V. Here, we simply assume that there is a method that can be used to determine the global solution z^α.

Algorithm 2 The on-policy subroutine of GPI.

Input: a feedback control policy α; an initial point (t, x).
Output: a feedback control policy α' not worse than α.

1: Find the solution X^α to the forward SDE (1).
2: Find an optimal solution z^α to the optimization problem

$$\min_{z \in C_b^{\mathrm{UniLip}}} \epsilon^\alpha := \mathbb{E} \int_t^T \|z(s, X_s^\alpha) - Z_s^\alpha\|^2 \, ds, \qquad (16)$$

where Z^α is a part of the solution to the BSDE in Eq. (8).
3: Construct the output policy by Eq. (14) with $z \leftarrow z^\alpha$.

Comparing the policies returned by Algorithm 2 and Algorithm 1, it can be seen that z^α plays the role of $\sigma^\intercal \partial_x v^\alpha$. According to Lemma 1, $\sigma^\intercal \partial_x v^\alpha$ is indeed a global solution to that optimization problem. Noting that $Z_s^\alpha = \sigma^\intercal \partial_x v^\alpha(s, X_s^\alpha)$ holds almost everywhere on the product space $[t, T] \times \Omega$, we can rewrite the objective function Eq. (16) as

$$\epsilon^\alpha(z) = \mathbb{E} \int_t^T h(s, X_s^\alpha) \, ds, \qquad (17)$$

where $h(\cdot, \cdot) := \|z(\cdot, \cdot) - \sigma^\intercal \partial_x v^\alpha(\cdot, \cdot)\|^2 \geq 0$. Hence, we have $\epsilon^\alpha(z^\alpha) = 0$. In the opposite direction, however, one cannot say that $\sigma^\intercal \partial_x v^\alpha$ is the unique optimal solution in C_b^{UniLip}, since $h \equiv 0$ is not the necessary condition of $\epsilon^\alpha = 0$. In fact, the necessary and sufficient condition is h equals zero almost everywhere on the product space under the measure induced by $X^\alpha(s, \omega)$. To put it another way, we can only say that $z^\alpha(\cdot, \cdot)$ equals $\sigma^\intercal \partial_x v^\alpha(\cdot, \cdot)$ almost everywhere along the process $X^{\alpha, t, x}$. Fortunately, Lemma 2 below suggests that this almost everywhere identity is enough to guarantee that Algorithm 1 and Algorithm 2 are equivalent, in the sense that the returned policies have the same cost value.

Before proceeding, we would like to clarify one more point regarding this algorithm. The first two steps for obtaining z^α can be implemented in a pure data-driven fashion. The forward state process $\{X_s^\alpha\}_{t \leq s \leq T}$ can be sampled by sending the current policy α to the dynamic system and observe the state trajectory. Furthermore, it is possible to solve that optimization problem using only samples without knowing the exact solution (Y^α, Z^α). This is the reason why we call Algorithm 2 on-policy. In the next subsection, we introduce the off-policy subroutine, where the forward SDE is driven by a fixed behavior policy α^b instead of the current policy α.

Lemma 2. *Let Assumption 1 hold. For any $\alpha^1, \alpha^2 \in C_b^{\mathrm{UniLip}}([0, T] \times \mathbb{R}^n)$, let X^1, X^2 be their state processes, respectively. Then, for any nonnegative measurable function $h(\cdot, \cdot) \geq 0$, the following statements are equivalent:*

1) *$h(s, X_s^1) = 0$ holds $ds \otimes d\mathbb{P}$-a.e. on $[t, T] \times \Omega$;*
2) *$h(s, X_s^2) = 0$ holds $ds \otimes d\mathbb{P}$-a.e. on $[t, T] \times \Omega$.*

Proof: Consider the following two auxiliary processes

$$W_s^i = W_s + \int_t^s \hat{b}^{\alpha^i}(\tau, X_\tau^i) \, d\tau, \quad s \in [t, T], \quad i = 1, 2.$$

Noting that $\{\hat{b}^{\alpha^i}(s, X_s^i); t \leq s \leq T\}$ is bounded and thus satisfies Novikov condition, there exists probability measure $\mathbb{P}^i$, equivalent to $\mathbb{P}$, such that W^i becomes a standard Brownian motion under $\mathbb{P}^i$. This is known as the Girsanov's theorem [23, Chapter 3]. Therefore, $(X^1, W^1, \mathbb{P}^1)$ and $(X^2, W^2, \mathbb{P}^2)$ are two weak solutions to the following SDE:

$$X_s = x + \int_t^s \bar{b}(\tau, X_\tau) \, d\tau + \int_t^s \sigma(\tau, X_\tau) \, dW_\tau.$$

By the uniformly Lipschitz continuity and boundness of $\bar{b}$ and σ, the strong existence and uniqueness hold for this SDE. Then the weak uniqueness in the sense of probability law holds too, namely, X^1 and X^2 have the same law. Thus, the integral of

$h(s, X_s^1)$ equals the integral of $h(s, X_s^2)$:

$$\int_t^T \left(\int h(s, X_s^1)\, d\mathbb{P}^1 \right) ds = \int_t^T \left(\int h(s, X_s^2)\, d\mathbb{P}^2 \right) ds.$$

We conclude that $h(s, X_s^1) = 0$ holds $ds \otimes d\mathbb{P}^1$-a.e. if and only if $h(s, X_s^2) = 0$ holds $ds \otimes d\mathbb{P}^2$-a.e.. The proof is finished by noting that $\mathbb{P}, \mathbb{P}^1, \mathbb{P}^2$ are equivalent to each other. ■

This lemma offers the freedom of changing the underlying process in the optimization problem of the on-policy subroutine. By setting $h(\cdot,\cdot)$ as Eq. (17), this lemma suggests that minimizing the $\mathbb{E}\int_t^T h(s, X_s^\alpha)$ to zero is equivalent to minimizing $\mathbb{E}\int_t^T h(s, X_s^{\alpha^b})$ to zero for any $\alpha^b \in C_b^{\text{UniLip}}$. Thus, it is also reasonable to choose a policy α^b different from α and optimize the integral of h along X^{α^b}. On the other hand, let $\mathbb{E}\int_t^T h(s, X_s^\alpha) = 0$ hold and α' be the policy returned by Algorithm 1. Then, $h(s, X_s^{\alpha'}) = 0$ holds almost everywhere on the product space $[t, T] \times \Omega$. This argument is also the key to prove the following equivalence between the PDE-based subroutine and the on-policy subroutine.

Theorem 2. *Let Assumption 1 hold. For an input policy $\alpha \in C_b^{\text{UniLip}}([0, T] \times \mathbb{R}^n)$ valued in A, let α_1' and α_2' denote the outputs of Algorithm 1 and Algorithm 2, respectively. Then, α_1' and α_2' generate the "same" trajectory starting at (t, x):*

$$X_s^{\alpha_1', t, x} = X_s^{\alpha_2', t, x}, \quad ds \otimes d\mathbb{P}\text{-a.e. on } [t, T] \times \Omega.$$

Moreover, $v^{\alpha_1'}(t, x) = v^{\alpha_2'}(t, x)$.

Proof: Let v^α and z^α denote the same objects in Algorithm 1 and Algorithm 2, repsectively. We write down the explicit expression of α_1', α_2':

$$\alpha_1'(\cdot,\cdot) = \mu(\cdot,\cdot,\sigma^\intercal \partial_x v^\alpha(\cdot,\cdot)), \quad \alpha_2'(\cdot,\cdot) = \mu(\cdot,\cdot, z^\alpha(\cdot,\cdot)),$$

and denote by $h(\cdot,\cdot) = \|z^\alpha(\cdot,\cdot) - \sigma^\intercal \partial_x v^\alpha(\cdot,\cdot)\|^2$.

According to Remark III.4, α_1', α_2' are admissible. Consider the forward SDEs satisfied by $X^{\alpha_1'}, X^{\alpha_2'}$:

$$X_s^{\alpha_1'} = x + \int_t^s b^{\alpha_1'}(\tau, X_\tau^{\alpha_1'})\, d\tau + \int_t^s \sigma(\tau, X_\tau^{\alpha_1'}) dW_s,$$
$$X_s^{\alpha_2'} = x + \int_t^s b^{\alpha_2'}(\tau, X_\tau^{\alpha_2'})\, d\tau + \int_t^s \sigma(\tau, X_\tau^{\alpha_2'}) dW_s.$$

We claim that

$$\alpha_1'(s, X_s^{\alpha_1'}) = \alpha_2'(s, X_s^{\alpha_1'}), \quad ds \otimes d\mathbb{P}\text{-a.e. on } [t, T]. \tag{18}$$

Indeed, it can be concluded from Lemma 1 that $h(s, X_s^\alpha) = 0$ almost everywhere on $[t, T] \times \Omega$. Then, applying Lemma 2 yields $h(s, X_s^{\alpha_1'}) = 0$ almost everywhere. Denote by

$$\widetilde{X}_s^{\alpha_1'} = x + \int_t^s b^{\alpha_2'}(\tau, X_\tau^{\alpha_1'})\, d\tau + \int_t^s \sigma(\tau, X_\tau^{\alpha_1'}) dW_s$$

and $\phi(u) := \mathbb{E}\int_t^{t+u} \|X_\tau^{\alpha_1'} - X_\tau^{\alpha_2'}\|^2\, d\tau$. Noting Eq. (18) and the Lipschitz continuity of $b^{\alpha_2'}$ and σ, we have

$$\begin{aligned}
\phi(u) &= \mathbb{E}\int_t^{t+u} \|\widetilde{X}_s^{\alpha_1'} - X_s^{\alpha_2'}\|^2\, ds \\
&\le \mathbb{E}\int_t^{t+u} \left\{ 2\left[\int_t^s (b^{\alpha_2'}(\tau, X_\tau^{\alpha_1'}) - b^{\alpha_2'}(\tau, X_\tau^{\alpha_2'}))\, d\tau \right]^2 \right. \\
&\quad \left. + 2\left[\int_t^s (\sigma(\tau, X_\tau^{\alpha_1'}) - \sigma(\tau, X_\tau^{\alpha_2'}))\, dW_\tau \right]^2 \right\} ds \\
&\le \mathbb{E}\int_t^{t+u} 2(s - t + 1)L^2 \int_t^s \|X_\tau^{\alpha_1'} - X_\tau^{\alpha_2'}\|^2\, d\tau\, ds \\
&\le 2(T+1)L^2 \int_0^u \phi(s)\, ds, \quad \forall u \in [0, T - t].
\end{aligned}$$

Hence, by Grönwall's inequality, there is $\phi(T - t) = 0$. This proves that $X_s^{\alpha_1'} = X_s^{\alpha_2'}$ almost everywhere on $[t, T] \times \Omega$. Moreover, the cost of α_1' and α_2' at (t, x) is equal. ■

Remark IV.1. This result reveals that there is no difference between the cost sequence produced by GPI using the PDE-based subroutine and the on-policy subroutine. Thus, all the convergence properties of the standard PI is preserved in our probabilistic framework.

Corollary 1. *For any fixed $(t, x) \in [0, T] \times \mathbb{R}^n$, the conclusions of Proposition 1 hold if Algorithm 1 is replaced by Algorithm 2.*

Remark IV.2. Because the output of Algorithm 2 may depend on the argument (t, x), we cannot make a conclusion that $\{v^{\alpha^n}(t', x')\}$ is monotone at any (t', x') as in Proposition 1. Nevertheless, the cost sequence $\{v^{\alpha^n}(t, x)\}$ is still monotonically decreasing, where (t, x) is the argument passed into Algorithm 2.

B. The off-policy subroutine

On-policy and off-policy are terminologies in reinforcement learning [5]. They are different in the way of collecting data. In an on-policy algorithm, a value function of a policy α is evaluated with data collected by itself. This corresponds to FBSDE (8), where the forward SDE is driven by α and the solution to the backward SDE is related to v^α too. However, in an off-policy algorithm, v^α is generally evaluated with data collected by a different policy, called the behavior policy α^b usually. The advantage of off-policy algorithms is the high data efficient. If we adopt the on-policy subroutine Algorithm 2 in GPI, then the current policy α generally changes during the iteration. Therefore, we have to resample data at the beginning of each iteration, i.e., solving a new forward SDE in our case.

With the help of nonlinear Feynman-Kac's formula, it is straightforward to extend the on-policy FBSDE characterization of value function to the off-policy case.

Lemma 3. *Let the condition of Lemma 1 hold and use the same notation. For any policy $\alpha^b \in C_b^{\text{UniLip}}([0, T] \times \mathbb{R}^n)$ valued in A, the value function v^α admits the following*

stochastic representation:

$$\begin{cases} X_s^b = x + \int_t^s b^{\alpha^b}(\tau, X_\tau^b)\, d\tau + \int_t^s \sigma(\tau, X_\tau^b)\, dW_\tau, \\ Y_s = g(X_T^b) + \int_s^T f^\alpha(\tau, X_\tau^b)\, d\tau - \int_s^T \langle Z_\tau, dW_\tau \rangle \\ \qquad + \int_s^T \langle \hat{b}^\alpha(\tau, X_\tau^b) - \hat{b}^{\alpha^b}(\tau, X_\tau^b), Z_\tau \rangle\, d\tau, \\ Y_s = v^\alpha(s, X_s^b), \quad \forall s \in [t,T], \quad d\mathbb{P}\text{-}a.s., \\ Z_s = \sigma^\intercal \partial_x v^\alpha(s, X_s^b), \quad ds \otimes d\mathbb{P}\text{-}a.e.\ on\ [t,T]\times\Omega. \end{cases} \tag{19}$$

Proof: By the definitions of $\bar{b}, \hat{b}$ and μ, we can rewrite the PDE satisfied by v^α as follows:

$$\begin{cases} 0 = \langle \hat{b}^\alpha(t,x) - \hat{b}^{\alpha^b}(t,x), \sigma^\intercal \partial_x v^\alpha(t,x) \rangle \\ \qquad + \mathscr{L}^{\alpha^b} v^\alpha(t,x) + f^\alpha(t,x), \quad \forall (t,x) \in [0,T)\times\mathbb{R}^n, \\ v^\alpha(T,x) = g(x), \quad \forall x \in \mathbb{R}^n. \end{cases}$$

Applying the nonlinear Feynman-Kac's formula [30, Theorem 7.4.5] to this leads to the desired representation. ∎

Remark IV.3. If $\alpha^b \equiv \alpha$, this degenerates to Lemma 1. It is important to note that the forward process X^b is independent of α, which is the key difference between the on-policy and off-policy methods. GPI equipped with the off-policy subroutine and a fixed α^b should be viewed as an iteration of BSDEs, while that equipped with the on-policy subroutine should be viewed as an iteration of FBSDEs.

Based on Lemma 3, we propose Algorithm 3, in which the optimization problem is modified according to Eq. (19). It can be concluded from Lemma 2 that the optimization problems in Algorithm 2 and Algorithm 3 have the same solutions. In light of this observation, we are able to prove that the returned polices of on-policy and off-policy subroutines are equivalent.

Algorithm 3 The off-policy subroutine of GPI.

Input: policies α, α^b; an initial condition (t,x).
Output: a policy α' not worse than α.

1: Find the solution X^b to the forward SDE (1) with $\alpha \leftarrow \alpha^b$.
2: Find an optimal solution z^α to the optimization problem

$$\min_{z \in C_b^{\mathrm{UniLip}}} \epsilon^\alpha := \mathbb{E}\int_t^T \|z(s, X_s^b) - Z_s^{\alpha,b}\|^2\, ds, \tag{20}$$

where $Z^{\alpha,b}$ is a part of the solution to the BSDE in Eq. (19).
3: Construct the output policy by Eq. (14) with $z \leftarrow z^\alpha$.

Theorem 3. *Let the condition of Theorem 2 hold and use the same notation. If $\alpha^b \in C_b^{\mathrm{UniLip}}([0,T]\times\mathbb{R}^n)$ is valued in A, then the output policies of Algorithm 2, 3, denoted by α_2', α_3', generate the "same" trajectory starting at (t,x):*

$$X_s^{\alpha_2', t, x} = X_s^{\alpha_3', t, x}, \quad ds \otimes d\mathbb{P}\text{-}a.e.\ on\ [t,T]\times\Omega.$$

Moreover, $v^{\alpha_2'}(t,x) = v^{\alpha_3'}(t,x)$.

Proof: The proof is similar to the proof of Theorem 2 except that we need to show

$$\alpha_2'(s, X_s^{\alpha_2'}) = \alpha_3'(s, X_s^{\alpha_2'}), \quad ds \otimes d\mathbb{P}\text{-a.e. on } [t,T]\times\Omega.$$

Let $z_i^\alpha (i = 2,3)$ be the term z^α in Algorithm 2 and Algorithm 3, respectively. Using Lemma 1–3, we have

$$z_i^\alpha(s, X_s^{\alpha_2'}) = \sigma^\intercal \partial_x v^\alpha(s, X_s^{\alpha_2'}), \quad ds \otimes d\mathbb{P}\text{-a.e. on } [t,T]\times\Omega.$$

Substituting this into the definition of α_i' finishes our proof. ∎

In view of Theorem 2 and Theorem 3, we conclude that these three subroutines are equivalent to each other. Consequently, the following convergence result for Algorithm 3 holds.

Corollary 2. *For any fixed $(t,x) \in [0,T]\times\mathbb{R}^n$ and $\alpha^b \in C^{\mathrm{UniLip}}([0,T]\times\mathbb{R}^n)$ valued in A, the conclusions of Proposition 1 hold if Algorithm 1 is replaced by Algorithm 3.*

We conclude the discussion on the proposed on-policy and off-policy subroutines by a comment on their partially model-free property. In this work, $\bar{b}, \hat{b}, \sigma, g, f$ are recognized as the model knowledge, and the minimizer μ as a combination knowledge of $\hat{b}$ and f, as shown in Eq. (10). However, it is easily to see that the on-policy method requires only f, g and μ to improve a given policy, while the off-policy method uses an additional term $\hat{b}$. Both methods can work without knowing all system dynamics if desired state trajectories for training are available. This enables the learning and improvement of control policies in a partially model-free setting.

C. A robust convergence result

Consider the optimization problem in Algorithm 3. In view of Lemma 3, there exists a z^α with $\epsilon^\alpha(z^\alpha) = 0$. In practice, however, it is usually the case that we can only find a suboptimal solution $\hat{z}$ and thus $\epsilon^\alpha(\hat{z}) > 0$. If we construct a policy by Eq. (14) with $z \leftarrow \hat{z}$, then there is no guarantee that this new policy $\hat{\alpha}$ performs better than the current policy α. To see this, we apply Itô's formula to obtain (noting the PDEs satisfied by value functions)

$$\begin{aligned} & v^\alpha(t,x) - v^{\hat{\alpha}}(t,x) \\ &= \mathbb{E}\int_t^T (\mathscr{L}^{\hat{\alpha}} v^{\hat{\alpha}} - \mathscr{L}^{\hat{\alpha}} v^\alpha)(s, X_s^{\hat{\alpha}})\, ds \\ &= \mathbb{E}\int_t^T \big(\mathscr{L}^\alpha v^\alpha + f^\alpha - \mathscr{L}^{\hat{\alpha}} v^\alpha - f^{\hat{\alpha}}\big)(s, X_s^{\hat{\alpha}})\, ds. \end{aligned}$$

If $\hat{z} = z^\alpha$, then $\hat{\alpha}(s, X_s^{\hat{\alpha}}) = \mu(s, X_s^{\hat{\alpha}}, \sigma^\intercal \partial_x v^\alpha(s, X_s^{\hat{\alpha}}))$ almost everywhere on $[t,T]\times\Omega$, and thus, $v^\alpha(t,x) - v^{\hat{\alpha}}(t,x)$ equals

$$\mathbb{E}\int_t^T \big(\mathscr{L}^\alpha v^\alpha + f^\alpha - \inf_{a\in A}\{\mathscr{L}^a v^\alpha + f^a\}\big)(s, X_s^{\hat{\alpha}})\, ds \geq 0.$$

If $\epsilon^\alpha(\hat{z}) > 0$, then generally $v^\alpha(t,x) \geq v^{\hat{\alpha}}(t,x)$ does not hold, and thus the monotonicity of policy improvement is broken.

Below, we study the case in which the objective function in the off-policy subroutine does not reach zero during policy iterations. Though the cost sequence $\{v^{\alpha^n}(t,x)\}$ may be not monotone, we show that it still converges to the optimal cost if the n-th objective value ϵ_n converges to zero. To make it more clear, we spell the policy iteration procedure in Algorithm 4. In comparison to the GPI that is equipped with the off-policy subroutine, Algorithm 4 contains two important differences.

The first difference is that the behavior policy α^b is fixed during iteration. This is not the only way to apply the off-policy BSDE subroutine in GPI, as it can be proved that the cost of the output policy does not change if α^b is different. In order to view the whole algorithm as the iteration of BSDEs, however, we do not allow the forward SDE changes during iteration. The second difference is that z^n is not necessarily an optimal solution of Eq. (20). Also, ϵ_n is not necessarily equal to 0.

Algorithm 4 A BSDE-based Policy Iteration Algorithm.

Input: policies α^0, α^b; an initial condition (t, x).
Output: a sequence of policies $\{\alpha^n\}$.

1: Find the solution X^b to the forward SDE (1) with $\alpha \leftarrow \alpha^b$.
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: Run a numerical method to solve the optimization problem (20) with $\alpha \leftarrow \alpha^n$. Denote by z^n the returned solution and ϵ_n the associated objective value.
4: Construct α^{n+1} by Eq. (14) with $z \leftarrow z^n$.
5: **end for**

With notations defined in Algorithm 4, we can state our robust convergence result as follows.

Theorem 4. *Let Assumption 1 hold and use notations in Algorithm 4. If $\alpha^0, \alpha^b \in C_b^{\mathrm{UniLip}}([0,T] \times \mathbb{R}^n)$ are policies valued in A, then α^n is admissible for any $n \geq 0$. Moreover, there exist constants $q \in (0,1)$ and $\gamma > 0$, both independent of (t,x), such that the following inequality holds*

$$\limsup_{n\to\infty} \left|v^{\alpha^n}(t,x) - v^*(t,x)\right|^2 \leq \frac{q e^{\gamma(T-t)}}{1-q} \cdot \limsup_{n\to\infty} \epsilon_n.$$

Proof: See Appendix I. ∎

Remark IV.4. If $\alpha^b = \alpha^0$ and $\epsilon_n = 0$ for any n, then $c_n = 0$ for any n, and Eq. (33) is reduced to $a_n + b_n \leq q^n b_0$. Dropping b_n and expanding the definition of a_n yield the Eq. (15). This shows that our proof can be adopted to prove Proposition 1.

V. Solving FBSDEs by Optimization

In this section, we discuss how to solve the optimization problems encountered in the FBSDE-based subroutines, in which we propose a novel criterion, called the (general) BML criterion. Due to the uncoupling nature of the FBSDEs in our policy iteration algorithms, we focus on solving BSDEs.

A. A practical objective function

The on-policy subroutine involves a BSDE in the form

$$Y_s = \xi + \int_s^T f_\tau \, d\tau - \int_s^T \langle Z_\tau, dW_\tau \rangle, \quad \forall s \in [t,T]. \tag{21}$$

Specifically, for a trial process $z \in \mathbb{H}^2$, we are interested in calculating the distance $\mathbb{E}\int_t^T \|Z_s - z_s\|^2 \, ds$ between z and the true solution Z. The difficulty is that Z is not known and goes into the equation. Hence, we need to find practical objective functions that do not explicitly contain Z. For this purpose, the following theorem provides useful insights.

Theorem 5. *Suppose that $\xi \in L^2_{\mathcal{F}_T}$ and $f \in \mathbb{H}^2$. Then, BSDE* (21) *admits a unique adapted solution $(Y,Z) \in \mathbb{S}^2 \times \mathbb{H}^2$. For adapted process $z \in \mathbb{H}^2$, let $\widetilde{Y}^z_s$ denote the process (not necessarily adapted)*

$$\widetilde{Y}^z_s = \xi + \int_s^T f_\tau \, d\tau - \int_s^T \langle z_\tau, dW_\tau \rangle, \quad \forall s \in [t,T].$$

Then, it holds that

$$\mathbb{E}\,|\widetilde{Y}^z_t - \mathbb{E}\,\widetilde{Y}^z_t|^2 = \mathbb{E}\int_t^T \|Z_s - z_s\|^2 \, ds. \tag{22}$$

Proof: The uniqueness and existence are standard results for BSDEs; see [40, Chapter 6] for example. We rewrite the left-hand side of Eq. (22) as

$$\mathbb{E}\,|\widetilde{Y}^z_t - Y_t|^2 + 2\,\mathbb{E}[(\widetilde{Y}^z_t - Y_t)(Y_t - \mathbb{E}\,\widetilde{Y}^z_t)] + \mathbb{E}\,|Y_t - \mathbb{E}\,\widetilde{Y}^z_t|^2.$$

Due to the fact that $\mathcal{F}_t$ contains only $\mathbb{P}$-null sets, we know that $Y_t = \mathbb{E}\,Y_t$ holds almost surely. Moreover,

$$\mathbb{E}\,\widetilde{Y}^z_t = \mathbb{E}\left[\xi + \int_t^T f_s \, ds\right] = \mathbb{E}\,Y_t.$$

Thus, $Y_t - \mathbb{E}\,\widetilde{Y}^z_t$ is almost surely zero and

$$\begin{aligned}\mathbb{E}\,|\widetilde{Y}^z_t - \mathbb{E}\,\widetilde{Y}^z_t|^2 &= \mathbb{E}\,|\widetilde{Y}^z_t - Y_t|^2 \\ &= \mathbb{E}\left[-\int_t^T \langle z_s, dW_s\rangle + \int_t^T \langle Z_s, dW_s\rangle\right]^2.\end{aligned}$$

Thus, the desired equality holds due to Itô's isometry. ∎

Remark V.1. By Remark III.2, the BSDE in the on-policy subroutine satisfies the conditions here. Thus, $\mathbb{E}\,|\widetilde{Y}^z_t - \mathbb{E}\,\widetilde{Y}^z_t|^2$ can be used in the place of the objective function. We call this the special BML criterion, where its general form is discussed in the next subsection.

An intuitive explanation of the BML criterion is based on the measurability. By definition, $(\widetilde{Y}^z, z)$ has already satisfied the stochastic integral relationship as (Y,Z). Not surprisingly, this is not sufficient to conclude that it is a solution, as z is just arbitrarily selected. The key is that a true pair of solution (Y,Z) should also be adapted. That is to say, $\widetilde{Y}^z_s$ should be $\mathcal{F}_s$-measurable for any $s \in [t,T]$. This is not a trivial matter since the definition of $\widetilde{Y}^z_s$ involves the "future" information, particularly the $\{W_\tau\}_{s\leq\tau\leq T}$. Assume, however, that $\widetilde{Y}^z_{t'}$ has been proved to be $\mathcal{F}_{t'}$-measurable. Then it is safe to conclude that $\widetilde{Y}^z_s$ is $\mathcal{F}_s$-measurable for any $s \in [t',T]$. This is because for any $s \in [t',T]$, we have

$$\begin{aligned}\widetilde{Y}^z_{t'} &= \xi + \int_{t'}^T f_\tau \, d\tau - \int_{t'}^T \langle z_\tau, dW_\tau\rangle \\ &= \widetilde{Y}^z_s + \int_{t'}^s f_\tau \, d\tau - \int_{t'}^s \langle z_\tau, dW_\tau\rangle.\end{aligned}$$

Clearly, the integral part is $\mathcal{F}_s$-measurable. As a result, $\widetilde{Y}^z_s$ is $\mathcal{F}_s$-measurable because $\widetilde{Y}^z_{t'}$ is $\mathcal{F}_s$-measurable (recall that $\mathcal{F}_{t'} \subset \mathcal{F}_s$ if $t' \leq s$).

The left-hand side of Eq. (22) serves as a criterion of the measurability loss of $\widetilde{Y}^z_t$ with respect to $\mathcal{F}_t$. Recall that $\mathcal{F}_t = \sigma(\mathcal{N} \cup \sigma(W_t))$, where $\mathcal{N}$ is the collection of $\mathbb{P}$-null sets and $\sigma(W_t)$ is the trivial σ-algebra with $W_t = 0$. $\widetilde{Y}^z_t$ is $\mathcal{F}_t$-measurable if and only if $\widetilde{Y}^z_t$ is a constant almost surely. To

put it in another way, $\widetilde{Y}_t^z$ should be equal to the expectation almost surely. This is exactly the case that Eq. (22) equals 0.

B. The BML Criterion

According to Theorem 5, the distance $\mathbb{E}\int_t^T \|Z_s - z_s\|^2$ can be calculated with only samples of ξ, f and W in BSDE (21). This allows an optimization-based approach to solving the Z part of solutions by parameterizing the trial process z, and then minimizing the practical objective function. However, in many applications, obtaining the Y part of solutions may be appealing as well. Indeed, according to Feynman-Kac's formula, Y_t is the value function at (t, x). If we manage to find the exact or an approximated solution of Y, then we also find a method to solve PDEs in the form of Eq. (7).

In the proof of Theorem 5, we utilize the fact that $\mathbb{E}\,\widetilde{Y}_t^z = \mathbb{E}\,Y_t = Y_t$ holds almost surely. Unfortunately, $\widetilde{Y}^z$ is not a suitable replacement for Y in applications. The major issue is that the definition of $\widetilde{Y}^z$ is "anticipated". Even if $z \equiv Z$, calculating the value of $\widetilde{Y}_t^z$ by its definition requires samples of $\{W_s; t \le s \le T\}$ and $\{f_s; t \le s \le T\}$, which are not available at the time instant t. Nevertheless, $\widetilde{Y}^z$ differs from the true solution only by a martingale term, and this difference can be eliminated by taking conditional expectation:

$$\mathbb{E}[\widetilde{Y}_s^z \,|\, \mathcal{F}_s] = \mathbb{E}[Y_s \,|\, \mathcal{F}_s] = Y_s, \quad \mathbb{P}\text{-a.s.}, \quad \forall s \in [t, T]. \tag{23}$$

In light of this, we extend Theorem 5 by adding the distance between a trial solution $\tilde{v} \in \mathbb{S}^2$ and the true solution Y.

Theorem 6. *Let the condition of Theorem 5 hold and use the same notation. Then, for any adapted process $\tilde{v} \in \mathbb{S}^2$, there is*

$$\begin{aligned}\mathbb{E}\int_t^T |\widetilde{Y}_s^z - \tilde{v}_s|^2\,\nu(ds) = \mathbb{E}\int_t^T \int_s^T \|Z_\tau - z_\tau\|^2\,d\tau\,\nu(ds) \\ + \mathbb{E}\int_t^T |Y_s - \tilde{v}_s|^2\,\nu(ds),\end{aligned} \tag{24}$$

where ν is an arbitrary σ-finite measure on $[t, T]$.

Proof: Similarly, we prove Eq. (24) by splitting the square term into three terms and showing that the expectation of the cross term is zero. As ν is σ-finite, we are able to change the order of expectation and integration, and thus, the left-hand side of Eq. (24) equals

$$\int_t^T \left[\mathbb{E}\,|\widetilde{Y}_s^z - Y_s|^2 + 2\,\mathbb{E}[(\widetilde{Y}_s^z - Y_s)(Y_s - \tilde{v}_s)] + \mathbb{E}\,|Y_s - \tilde{v}_s|^2\right]\nu(ds).$$

The first term can be transformed with Itô's isometry:

$$\begin{aligned}\mathbb{E}\int_t^T |\widetilde{Y}_s^z - Y_s|^2\,\nu(ds) &= \mathbb{E}\int_t^T \left|\int_s^T \langle z_\tau - Z_\tau, dW_\tau\rangle\right|^2 \nu(ds) \\ &= \mathbb{E}\int_t^T \int_s^T \|z_\tau - Z_\tau\|^2\,d\tau\,\nu(ds).\end{aligned}$$

The second term vanishes according to the tower property of conditional expectation:

$$\begin{aligned}\mathbb{E}[(\widetilde{Y}_s^z - Y_s)(Y_s - \tilde{v}_s)] &= \mathbb{E}\big[\mathbb{E}[(\widetilde{Y}_s^z - Y_s)(Y_s - \tilde{v}_s) \,|\, \mathcal{F}_s]\big] \\ &= \mathbb{E}\big[(Y_s - \tilde{v}_s)\,\mathbb{E}[(\widetilde{Y}_s^z - Y_s) \,|\, \mathcal{F}_s]\big] \\ &= 0.\end{aligned}$$

The last equality comes from the fact that $\mathbb{E}[(\widetilde{Y}_t^z - Y_t) \,|\, \mathcal{F}_s]$ is zero almost surely. ■

Remark V.2. We call Eq. (24) the general BML criterion. While the special BML criterion focuses solely on the Z part, its generalization takes the Y part into account as well. We do this by replacing $\mathbb{E}\,Y_t^z$ with $\tilde{v}_s$. Moreover, Eq. (24) introduces a measure on the time space $[t, T]$. The left-hand side of Eq. (24) actually describes the distance between $\widetilde{Y}^z$ and $\tilde{v}$ on the product space $(\Omega \times [t, T], \mathbb{P} \otimes \nu)$. On the other hand, this practical objective function can also be interpreted as the distance between $(\tilde{v}, z)$ and (Y, Z) using this product measure. Under this generalization, we are given the freedom of choosing ν when comparing the trial solution with the true solution. In particular, if ν is set to the Dirac measure centered on t and $\tilde{v}$ to $\mathbb{E}[\widetilde{Y}_s^z \,|\, \mathcal{F}_s]$, then it comes to the special BML criterion. It is also possible to choose different settings of ν and $(\tilde{v}, z)$. It will be discussed shortly and how the general BML criterion degenerates into existing methods.

Remark V.3. It is worth noting that if the choice of $\tilde{v}$ does not rely on z, then the two terms in Eq. (24) are decoupled. This means that the gradient with respect to $\tilde{v}$ is independent of the gradient with respect to z. Therefore, z and $\tilde{v}$ can be optimized independently. In this case, our estimation of Z does not affect the estimation of Y, and vice versa. One advantage of this property is that even if z is actually far from the true solution Z, it is still possible to have a good estimation of Y that is fairly accuracy. As an application, we could fix $z \equiv 0$ and focus solely on estimating of Y by optimizing only $\tilde{v}$. According to our analysis, this simply results in the distance between z and Z remaining constant, and we may still be able to obtain a reasonable estimation of Y if the general BML criterion reaches its minimum.

By choosing $\nu = \delta_t$ and $\tilde{v}(s, \omega) \equiv y_0$, we recover the popular Deep BSDE method proposed in [24]. There, δ_t is the Dirac measure centered at t and $y_0 \in \mathbb{R}$ does not change along with time s and the sample event ω. The general BML criterion is then reduced to $\mathbb{E}\,|\widetilde{Y}_t - y_0|^2$, which can be interpreted as

$$\begin{aligned}&\mathbb{E}\left|\xi - \left(y_0 - \int_t^T f_s\,ds + \int_t^T \langle z_s, dW_s\rangle\right)\right|^2 \\ &= \mathbb{E}\int_t^T \|z_s - Z_s\|^2\,ds + \mathbb{E}\,|Y_t - y_0|^2\end{aligned} \tag{25}$$

by Theorem 6. The original motivation of Deep BSDE method is to examine the process

$$\widetilde{Y}_s^{z,y_0} = y_0 - \int_t^s f_\tau\,d_\tau, + \int_t^s \langle z_\tau, dW_\tau\rangle.$$

In fact, this is a forward stochastic differential equation. One can relate it to BSDE (21) by requiring $Y_T^{z,y_0} = \xi$ holds almost surely, i.e., forcing $\mathbb{E}\,|\xi - Y_T^{z,y_0}|^2 = 0$. This is exactly the criterion used in Deep BSDE method. If the choices of y_0 and z do not depend on each other, Remark V.3 reveals that this criterion is equivalent to $\mathbb{E}\,|Y_t - y_0|^2$ when one is only interested in estimating the value of Y_t. We should also mention that Deep BSDE method applies for a wider class of BSDEs other than the simple form Eq. (21). There, the generator f_s is coupled with (Y_s, Z_s) by a nonlinear function

f. In that case, Eq. (22) and Eq. (24) are no longer valid. We will briefly discuss that topic at the end of this section.

By choosing $\nu(ds) = ds$ and $z \equiv 0$, we recover the martingale approach proposed in [22]. The general BML criterion is then reduced to

$$\mathbb{E}\int_t^T \left| \left(\xi + \int_t^T f_\tau \, d\tau\right) - \left(\tilde{v}_s + \int_t^s f_\tau \, d\tau\right) \right|^2 ds$$
$$= \mathbb{E}\int_t^T \int_s^T \|Z_\tau\|^2 \, d\tau \, ds + \mathbb{E}\int_t^T |Y_s - \tilde{v}_s|^2 \, ds$$

by Theorem 6. In the martingale approach, one takes no care of the Z part of solution and just set the trial solution z to zero. This treatment is permitted by Remark V.3 as well. Minimizing the distance between $\widetilde{Y}^z$ and $\tilde{v}$ with $z \equiv 0$ is indeed equivalent to minimizing the distance between $\tilde{v}$ and the true solution Y. The similar result is reported along with the martingale approach in [22], but there is no discussion about its connection to Deep BSDE method.

Corollary 3. *Let the condition of Theorem 5 hold and use the same notation. For any $y_0 \in \mathbb{R}$ and $z \in \mathbb{H}^2$, let $\widehat{Y}_s^{z,y_0}$ denote the process*

$$\widehat{Y}_s^{z,y_0} = y_0 - \int_t^s f_\tau \, d\tau + \int_t^s \langle z_\tau, dW_\tau \rangle, \quad \forall s \in [t, T].$$

Then, it holds that

$$\min_{y_0 \in \mathbb{R}} \mathbb{E}\,|\widehat{Y}_T^{z,y_0} - \xi|^2 = \mathbb{E}\,|\widetilde{Y}_0^z - \mathbb{E}\,\widetilde{Y}_0^z|^2.$$

Proof: This is a direct consequence of Theorem 5 and Eq. (25). ∎

Remark V.4. In general, the criterion $\mathbb{E}\,|\widehat{Y}_s^{z,y_0} - \xi|^2$, used in Deep BSDE method, depends on both z and y_0. If y_0 is optimized with fixed z, it comes to the special BML criterion.

C. Optimize with the proposed criterion

In this subsection, we illustrate how to solve a BSDE by optimizing the proposed criterion. As discussed at the end of the last subsection, the general BML criterion is a class of objective functions and choosing different $(\nu, \tilde{v}, z)$ leads to different specific objective functions. We summarize four sets of $(v, \tilde{v}, z)$ in Table I and refer to them as Set (a)–(d). It should be noted that Set (a) and Set (c) are used in Deep BSDE method and the martingale approach, respectively. Set (b) corresponds to the special BML criterion proposed in Theorem 5, while Set (d) is considered here to show the general form helps us in finding new objective functions. It should be pointed out that $\tilde{v}_s = \mathbb{E}[\widetilde{Y}_s^z \,|\, \mathcal{F}_s]$in Set (b) is merely provided for completeness, and is not required for calculations. We stress that these four sets cover only a small part of the general BML criterion, and it is always possible to design appropriate forms of $\nu, \tilde{v}$ and z based on specific requirements. In order to focus on ideas, we test these four criteria on the following toy example. A more involved example will be discussed in the next section.

Example 1. *Solve the BSDE* (21) *with $t = 0, T = 1, f(s, \omega) \equiv -1, \xi = \langle W_T, W_T \rangle / n$, where n is the dimension of the Brownian motion and is set to 100.*

TABLE I
FOUR SPECIAL CASES OF THE GENERAL BML CRITERION.

Name	$d\nu/ds$	$\tilde{v}$	z	Practical objective function
Set (a)	δ_t	y_0	z_s	$\mathbb{E}\,\lvert\widetilde{Y}_0^z - y_0\rvert^2$
Set (b)	δ_t	$\mathbb{E}[\widetilde{Y}_s^z \,\lvert\, \mathcal{F}_s]$	z_s	$\mathbb{E}\,\lvert\widetilde{Y}_0^z - \mathbb{E}\,\widetilde{Y}_0^z\rvert^2$
Set (c)	1	$\tilde{v}_s$	0	$\mathbb{E}\int_t^T \lvert\widetilde{Y}_s^0 - \tilde{v}_s\rvert^2 \, ds$
Set (d)	1	$\tilde{v}_s$	z_s	$\mathbb{E}\int_t^T \lvert\widetilde{Y}_s^z - \tilde{v}_s\rvert^2 \, ds$

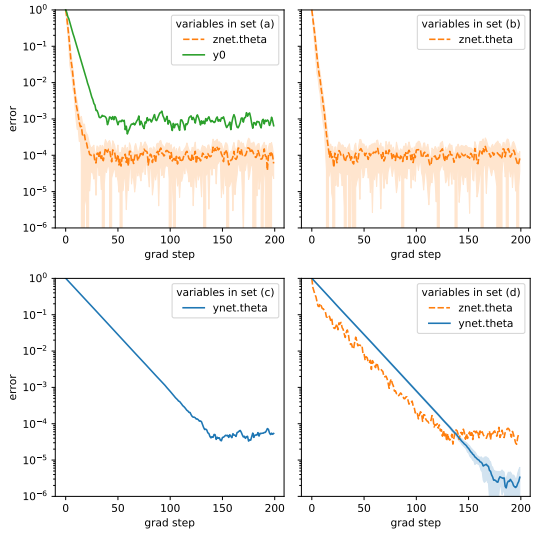


Fig. 2. The absolute errors of $\boldsymbol{\theta_y}, \boldsymbol{\theta_z}, \boldsymbol{y_0}$ at each gradient steps for Example 1. From left to right and from top to bottom, the subplots correspond to Set (a), (b), (c) and (d). The solid lines and shaded areas indicate the mean and standard deviation of absolute errors for 10 runs.

We parameterize the trial processes in Table I as $\tilde{v}_s = W_s^\intercal \theta_y W_s$, $z_s = 2\theta_z W_s$. Additionally, Set (a) involves optimizing a standalone variable y_0. The Brownian motion is simulated with time step $\Delta t = 0.01$. The expectation is estimated via Monte Carlo simulation with sample size $M = 16$. Integration is approximated with the Euler method. Optimization method is chosen as the standard stochastic gradient descent (SGD) method with different learning rates: 1.0×10^{-1} for y_0, 1.0×10^{-3} for θ_z and 1.0×10^{-5} for θ_y. The initial values of y_0, θ_y, θ_z are set to $1.0, -1.0, -1.0$, respectively. For each set, we perform 200 gradient steps and repeat the whole procedure 10 times with different random seeds. The true value of these variables are obtained via theoretical analysis. It can be verified by Itô's formula that $Y_s = \langle W_s, W_s \rangle / n, Z_s = 2W_s/n$ is a pair of adapted solution. This solution is also unique because $\xi \in L^2_{\mathcal{F}_T}$ and $f \in \mathbb{H}^2$. Thus, the optimal values are $\theta_y^* = \theta_z^* = 1/n$. Additionally, the y_0 in Set (a) is used to estimate the value of Y_0, and thus, has the optimal value $y_0^* = 0$. Results are reported in Figure 2.

Figure 2 plots the absolute errors of θ_y, θ_z, y_0 at each gradient steps in four subplots, corresponding to the four sets in Table I. It can be seen that all variables in these sets converge to their true values with fairly high accuracy in 200 gradient steps. There are two interesting phenomena of convergence trends. The first one is that θ_z converges very quickly in Set (a) and Set (b) with almost the same rate, but

is slightly slower in Set (d). The second one is that the θ_y in Set (d) converges to a better value than that in Set (c). We can explain them with the help Theorem 6.

According to Eq. (24), the objective functions in these four sets can be interpreted as follows. Set (a) minimizes $\mathbb{E}\int_t^T \|z_s - Z_s\|^2\,ds$, which also is the term to be minimized in Set (b) plus an additional term $\mathbb{E}\,|y_0 - Y_0|^2$. Therefore, the gradients for θ_z computed in Set (a) and Set (b) should be identical except for the noise introduced by Monte Carlo sampling. This is the reason why the convergent behavior of θ_z is similar in these two sets. On the other hand, the θ_z in Set (d) appears in a double integration $\mathbb{E}\int_t^T\int_s^T \|z_\tau - Z_\tau\|^2\,d\tau\,ds$ due to the choice of ν. In order to explain the second phenomena, we need to review the proof of Theorem 6. There, the cross-term is eliminated by taking expectation. However, in practice, this term does not vanish if we use Monte Carlo estimation. A simple analysis shows that its variance is proportional to $|Y_s - \tilde{v}_s|^2$, which is also minimized in Set (d), but not in Set (c). Therefore, a slight performance improvement in Set (d) compared to Set (c) is expected.

In addition, Theorem 6 gives us the hint of choosing better learning rates. Take y_0 as an example at first. In Set (a), y_0 appears in the term $\mathbb{E}\,|y_0 - Y_0|^2$. In optimization theory, the optimal learning rate for quadratic function $a\|x - x^*\|^2$ is $\frac{1}{2a}$; see, for example, Nesterov *et al.* [41]. Thus, the optimal learning rate for y_0 is 0.5. Considering the noise effect, we select a much smaller and thus safer value 0.1. For θ_y, the analysis becomes a little more complicated. Eq. (24) tells us that θ_y appears in the term $\mathbb{E}\int_t^T |Y_s - \tilde{v}_s|^2\,ds$. Substituting $Y_s = \theta_y^*\|W_s\|^2$ and $\tilde{v}_s = \theta_y\|W_s\|^2$ into it yields $(\theta_y - \theta_y^*)\int_t^T \mathbb{E}\,\|W_s\|^4\,ds$. By integrating on a sphere, we can calculate that $\int_t^T \mathbb{E}\,\|W_s\|^4\,ds = n(n+2)(T-t)^3/3$. Thus, the optimal learning rate for θ_y is in the order of 10^{-4}. Based on this, we select the value 1×10^{-5}.

D. Other type of BSDEs

The BSDE (21) considered before is only a basic type of general BSDEs. In many applications, for example in our off-policy subroutine, the generator f may be unknown and is expressed as $f(s, Y_s, Z_s)$ with a deterministic (or even random) coefficient $f(\cdot,\cdot,\cdot)$. Elementary extensions of Theorem 5 and Theorem 6 in this line are provided below.

Consider the following BSDE

$$Y_s = \xi + \int_s^T f(\tau, Z_\tau)\,d\tau - \int_s^T \langle Z_\tau, dW_\tau\rangle,\quad \forall s\in[t,T], \tag{26}$$

where the generator f is only coupled with Z. For any $z\in\mathbb{H}^2$, we can still introduce the process $\widetilde{Y}^z$ by replacing Z with z. Sadly, Eq. (23) fails to hold because $f(s, z_s)$ may be not equal to $f(s, Z_s)$. As a result, Theorem 5 and Theorem 6 no longer hold. Nevertheless, The general BML criterion for trial solutions $(\tilde{v}, z)$ can still be calculated and optimized, and obviously the true (Y, Z) is a global minimum of this criterion. Thus, the proposed criterion equals zero is a necessary condition for solving such a BSDE. Moreover, we are able to say it is also a sufficient condition to some extent.

Proposition 7. *Suppose that $\xi\in L^2_{\mathcal{F}_T}$ and $f:\Omega\times[t,T]\times\mathbb{R}^d\to\mathbb{R}$ satisfies the following conditions: 1) for any $z\in\mathbb{R}^d$, $f(s,z)$ is adapted; 2) $f(s,0)\in\mathbb{H}^2$; 3) there exists a constant L such that for any $z^1, z^2\in\mathbb{R}^d$,*

$$|f(s,z^1) - f(s,z^2)| \le L|z^1 - z^2|,\quad ds\otimes d\mathbb{P}\text{-a.e.}$$

on $[t,T]\times\Omega$. Then, BSDE (26) admits a unique adapted solution $(Y,Z)\in\mathbb{S}^2\times\mathbb{H}^2$. For any adapted process $z\in\mathbb{H}^2$, let $\widetilde{Y}^z_s$ denote the process (not necessarily adapted)

$$\widetilde{Y}^z_s = \xi + \int_s^T f(\tau, z_\tau)\,d\tau - \int_s^T \langle z_\tau, dW_\tau\rangle,\quad \forall s\in[t,T].$$

Then, $\mathbb{E}\int_t^T \|Z_s - z_s\|^2\,ds$ equals zero if and only if $\mathbb{E}\,|\widetilde{Y}^z_t - \mathbb{E}\,\widetilde{Y}^z_t|^2$ equals zero.

Proof: The uniqueness and existence are standard results for BSDEs; see [40, Chapter 6] for example.

Let $\mathbb{E}\int_t^T \|Z_s - z_s\|^2\,ds = 0$ be true. Noting the assumptions on f, for any $s\in[t,T]$, we have

$$\begin{aligned}&\mathbb{E}\bigg[\int_s^T f(\tau,Z_\tau)\,d\tau - \int_s^T f(\tau,z_\tau)\,d\tau\bigg]^2\\ &\le (T-s)\,\mathbb{E}\int_s^T |f(\tau,Z_\tau) - f(\tau,z_\tau)|^2\,d\tau\\ &\le L^2(T-s)\,\mathbb{E}\int_s^T \|Z_\tau - z_\tau\|^2\,d\tau = 0.\end{aligned}$$

Furthermore, according to Itô's isometry, there is

$$\mathbb{E}\bigg[\int_s^T \langle Z_\tau, dW_\tau\rangle - \int_s^T \langle z_\tau, dW_\tau\rangle\bigg]^2 = \mathbb{E}\int_s^T \|Z_\tau - z_\tau\|^2 = 0.$$

Hence, $\widetilde{Y}^z_s = Y_s$ holds almost surely for any $s\in[t,T]$. In particular, $\mathbb{E}\,|\widetilde{Y}^z_t - \mathbb{E}\,\widetilde{Y}^z_t|^2 = \mathbb{E}\,|Y_t - \mathbb{E}\,Y_t|^2 = 0$. This proves the "only if" part.

In order to prove the "if" part, we consider the BSDE

$$\widehat{Y}_s = \xi + \int_s^T \hat{f}_\tau\,d\tau - \int_s^T \langle \widehat{Z}_\tau, dW_\tau\rangle,\quad \forall s\in[t,T], \tag{27}$$

where $\hat{f}_\tau := f(\tau, z_\tau)$. This is the type of BSDE studied in previous subsections. By assumptions on f, the process $\hat{f}\in\mathbb{H}^2$. Applying Theorem 5 to BSDE (27) concludes that the solution $(\widehat{Y},\widehat{Z})\in\mathbb{S}^2\times\mathbb{H}^2$ uniquely exists and

$$\mathbb{E}\int_t^T \|\widehat{Z}_s - z_s\|^2\,ds = \mathbb{E}\,|\widetilde{Y}^z_t - \mathbb{E}\,\widetilde{Y}^z_t|^2 = 0.$$

Therefore, $z_s = \widehat{Z}_s$ holds $ds\otimes d\mathbb{P}$ almost everywhere.

In view of BSDE (26) and BSDE (27), we denote

$$\overline{Y} := Y - \widehat{Y},\quad \overline{Z} := Z - \widehat{Z},\quad \bar{f}_s := f(s,Z_s) - \hat{f}_s.$$

Let γ be a positive constant such that $\gamma > 2L^2$. By applying Itô's formula to $e^{\gamma s}|\overline{Y}_s|^2$, we obtain

$$\begin{aligned}&\mathbb{E}\,e^{\gamma t}|\overline{Y}_t|^2 + \mathbb{E}\int_t^T e^{\gamma s}(\gamma|\overline{Y}_s|^2 + \|\overline{Z}_s\|^2)\,ds\\ &= 2\,\mathbb{E}\int_t^T e^{\gamma s}\overline{Y}_s\bar{f}_s\,ds - 2\,\mathbb{E}\int_t^T e^{\gamma s}\overline{Y}_s\langle\overline{Z}_s, dW_s\rangle.\end{aligned} \tag{28}$$

A standard analysis based on Burkholder-Davis-Gundy inequality shows that the second term vanishes; see the proof of [40, Theorem 6.2.1]. On the other hand, for any $s \in [t,T]$,

$$2\overline{Y}_s \bar{f}_s \le \gamma|\overline{Y}_s|^2 + \frac{1}{\gamma}|\bar{f}_s|^2 \le \gamma|\overline{Y}_s|^2 + \frac{L^2}{\gamma}\|Z_s - z_s\|^2. \quad (29)$$

Noting $L^2/\gamma < 1/2$, Eq. (28) and Eq. (29), there is

$$\begin{aligned}\mathbb{E}\int_t^T e^{\gamma s}\|Z_s - \widehat{Z}_s\|^2\,ds &\le \frac{1}{2}\,\mathbb{E}\int_t^T e^{\gamma s}\|Z_s - z_s\|^2\,ds \\ &= \frac{1}{2}\,\mathbb{E}\int_t^T e^{\gamma s}\|Z_s - \widehat{Z}_s\|^2\,ds.\end{aligned}$$

The last equality comes from the fact that $z_s = \widehat{Z}_s$ holds $ds \otimes d\mathbb{P}$ almost everywhere. Hence, $\mathbb{E}\int_t^T e^{\gamma s}\|Z_s - \widehat{Z}_s\|^2\,ds = 0$. Replacing $\widehat{Z}_s$ with z_s again finishes our proof. ∎

Remark V.5. Under Assumptions 1, the BSDE in the off-policy subroutine satisfies the conditions here.

Proposition 8. *Let the condition of Proposition 7 hold and use the same notation. Let $\tilde{v}_s$ be an adapted process in $\mathbb{S}^2$ and ν be a σ-finite measure on $[t,T]$. Then,*

$$\mathbb{E}\int_s^T \|Z_\tau - z_\tau\|^2\,d\tau = \mathbb{E}\,|Y_s - \tilde{v}_s|^2 = 0, \quad \nu\text{-a.e.}, \quad \forall s \in [t,T]$$

if and only if $\mathbb{E}\int_t^T |\widetilde{Y}_s^z - \tilde{v}_s|^2\,\nu(ds) = 0$.

Proof: The sketch of this proof is similar to that of Proposition 7 except for a few minor differences concerning the additional $\tilde{v}$ and ν. A brief description of it is provided below, and readers may refer to Proposition 7's proof for more explanations.

We prove the "only if" part at first. By the assumption on f, we are able to show that $\widetilde{Y}_s^z = Y_s$ holds $d\nu \times d\mathbb{P}$-a.e.. Hence,

$$\mathbb{E}\int_t^T |\widetilde{Y}_s^z - \tilde{v}_s|^2\,\nu(ds) \le 2\,\mathbb{E}\int_t^T |\widetilde{Y}_s^z - Y_s|^2 + |Y_s - \tilde{v}_s|^2\,\nu(ds),$$

which equals zero by assumptions.

Then we prove the "if" part. Consider BSDE (27) with $\hat{f}_\tau := f(\tau, z_\tau)$. Applying Theorem 6 to that BSDE, we conclude that the solution $(\widehat{Y}, \widehat{Z}) \in \mathbb{S}^2 \times \mathbb{H}^2$ uniquely exists and that for any $s \in [t,T]$,

$$\mathbb{E}\int_s^T \|\widehat{Z}_\tau - z_\tau\|^2\,d\tau = \mathbb{E}\,|\widehat{Y}_s - \tilde{v}_s|^2 = 0, \quad \nu\text{-a.e.}. \quad (30)$$

Moreover, in view of BSDE (26) and BSDE (27), we have

$$\begin{aligned}&\mathbb{E}\,e^{4L^2 s}|Y_s - \widehat{Y}_s|^2 + \mathbb{E}\int_s^T e^{4L^2\tau}\|Z_\tau - \widehat{Z}_\tau\|^2\,d\tau \\ &\le \frac{1}{4}\,\mathbb{E}\int_s^T e^{4L^2\tau}\|Z_\tau - z_\tau\|^2\,d\tau.\end{aligned}$$

Integrating on $([t,T],\nu)$ and noting Eq. (30) yield

$$\begin{aligned}&\mathbb{E}\int_t^T\int_s^T e^{4L^2\tau}\|Z_\tau - \widehat{Z}_\tau\|^2\,d\tau\,\nu(ds) \\ &\le \frac{1}{2}\,\mathbb{E}\int_t^T\int_s^T e^{4L^2\tau}\|Z_\tau - \widehat{Z}_\tau\|^2\,d\tau\,\nu(ds).\end{aligned}$$

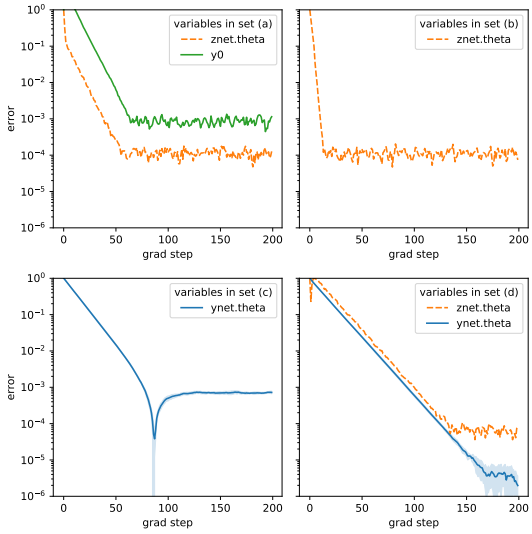


Fig. 3. The absolute errors of $\boldsymbol{\theta_y}, \boldsymbol{\theta_z}, \boldsymbol{y_0}$ at each gradient steps for Example 2. From left to right and from top to bottom, the subplots correspond to Set (a), (b), (c) and (d). The solid lines and shaded areas indicate the mean and standard deviation of absolute errors for 10 runs.

Hence, for any $s \in [t,T]$,

$$\mathbb{E}\int_s^T e^{4L^2\tau}\|\widehat{Z}_\tau - Z_\tau\|^2\,d\tau = \mathbb{E}\,e^{4L^2 s}|\widehat{Y}_s - Y_s|^2 = 0, \quad \nu\text{-a.e.}.$$

Using Eq. (30) again finishes our proof. ∎

The BSDE encountered in the off-policy subroutine is a special case of the BSDE considered in this subsection, where the generator $f(s,Z)$ is linear to Z. While Proposition 7 and Proposition 8 provide general treatments for nonlinear generator, a generator linearly coupled in Z can also be transformed into a decoupled generator by absorbing the linear coupling term into the Brownian motion using Girsanov's transformation. However, this treatment involves a change of probability measure [42] and is left for future discussion.

In order to verify our theory, we test the four realizations of the proposed general criterion listed in Table I by the following example, which is modified based on Example 1.

Example 2. *Solve the BSDE* (26) *with* $t = 0, T = 1, f(\omega, s, z) = -1 + \langle b_0 X_s, Z_s\rangle, \xi = \langle X_T, X_T\rangle/n$, *where* $n = 100$ *is the dimension of the process* X *and Brownian motion* W. *The process* X *satisfies the stochastic differential equation:* $X_s = W_s - \int_t^s b_0 X_s\,ds$ *with* $b_0 = -0.1$.

We parameterize the trial processes as $\tilde{v}_s = X_s^\intercal \theta_y X_s$, $z_s = 2\theta_z X_s$. Other treatments remain unchanged from Example 1. The true values can be verified by Itô's formula as well: $\theta_y^* = \theta_z^* = 1/n$. Results are reported in Figure 3.

VI. Simulation Results

In this section, we test our on-policy and off-policy subroutines on a 100 dimensional optimal control problem. We obtain the z function in these subroutines via optimizing the general BML criterion discussed in the last section. Specifically, we consider the four cases listed in Table I.

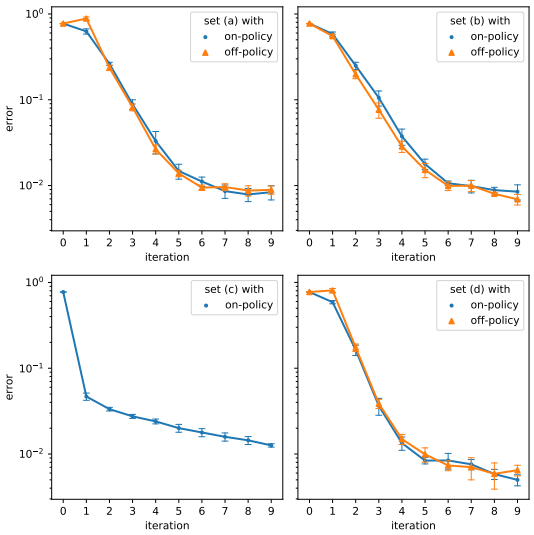


Fig. 4. The absolute error between the optimal cost and i-th policy's cost for Example 3. From left to right and from top to bottom, the subplots correspond to Set (a), (b), (c) and (d). Each subplot, except for Set (c), contains two lines representing the on-policy and the off-policy subroutines. The data points and error bars represent mean and standard deviation of 5 independent runs.

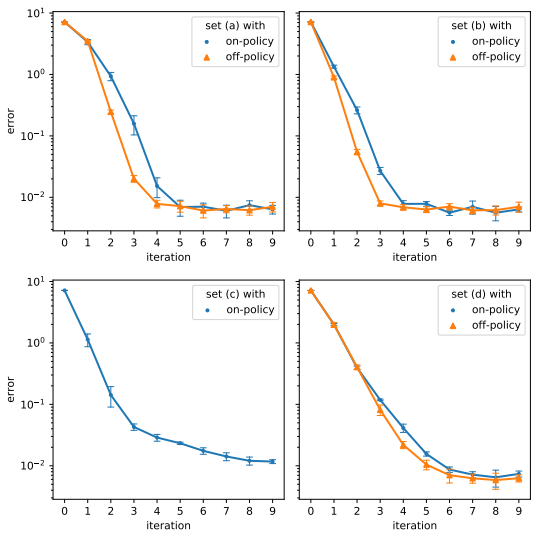


Fig. 5. The absolute error between the optimal cost and i-th policy's cost for Example 4. See Figure 4 for the explanations of elements in figures.

Example 3. *Consider the following stochastic optimal control problem, which is an extension of the example in [43]:*

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\left[\log \frac{1+\|X_T\|^2}{2} + \int_t^T \|\alpha_s\|^2 \, ds\right], \\ \text{subject to} \quad & X_s = x + \int_t^s \sigma_0(\hat{b}_0 \alpha_\tau \, d\tau + dW_\tau), \quad s \in [t,T], \end{aligned}$$

where W is a standard 100 dimensional Brownian motion with $W_t = 0$, and $\sigma_0, \hat{b}_0 \in \mathbb{R}$ are positive constants. Determine the optimal cost when $x = 0, t = 0, T = 1, \hat{b}_0 = 1$ and $\sigma_0 = \sqrt{2}$.

We run the GPI equipped with Algorithm 2 and Algorithm 3. The initial policy is chosen to be $\alpha^0(t,x) = -0.1x$ and the behavior policy α^b is fixed to α^0. In order to satisfy Assumption 1.2, we manually force control to set $A = [-a_{\max}, a_{\max}]^{100} \subset \mathbb{R}^{100}$ with $a_{\max} = 100$. Euler-Maruyama method with time step size $\Delta t = 0.01$ [44] is used for time discretization. The proposed criterion is optimized with SGD on PyTorch platform [45]. Table I is implemented with $\tilde{v}_s = \tilde{v}(s, X_s; \theta_y)$ and $z_s = z(s, X_s; \theta_z)$, where functions $\tilde{v}$ and z are feed-forward neural networks with a single hidden layer with 16 neurons. The SGD optimizer uses Nesterov acceleration technique with momentum 1×10^{-3} [46]. The optimization procedure is terminated after 75 gradient steps, and in each gradient step, the standard Euclidean norm of the total gradient is clipped to 10, and the learning rates are multiplied by a factor of 0.99. Learning rates for y_0, θ_y, θ_z are $0.5, 0.1, 0.1$, respectively. The sample size for estimating expectations is 16. For each criterion, we call the on-policy subroutine or the off-policy subroutine 9 consecutive times starting at α^0. Results are reported in Figure 4.

Figure 4 plots the absolute error between the theoretical optimal cost and i-th policy's cost. The theoretical optimal cost is [24]

$$v^*(t,x) = -\frac{2}{\hat{b}_0^2} \log \mathbb{E}\left[\exp\left(-\frac{\hat{b}_0^2}{2} \log \frac{1+\|x+\sigma_0 \epsilon\|^2}{2}\right)\right],$$

where $\epsilon \in \mathbb{R}^{100}$ and is normally distributed with mean 0 and covariance matrix $(T-t)I$. We estimate this expectation by Monte Carlo with sample size $M = 12800$. Figure 4 shows that both the on-policy and off-policy subroutines and the four specific criteria can produce a good enough policy after 9 policy iteration steps. It is worth noting that there is no suitable off-policy method for the criterion of Set (c). This is due to the fact that the generator of the BSDE in Algorithm 3 is explicitly coupled with Z, and thus, the optimization of z and $\tilde{v}$ is not independent, cf. Remark V.3. Despite this, we construct the improved policy by setting $z^\alpha = \sigma_0 \partial_x \tilde{v}(\cdot, \cdot; \theta_y)$ in the on-policy subroutine for Set (c).

Example 4. *Determine the optimal cost of Example 3 with $\sigma_0 = 20$.*

Compared with the previous example, this only changes the system dynamics. Benefited from the data-driven nature of our algorithms, we can rerun the program with the only difference that trajectories are now sampled from this new system. Results are reported in Figure 5.

REFERENCES

[1] R. A. Howard, *Dynamic programming and markov processes.* John Wiley, 1960.

[2] 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.

[3] 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.

[4] J. Lee and R. S. Sutton, "Policy iterations for reinforcement learning problems in continuous time and space—fundamental theory and methods," *Automatica*, vol. 126, p. 109421, 2021.

[5] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[6] Q. Wei and D. Liu, "A novel policy iteration based deterministic q-learning for discrete-time nonlinear systems," *Science China Information Sciences*, vol. 58, no. 12, pp. 1–15, 2015.

[7] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[8] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*. PMLR, 2014, pp. 387–395.

[9] S. M. Kakade, "A natural policy gradient," *Advances in neural information processing systems*, vol. 14, 2001.

[10] D. Kleinman, "On an iterative technique for riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114–115, 1968.

[11] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE circuits and systems magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[12] Y.-H. Ni and H.-T. Fang, "Policy iteration algorithm for singular controlled diffusion processes," *SIAM Journal on Control and Optimization*, vol. 51, no. 5, pp. 3844–3862, 2013.

[13] X. Li, Z. Peng, L. Liang, and W. Zha, "Policy iteration based q-learning for linear nonzero-sum quadratic differential games," *Science China Information Sciences*, vol. 62, no. 5, pp. 1–19, 2019.

[14] X. Li, Z. Peng, L. Jiao, L. Xi, and J. Cai, "Online adaptive q-learning method for fully cooperative linear quadratic dynamic games," *Science China Information Sciences*, vol. 62, no. 12, pp. 1–14, 2019.

[15] 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.

[16] B. Pang, T. Bian, and Z.-P. Jiang, "Robust policy iteration for continuous-time linear quadratic regulation," *IEEE Transactions on Automatic Control*, vol. 67, no. 1, pp. 504–511, 2022.

[17] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[18] R. Leake and R.-W. Liu, "Construction of suboptimal control sequences," *SIAM Journal on Control*, vol. 5, no. 1, pp. 54–63, 1967.

[19] W. E, J. Han, and A. Jentzen, "Algorithms for solving high dimensional pdes: from nonlinear monte carlo to machine learning," *Nonlinearity*, vol. 35, no. 1, p. 278, 2021.

[20] K. Doya, "Reinforcement learning in continuous time and space," *Neural computation*, vol. 12, no. 1, pp. 219–245, 2000.

[21] N. Frémaux, H. Sprekeler, and W. Gerstner, "Reinforcement learning using a continuous time actor-critic framework with spiking neurons," *PLoS computational biology*, vol. 9, no. 4, p. e1003024, 2013.

[22] Y. Jia and X. Y. Zhou, "Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach," *Journal of Machine Learning Research*, vol. 23, no. 154, pp. 1–55, 2022.

[23] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, ser. Graduate Texts in Mathematics. New York, NY: Springer New York, 1998, vol. 113.

[24] 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.

[25] B. Bouchard and N. Touzi, "Discrete-time approximation and monte-carlo simulation of backward stochastic differential equations," *Stochastic Processes and their applications*, vol. 111, no. 2, pp. 175–206, 2004.

[26] E. Gobet, J.-P. Lemor, and X. Warin, "A regression-based Monte Carlo method to solve backward stochastic differential equations," *The Annals of Applied Probability*, vol. 15, no. 3, pp. 2172 – 2202, 2005.

[27] Q. Chan-Wai-Nam, J. Mikael, and X. Warin, "Machine learning for semi linear pdes," *Journal of scientific computing*, vol. 79, no. 3, pp. 1667–1712, 2019.

[28] W. E, M. Hutzenthaler, A. Jentzen, and T. Kruse, "On multilevel picard numerical approximations for high-dimensional nonlinear parabolic partial differential equations and high-dimensional nonlinear backward stochastic differential equations," *Journal of Scientific Computing*, vol. 79, no. 3, pp. 1534–1571, 2019.

[29] C. Bender and J. Zhang, "Time discretization and Markovian iteration for coupled FBSDEs," *The Annals of Applied Probability*, vol. 18, no. 1, 2008.

[30] J. Yong and X. Y. Zhou, *Stochastic controls*. New York, NY: Springer New York, 1999.

[31] P. Meyer and G. Denzel, *Probability and Potentials*. Blaisdell Publishing Company, 1966.

[32] M. G. Crandall and P.-L. Lions, "Viscosity solutions of hamilton-jacobi equations," *Transactions of the American mathematical society*, vol. 277, no. 1, pp. 1–42, 1983.

[33] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 39–47, 2009.

[34] E. Theodorou, J. Buchli, and S. Schaal, "A generalized path integral control approach to reinforcement learning," *The Journal of Machine Learning Research*, vol. 11, pp. 3137–3181, 2010.

[35] R. Song, W. Xiao, and C. Sun, "A new self-learning optimal control laws for a class of discrete-time nonlinear systems based on esn architecture," *Science China Information Sciences*, vol. 57, no. 6, pp. 1–10, 2014.

[36] Y. Jiang and Z.-P. Jiang, "Global adaptive dynamic programming for continuous-time nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2917–2929, 2015.

[37] S. Satoh, H. J. Kappen, and M. Saeki, "An iterative method for nonlinear stochastic optimal control based on path integrals," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 262–276, 2016.

[38] I. Exarchos, M. A. Pereira, Z. Wang, and E. Theodorou, "Novas: Non-convex optimization via adaptive stochastic search for end-to-end learning and control," in *International Conference on Learning Representations*, 2021.

[39] R. S. Liptser and A. N. Shiriaev, *Statistics of random processes I: General theory*. Springer, 1977, vol. 394.

[40] H. Pham, *Continuous-time stochastic control and optimization with financial applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 61.

[41] Y. Nesterov *et al.*, *Lectures on convex optimization*. Springer, 2018, vol. 137.

[42] I. Exarchos and E. A. Theodorou, "Stochastic optimal control via forward and backward stochastic differential equations and importance sampling," *Automatica*, vol. 87, pp. 159–165, 2018.

[43] S. Ji, S. Peng, Y. Peng, and X. Zhang, "Solving stochastic optimal control problem via stochastic maximum principle with deep learning method," *arXiv preprint arXiv:2007.02227*, 2020.

[44] D. J. Higham, "An algorithmic introduction to numerical simulation of stochastic differential equations," *SIAM review*, vol. 43, no. 3, pp. 525–546, 2001.

[45] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[46] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*. PMLR, 2013, pp. 1139–1147.

[47] N. V. Krylov, *Controlled diffusion processes*. Springer Science & Business Media, 2008, vol. 14.

APPENDIX I
PROOF TO THEOREM 4

Proof: Throughout this proof, we fix the forward state to X^b, and use F_s to denote $F(s, X_s^b)$ for any function $F(\cdot,\cdot)$.

The admissibility is a direct consequence of Remark III.2. According to Lemma 3, for $n \geq 1$, we have

$$Y_s^n = g(X_T^b) + \int_s^T f_\tau^{\alpha^n} + (\hat{b}_\tau^{\alpha^n} - \hat{b}_\tau^{\alpha^b})^\intercal Z_\tau^n \, d\tau - \int_s^T (Z_\tau^n)^\intercal dW_\tau,$$

where $Y_s^n = v^{\alpha^n}(s, X_s^b)$, $Z_s^n = \sigma^\intercal \partial_x v^{\alpha^n}(s, X_s^b)$. Similarly,

$$Y_s^* = g(X_T^b) + \int_s^T f_\tau^{\alpha^*} + (\hat{b}_\tau^{\alpha^*} - \hat{b}_\tau^{\alpha^b})^\intercal Z_\tau^* \, d\tau - \int_s^T (Z_\tau^*)^\intercal dW_\tau,$$

where $Y_s^* = v^*(s, X_s^b)$, $Z_s^* = \sigma^\intercal \partial_x v^*(s, X_s^b)$.

Define $h : \Omega \times [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$h(s, z, Z) \coloneqq f_s^{\mu_s(z)} + \langle \hat{b}_s^{\mu_s(z)} - \hat{b}_s^{\alpha^b}, Z \rangle.$$

Then, we can verify that under Assumption 1 there is a constant L such that for any $(s, z, Z) \in [t,T] \times \mathbb{R}^d \times \mathbb{R}^d$,

$$|h(s,z,Z) - h(s,0,0)| \leq L\|z\| + L\|Z\|, \quad \mathbb{P}\text{-a.s.},$$

and that $\mathbb{E}\int_t^T \|h(s,0,0)\|^2\,ds < \infty$. Moreover, it can be proved that $\|Z_s^*\| \le L\|\partial_x v^*(s, X_s^b)\|$ can be further bounded by some constant K [30, Proposition 4.3.1]; see also [47, Chapter 4] for more general discussions on the properties of $\partial_x v^\alpha$. Hence, we have

$$\begin{aligned}
&|h(s, z_s^{n-1}, Z_s^n) - h(s, Z_s^*, Z_s^*)| \\
&\le |f_s^{\mu_s(z_s^{n-1})} - f_s^{\mu_s(Z_s^*)}| + |\langle \hat{b}_s^{\mu_s(z_s^{n-1})} - \hat{b}_s^{\alpha^b}, Z_s^n - Z_s^*\rangle| \\
&\quad + |\langle \hat{b}_s^{\mu_s(z_s^{n-1})} - \hat{b}_s^{\mu_s(Z_s^*)}, Z_s^*\rangle| \\
&\le L\|\mu_s(z_s^{n-1}) - \mu_s(Z_s^*)\| + 2L\|Z_s^n - Z_s^*\| \\
&\quad + K\|\hat{b}_s^{\mu_s(z_s^{n-1})} - \hat{b}_s^{\mu_s(Z_s^*)}\| \\
&= (L^2 + KL)\|z_s^{n-1} - Z_s^*\| + 2L\|Z_s^n - Z_s^*\|.
\end{aligned}$$

To this end, all conditions of [15, Lemma A.5] are verified, and thus, the following estimation holds for any $n \ge 1$:

$$\begin{aligned}
&\mathbb{E}\,|Y_t^n - Y_t^*|^2 + \mathbb{E}\int_t^T e^{\gamma(s-t)}\|Z_s^n - Z_s^*\|^2\,ds \\
&\le \tilde{q}\,\mathbb{E}\int_t^T e^{\gamma(s-t)}\|z_s^{n-1} - Z_s^*\|^2\,ds,
\end{aligned} \tag{31}$$

where $\gamma > 0$ and $\tilde{q} \in (0, 1/2)$ depend only on the Lipschitz constant in Assumption 1. Introducing the following notations

$$\begin{aligned}
a_n &\coloneqq \mathbb{E}\,|Y_t^n - Y_t^*|^2 = |v^{\alpha^n}(t,x) - v^*(t,x)|^2, \\
b_n &\coloneqq \mathbb{E}\int_t^T e^{\gamma(s-t)}\|Z_s^n - Z_s^*\|^2\,ds, \\
c_n &\coloneqq \mathbb{E}\int_t^T e^{\gamma(s-t)}\|z_s^n - Z_s^n\|^2\,ds \le e^{\gamma(T-t)}\epsilon_n,
\end{aligned}$$

we further relax the inequality Eq. (31) to (letting $q = 2\tilde{q}$)

$$a_n + b_n \le q(b_{n-1} + c_{n-1}), \quad \forall n \ge 1. \tag{32}$$

Noting that $a_n \ge 0$, we substitute $b_n \le q(b_{n-1} + c_{n-1})$ into the right-hand side of Eq. (32) repeatedly:

$$\begin{aligned}
a_n + b_n &\le qc_{n-1} + q^2(b_{n-2} + c_{n-2}) \\
&\le qc_{n-1} + q^2c_{n-2} + q^3(b_{n-3} + c_{n-3}) \\
&\le qc_{n-1} + \cdots + q^{n-1}c_1 + q^n(b_0 + c_0) =: S_n.
\end{aligned} \tag{33}$$

Without loss of generality, we assume $\operatorname{limsup}\epsilon_n < \infty$. Otherwise, the equality to be proved holds trivially. Then, we have $\operatorname{limsup} c_n \le e^{\gamma(T-t)}\operatorname{limsup}\epsilon_n < \infty$. This means there is a positive integer M such that c_n is bounded by some $c < \infty$ for any $n \ge M$. Hence,

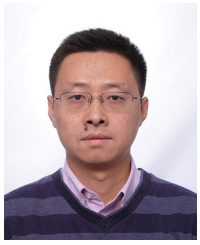
$$\begin{aligned}
S_n &= qc_{n-1} + \cdots + q^{n-M}c_M + \cdots + q^nc_0 + q^nb_0 \\
&\le (q + q^2 + \cdots + q^{n-M})c \\
&\quad + q^{n-M+1}\max\{c_k : 0 \le k \le M-1\} + q^nb_0 \\
&\le \frac{q}{1-q}c + q^{n-M+1}\max\{c_k : 0 \le k \le M-1\} + q^nb_0.
\end{aligned}$$

This implies that S_n is also bounded for sufficient large n, i.e., $\operatorname{limsup} S_n < \infty$. Observing that S_n satisfies the recurrence equation $S_n = q(S_{n-1} + c_{n-1})$, we can conclude that $\operatorname{limsup} S_n < \frac{q}{1-q}\operatorname{limsup} c_n$ by taking limsup on both sides. Noting $a_n \le S_n$, we have

$$\limsup_{n\to\infty} a_n \le \frac{q}{1-q}\limsup_{n\to\infty} c_n.$$

Expanding the definitions of a_n and c_n finishes the proof. ■

Yutian Wang received the B.S. degree in physics and M.S. degree in artificial intelligence from Nankai University, Tianjin, China, in 2020 and 2023. He is currently working toward the Ph.D. degree in applied optimization and optimal control in the Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong, China. His current research interests include stochastic optimal control, Hamilton-Jacobi theory and reinforcement learning.



Yuan-Hua Ni received the Ph.D. degree in operation research and cybernetics from the Chinese Academy of Sciences, Beijing, China, in 2010. He is currently with the College of Artificial Intelligence, Nankai University (NKU), Tianjin, China, as an Associate Professor. From April 2014 to May 2015 and from Jan 2016 to Jan 2017, he was a Visiting Scholar with the University of California, San Diego, USA, and with the Hong Kong Polytechnic University, Hong Kong, respectively. His current research interests include stochastic control, optimal control, reinforcement learning and distributed optimization and control. He has served as Associate Editor of *System & Control Letters*.

Zengqiang Chen received the B.S. degree in mathematics, M.S. and Ph.D. degrees in control theory and control engineering from the Nankai University, Tianjin, China, in 1987, 1990 and 1997, respectively. He is currently a Professor in the Department of Automation. His research interests include neural network control, complex networks and multi-agents system.

Ji-Feng Zhang received the B.S. degree in mathematics from Shandong University, China, in 1985 and the Ph.D. degree from the Institute of Systems Science (ISS), Chinese Academy of Sciences (CAS), China, in 1991. He is now with the ISS, Academy of Mathematics and Systems Science, CAS. His current research interests include system modeling, adaptive control, stochastic systems, and multi-agent systems. He is an IEEE Fellow, IFAC Fellow, CAA Fellow, CSIAM Fellow, member of the European Academy of Sciences and Arts, and Academician of the International Academy for Systems and Cybernetic Sciences. He received the second prize of the State Natural Science Award of China in 2010 and 2015, respectively. He was a Vice-Chair of the IFAC Technical Board, member of the Board of Governors, IEEE Control Systems Society; Convenor of Systems Science Discipline, Academic Degree Committee of the State Council of China; Vice-President of the Systems Engineering Society of China, Chinese Mathematical Society, and the Chinese Association of Automation. He has served as Editor-in-Chief, Deputy Editor-in-Chief, Senior Editor or Associate Editor for more than 10 journals, including *Science China Information Sciences*, *National Science Review*, *IEEE Transactions on Automatic Control*, and *SIAM Journal on Control and Optimization* etc.