Information Relaxation and Dual Formulation of Controlled Markov Diffusions

Fan Ye and Enlu Zhou

Abstract

Information relaxation and duality in Markov decision processes have been studied recently by several researchers with the goal to derive dual bounds on the value function. In this paper we extend this dual formulation to controlled Markov diffusions: in a similar way we relax the constraint that the decision should be made based on the current information and impose penalty to punish the access to the information in advance. We establish the weak duality, strong duality and complementary slackness results in a parallel way as those in Markov decision processes. We explore the structure of the optimal penalties and expose the connection between Markov decision processes and controlled Markov diffusions. We demonstrate the use of the dual representation for controlled Markov diffusions in a classic dynamic portfolio choice problem. We evaluate the lower bounds on the expected utility by Monte Carlo simulation under a sub-optimal policy, and we propose a new class of penalties to derive upper bounds with little extra computation. The small gaps between the lower bounds and upper bounds indicate that the available policy is near optimal as well as the effectiveness of our proposed penalty in the dual method.

I. Introduction

Markov decision processes (MDPs) and controlled Markov diffusions play a central role respectively in modeling discrete-time and continuous-time dynamic decision making problems under uncertainty, and hence have wide applications in diverse fields such as engineering, operations research and economics. However, the standard approach of solving for optimal polices via dynamic programming and Hamlton-Jacobi-Bellman (HJB) equation suffers from the "curse of dimensionality"- the size of the state space increases exponentially with the dimension of the state. Many modern approximate dynamic programming methods have been proposed for solving MDPs in recent years to combat this curse of dimensionality, such as [1], [2], [3], [4]. These methods often generate sub-optimal policies, simulation under which leads to lower bounds (or upper bounds) on the optimal expected reward (or cost). Though the accuracy of the sub-optimal policies is generally unknown, the lack of performance guarantee on sub-optimal policies can be potentially addressed by providing a dual bound, i.e., an upper bound (or lower bound) on the optimal expected reward (or cost). Valid and tight dual bounds based on a dual representation of MDPs were recently developed by [5] and [6]. The main idea of this duality approach is to allow the decision maker to foresee the future uncertainty but impose a penalty for getting access to the information in advance. In

F. Ye and E. Zhou are with the Department of Industrial & Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA {fanye2,enluzhou}@illinois.edu

addition, this duality approach only encompasses pathwise deterministic optimization problems and therefore is well-suited to Monte Carlo simulation, making it useful to evaluate the quality of sub-optimal policies in complex dynamic systems.

This dual formulation of MDPs is attractive in both theoretical and practical aspects. On one hand, the idea of relaxing the constraint on the non-anticipative policies in the setting of MDPs at least dates back to [7], as exposed by [8]. In addition, the optimal penalty is not unique: for general problems we have the value function-based penalty developed by [5] and [6]; for problems with convex structure there is an alternative optimal penalty, that is, the gradient-based penalty, as pointed out by [9]. On the other hand, in order to derive tight dual bounds, various approximation schemes based on different optimal penalties have been proposed including [6], [9], [10], [11]. We notice that the real implementation of computing the dual bounds based on this dual framework has just begun, and it has found increasing applications in different fields of problems such as [12], [9], [13], [14], [15].

The goal of this paper is to extend the information relaxation approach and the dual representation of MDPs to controlled Markov diffusions. The motivation is that the HJB equation rarely allows a closed-form value function, especially when the dimension of the state space is high or there are constraints on the control space. Many numerical methods have been developed based on different approximation schemes: [16] considered the Markov chain approximation method by discretizing the HJB equation; [17] extended the approximate linear programming method to controlled Markov diffusions. Another standard numerical approach is to discretize the time space, which reduces the original continuous-time problem to an MDP and hence the techniques of approximate dynamic programming can be applied. Since the quality of a numerical solution is hard to justify in many problems, we are interested in deriving a tight dual bound on the value function of a controlled Markov diffusion by formulating its dual representation. Around this topic some central questions are

- Can we establish a similar framework of dual formulation for controlled Markov diffusions based on information relaxation as that for MDPs?
- If the answer is yes, what is the form of the optimal penalty in the setting of controlled Markov diffusions? Is the optimal penalty unique?
- If certain optimal penalty exists, does it help to facilitate the computation of the dual bound on the value function?

The answer to the first question is yes, at least for a wide class of controlled Markov diffusions. To fully answer all the questions we should employ the technical machinery "anticipating stochastic calculus" (see, e.g., [18], [19]). To ease reading we first present the information relaxation-based dual formulation of controlled Markov diffusions without using the heavy machinery. We establish the weak duality, strong duality and complementary slackness results in a parallel way as those in the dual formulation of MDPs. The complete answers to the second question are postponed to Appendix D, where we develop all the needed technical machinery and investigate one type of the optimal penalties, the so-called "value function-based penalty". Then we emphasize on the computational aspect using the result of this dual representation so as to answer the third question. One key feature of the value

function-based optimal penalty is that it can be written compactly as an Ito stochastic integral under the natural filtration generated by the Brownian motions. This compact expression potentially enables us to design sub-optimal penalties in simple forms and also facilitates the computation of the dual bound. A direct application is illustrated by a classic dynamic portfolio choice problem with predictable returns and intermediate consumptions: we consider the numerical solution to a discrete-time model that is discretized from a continuous-time model; an effective class of penalties that are easy to evaluate is proposed to derive dual bounds on the value function for the discrete-time model.

It turns out that [20], [21], [22] have pioneered a series of related work for controlled Markov diffusions. They also adopted the approach of relaxing the future information and penalizing, which is much earlier than the dual framework of MDPs established. In particular, [20] proposed a Lagrangian approach for penalization, where the Lagrangian term plays essentially the same role of a penalty in our dual framework; in addition, we find that the Lagrangian term has a similar flavor of the gradient-based penalty proposed by [9] for MDPs. The early work of [20] is not widely known maybe due to its technical complication. The main difference of their work compared with ours is that we propose a more general framework that may incorporate their Lagrangian approach as a special case; the optimal penalty we develop in this paper is value function-based while their Lagrangian approach behaves like a gradient-based penalty. In addition, their work is purely theoretical and does not suggest any computational method. In contrast, we provide a numerical example to demonstrate the practical use of our dual formulation. We summarize our contributions as follows:

- We establish a dual representation of controlled Markov diffusions based on information relaxation. We also explore the structure of the optimal penalty and expose the connection between MDPs and controlled Markov diffusions.
- Based on the result of the dual representation of controlled Markov diffusions, we demonstrate its practical use in a dynamic portfolio choice problem. In many cases the numerical results of the upper bounds on the expected utility show that our proposed penalties are near optimal, comparing with the lower bounds induced by sub-optimal policies for the same problem.

The rest of the paper is organized as follows. In Section II, we review the dual formulation of MDPs and derive the dual formulation of controlled Markov diffusions. In Section III, we illustrate the dual approach and carry out numerical studies in a dynamic portfolio choice problem. Finally, we conclude with future directions in Section IV. We leave most of the proofs and discuss the connection between [20], [9] and our work in Appendix.

II. Controlled Markov Diffusions and Its Dual Representation

We begin with a brief review of the dual framework on Markov Decision Processes that is first developed by [5] and [6] in Section II-A. We then give the basic setup of the controlled Markov diffusion and its associated Hamilton-Jacobi-Bellman equation in Section II-B. We develop the dual representation of controlled Markov diffusions and present the main results in Section II-C.

A. Review of Dual Formulation of Markov Decision Processes

Consider a finite-horizon MDP on the probability space $(\Omega, \mathscr{G}, \mathbb{P})$. Time is indexed by $\mathscr{K} = \{0, 1, \cdots, K\}$. Suppose $\mathscr{X}$ is the state space and $\mathscr{A}$ is the control space. The state $\{x_k\}$ follows the equation

$$x_{k+1} = f(x_k, a_k, v_{k+1}),\ k = 0, 1, \cdots, K-1, \tag{1}$$

where $a_k \in \mathscr{A}_k$ is the control whose value is decided at time k, and v_k is a random variable taking values in the set $\mathscr{V}$ with a known distribution. The evolution of the information is described by the filtration $\mathbb{G} = \{\mathscr{G}_0, \cdots, \mathscr{G}_K\}$ with $\mathscr{G} = \mathscr{G}_K$. In particular, each v_k is $\mathscr{G}_k$-adapted.

Denote by $\mathbb{A}$ the set of all control strategies $\mathbf{a} \triangleq (a_0, \cdots, a_{K-1})$, i.e., each a_k takes value in $\mathscr{A}$. Let $\mathbb{A}_{\mathbb{G}}$ be the set of control strategies that are adapted to the filtration $\mathbb{G}$, i.e., each a_k is $\mathscr{G}_k$-adapted. We also call $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$ a *non-anticipative* policy. Given an $x_0 \in \mathscr{X}$, the objective is to maximize the expected reward by selecting a non-anticipative policy $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$:

$$\begin{aligned} V_0(x_0) &= \sup_{\mathbf{a} \in \mathbb{A}_{\mathbb{G}}} J_0(x_0; \mathbf{a}), \\ \text{where}\quad J_0(x_0; \mathbf{a}) &\triangleq \mathbb{E}\left[\sum_{k=0}^{K-1} g_k(x_k, a_k) + \Lambda(x_K) | x_0\right]. \end{aligned} \tag{2}$$

The expectation in (2) is taken with respect to the random sequence $\mathbf{v} = (v_1, \cdots, v_K)$. The value function V_0 is a solution to the following dynamic programming recursion:

$$\begin{aligned} &V_K(x_K) \triangleq \Lambda(x_K); \\ &V_k(x_k) \triangleq \sup_{a_k \in \mathscr{A}} \{g_k(x_k, a_k) + \mathbb{E}[V_{k+1}(x_{k+1}) | x_k, a_k]\},\ k = K-1, \cdots, 0. \end{aligned}$$

Next we describe the dual formulation of the value function $V_0(x_0)$. Here we only consider the *perfect information relaxation*, i.e., we have full knowledge of the future randomness, since this relaxation is usually more applicable in practice.

Define $\mathbb{E}_{0,x}[\cdot] \triangleq \mathbb{E}[\cdot | x_0 = x]$. Let $\mathscr{M}_{\mathbb{G}}(0)$ denote the set of *dual feasible penalties* $M(\mathbf{a}, \mathbf{v})$, which do not penalize non-anticipative policies in expectation, i.e.,

$$\mathbb{E}_{0,x}[M(\mathbf{a}, \mathbf{v})] \leq 0 \quad \text{for all } x \in \mathscr{X} \text{ and } \mathbf{a} \in \mathbb{A}_{\mathbb{G}}.$$

Denote by $\mathscr{D}$ the set of real-valued functions on $\mathscr{X}$. Then we define an operator $\mathscr{L} : \mathscr{M}_{\mathbb{G}}(0) \to \mathscr{D}$:

$$(\mathscr{L}M)(x) = \mathbb{E}_{0,x}\left[\sup_{\mathbf{a} \in \mathbb{A}} \{\sum_{k=0}^{K-1} g_k(x_k, a_k) + \Lambda(x_K) - M(\mathbf{a}, \mathbf{v})\}\right]. \tag{3}$$

Note that the supremum in (3) is over the set $\mathbb{A}$ not the set $\mathbb{A}_{\mathbb{F}}$, i.e., the control a_k is made based on the exposed future information. The optimization problem inside the expectation in (3) is usually referred to as the *inner optimization problem*. In particular, the right hand side of (3) is well-suited to Monte Carlo simulation: we can simulate a realization of $\mathbf{v} = \{v_1, \cdots, v_K\}$ and solve the inner optimization problem:

$$I(x,M,\mathbf{v}) \triangleq \max_{\mathbf{a}} \quad \sum_{k=0}^{K-1} g_k(x_k,a_k) + \Lambda(x_K) - M(\mathbf{a},\mathbf{v}) \tag{4a}$$

$$\text{s.t. } x_0 = x,$$

$$x_{k+1} = f(x_k,a_k,v_{k+1}),\ k=0,\cdots,K-1, \tag{4b}$$

$$a_k \in \mathscr{A}_k,\ k=0,\cdots,K-1, \tag{4c}$$

which is in fact a *deterministic* dynamic program. The optimal value $I(x,M,\mathbf{v})$ is an unbiased estimator of $(\mathscr{L}M)(x)$.

Theorem 1 below establishes a strong duality in the sense that for all $x_0 \in \mathscr{X}$,

$$\sup_{\mathbf{a}\in\mathbb{A}_{\mathbb{G}}} J_0(x_0;\mathbf{a}) = \inf_{M\in\mathscr{M}_{\mathbb{G}}(0)} (\mathscr{L}M)(x_0).$$

In particular, Theorem 1(a) suggests that $\mathscr{L}M(x_0)$ can be used to derive an upper bound on the value function $V_0(x_0)$ given any $M \in \mathscr{M}_{\mathbb{G}}(0)$, i.e., $I(x_0,M,\mathbf{v})$ is a high-biased estimator of $V_0(x_0)$ for all $x_0 \in \mathscr{X}$; Theorem 1(b) states that the duality gap vanishes if the dual problem is solved by choosing M in the form of (5).

Theorem 1 (Theorem 2.1 in [6])

1) *(Weak Duality) For all $M \in \mathscr{M}_{\mathbb{G}}(0)$ and all $x \in \mathscr{X}$, $V_0(x) \le (\mathscr{L}M)(x)$.*
2) *(Strong Duality) For all $x \in \mathscr{X}$, $V_0(x) = (\mathscr{L}M^*)(x)$, where*

$$M^*(\boldsymbol{a},\boldsymbol{v}) = \sum_{k=0}^{K-1} (V_{k+1}(x_{k+1}) - \mathbb{E}[V_{k+1}(x_{k+1})|x_k,a_k]). \tag{5}$$

Remark 1

1) *Note that the right hand side of (5) is a function of $(\boldsymbol{a},\boldsymbol{v})$, since $\{x_k\}$ depend on $(\boldsymbol{a},\boldsymbol{v})$ through the state equation (1).*
2) *The reason that $M \in \mathscr{M}_{\mathbb{G}}(0)$ is called a (dual feasible) penalty function becomes clear: if the relaxation of the requirement on the non-anticipative policies is penalized by using a proper function in $\mathscr{M}_{\mathbb{G}}(0)$, then the value function V_0 can be recovered via the dual approach due to the strong duality result.*
3) *Note that the optimal penalty $M^*(\boldsymbol{a},\boldsymbol{v})$ is the sum of a $\mathbb{G}$-martingale difference sequence when $\boldsymbol{a} \in \mathbb{A}_{\mathbb{G}}$; therefore, $M^*(\boldsymbol{a},\boldsymbol{v}) \in \mathscr{M}_{\mathbb{G}}(0)$. Since M^* depends on the value function $\{V_k\}$, it is refereed to as the* value function-based penalty.

The optimal penalty (5) that achieves the strong duality involves the value function $\{V_k\}$, and hence is intractable in practical problems. In order to obtain tight dual bounds, a natural idea is to derive sub-optimal penalty functions based on a good approximate value function $\{\hat{V}_k\}$ or some sub-optimal policy $\hat{\mathbf{a}}$. Methods based on these ideas have been successfully implemented in the American option pricing problems by [23], [24], [25], and also in [6], [12], [13]. However, these approaches cannot be extended immediately to general continuous-state MDPs in parallel with the American options pricing problem. The first difficulty, as pointed out by [8], is that $\mathbb{E}[\hat{V}_{k+1}(x_{k+1})|x_k,a_k]$

usually cannot be written as an analytic function of x_k and a_k. Though this conditional expectation may be evaluated approximately after discretizing the state space, it could be time-consuming when $\mathscr{V}$ is of high dimension. Second, even if the penalty $\hat{V}_{k+1}(x_{k+1}) - \mathbb{E}[\hat{V}_{k+1}(x_{k+1})|x_k, a_k]$ can be computed analytically, the inner optimization problem (4) may still be difficult to solve since no convex structure can be guaranteed (even assuming that (4) is convex with $M = 0$). To overcome these difficulties, [9] introduced the gradient-based penalty in the context of dynamic portfolio optimization with transaction costs, and [10] derived a penalty by employing a parameterized quadratic function to approximate the value function in linear systems with convex costs. [11] proposed the idea of parametrization on penalties directly to avoid starting with approximate value functions. Furthermore, [10] explored the connection between information relaxation duality method and the approximate linear programming approach proposed by [4], and [8] revealed that in linear-quadratic problems the value function-based penalty and gradient-based penalty are both optimal, but in different senses.

B. Controlled Markov Diffusions and Hamilton-Jacobi-Bellman Equation

This subsection is concerned with the control of Markov diffusion processes. Applying the Bellman's principle of dynamic programming leads to a second-order nonlinear partial differential equation, which is referred to as the Hamilton-Jacobi-Bellman equation. For a comprehensive treatment on this topic we refer the readers to [26].

Let us consider an $\mathbb{R}^n$-valued controlled Markov diffusion process $(x_t)_{0\le t\le T}$ driven by an m-dimensional Brownian motion $(w_t)_{0\le t\le T}$ on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, following the stochastic differential equation (SDE):

$$dx_t = b(t, x_t, u_t)dt + \sigma(t, x_t, u_t)dw_t, \quad 0 \le t \le T, \tag{6}$$

where $u_t \in \mathscr{U} \subset \mathbb{R}^{d_u}$ is the control applied at time t, and b and σ are functions $b : [0,T] \times \mathbb{R}^n \times \mathscr{U} \to \mathbb{R}^n$ and $\sigma : [0,T] \times \mathbb{R}^n \times \mathscr{U} \to \mathbb{R}^{n\times m}$. The natural (augmented) filtration generated by the Brownian motions is denoted by $\mathbb{F} = \{\mathscr{F}_t, 0 \le t \le T\}$ with $\mathscr{F} = \mathscr{F}_T$. In the following $\|\cdot\|$ denotes the Euclidean norm.

Definition 1 *A control strategy $\boldsymbol{u} = (u_s)_{s\in[t,T]}$ is called an admissable strategy at time t if*

1) *$\boldsymbol{u} = (u_s)_{s\in[t,T]}$ is an $\mathbb{F}$-progressively measurable process taking values in $\mathscr{U}$ (i.e., $\boldsymbol{u}$ is a non-anticipative policy), and satisfying $\mathbb{E}[\int_t^T ||u_s||^2 ds] < \infty$;*
2) *$\mathbb{E}_{t,x}[\sup_{s\in[t,T]} ||x_s||^2] < \infty$, where $\mathbb{E}_{t,x}[\cdot] \triangleq \mathbb{E}[\cdot|x_t = x]$.*

The set of admissible strategies at time t is denoted by $\mathscr{U}_{\mathbb{F}}(t)$. With the standard technical conditions imposed on b and σ (specified in Appendix A), the SDE (6) admits a unique pathwise solution when $\boldsymbol{u} \in \mathscr{U}_{\mathbb{F}}(0)$.

Let $Q = [0,T) \times \mathbb{R}^n$ and $\bar{Q} = [0,T] \times \mathbb{R}^n$. We define the functions $\Lambda : \mathbb{R}^n \to \mathbb{R}$ and $g : \bar{Q} \times \mathscr{U} \to \mathbb{R}$ as the final reward and intermediate reward, respectively. Then we introduce the reward functional

$$J(t, x; \mathbf{u}) \triangleq \mathbb{E}_{t,x}[\Lambda(x_T) + \int_t^T g(s, x_s, u_s)ds].$$

The final reward Λ and the intermediate reward g satisfy the polynomial growth conditions that are specified in Appendix A. Given an initial condition $(t,x) \in Q$, the objective is to maximize $J(t,x,u)$ over all the control $\mathbf{u}$ in

$\mathscr{U}_{\mathbb{F}}(t)$:

$$V(t,x) = \sup_{\mathbf{u}\in\mathscr{U}_{\mathbb{F}}(t)} J(t,x;\mathbf{u}). \tag{7}$$

Here we abuse the notation of the state x, the rewards Λ and g, and the value function V, since they play the same roles as those in MDPs.

Let $C^{1,2}(Q)$ denote the space of function $L(t,x): Q \to \mathbb{R}$ that is C^1 in t and C^2 in x on Q. For $L \in C^{1,2}(Q)$, define a partial differential operator A^u by

$$A^u L(t,x) \triangleq L_t(t,x) + L_x(t,x)^\top b(t,x,u) + \frac{1}{2}\mathrm{tr}(L_{xx}(t,x)(\sigma\sigma^\top)(t,x,u)),$$

where L_t, L_x, and L_{xx} denote the t-partial derivative, the gradient and the Hessian with respect to x respectively, and $(\sigma\sigma^\top)(t,x,u) \triangleq \sigma(t,x,u)\sigma(t,x,u)^\top$. Let $C_p(\bar{Q})$ denote the set of function $L(t,x): \bar{Q} \to \mathbb{R}$ that is continuous on $\bar{Q}$ and satisfies a polynomial growth condition in x, i.e.,

$$|L(t,x)| \le C_L(1 + \| x \|^{c_L})$$

for some constants C_L and c_L. The following well-known verification theorem provides a sufficient condition for the value function and an optimal control strategy using Bellman's principle of dynamic programming.

Theorem 2 (Verification Theorem, Theorem 4.3.1 in [26]) *Suppose that $\bar{V} \in C^{1,2}(Q) \cap C_p(\bar{Q})$ satisfies*

$$\sup_{u\in\mathscr{U}} \{g(t,x,u) + A^u \bar{V}(t,x)\} = 0, \quad (t,x) \in Q \tag{8}$$

and $\bar{V}(T,x) = \Lambda(x)$. Then

(a) $J(t,x;\boldsymbol{u}) \le \bar{V}(t,x)$ for any $\boldsymbol{u} \in \mathscr{U}_{\mathbb{F}}(t)$ and any $(t,x) \in \bar{Q}$.

(b) If there exists a function $u^* : \bar{Q} \to \mathscr{U}$ such that

$$g(t,x,u^*(t,x)) + A^{u^*(t,x)}\bar{V}(t,x) = \max_{u\in\mathscr{U}} \{g(t,x,u) + A^u \bar{V}(t,x)\} = 0 \tag{9}$$

for all $(t,x) \in Q$ and if the control strategy defined as $\boldsymbol{u}^* = (u_t^*)_{t\in[0,T]}$ with $u_t^* \triangleq u^*(t,x_t)$ is admissible at time 0 (i.e., $\boldsymbol{u}^* \in \mathscr{U}_{\mathbb{F}}(0)$), then

1) *$\bar{V}(t,x) = V(t,x) = \sup_{\boldsymbol{u}\in\mathscr{U}_{\mathbb{F}}(t)} J(t,x;\boldsymbol{u})$. for all $(t,x) \in \bar{Q}$.*
2) *$\boldsymbol{u}^*$ is an optimal control strategy, i.e., $V(0,x) = J(0,x;\boldsymbol{u}^*)$.*

Equation (8) is the well-known HJB equation associated with the stochastic optimal control problem (6)-(7). However, the existence of $\bar{V} \in C^{1,2}(Q)$ in Theorem 2 requires many technical assumptions that might not be true in practice. For example, the HJB equation is usually assumed to be of *uniformly parabolic* type if there exists $c_\sigma > 0$ such that for all $(t,x,u) \in Q \times \mathscr{U}$ and $\xi \in \mathbb{R}^n$,

$$\xi^\top(\sigma\sigma^\top)(t,x,u)\xi \ge c_\sigma \| \xi \|^2 .$$

Otherwise, a classic solution $\bar{V} \in C^{1,2}(Q)$ may not be expected and we need to interpret the value function as a viscosity solution to the HJB equation (see, e.g., [26]).

C. Dual Representation of Controlled Markov Diffusions

In this subsection we present the information relaxation-based dual formulation of controlled Markov diffusions. In a similar way we relax the constraint that the decision at every time instant should be made based on the current information and impose penalty to punish the access to the future information. We will establish the weak duality, strong duality and complementary slackness results for controlled Markov diffusions, which parallel the results in MDPs. The value function-based optimal penalty is also characterized to motivate the practical use of our dual formulation, which will be demonstrated in Section III.

We consider the perfect information relaxation, i.e., we can foresee all the future randomness generated by the Brownian motion so that the decision made at any time $t \in [0,T]$ is based on the information set $\mathscr{F} = \mathscr{F}_T$. To expand the set of the feasible controls, we use $\mathscr{U}(t)$ to denote the set of measurable $\mathscr{U}$-valued control strategies at time t, i.e., $\mathbf{u} = (u_s)_{s\in[t,T]} \in \mathscr{U}(t)$ if $\mathbf{u}$ is $\mathscr{B}([t,T]) \times \mathscr{F}$-measurable and u_s takes value in $\mathscr{U}$ for $s \in [t,T]$, where $\mathscr{B}([t,T])$ is the Borel σ-algebra on $[t,T]$. In particular, $\mathscr{U}(0)$ can be viewed as the counterpart of $\mathbb{A}$ introduced in Section II-A for MDPs.

Unlike the case of MDPs, the first technical problem we have to face with is to define a solution of (6) with an anticipative control $\mathbf{u} \in \mathscr{U}(0)$. Since it involves the concept of "anticipating stochastic calculus", we postpone the relevant details to Appendix D-A, where we use the decomposition technique to define the solution of an anticipating SDE following [20], [18]. For this purpose we need to restrict the dependence of $\sigma(t,x,u)$ in (6) only on t and x, i.e., we have $\sigma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$ in this subsection.

Right now we assume that given a control strategy $\mathbf{u} \in \mathscr{U}(0)$ there exists a unique solution $(x_t)_{t\in[0,T]}$ to (6) that is $\mathscr{B}([0,T]) \times \mathscr{F}$-measurable. Next we consider the set of penalty functions in the setting of controlled Markov diffusions. Suppose $h(\mathbf{u},\mathbf{w})$ is a function depending on a control strategy $\mathbf{u} \in \mathscr{U}(0)$ and a sample path of Brownian motion $\mathbf{w} \triangleq (w_t)_{t\in[0,T]}$. We define the set $\mathscr{M}_{\mathbb{F}}(0)$ of *dual feasible penalties* $h(\mathbf{u},\mathbf{w})$ that do not penalize non-anticipative policies in expectation, i.e.,

$$\mathbb{E}_{0,x}[h(\mathbf{u},\mathbf{w})] \leq 0 \quad \text{for all } x \in \mathbb{R}^n \text{ and } \mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0).$$

In the following we will show $\mathscr{M}_{\mathbb{F}}(0)$ parallels the role of $\mathscr{M}_{\mathbb{G}}(0)$ for MDPs in the dual formulation of controlled Markov diffusions.

With an arbitrary choice of $h \in \mathscr{M}_{\mathbb{F}}(0)$, we can determine an upper bound on (7) with $t = 0$ by relaxing the constraint on the adaptiveness of control strategies.

Proposition 1 (Weak Duality) *If $h \in \mathscr{M}_{\mathbb{F}}(0)$, then for all $x \in \mathbb{R}^n$,*

$$\sup_{\boldsymbol{u}\in\mathscr{U}_{\mathbb{F}}(0)} J(0,x;\boldsymbol{u}) \leq \mathbb{E}_{0,x}\Big[\sup_{\boldsymbol{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h(\boldsymbol{u},\boldsymbol{w})\}\Big]. \tag{10}$$

Proof: Proof. For any $\bar{\mathbf{u}} \in \mathscr{U}_{\mathbb{F}}(0)$,

$$\begin{aligned} J(0,x;\bar{\mathbf{u}}) =& \mathbb{E}_{0,x}[\Lambda(x_T) + \int_0^T g(t,x_t,\bar{u}_t)dt] \\ \leq& \mathbb{E}_{0,x}[\Lambda(x_T) + \int_0^T g(t,x_t,\bar{u}_t)dt - h(\bar{\mathbf{u}},\mathbf{w})] \\ \leq& \mathbb{E}_{0,x}[\sup_{\mathbf{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h(\mathbf{u},\mathbf{w})\}]. \end{aligned}$$

Then inequality (10) can be obtained by taking the supremum over $\bar{\mathbf{u}} \in \mathscr{U}_{\mathbb{F}}(0)$ on the left hand side of the last inequality. ■

The optimization problem inside the conditional expectation in (10) is the counterpart of (4) in the context of controlled Markov diffusions: an entire path of $\mathbf{w}$ is known beforehand (i.e., *perfect information relaxation*), and the objective function depends on a specific trajectory of $\mathbf{w}$. Therefore, it is a deterministic and path-dependent optimal control problem parameterized by $\mathbf{w}$. We also call it an *inner optimization problem*, and the expectation term on the right hand side of (10) is a *dual bound* on the value function $V(0,x)$. [20], [22], [21] have conducted a series of research on this problem under the name "anticipative stochastic control"; in particular, one of the special cases they have considered is $h = 0$, which means the future information is accessed without any penalty. [20] characterized this reward due to the perfect information relaxation by a PDE. We would expect that the dual bound associated with zero penalty can be very loose as that in MDPs. Suppose the inner optimization problem can be solved by some technique. Then the evaluation of the dual bound is well suited to Monte Carlo simulation: we can generate a sample path of $\mathbf{w}$ and solve the inner optimization problem in (10), the solution of which is a high-biased estimator of $V(0,x)$.

An interesting case is when we choose

$$h^*(\mathbf{u},\mathbf{w}) = \Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - V(0,x). \tag{11}$$

Note that $h^* \in \mathscr{M}_{\mathbb{F}}(0)$, since

$$\mathbb{E}_{0,x}[\Lambda(x_T) + \int_0^T g(s,x_s,u_s)ds] \leq V(0,x) \quad \text{for all } x \in \mathbb{R}^n \text{ and } \mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0),$$

by the definition of $V(0,x)$. We also note that by plugging $h = h^*$ in the inner optimization problem in (10), the objective value of which is independent of $\mathbf{u}$ and it is always equal to $V(0,x)$. So the following strong duality result is obtained.

Theorem 3 (Strong Duality) *For all $x \in \mathbb{R}^n$,*

$$\sup_{\boldsymbol{u}\in\mathscr{U}_{\mathbb{F}}(0)} J(0,x;\boldsymbol{u}) = \inf_{h\in\mathscr{M}_{\mathbb{F}}(0)} \left\{ \mathbb{E}_{0,x}\big[\sup_{\boldsymbol{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h(\boldsymbol{u},\boldsymbol{w})\}\big] \right\}. \tag{12}$$

The minimum of the right hand side of (12) can always be achieved by choosing an $h \in \mathscr{M}_{\mathbb{F}}(0)$ in the form of (11).

Due to the strong duality result, the left hand side problem of (12) is referred to as the *primal problem* and the right hand side problem of (12) is referred to as the *dual problem*. Since the relaxation of the requirement on admissible strategies is penalized and compensated by using a proper function in $\mathscr{M}_{\mathbb{F}}(0)$, we can see why $h \in \mathscr{M}_{\mathbb{F}}(0)$

is called a (dual feasible) penalty function. If $\mathbf{u}^\star$ is a control strategy that achieves the supremum on the left side of (12), and $h^\star$ is a dual feasible penalty that achieves the infimum on the right side of (12), then they are referred to as the optimal solutions to the primal and dual problem, respectively. The "complementary slackness condition" in the next theorem characterizes such a pair $(u^\star, h^\star)$, which parallels the discrete-time problem (Theorem 2.2 in [6]).

Theorem 4 (Complementary Slackness) *Given $\boldsymbol{u}^\star \in \mathscr{U}_{\mathbb{F}}(0)$ and $h^\star \in \mathscr{M}_{\mathbb{F}}(0)$, a sufficient and necessary condition for $\boldsymbol{u}^\star$ and $h^\star$ being optimal to the primal and dual problem respectively is that*

$$\mathbb{E}_{0,x}[h^\star(\boldsymbol{u}^\star, \boldsymbol{w})] = 0,$$

and

$$\begin{aligned} &\mathbb{E}_{0,x}[\Lambda(x_T^\star) + \int_t^T g(s, x_s^\star, u_s^\star)ds - h^\star(\boldsymbol{u}^\star, \boldsymbol{w})] \\ =&\mathbb{E}_{0,x}\big[\sup_{\boldsymbol{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(s, x_s, u_s)ds - h^\star(\boldsymbol{u}, \boldsymbol{w})\}\big], \end{aligned} \tag{13}$$

where $x_t^\star$ is the solution of (6) using the control strategy $\boldsymbol{u}^\star = (u_t^\star)_{t\in[0,T]}$ on $[0,t)$ with the initial condition $x_0^\star = x$.

Here, we have the same interpretation on complementary slackness condition as that in the dual formulation of MDPs: if the penalty is optimal to the dual problem, the decision maker will be satisfied with an optimal non-anticipative control strategy even if she is able to choose any anticipative control strategy. Clearly, if an optimal control strategy $\mathbf{u}^*$ to the primal problem (6)-(7) does exist (see, e.g., Theorem 2(b)), then $\mathbf{u}^*$ and $h^*(\mathbf{u}, \mathbf{w})$ defined in (11) is a pair of the optimal solutions to the primal and dual problem. However, we note that the optimal penalty in the form of (11) is intractable as it depends on the exact value of $V(0,x)$. The next proposition provides some motivation to design good penalties.

Proposition 2 *Suppose the value function $V(t,x)$ and the optimal control $\boldsymbol{u}^*$ satisfy all the assumptions in Theorem 2(b). Then $h^*(\boldsymbol{u}^*, \boldsymbol{w})$ has the following equivalent form:*

$$h^*(\boldsymbol{u}^*, \boldsymbol{w}) = \int_0^T V_x(t, x_t^*)^\top \sigma(t, x_t^*) dw_t,$$

where x_t^* is the solution of (6) using the optimal control $\boldsymbol{u}^* = (u_t^*)_{t\in[0,T]}$ on $[0,t)$ with the initial condition $x_0^* = x$.

Proof: Proof. Since the value function $V(t,x) \in C^{1,2}(Q) \cap C_p(\bar{Q})$, we can apply the Ito differential rule on

$V(t,x_t^*)$ given $\mathbf{u}^* = (u^*)_{0\le t\le T}$ (note that $V(T,x_T^*) = \Lambda(x_T^*)$):

$$\begin{aligned} h^*(\mathbf{u}^*,\mathbf{w}) =& \Lambda(x_T^*) + \int_0^T g(t,x_t^*,u_t^*)dt - V(0,x) \\ =& V(0,x) + \int_0^T A^{u_t^*}V(t,x_t^*)dt + \int_0^T V_x(t,x_t^*)^\top \sigma(t,x_t^*)dw_t \\ &+ \int_0^T g(t,x_t^*,u_t^*)dt - V(0,x) \\ =& \int_0^T (A^{u_t^*}V(t,x_t^*) + g(t,x_t^*,u_t^*))dt + \int_0^T V_x(t,x_t^*)^\top \sigma(t,x_t^*)dw_t \\ =& \int_0^T V_x(t,x_t^*)^\top \sigma(t,x_t^*)dw_t, \end{aligned}$$

where the last equality holds due to (9). ■

The optimal penalty $h^*(\mathbf{u},\mathbf{w})$ can be written compactly as an Ito stochastic integral, when it is evaluated at $\mathbf{u} = \mathbf{u}^*$. A natural question would be whether $\int_0^T V_x(t,x_t)^\top \sigma(t,x_t)dw_t$ plays the role of an optimal penalty in (12) as $M^*(\mathbf{a},\mathbf{v})$ does in Theorem 1 achieving the strong duality. Unfortunately, $\int_0^T V_x(t,x_t)^\top \sigma(t,x_t)dw_t$ is not even a well-defined object in terms of an Ito stochastic integral, when $\mathbf{u}$ is not adapted to $\mathbb{F}$. To fix this problem, we also need the machinery of "anticipating stochastic calculus". However, we can still provide a concise answer here, that is, there exists an alternative optimal penalty that coincides with $\int_0^T V_x(t,x_t)^\top \sigma(t,x_t)dw_t$ when $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$. We fully develop the relevant results in Theorem 5 in Appendix D-B. In the following proposition we formalize one of the main results in Theorem 5, which also guides the numerical approximation scheme that will be illustrated in Section III.

Proposition 3 *Suppose the value function $V(t,x)$ defined in (7) satisfies all the assumptions in Theorem 2(b). Then under some technical conditions, there is an optimal solution to the dual problem, i.e., an optimal penalty $h_v^*(\boldsymbol{u},\boldsymbol{w}) \in \mathscr{M}_{\mathbb{F}}(0)$ in the form of*

$$h_v^*(\boldsymbol{u},\boldsymbol{w}) = \int_0^T V_x(t,x_t)^\top \sigma(t,x_t)dw_t \quad \textit{for } \boldsymbol{u} \in \mathscr{U}_{\mathbb{F}}(0), \tag{14}$$

where x_t is the solution of (6) using the control $\boldsymbol{u} = (u_t)_{t\in[0,T]}$ on $[0,t)$ with the initial condition $x_0 = x$.

Since the value functions $\{V(t,x), 0 \le t \le T\}$ are unknown in real applications, how does Proposition 3 guide us to generate a suboptimal penalty given approximate value functions $\{\hat{V}(t,x), 0 \le t \le T\}$ that are of sufficient regularity? The form of $h_v^*(\mathbf{u},\mathbf{w})$ implies that it can be approximated by $\hat{h}_v(\mathbf{u},\mathbf{w}) \triangleq \int_0^T \hat{V}_x(t,x_t)^\top \sigma(t,x_t)dw_t$ at least for $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$. If we further assume $\hat{V}_x(t,x)^\top \sigma(t,x)$ satisfies the polynomial growth condition in x, then $\mathbb{E}_{0,x}[\hat{h}_v(\mathbf{u},\mathbf{w})] = 0$ for all $x \in \mathbb{R}^n$ and $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$. As a result, $\hat{h}_v(\mathbf{u},\mathbf{w}) \in \mathscr{M}_{\mathbb{F}}(0)$, which means that $\hat{h}_v$ can be used to derive an upper bound on the value function $V(0,x)$ through (10). Therefore, in terms of the approximation scheme implied by the form of the optimal penalty, Proposition 3 presents a *value function-based penalty* that can be viewed as the continuous-time analogue of $M^*(\mathbf{a},\mathbf{v})$ in (5).

It is revealed by the complementary slackness condition in both discrete-time (Theorem 2.2 in [6]) and continuous-time (Theorem 4) cases that any optimal penalty has zero expectation evaluating at an optimal policy; as a stronger

version, the value function-based optimal penalty in both cases assign zero expectation to all non-anticipative polices (note that M^* in (5) is a sum of martingale differences under the original filtration $\mathbb{G}$).

Intuitively, we can interpret the strong duality achieved by the value function-based penalty as to offset the path-dependent randomness in the inner optimization problem; then the optimal control to the inner optimization problem coincides with that to the original stochastic control problem in the expectation sense, which is reflected by the proof of Theorem 5 in Appendix D-B for controlled Markov diffusions (resp., see the proof of Theorem 1(b) in [11] for MDPs). This idea should also apply to other continuous-time controlled Markov processes, for example, we can directly formulate the dual representation of controlled jump diffusions in a parallel way, and a similar result of Proposition 2 can be derived after applying the Ito formula with jumps.

In addition to the value function-based penalty, [20] (see its Theorem 2.1 and Theorem 2.2) proposed a Lagrangian approach that falls into our dual framework of controlled Markov diffusions, where the Lagrangian term behaves like a "penalty" function in the sense that it satisfies the complementary slackness condition developed in Theorem 4. We find that the derivation of this Lagrangian term is analogous to that of the gradient-based penalty proposed by [9] for MDPs; therefore, we review these results in Appendix D-C for reference.

We note that most of the numerical methods proposed so far for solving (7) focus on the operator of the HJB equation (see, for example, [16] and [17]). We will show the practical use of the dual formulation of controlled Markov diffusions, especially the value function-based penalty in the form of (14), in solving a dynamic portfolio choice problem in the next section.

Finally, we should point out that though the dual formulation of controlled Markov diffusions established in this subsection is valid provided that σ is a function of t and x, its validity only relies on the existence of a unique pathwise solution $(x_t)_{t\in[0,T]}$ to (6) with $\mathbf{u} \in \mathscr{U}(0)$. In other words, the dual formulation remains valid if such a solution can be properly defined for a general σ that also depends on u.

III. Dynamic Portfolio Choice Problem

In this section we will show how the value function-based optimal penalty helps to solve a classic dynamic portfolio choice problem with predictable returns and intermediate consumptions. Dating back to [27], [28], [29], dynamic portfolio choice problems have become computationally intensive due to more model features incorporated, such as position constraints, transaction costs and risk measures. Some recent works along this line include [30], [31], [32], [14]. Since most portfolio choice problems of practical interest cannot be solved analytically, various numerical methods have been developed to address this problem. These approximation schemes include but not limited to the martingale approach [33], [34], state-space discretization methods [35], [36], and approximate dynamic programming methods [37], [17]. These methods all induce sub-optimal policies, by performing which it is straightforward to obtain a lower bound on the optimal expected utility. However, it is often hard to tell how far the induced policy is from the optimal ones. Though some methods bear the property of asymptotic convergence, its accuracy with limited computational power cannot be measured. To overcome this problem, [38] and [9] constructed an upper bound on the expected utility based on the dual formulation of the constrained portfolio choice problem proposed

by [39] and the information relaxation duality method proposed by [6], [5], respectively. The gap between the lower bound and the upper bound can be used to justify the performance of a candidate policy.

We focus on solving a discrete-time dynamic portfolio choice problem that is discretized from a continuous-time model, which is similar to the one considered in [40] and [39]. We evaluate the lower bound on the optimal expected utility for this discrete-time model by performing sub-optimal policies using Monte Carlo simulation. To obtain an upper bound on the optimal expected utility simultaneously, we apply the information relaxation dual approach and propose a new class of penalties based on the time discretization of the value function-based optimal penalties of the continuous-time model; these penalties make the inner optimization problem much easier to solve in terms of computation compared with the penalties directly derived from the discrete-time model. We demonstrate the effectiveness of our method in computing dual bounds through numerical experiments.

A. The Portfolio Choice Model

We first consider a continuous-time financial market with finite horizon $[0,T]$, which is built on the probability space $(\Omega,\mathscr{F},\mathbb{P})$. There are one risk-free asset(cash) and n risky assets the investor can invest on. The risk-free asset is denoted by S_t^0 and the instantaneous risk-free rate of return is denoted by r_f. Then S_t^0 follows the process

$$\frac{dS_t^0}{S_t^0} = r_f dt.$$

The vector of risky assets is denoted by $S_t = (S_t^1,\cdots,S_t^n)^\top$ and it follows a geometric Brownian motion

$$\frac{dS_t}{S_t} = \mu_t dt + \sigma_t dz_t, \tag{15}$$

where $\frac{dS_t}{S_t}$ denotes $(\frac{dS_t^1}{S_t^1},\cdots,\frac{dS_t^n}{S_t^n})^\top$, and $\mathbf{z} \triangleq (z_t)_{0\le t\le T}$ is an n-dimensional standard Brownian motion. The drift $\mu_t = \mu(t,\phi_t)$ and the volatility $\sigma_t = \sigma(t,\phi_t)$ are of proper dimensions and are functions of an m-dimensional market state variable ϕ_t that follows another diffusion process

$$d\phi_t = \mu_t^\phi dt + \sigma_t^{\phi,1} dz_t + \sigma_t^{\phi,2} d\tilde{z}_t, \tag{16}$$

where $\mu_t^\phi = \mu^\phi(t,\phi_t)$, $\sigma_t^{\phi,1} = \sigma^{\phi,1}(t,\phi_t)$, $\sigma_t^{\phi,2} = \sigma^{\phi,2}(t,\phi_t)$, and $\tilde{\mathbf{z}} \triangleq (\tilde{z}_t)_{0\le t\le T}$ is another d-dimensional standard Brownian motion independent of $\mathbf{z}$. Denote the filtration by $\mathbb{F} = \{\mathscr{F}_t, 0\le t\le T\}$, where $\mathscr{F}_t$ is generated by $\{(z_s,\tilde{z}_s), 0\le s\le t\}$. The covariance matrices $\sigma_t\sigma_t^\top, \sigma_t^{\phi,1}(\sigma_t^{\phi,1})^\top, \sigma_t^{\phi,2}(\sigma_t^{\phi,2})^\top$ are denoted by $\Sigma_t, \Sigma_t^{\phi,1}$, and $\Sigma_t^{\phi,2}$, respectively.

Let $\pi_t = (\pi_t^1,\cdots,\pi_t^n)^\top$ denote the fraction of wealth invested in the risky assets. The instantaneous rate of consumption is denoted by $\tilde{c}_t$. The total wealth W_t of a portfolio that consists of n risky assets and one risk-free asset evolves according to

$$\begin{aligned} dW_t =& W_t\left[\pi_t^\top(\mu_t dt + \sigma_t dz_t) + r_f(1-\pi_t^\top \mathbf{1}_n)dt - \tilde{c}_t dt\right] \\ =& W_t\left(\pi_t^\top(\mu_t - r_f\mathbf{1}_n) + r_f - \tilde{c}_t\right)dt + W_t\pi_t^\top\sigma_t dz_t, \end{aligned} \tag{17}$$

where $\mathbf{1}_n$ is the n-dimensional all-ones vector. The control process $\mathbf{u} \triangleq (u_t)_{0\le t\le T}$ with $u_t \triangleq (\pi_t,\tilde{c}_t)$ is assumed to be an admissible strategy in the sense that

1) The control $\mathbf{u}$ is $\mathbb{F}$-progressively measurable and $\mathbb{E}[\int_0^T ||u_t||^2 dt] < \infty$;

2) We require $W_t > 0$, $\tilde{c}_t \geq 0$, and $\int_0^T W_t \tilde{c}_t dt < \infty$ a.s.;

3) We may restrict $u_t \in \mathscr{U}$, where $\mathscr{U}$ is a closed convex set in $\mathbb{R}^{n+1}$.

We still use $\mathscr{U}_{\mathbb{F}}(t)$ to denote the set of admissible strategies at time t and we will specify the control space $\mathscr{U}$ later. Suppose U_1 and U_2 are two strictly increasing and concave utility functions. The investor's objective is to maximize the weighted sum of the expected utility of the intermediate consumption and the final wealth :

$$V(t,\phi_t,W_t) = \sup_{\mathbf{u}\in\mathscr{U}_{\mathbb{F}}(t)} \mathbb{E}\big[\int_t^T \alpha\beta^s U_1(\tilde{c}_s W_s)ds + (1-\alpha)\beta^T U_2(W_T)\big|\phi_t, W_t\big], \tag{18}$$

where β is the discount rate, and α indicates the relative importance of the intermediate consumption. Since the utility gain caused by the intermediate consumption generally means a potential loss of the utility from the final wealth, the investor seeks to balance her dynamic portfolio strategy at every time instant.

The value function (18) sometimes admits an analytic solution, for example, under the assumption that μ_t and σ_t are constants and there is no constraint on $u_t = (\pi_t, \tilde{c}_t)$. A recent progress on the analytic tractability of (18) can be found in [40]. However, (18) usually does not have an analytic result when there is a position constraint on π_t.

Considering that the investment and consumption can only take place in a finite number of times in the real world, we solve the discrete-time counterpart of the continuous-time problem (16)-(18) by discretizing its time space. Suppose the decision takes place at equally spaced times $\{0 = t_0, t_1 \cdots, t_K\}$ such that $K = T/\delta$, where $\delta = t_{k+1} - t_k$ for $k = 0, 1, \cdots, K-1$. We simply denote the time grids by $\{0, 1, \cdots, K\}$. Note that (15) is equivalent to

$$d\log(S_t) = (\mu_t - \frac{1}{2}\sigma_t^2)dt + \sigma_t dz_t,$$

where σ_t^2 denotes the vector that consists of the diagonal of Σ_t. That is to say, $S_{k+1} = R_{k+1} S_k$ with $\log(R_{k+1}) \sim N((\mu_k - \frac{1}{2}\sigma_k^2)\delta, \Sigma_k\delta)$, or more precisely, $\log(R_{k+1}) \sim N(\int_{k\delta}^{(k+1)\delta}(\mu_s - \frac{1}{2}\sigma_s^2)ds, \int_{k\delta}^{(k+1)\delta}\Sigma_s ds)$. Hence, we can discretize (16),(15), and (17) as follows:

$$\phi_{k+1} = \phi_k + \mu_k^{\phi}\delta + \sigma_k^{\phi,1}\sqrt{\delta}Z_{k+1} + \sigma_k^{\phi,2}\sqrt{\delta}\tilde{Z}_{k+1}, \tag{19a}$$

$$\log(R_{k+1}) = (\mu_k - \frac{1}{2}\sigma_k^2)\delta + \sigma_k\sqrt{\delta}Z_{k+1}, \tag{19b}$$

$$\begin{aligned} W_{k+1} &= W_k(R_{k+1}^{\top}\pi_k) + W_k(1 - \mathbf{1}_n^{\top}\pi_k)R_f - W_k c_k, \\ &= W_k(R_f + (R_{k+1} - R_f\mathbf{1}_n)^{\top}\pi_k - c_k), \end{aligned} \tag{19c}$$

where $\{(Z_k, \tilde{Z}_k), k = 1, \cdots, K\}$ is a sequence of identically and independently distributed standard Gaussian random vectors. In particular, we use $R_f \triangleq 1 + r_f\delta$ and the decision variable c_k to approximate $e^{r_f\delta}$ and $\tilde{c}_k\delta$ due to the discretization procedure.

Here we abuse the notations ϕ, W, and π in the continuous-time and discrete-time settings. However, the subscripts make them easy to distinguish: the subscript $t \in [0,T]$ is used in the continuous-time model, while $k = 0, \cdots, K$ is used in the discrete-time model.

Denote the filtration of the process (19) by $\mathbb{G} = \{\mathscr{G}_0, \cdots, \mathscr{G}_K\}$, where $\mathscr{G}_k$ is generated by $\{(Z_j, \tilde{Z}_j), j = 0, \cdots, k\}$. In our numerical examples we assume that short sales and borrowing are not allowed, and the consumption cannot exceed the amount of the risky-free asset. Then the constraint on the control $a_k \triangleq (\pi_k, c_k)$ for the discrete-time problem can be defined as

$$\mathscr{A} \triangleq \{(\pi, c) \in \mathbb{R}^{n+1} | \pi \geq 0, c \geq 0, c \leq R_f(1 - \mathbf{1}_n^\top \pi)\}. \tag{20}$$

Since c_k is used to approximate $\tilde{c}_k \delta$, (20) corresponds to a control set for the continuous-time model, which is defined as

$$\mathscr{U} \triangleq \{(\pi, \tilde{c}) \in \mathbb{R}^{n+1} | \pi \geq 0, \tilde{c} \geq 0, \tilde{c} \leq R_f(1 - \mathbf{1}_n^\top \pi)/\delta\}.$$

Let $\mathbb{A}_\mathbb{G}$ again denote the set of $\mathscr{A}$-valued control strategies $\mathbf{a} \triangleq (a_1, \cdots, a_{K-1})$ that are adapted to the filtration $\mathbb{G}$. The discretization of (18) serves as the value function to the discrete-time problem:

$$H_0(\phi_0, W_0) = \sup_{\mathbf{a} \in \mathbb{A}_\mathbb{G}} \mathbb{E}_0[\sum_{k=0}^{K-1} \alpha \beta^{k\delta} U_1(c_k W_k)\delta + (1-\alpha)\beta^{K\delta} U_2(W_K)], \tag{21}$$

which can be solved via dynamic programming:

$$\begin{aligned} H_K(\phi_K, W_K) &= (1-\alpha)\beta^{K\delta} U_2(W_K); \\ H_k(\phi_k, W_k) &= \sup_{a_k \in \mathscr{A}} \{\alpha \beta^{k\delta} U_1(c_k W_k)\delta + \mathbb{E}[H_{k+1}(\phi_{k+1}, W_{k+1}) | \phi_k, W_k]\}. \end{aligned} \tag{22}$$

We will focus on solving the discrete-time model (19)-(21), which is discretized from the continuous-time model (16)-(18). Though our methods proposed later can apply on general utility functions, for the purpose of illustration we consider the utility functions of the constant relative risk aversion (CRRA) type with coefficient $\gamma > 0$, *i.e*, $U(x) \triangleq U_1(x) = U_2(x) = \frac{1}{1-\gamma} x^{1-\gamma}$, which are widely used in economics and finance. Due to the utility functions of CRRA type, both value functions (18) and (21) have simplified structures. To be specific, the value function to the continuous-time problem can be written as

$$V(t, \phi_t, W_t) = \beta^t W_t^{1-\gamma} \tilde{J}(t, \phi_t), \tag{23}$$

where $\tilde{J}(T, \phi_T) = (1-\alpha)/(1-\gamma)$, and

$$\tilde{J}(t, \phi) = \sup_{\mathbf{u} \in \mathscr{U}_\mathbb{F}(t)} \mathbb{E}[\int_t^T \beta^{s-t} \frac{\alpha}{1-\gamma} (\tilde{c}_s W_s)^{1-\gamma} ds + \beta^{T-t} \frac{1-\alpha}{1-\gamma} W_T^{1-\gamma} | \phi_t = \phi, W_t = 1];$$

and the value function to the discrete-time problem is

$$H_k(\phi_k, W_k) = \beta^{k\delta} W_k^{1-\gamma} J_k(\phi_k), \tag{24}$$

where J_k is defined recursively as $J_K(\phi_K) = (1-\alpha)/(1-\gamma)$ and

$$J_k(\phi_k) = \sup_{(\pi_k, c_k) \in \mathscr{A}} \{\frac{\alpha}{1-\gamma} c_k^{1-\gamma} \delta + \beta^\delta \mathbb{E}\big[\big(R_f + (R_{k+1} - R_f)^\top \pi_k - c_k\big)^{1-\gamma} J_{k+1}(\phi_{k+1}) | \phi_k\big]\}. \tag{25}$$

It can be seen that the structure of the value functions to both continuous-time model and discrete-time model are similar: they can be decomposed as a product of a function of the wealth W and a function of the market state variable ϕ. If δ is small, $\tilde{J}(k\delta, \phi)$ and $J_k(\phi)$ may be close to each other. As a byproduct of this decomposition, another

feature of the dynamic portfolio choice problem with CRRA utility function is that the optimal asset allocation and consumption $(\pi_t, \tilde{c}_t)$ in continuous-time model is independent of the wealth W_t given ϕ_t (respectively, the optimal (π_k, c_k) in discrete-time model is independent of the wealth W_k given ϕ_k). So the dimension of the state space in (22) is actually the dimension of ϕ_k. A number of numerical methods have been developed to solve the discrete-time model based on the recursion (25) including the state-space discretization approach [35], [36], and a simulation-based method [37].

B. Penalties and Dual Bounds

Since we can use Monte Carlo simulation to evaluate the expected utility under any admissible strategy, we are interested in how good the strategy is and how much better we could do. The purpose of this subsection is to provide a way to evaluate the quality of the strategies developed for the discrete-time (continuous-state) model (19)-(21). Note that this problem falls in the framework of the dual approach for MDPs introduced in Theorem 1, which can be used to complement the lower bound on the value function H_0 by an upper bound in principle. However, Theorem 1 does not directly suggest a tractable approximation scheme on penalty functions for continuous-state problems. According to the numerical experiments reported in [6], [9], the choices of the penalties significantly influence the quality of the dual bounds. Therefore, we are aiming to accomplish two goals in this subsection:

- We want to design appropriate penalties that can help to achieve tight dual bounds.
- We want to keep the computational cost of the inner optimization problem at a reasonable level.

Throughout this subsection we assume that an approximate function of $J_k(\phi)$, say $\hat{J}_k(\phi)$, and an approximate policy $\hat{\mathbf{a}} \in \mathbb{A}_{\mathbb{G}}$ are available. We do not require that $\hat{\mathbf{a}}$ should be derived based on $\hat{J}_k(\phi)$ and vice versa; in other words, they can be obtained using different approaches. We first describe the information relaxation dual approach of MDPs in the context of our portfolio choice problem. We focus on the perfect information relaxation that assumes the investor can foresee the future uncertainty $\mathbf{Z} = (Z_1, \cdots, Z_K)$ and $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \cdots, \tilde{Z}_K)$, i.e., all the market states and returns of the risky assets. A function $M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})$ is a *dual feasible penalty* in the setting of dynamic portfolio choice problem if for any (ϕ_0, W_0),

$$\mathbb{E}[M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) | \phi_0, W_0] \leq 0 \quad \text{for all} \quad \mathbf{a} \in \mathbb{A}_{\mathbb{G}}. \tag{26}$$

Let $\mathscr{M}_{\mathbb{G}}(0)$ denote the set of all dual feasible penalties. For $M \in \mathscr{M}_{\mathbb{G}}(0)$ we define $\mathscr{L}M$ as a function of (ϕ_0, W_0):

$$(\mathscr{L}M)(\phi_0, W_0) = \mathbb{E}\left[\sup_{\mathbf{a}\in\mathbb{A}}\{\sum_{k=0}^{K-1} \alpha\beta^{k\delta} U(c_k W_k)\delta + (1-\alpha)\beta^{K\delta} U(W_K) - M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})\} | \phi_0, W_0\right]. \tag{27}$$

Based on Theorem 1(a), $(\mathscr{L}M)(\phi_0, W_0)$ is an upper bound on $H_0(\phi_0, W_0)$ for any $M \in \mathscr{M}_{\mathbb{G}}(0)$.

To ease the inner optimization problem, we introduce equivalent decision variables $\Pi_k = W_k \pi_k$ and $C_k = W_k c_k$, which can be interchangeably used with π_k and c_k. We still use $\mathbf{a}$ to denote an admissable strategy, though in terms of (Π_k, C_k) now. Then we can rewrite the inner optimization problem in the conditional expectation in (27) as follows:

$$I(\phi_0, W_0, M, \mathbf{Z}, \tilde{\mathbf{Z}}) \triangleq \max_{\mathbf{a}} \{ \sum_{k=0}^{K-1} \alpha \beta^{k\delta} U(C_j)\delta + (1-\alpha)\beta^{K\delta} U(W_K) - M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \} \tag{28a}$$

$$\text{s.t. } \phi_{k+1} = \phi_k + \mu_k^{\phi}\delta + \sigma_k^{\phi,1}\sqrt{\delta} Z_{k+1} + \sigma_k^{\phi,2}\sqrt{\delta}\tilde{Z}_{k+1},$$

$$\log(R_{k+1}) = (\mu_k - \frac{1}{2}\sigma_k^2)\delta + \sigma_k\sqrt{\delta} Z_{k+1},$$

$$W_{k+1} = W_k R_f + (R_{k+1} - R_f \mathbf{1}_n)^{\top}\Pi_k - C_k, \tag{28b}$$

$$\Pi_k \geq 0, \quad C_k \geq 0, \tag{28c}$$

$$C_k \leq R_f(W_k - 1_n^{\top}\Pi_k), \text{ for } k = 0, \cdots, K-1. \tag{28d}$$

Note that (28b) is equivalent to (19c), and(28c)-(28d) are equivalent to (20). The advantage of this reformulation is that the inner optimization problem (28) has linear constraints. Therefore, we may find the global maximizer of (28) as long as the objective function in (28a) is jointly concave in $\mathbf{a}$.

Heuristically, we need to design near-optimal penalty functions in order to obtain tight dual bounds on H_0. A natural approach is to construct a penalty function by replacing the value function in the optimal penalty for the discrete-time problem (see (5)) with its approximation. In the following we will see why deriving penalty in this way will result in an intractable inner optimization problem. Alternatively, the inner optimization problem becomes tractable when we construct a heuristic penalty according to the form of the optimal penalty for the continuous-time model. We first investigate the optimal penalty for the discrete-time problem according to (5):

$$M^*(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) = \sum_{k=0}^{K-1} \Delta H_{k+1}(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}).$$

Based on the factorized structure of H_k in (24), we can obtain

$$\begin{aligned}\Delta H_{k+1}(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) =& H_{k+1}(\phi_{k+1}, W_{k+1}) - \mathbb{E}[H_{k+1}(\phi_{k+1}, W_{k+1})|\phi_k, W_k, a_k] \\ =& \beta^{(k+1)\delta}\big(\kappa(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) - \mathbb{E}[\kappa(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})|\phi_k, W_k, a_k]\big),\end{aligned} \tag{29}$$

where

$$\kappa(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \triangleq (W_k R_f + (R_{k+1} - R_f \mathbf{1}_n)^{\top}\Pi_k - C_k)^{1-\gamma} J_{k+1}(\phi_{k+1}).$$

We can explicitly write out $\phi_{k+1} = \phi_{k+1}(\phi_0, \tilde{\mathbf{Z}}_{1:k+1}, \mathbf{Z}_{1:k+1})$ and $R_{k+1} = R_{k+1}(\phi_0, \tilde{\mathbf{Z}}_{1:k+1}, \mathbf{Z}_{1:k+1})$ via (19a)-(19b) to emphasize the dependence on the randomness sequence $(\mathbf{Z}, \tilde{\mathbf{Z}})$, where $\mathbf{Z}_{1:k}$ (resp., $\tilde{\mathbf{Z}}_{1:k}$) denotes the first k entries of $\mathbf{Z}$ (resp., $\tilde{\mathbf{Z}}$). In practice we can approximate $J_k(\phi_k)$ by $\hat{J}_k(\phi_k)$; however, it does not mean that an approximation of ΔH_{k+1} can be easily computed, since an intractable conditional expectation over $(m+n)$-dimensional space is involved in (29). Another difficulty is that $M^* = \sum_{k=0}^{K-1} \Delta H_{k+1}$ enters into (28a) with possibly positive or negative signs for different realizations of $(\mathbf{Z}, \tilde{\mathbf{Z}})$, making the objective function of (28) nonconcave, even if U_1 and U_2 are concave functions. Therefore, it might be extremely hard to locate the global maximizer of (28).

To address these problems, we exploit the value function-based optimal penalty h_v^* for the continuous-time problem (16)-(18), assuming that all the technical conditions are satisfied. If we disregard the dependence of the

diffusion coefficient of W_t (the second term on the right side of (17)) on π_t, then according to Proposition 3 we can formally write h_v^* as

$$\begin{aligned}
h_v^*(\mathbf{u},\mathbf{z},\tilde{\mathbf{z}}) &= \int_0^T \begin{pmatrix} V_\phi(t,\phi_t,W_t) \\ V_W(t,\phi_t,W_t) \end{pmatrix}^\top \begin{pmatrix} \sigma_t^{\phi,1} & \sigma_t^{\phi,2} \\ W_t\pi_t\sigma_t & 0 \end{pmatrix} \begin{pmatrix} dz_t \\ d\tilde{z}_t \end{pmatrix} \\
&= \sum_{k=0}^{K-1} \int_{k\delta}^{(k+1)\delta} \big[V_\phi(t,\phi_t,W_t)^\top \sigma_t^{\phi,1} dz_t + V_\phi(t,\phi_t,W_t)^\top \sigma_t^{\phi,2} d\tilde{z}_t \\
&\quad + V_W(t,\phi_t,W_t) W_t \pi_t \sigma_t dz_t\big], \\
&= \sum_{k=0}^{K-1} \int_{k\delta}^{(k+1)\delta} \beta^t \big[W_t^{1-\gamma} \nabla_\phi \tilde{J}(t,\phi_t)^\top \sigma_t^{\phi,1} dz_t + W_t^{1-\gamma} \nabla_\phi \tilde{J}(t,\phi_t)^\top \sigma_t^{\phi,2} d\tilde{z}_t \\
&\quad + (1-\gamma) W_t^{1-\gamma} \tilde{J}(t,\phi_t) \pi_t \sigma_t dz_t\big],
\end{aligned} \tag{30}$$

for $\mathbf{u} = (\pi_t, \tilde{c}_t)_{0\le t\le T} \in \mathscr{U}_{\mathbb{F}}(0)$, and the last equality holds due to structure of the value function (23). In particular, we use ∇_ϕ to denote the gradient of the function $\tilde{J}$ with respect to ϕ. Motivated by the fact that our discrete-time model is discretized from the continuous-time model, we propose the following function that approximates each ΔH_{k+1} in M^*, that is,

$$\beta^{k\delta}\big[W_k^{1-\gamma} \nabla_\phi J_k(\phi_k)^\top \sigma_k^{\phi,1} \sqrt{\delta} Z_{k+1} + W_k^{1-\gamma} \nabla_\phi J_k(\phi_k)^\top \sigma_k^{\phi,2} \sqrt{\delta} \tilde{Z}_{k+1} + (1-\gamma) W_k^{-\gamma} J_k(\phi_k) \Pi_k^\top \sigma_k \sqrt{\delta} Z_{k+1}\big]. \tag{31}$$

It is obvious that this term is derived by discretizing the Ito stochastic integrals in the $(k+1)$-th term of the summation in (30), and we will justify that it is a dual feasible penalty for the discrete-time problem later. For the purpose of incorporating this term to the dual approach in the numerical implementation, we should note that the value of $J_k(\phi_k)$ and the differentiability of $J_k(\phi)$ in ϕ are not known yet. However, as mentioned in the beginning of Section III, we can approximate $J_k(\phi_k)$ using different approaches that lead to piecewise linear functions (by state-space discretization method) or smooth functions (by approximate dynamic programming method), denoted by $\hat{J}_k(\phi_k)$. Hence, $\nabla_\phi J_k(\phi_k)$ can be formally approximated by the gradient of these approximate functions, namely, $\nabla_\phi \hat{J}_k(\phi_k)$. We will formalize in Proposition 4 that the approximations in $J_k(\phi_k)$ and $\nabla_\phi J_k(\phi_k)$ will not influence the validity of (31) being a dual feasible penalty for the discrete-time problem (i.e., condition (26) is satisfied).

We describe the procedure of evaluating an applicable penalty function using simulation. We first generate a realization of $(\mathbf{Z},\tilde{\mathbf{Z}})$ and consequently we can obtain $\bar{\phi}_k \triangleq \phi_k(\phi_0,\mathbf{Z}_k,\tilde{\mathbf{Z}}_k)$, $\bar{\sigma}_k \triangleq \sigma(\bar{\phi}_k)$, $\bar{\sigma}_k^{\phi,1} \triangleq \sigma^{\phi,1}(k,\bar{\phi}_k)$, $\bar{\sigma}_k^{\phi,2} \triangleq \sigma^{\phi,2}(k,\bar{\phi}_k)$, $\hat{J}_k(\bar{\phi}_k)$, and $\nabla_\phi \hat{J}_k(\bar{\phi}_k)$; with an admissible strategy $\hat{\mathbf{a}} = (\hat{a}_0,\cdots,\hat{a}_K)$, we can also obtain $\bar{W}_k \triangleq W_k(W_0,\hat{\mathbf{a}}(\phi_0,W_0,\mathbf{Z}_k,\tilde{\mathbf{Z}}_k),\mathbf{Z}_k,\tilde{\mathbf{Z}}_k)$ via (19c) as an approximation to the wealth under the optimal policy. Then we can approximate $M^*(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}})$ by

$$M_1(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) \triangleq \sum_{k=0}^{K-1} \beta^{k\delta}\big[\Psi_{k+1}^1(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) + \Psi_{k+1}^2(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) + \Psi_{k+1}^3(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}})\big], \tag{32}$$

where

$$\begin{aligned}
\Psi_{k+1}^1(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) &= \bar{W}_k^{1-\gamma} \nabla_\phi \hat{J}_k(\bar{\phi}_k)^\top \bar{\sigma}_k^{\phi,1} \sqrt{\delta} Z_{k+1}, \\
\Psi_{k+1}^2(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) &= \bar{W}_k^{1-\gamma} \nabla_\phi \hat{J}_k(\bar{\phi}_k)^\top \bar{\sigma}_k^{\phi,2} \sqrt{\delta} \tilde{Z}_{k+1}, \\
\Psi_{k+1}^3(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) &= (1-\gamma) \bar{W}_k^{-\gamma} \hat{J}_k(\bar{\phi}_k) \Pi_k^\top \bar{\sigma}_k \sqrt{\delta} Z_{k+1}.
\end{aligned}$$

Note that when a realization of $(\mathbf{Z},\tilde{\mathbf{Z}})$ is fixed, Ψ^1_{k+1} and Ψ^2_{k+1} are constants with respect to $\mathbf{a}$ (but varies across sample paths), which can be seen as control variates; Ψ^3_{k+1} depends on Π_k (hence, on $\mathbf{a}$), and thus is the only term that contributes to the inner optimization problem (28). Since Ψ^3_{k+1} is affine in Π_k, the objective function (28a) is jointly concave in $\mathbf{a}$ with $M = M_1$. As a result, the inner optimization problem (28) remains a convex optimization problem and can be easily solved. In our numerical experiments, we will consider dual bounds generated by this penalty.

To find some variants of the penalties while still keeping the inner optimization problem convex, we also generate $\breve{\Psi}^1_{k+1}$ based on a first-order Taylor expansion of Ψ^1_{k+1} around the strategy $\hat{a}_{k-1} = (\hat{\Pi}_{k-1}, \hat{C}_{k-1})$:

$$\begin{aligned}\breve{\Psi}^1_{k+1}(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) = \big[\bar{W}_k^{1-\gamma} + (1-\gamma)\bar{W}_k^{-\gamma}\big((\bar{R}_k - R_f\mathbf{1}_n)^\top(\Pi_{k-1} - \bar{\Pi}_{k-1}) \\ - (C_{k-1} - \bar{C}_{k-1})\big)\big] \cdot \nabla_\phi \hat{J}_k(\bar{\phi}_k)^\top \bar{\sigma}_k^{\phi,1}\sqrt{\delta} Z_{k+1},\end{aligned} \tag{33}$$

where $\bar{R}_k \triangleq R_k(\phi_0, \mathbf{Z}_k, \tilde{\mathbf{Z}}_k)$, $\bar{\Pi}_{k-1} \triangleq \hat{\Pi}_{k-1}(\phi_0, W_0, \mathbf{Z}_{k-1}, \tilde{\mathbf{Z}}_{k-1})$, and $\bar{C}_{k-1} \triangleq \hat{C}_{k-1}(\phi_0, W_0, \mathbf{Z}_{k-1}, \tilde{\mathbf{Z}}_{k-1})$. Then $\breve{\Psi}^1_{k+1}$ is linear in Π_{k-1} and C_{k-1}. We can also obtain a variant of Ψ^2_{k+1}, say $\breve{\Psi}^2_{k+1}$, in exactly the same way. Since Ψ^3_{k+1} is already linear in Π_k, we do not linearize it with respect to $\hat{a}_{k-1}$. In our numerical experiments we will also consider dual bounds generated by

$$M_2(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) \triangleq \sum_{k=0}^{K-1} \beta^{k\delta}\big[\breve{\Psi}^1_{k+1}(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) + \breve{\Psi}^2_{k+1}(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) + \Psi^3_{k+1}(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}})\big]. \tag{34}$$

To go further, we can also generate a penalty function by linearizing Ψ^1_{k+1} around $(\hat{a}_0, \cdots, \hat{a}_{k-1})$. Finally, we justify the validity of M_1 and M_2 being dual feasible penalties in Proposition 4.

Proposition 4 *The functions M_1 in (32) and M_2 in (34) are dual feasible penalties, i.e., $M_1, M_2 \in \mathscr{M}_{\mathbb{G}}(0)$. Hence, $\mathscr{L}M_1(\phi_0, W_0)$ and $\mathscr{L}M_2(\phi_0, W_0)$ are upper bounds on the value function $H_0(\phi_0, W_0)$.*

Proof: Proof. We observe that with a fixed non-anticipative policy $\hat{\mathbf{a}} \in \mathbb{A}_{\mathbb{G}}$, it is obvious that $\bar{\phi}_k$, $\bar{W}_k$, $\hat{J}_k(\bar{\phi}_k)$, $\nabla_\phi \hat{J}_k(\bar{\phi}_k)$, $\bar{\sigma}_k$, and $\bar{\sigma}_k^{\phi,j}$, $j = 1,2$, are naturally $\mathscr{G}_k$-adapted for $k = 0, \cdots, K-1$. We also note that Π_k is $\mathscr{G}_k$-adapted due to $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$. Since Z_{k+1} and $\tilde{Z}_{k+1}$ have zero means and are independent of $\mathscr{G}_k$ and (ϕ_0, W_0), we have for any (ϕ_0, W_0),

$$\mathbb{E}[\Psi^i_{k+1}(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) | \phi_0, W_0] = 0 \quad \text{for all} \quad \mathbf{a} \in \mathbb{A}_{\mathbb{G}},$$

for $i = 1,2,3$. So $\mathbb{E}[M_1(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) | \phi_0, W_0] = 0$ for all $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$, and hence $M_1 \in \mathscr{M}_{\mathbb{G}}(0)$. Since the same argument can apply on $\breve{\Psi}^i_{k+1}(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}})$ for $i = 1,2$, it can be concluded that $M_2 \in \mathscr{M}_{\mathbb{G}}(0)$. ■

The penalty function in the form of (32) or (34) bear several advantages: first, it can be evaluated without computing any conditional expectation, i.e., a substantial computational work can be avoided; second, the design of the penalty function is quite flexible: we can use any admissible policy to obtain a valid penalty, and we can choose to do a linearization around this policy, which makes the inner optimization problem (28) convex and computationally tractable.

C. Numerical Experiments

In this section we discuss the use of Monte Carlo simulation to evaluate the performance of the suboptimal policies and the dual bounds on the expected utility (21). We consider a model with three risky assets $(n = 3)$ and one market state variable $(m = 1)$. We choose $T = 1$ year and $\delta = 0.1$ year in our numerical experiments. In addition, we use $\alpha = 0.5$ for the weight of the intermediate utility function and use $\beta = 1$ as the discount factor. Other information on the state equation (19) can be found in Appendix E. In particular, the market state variable $\{\phi_k\}$ follows a mean-reverting Ornstein-Uhlenbeck process: it has relatively small mean reversion rate and volatility in parameter sets 1 and 3, while it has relatively large mean reversion rate and volatility in parameter sets 2 and 4. We assume $\phi_0 = 0$ and $W_0 = 1$ as the initial condition and impose the constraint (20) on the control space $\mathscr{A}$ in the following numerical tests.

For each parameter set we first use the discrete state-space approximation method to solve the recursion (25). In particular, we approximate the market state variable ϕ_k using a grid with 21 equally spaced girds from -2 to 2, and the transition between these grid points is determined by (19a) noting that $\phi_{k+1} \sim N\big(\phi_k + \mu_k^{\phi}\delta, (\| \sigma_k^{\phi,1} \|^2 + \| \sigma_k^{\phi,2} \|^2)\delta\big)$; the random variables Z_k and $\tilde{Z}_k$ are approximated by Gaussian quadrature method with 3 points for each dimension (see, e.g., [41]). So the joint distribution of the market state and the returns are approximated by a total of $3^3 \times 21 = 567$ grid points, which are used to compute the conditional expectation in (25): we assume ϕ_{k+1} and R_{k+1} are independent conditioned on ϕ_k, then the conditional expectation reduces to a finite weighted sum. For the optimization problem in (25) we use CVX ([42]), a package to solve convex optimization problems in Matlab, to determine the optimal consumption and investment policy on each grid of ϕ_k at time k. We record the value function and the corresponding policy on this grid at each time $k = 0, \cdots, K$. Note that the market state variable ϕ_k is one dimensional, so the value function and the policy can be naturally defined on the market state ϕ_k that is outside the grid by piecewise linear interpolation. In our numerical implementation the extended value function and the extended policy play the roles of the approximate value function $\hat{J}_k(\phi_k)$ and the approximate policy $\hat{\mathbf{a}}$ to the discrete-time problem (19)-(21); and we take the slope of the piecewise linear function $\hat{J}_k(\phi)$ as $\nabla_\phi \hat{J}_k(\phi)$, if ϕ is between the grid points; otherwise, we can use the average slope of two consecutive lines as $\nabla_\phi \hat{J}_k(\phi)$.

We then repeatedly generate random sequences of $(\mathbf{Z}, \tilde{\mathbf{Z}})$, based on which we generate the sequences of market states and returns according to their joint probability distribution (19)-(21). Then we apply the aforementioned policy $\hat{\mathbf{a}}$ on these sequences to get an estimate of the lower bound on the value function H_0; based on each random sequence we can also solve the inner optimization problem (28) with penalty M_1 in (32) or M_2 in (34), which leads to an estimate of the upper bound on H_0. We present our numerical results in the following tables: the lower bound, which is referred to as "Lower Bound", is obtained by generating 100 random sequences of $(\mathbf{Z}, \tilde{\mathbf{Z}})$ and their antithetic pairs (see [43] for an introduction on antithetic variates) in a single run and a total number of 10 runs; the upper bounds induced by penalties M_1 and M_2, which are referred to as "Dual Bound 1" and "Dual Bound 2" respectively, are obtained by generating 30 random sequences of $(\mathbf{Z}, \tilde{\mathbf{Z}})$ and their antithetic pairs in a single run and a total number of 10 runs. To see the effectiveness of these proposed penalties, we use zero penalty and

repeat the same procedure to compute the upper bounds that are referred to as "Zero Penalty" in the table. These bounds on the value function H_0 (i.e., the expected utility) are reported in the sub-column "Value", where each entry shows the sample average and the standard error (in parentheses) of the 10 independent runs. We also list the certainty equivalent of the expected utility in the sub-column "CE" (this is reported in the literature such as [33]), where "CE" is defined through $U(\text{CE}) = \text{Value}$. For ease of comparison, we compute the duality gap – the smaller differences of lower bounds with two upper bounds on the expected utility and its certainty equivalent – as a fraction of the lower bounds in the column "Duality Gap".

TABLE I

RESULTS WITH PARAMETER SET 1

	Lower Bound		Dual Bound 1		Dual Bound 2		Zero Penalty		Duality Gap	
γ	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE
1.5	-5.480	1.332	-5.391	1.376	-5.392	1.376	-4.861	1.693	1.61%	3.30%
	(0.003)	(0.001)	(0.008)	(0.004)	(0.007)	(0.004)	(0.012)	(0.008)		
3.0	-42.887	1.080	-39.227	1.129	-39.873	1.120	-27.562	1.347	7.53%	3.70%
	(0.036)	(0.001)	(0.164)	(0.002)	(0.317)	(0.004)	(0.252)	(0.006)		
5.0	-2445.9	1.005	-2066.5	1.049	-2025.5	1.054	-1105.7	1.226	15.51%	4.38%
	(1.635)	(0.001)	(22.019)	(0.003)	(17.833)	(0.002)	(16.438)	(0.004)		

TABLE II

RESULTS WITH PARAMETER SET 2

	Lower Bound		Dual Bound 1		Dual Bound 2		Zero Penalty		Duality Gap	
γ	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE
1.5	-5.466	1.339	-5.380	1.382	-5.381	1.381	-4.864	1.691	1.56%	3.14%
	(0.005)	(0.001)	(0.011)	(0.006)	(0.015)	(0.008)	(0.020)	(0.008)		
3.0	-42.585	1.084	-39.645	1.123	-39.690	1.122	-27.708	1.343	6.80%	3.51%
	(0.081)	(0.001)	(0.229)	(0.003)	(0.155)	(0.002)	(0.209)	(0.005)		
5.0	-2431.6	1.007	-2043.8	1.052	-2040.7	1.052	-1122.1	1.222	15.95%	4.47%
	(7.510)	(0.001)	(11.881)	(0.002)	(19.882)	(0.003)	(9.842)	(0.004)		

We consider utility functions with different relative risk aversion coefficients $\gamma = 1.5, 3.0$, and 5.0, which reflect low, medium and high degrees of risk aversions. The dual bounds induced by zero penalty perform poorly as we expected. On the other hand, it is hard to distinguish the performance of "Dual Bound 1" and "Dual Bound 2", which may imply that Ψ^3_{k+1} plays an essential role in the inner optimization problem in order to make the dual bounds tight in this problem. We observe that the duality gaps on the value function H_0 are generally smaller when γ is small, implying that both the approximate policy and penalties are near optimal. For example, when $\gamma = 1.5$, the duality gaps are within 2% of the optimal expected utility for all sets of parameters. As γ increases, the duality gaps generally become larger.

TABLE III

RESULTS WITH PARAMETER SET 3

	Lower Bound		Dual Bound 1		Dual Bound 2		Zero Penalty		Duality Gap	
γ	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE
1.5	-5.439	1.352	-5.376	1.384	-5.376	1.384	-4.904	1.663	1.16%	2.37%
	(0.002)	(0.001)	(0.008)	(0.004)	(0.007)	(0.003)	(0.014)	(0.009)		
3.0	-41.961	1.092	-38.221	1.144	-38.800	1.135	-28.5874	1.322	7.53%	3.94%
	(0.074)	(0.001)	(0.243)	(0.003)	(0.241)	(0.003)	(0.168)	(0.004)		
5.0	-2402.6	1.005	-1997.3	1.049	-2011.5	1.056	-1157.3	1.212	16.28%	4.38%
	(7.243)	(0.001)	(16.467)	(0.002)	(15.288)	(0.002)	(14.836)	(0.004)		

TABLE IV

RESULTS WITH PARAMETER SET 4

	Lower Bound		Dual Bound 1		Dual Bound 2		Zero Penalty		Duality Gap	
γ	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE(10^{-1})	Value	CE
1.5	-5.441	1.351	-5.384	1.380	-5.375	1.385	-4.895	1.669	1.05%	2.15%
	(0.001)	(0.001)	(0.004)	(0.002)	(0.008)	(0.004)	(0.008)	(0.005)		
3.0	-41.933	1.092	-38.432	1.141	-38.664	1.137	-28.661	1.321	7.80%	4.12%
	(0.088)	(0.001)	(0.380)	(0.006)	(0.218)	(0.003)	(0.127)	(0.003)		
5.0	-2394.2	1.011	-1984.6	1.059	-2005.9	1.057	-1135.2	1.218	16.22%	4.55%
	(5.130)	(0.001)	(12.951)	(0.002)	(13.758)	(0.002)	(14.353)	(0.004)		

There are several reasons to explain the enlarged duality gaps on the value function with increasing γ. Note that the utility function $U(x)$ is a power function (with negative power of $1-\gamma$) of x and it decreases at a higher rate with larger γ, as x approaches zero. This is reflected by the fact that both the lower and upper bounds on the value function H_0 decrease rapidly with higher value of γ. In the case of evaluating the upper bounds on H_0, it can be inferred that with larger γ the objective value (28a) is more sensitive to the solution of the inner optimization problem (28), and hence the quality of the penalty functions. In other words, even a small torsion of the optimal penalty will lead to a significant deviation of the dual bound. In our case the heuristic penalty is derived by discretizing the value function-based penalty for the continuous-time problem, however, this penalty may become far away from optimal for the discrete-time problem when γ increases. Similarly, obtaining tight lower bounds on the expected utility by simulation under a sub-optimal policy also suffers the same problem, that is, solving a sub-optimal policy based on a same approximation scheme of the recursion (25) may cause more utility loss with larger γ. The performance of the sub-optimal policy also influences the quality of the penalty function, since the penalties M_1 and M_2 involve the wealth $\bar{W}_k$ induced by the suboptimal policy and its error compared with the wealth under the optimal policy will be accumulated over time. Hence, the increasing duality gaps on the value function with larger risk aversion coefficients are contributed by both sub-optimal policies and sub-optimal penalties.

These numerical results provide us with some guidance in terms of computation when we apply the dual approach: we should be more careful with designing the penalty function if the objective value of the inner optimization problem is numerically sensitive either to its optimal solution or to the choice of the penalty function. Fortunately, the sensitivity of the expected utility with respect to γ in this problem is relieved to some extent by considering its certainty equivalent. We can see from the table that the differences between the lower bounds and the upper bounds in terms of "CE" are kept at a relatively constant range for different values of γ.

IV. Conclusion

In this paper we study the dual formulation of controlled Markov diffusions by means of information relaxation. This dual formulation provides new insights into seeking the value function: if we can find an optimal solution to the dual problem, i.e., an optimal penalty, then the value function can be recovered without solving a HJB equation. From a more practical point of view, this dual formulation can be used to find a dual bound on the value function. We explore the structure of the value function-based optimal penalty, which provides the theoretical basis for developing near-optimal penalties that lead to tight dual bounds. As in the case of MDPs, if we compare the dual bound on the value function of a controlled Markov diffusion with the lower bound generated by Monte Carlo simulation under a sub-optimal policy, the duality gap can serve as an indication on how well the sub-optimal policy performs and how much we can improve on our current policy. Furthermore, we also expose the connection of the gradient-based optimal penalty between controlled Markov diffusions and MDPs in Appendix.

We carried out numerical studies in a dynamic portfolio choice problem that is discretized from a continuous-time model. To derive tight dual bounds on the expected utility, we proposed a class of penalties that can be viewed as discretizing the value function-based optimal penalty of the continuous-time problem, and these new penalties make the inner optimization problem computationally tractable. This approach has potential use in many other interesting applications, where the system dynamic is modeled as a controlled Markov diffusion. Moreover, by examining the duality gaps on the expected utility with different parameters, we find that the objective function in the primal problem may largely influence the sensitivity of the optimal solution to the dual problem, and hence the quality of the dual bounds. These numerical studies complement the existing examples on applying the dual approach to continuous-state MDPs.

This dual approach also sheds light on some future directions. First, we attempt to study more practical methods that can apply the dual approach on general (continuous-state) MDPs. For example, a new type of the gradient-based penalty has been presented in [44]. Second, we would like to formulate the dual representation of other continuous-time controlled Markov processes. An analogue of Proposition 2 or Proposition 3 may be established as long as the evolution of the value function under the state dynamics can be explicitly represented; if the value function-based penalty admits simple structure (under natural filtration) as that in the setting of controlled Markov diffusions, it may have the potential to generate dual bounds easily in terms of computation.

References

[1] D. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, 2007.

[2] H. S. Chang, M. C. Fu, J. Hu, and S. I. Marcus, *Simulation-based Algorithms for Markov Decision Processes*, 1st ed., ser. Communications and Control Engineering Series. New York: Springer, 2007.

[3] W. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*, 2nd ed. John Wiley and Sons, 2011.

[4] D. de Farias and B. van Roy, "The linear programming approach to approximate dynamic programming," *Operations Research*, pp. 850–865, 2003.

[5] L. C. G. Rogers, "Pathwise stochastic optimal control," *SIAM J.Control Optimization*, vol. 46, no. 3, pp. 1116 – 1132, 2007.

[6] D. Brown, J. Smith, and P. Sun, "Information relaxations and duality in stochastic dynamic programs," *Operations Research*, vol. 58, no. 4, pp. 758 – 801, 2010.

[7] M. Davis and M. Zervos, "A new proof of the discrete-time LQG optimal control theorems," *Automatic Control, IEEE Transactions on*, vol. 40, no. 8, pp. 1450–1453, 1995.

[8] M. Haugh and A. Lim, "Linear-quadratic control and information relaxations," *Operations Research Letters*, vol. 40, no. 6, pp. 521 – 528, 2012.

[9] D. Brown and J. Smith, "Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds," *Management Science*, vol. 57, no. 10, pp. 1752–1770, 2011.

[10] V. Desai, V. Farias, and C. Moallemi, "Bounds for Markov decision processes," 2011, chapter in Reinforcement Learning and Approximate Dynamic Programming for Feedback Control (F. L. Lewis, D. Liu, eds.).

[11] F. Ye and E. Zhou, "Parameterized penalties in the dual representation of Markov decision processes," in *Proceedings of the 51st IEEE Conference on Decision and Control*, 2012, pp. 870–876.

[12] G. Lai, F. Margot, and N. Secomandi, "An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation," *Operations research*, vol. 58, no. 3, pp. 564–582, 2010.

[13] S. Devalkar, "Essays in optimization of commodity procurement, processing and trade operations," Ph.D. dissertation, The University of Michigan, 2011.

[14] C. Moallemi and M. Saglam, "Dynamic portfolio choice with linear rebalancing rules," *Available at SSRN 2011605*, 2012.

[15] M. Haugh and C. Wang, "Dynamic portfolio execution and martingale duality," *Available at SSRN 2144517*, 2012.

[16] H. Kushner and P. Dupuis, *Numerical methods for stochastic control problems in continuous time*. Springer Verlag, 2001, vol. 24.

[17] J. Han and B. van Roy, "Control of diffusions via linear programming," *Stochastic Programming*, pp. 329–353, 2011.

[18] D. Ocone and E. Pardoux, "A generalized it-ventzell formula. application to a class of anticipating stochastic differential equations," *Annales de l'institut Henri Poincar (B) Probabilits et Statistiques*, vol. 25, no. 1, pp. 39–71, 1989.

[19] D. Nualart, *The Malliavin calculus and related topics*, 2nd ed. Springer-Verlag, 2006.

[20] M. Davis and G. Burstein, "A deterministic approach to stochastic optimal control with application to anticipative control," *Stochastics: An International Journal of Probability and Stochastic Processes*, vol. 40, no. 3-4, pp. 203–256, 1992.

[21] ——, "Anticipative stochastic control," in *Proceedings of the 30th IEEE Conference on Decision and Control*, 1991, pp. 1830–1835.

[22] M. Davis, "Anticipative LQG control," *IMA Journal of Mathematical Control and Information*, vol. 6, no. 3, pp. 259–265, 1989.

[23] L. C. G. Rogers, "Monte Carlo valuation of American options," *Mathematical Finance*, vol. 12, no. 3, pp. 271 – 286, 2002.

[24] M. Haugh and L. Kogan, "Pricing American options: A duality approach," *Operations Research*, vol. 52, no. 2, pp. 258 – 270, 2004.

[25] L. Andersen and M. Broadie, "Primal-dual simulation algorithm for pricing multidimensional American options," *Management Science*, vol. 50, no. 9, pp. 1222 – 1234, 2004.

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

[27] P. Samuelson, "Lifetime portfolio selection by dynamic stochastic programming," *The Review of Economics and Statistics*, vol. 51, no. 3, pp. 239–246, 1969.

[28] R. Merton, "Lifetime portfolio selection under uncertainty: The continuous-time case," *The review of Economics and Statistics*, vol. 51, no. 3, pp. 247–257, 1969.

[29] ——, "Optimum consumption and portfolio rules in a continuous-time model," *Journal of Economic Theory*, vol. 3, no. 4, pp. 373–413, 1971.

[30] N. Garleanu and L. Pedersen, "Dynamic trading with predictable returns and transaction costs," National Bureau of Economic Research, Tech. Rep., 2009.

[31] S. Boyd, M. Mueller, B. O'Donoghue, and Y. Wang, "Performance bounds and suboptimal policies for multi-period investment," 2012.

[32] P. Glasserman and X. Xu, "Robust portfolio control with stochastic factor dynamics," *Available at SSRN 1960723*, 2011.

[33] J. Cvitanic, L. Goukasian, and F. Zapatero, "Monte Carlo computation of optimal portfolios in complete markets," *Journal of Economic Dynamics and Control*, vol. 27, no. 6, pp. 971–986, 2003.

[34] H. He and N. Pearson, "Consumption and portfolio policies with incomplete markets and short-sale constraints: The infinite dimensional case," *Journal of Economic Theory*, vol. 54, no. 2, pp. 259–304, 1991.

[35] G. Tauchen and R. Hussey, "Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models," *Econometrica: Journal of the Econometric Society*, pp. 371–396, 1991.

[36] P. Balduzzi and A. Lynch, "Transaction costs and predictability: Some utility cost calculations," *Journal of Financial Economics*, vol. 52, no. 1, pp. 47–78, 1999.

[37] M. Brandt, A. Goyal, P. Santa-Clara, and J. Stroud, "A simulation approach to dynamic portfolio choice with an application to learning about return predictability," *Review of Financial Studies*, vol. 18, no. 3, pp. 831–873, 2005.

[38] M. Haugh, L. Kogan, and J. Wang, "Evaluating portfolio policies: A duality approach," *Operations Research*, vol. 54, no. 3, pp. 405–418, 2006.

[39] J. Cvitanic and I. Karatzas, "Convex duality in constrained portfolio optimization," *The Annals of Applied Probability*, vol. 2, no. 4, pp. 767–818, 1992.

[40] J. Liu, "Portfolio selection in stochastic environments," *Review of Financial Studies*, vol. 20, no. 1, pp. 1–39, 2007.

[41] K. Judd, *Numerical Methods in Economics*. The MIT press, 1998.

[42] M. Grant and S. Boyd, "Cvx: Matlab software for disciplined convex programming, version 2.0 (beta)," *http://cvxr.com/cvx/*, 2013.

[43] P. Glasserman, *Monte Carlo Methods in Financial Engineering*. Springer, 2004.

[44] D. Brown and J. Smith, "Subgradient bounds for convex dynamic programs," 2013, working paper.

APPENDIX A
TECHNICAL ASSUMPTIONS ON CONTROLLED MARKOV DIFFUSIONS

In integral form, (6) can be written as

$$x_t = x_0 + \int_0^t b(s,x_s,u_s)ds + \int_0^t \sigma(s,x_s,u_s)dw_s, \quad 0 \le t \le T. \tag{35}$$

The second term on the right side of (35) is an Ito stochastic integral. We say that a process $(x_t)_{0\le t\le T}$ is a pathwise solution of (6) or (35) if it is $\mathbb{F}$-progressively measurable and has continuous sample paths almost surely (a.s.) given $x_0 = x \in \mathbb{R}^n$.

To guarantee a unique pathwise solution $(x_t)_{t\in[0,T]}$ to (6) or (35) when $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$, we impose the following standard assumptions:

Assumption 1 *b and σ are continuous on $\bar{Q}\times\mathscr{U}$, and for some constants $C_1, C_2 > 0$,*

1) $\| b(t,x,u) \| + \| \sigma(t,x,u) \| \le C_1(1+\| x \| + \| u \|)$ *for all* $(t,x,u) \in \bar{Q}\times\mathscr{U}$;
2) $\| b(t,x,u) - b(s,y,u) \| + \| \sigma(t,x,u) - \sigma(s,y,u) \| \le C_2(|t-s| + \| x-y \|)$ *for all* $(t,x),(s,y) \in \bar{Q}$.

To guarantee Theorem 2 to hold, we also assume that Λ and g satisfy the following polynomial growth conditions:

Assumption 2 *For some constants $C_\Lambda, c_\Lambda, C_g, c_g > 0$,*

1) $|\Lambda(x)| \le C_\Lambda(1+\| x \|^{c_\Lambda})$ *for all* $x \in \mathbb{R}^n$;
2) $|g(s,x,u)| \le C_g(1+\| x \|^{c_g} + \| u \|^{c_g})$ *for all* $(t,x) \in \bar{Q}$.

APPENDIX B
PROOF OF THEOREM 3

We note that the left side of (12) is the definition of $V(0,x)$. By weak duality result, the left side of (12) is less than or equal to the right side. We only need to show that with $h = h^*$ in (11), the right side of (12) is equal to the left side. This is done by the argument before the statement of Theorem 3.

APPENDIX C
PROOF OF THEOREM 4

We first consider sufficiency. Let $\mathbf{u}^\star \in \mathscr{U}_{\mathbb{F}}(0)$ and $h^\star \in \mathscr{M}_{\mathbb{F}}(0)$. We assume $\mathbb{E}_{0,x}[h^\star(\mathbf{u}^\star, \mathbf{w})] = 0$ and (13) holds. Then by weak duality, $\mathbf{u}^\star$ and $h^\star$ should be optimal to the primal and dual problem, respectively.

Next we consider necessity. Let $\mathbf{u}^\star \in \mathscr{U}_{\mathbb{F}}(0)$ and $h^\star \in \mathscr{M}_{\mathbb{F}}(0)$. Then we have

$$\begin{aligned}
&\mathbb{E}_{0,x}\big[\sup_{\mathbf{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h^\star(\mathbf{u},\mathbf{w})\}\big] \\
\geq &\mathbb{E}_{0,x}[\Lambda(x_T^\star) + \int_t^T g(t,x_t^\star,u_t^\star)dt - h^\star(\mathbf{u}^\star,\mathbf{w})] \\
\geq &J(0,x;\mathbf{u}^\star).
\end{aligned}$$

The last inequality holds due to $h^\star \in \mathscr{M}_{\mathbb{F}}(0)$. Since we know $\mathbf{u}^\star$ and $h^\star$ are optimal to the primal and dual problem respectively, by strong duality result, we have

$$J(0,x;\mathbf{u}^\star) = \mathbb{E}_{0,x}\big[\sup_{\mathbf{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h^\star(\mathbf{u},\mathbf{w})\}\big],$$

which implies all the inequalities above are equalities. Therefore, we know $\mathbb{E}_{0,x}[h^\star(\mathbf{u}^\star,\mathbf{w})] = 0$ and (13) holds.

APPENDIX D
COMPLEMENT OF SECTION 2.1

In this section we aim to develop the value function-based penalty as a solution to the dual problem on the right side of (12), which can be viewed as the counterpart of (5) in the setting of controlled Markov diffusions. For this purpose we introduce the anticipating stochastic calculus and anticipating stochastic differential equation in Appendix D-A, and present the value function-based optimal penalty in Appendix D-B. In Appendix D-C, we will review a Lagrangian approach proposed by [20], which falls in our dual framework of controlled Markov diffusions; the Lagrangian term they proposed satisfies the complementary slackness condition developed in Theorem 4, and hence it behaves like a "penalty" function. Moreover, we compare the procedure of deriving this Lagrangian term with that of the gradient-based penalty proposed by [9] for MDPs, in order to expose their similarities.

Throughout this section we assume that σ in (6) only depends on t and x. To be specific,

$$x_t = x + \int_0^t b(s,x_s,u_s)\,ds + \int_0^t \sigma(s,x_s)\,dw_s, \quad 0 \leq t \leq T, \tag{36}$$

where $b : [0,T] \times \mathbb{R}^n \times \mathscr{U} \to \mathbb{R}^n$ and $\sigma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$. Besides the technical assumption in Appendix A, we further assume the gradient of $\sigma(t,x)$ with respect to x exists and it is continuous and bounded on $\bar{Q}$. The reason

to suppress the dependence of σ on u is to extend the definition of a solution to (36) with anticipative controls. When the stochastic integral is defined in Ito sense, SDE (36) has a well-defined solution provided that the control strategy $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$; however, if an anticipative control strategy $\mathbf{u} \in \mathscr{U}(0)$ is considered, we need to first define a stochastic integral with respect to an anticipative process and then define a solution to an anticipating stochastic differential equation (see, e.g., [19], [18], [20]). This solution is an extension of that to the (regular) SDE in the Ito sense, i.e., it should coincide with the solution to the SDE in Ito sense when $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$. We follow [20] to define such a solution using the decomposition technique, and for this purpose we require that σ is a function of only t and x.

The reward functional to be maximized and the value function are the same as in (7):

$$V(t,x) = \sup_{\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(t)} J(t,x;\mathbf{u}), \tag{37}$$

where

$$J(t,x;\mathbf{u}) \triangleq \mathbb{E}_{t,x}[\Lambda(x_T) + \int_t^T g(s,x_s,u_s)\,ds].$$

The partial differential operator A^u is then redefined as

$$A^u L(t,x) \triangleq L_t(t,x) + L_x(t,x)^\top b(t,x,u) + \frac{1}{2}\mathrm{tr}(L_{xx}(t,x)(\sigma\sigma^\top)(t,x)),\ L \in C^{1,2}(Q).$$

A. Anticipating Stochastic Calculus and Anticipating Stochastic Differential Equation

There are several ways to integrate stochastic processes that are not adapted to Brownian motions such as Skorohod and (generalized) Stratonovich integrals (see, e.g, [19], [18]). In this subsection we present the Stratonovich integral and its associated Ito formula. Then we define the solution to the anticipating stochastic differential equation in the Stratonovich sense.

We first assume that $\mathbf{w} = (w_t)_{t\in[0,T]}$ is a one-dimensional Brownian Motion in the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We denote by I an arbitrary partition of the interval $[0,T]$ of the form $I = \{0 = t_0 < t_1 < \cdots < t_n = T\}$ and define $|I| = \sup_{0\le i\le n-1}(t_{i+1} - t_i)$.

Definition 2 *(Definition 3.1.1 in [19]) We say that a measurable process $\mathbf{y} = (y_t)_{t\in[0,T]}$ such that $\int_0^T |y_t|dt < \infty$ a.s. is Stratonovich integrable if the family*

$$S^I = \int_0^T y_t \sum_{i=0}^{n-1} \frac{w_{t_i+1} - w_{t_i}}{t_{i+1} - t_i} 1_{(t_i,t_{i+1}]}(t)\,dt$$

converges in probability as $|I| \to 0$, and in this case the limit will be denoted by $\int_0^T y_t \circ dw_t$.

Remark 2 *We can translate an Ito integral to a Stratonovich integral and vice versa. If $\mathbf{y} = (y_t)_{t\in[0,T]}$ is a continuous semimartingale of the form*

$$y_t = y_0 + \int_0^t \upsilon_s\,ds + \int_0^t \zeta_s\,dw_s,$$

where $(\upsilon_t)_{t\in[0,T]}$ *and* $(\zeta_t)_{t\in[0,T]}$ *are adapted processes taking value in* $\mathbb{R}^n$ *and* $\mathbb{R}^{n\times m}$ *such that* $\int_0^T \parallel \upsilon_s \parallel ds < \infty$ *and* $\int_0^T \parallel \zeta_s \parallel^2 ds < \infty$ *a.s.. Then* $\mathbf{y}$ *is Stratonovich integrable on any interval* $[0,t]$*, and*

$$\int_0^t y_s \circ dw_s = \int_0^t y_s dw_s + \langle y, w\rangle_t = \int_0^t y_s dw_s + \frac{1}{2}\int_0^t \zeta_s ds, \tag{38}$$

where $\langle y,w\rangle_t$ *denotes the joint quadrature variation of the semimartingale* $\mathbf{y}$ *and the Brownian motion* $\mathbf{w}$*. Definition 2 and the equality (38) can be naturally extended to the vector case.*

Then we present the Ito formula for Stratonovich integral in Proposition 5, the detail of which can be found in Section 3.2.3 of [19].

Proposition 5 (Theorem 3.2.6 in [19]) *Let* $\mathbf{w} = (w_t^1, \cdots, w_t^m)_{t\in[0,T]}$ *be an m-dimensional Brownian motion. Suppose that* $y_0 \in \mathbb{D}^{1,2}$*,* $\upsilon_s \in \mathbb{L}^{1,2}$*, and* $\zeta^i \in \mathbb{L}_S^{2,4}$*,* $i = 1, \cdots, m$*. Consider a process* $\mathbf{y} = (y_t)_{t\in[0,T]}$ *of the form*

$$y_t = y_0 + \int_0^t \upsilon_s ds + \sum_{i=1}^m \int_0^t \zeta_s^i \circ dw_s^i,\ 0 \le t \le T.$$

Assume that $(y_t)_{0\le t\le T}$ *has continuous paths. Let* $F : \mathbb{R}^n \to \mathbb{R}$ *be a twice continuously differentiable function. Then we have*

$$F(y_t) = F(y_0) + \int_0^t F_y(y_s)^\top \upsilon_s ds + \sum_{i=1}^m \int_0^t [F_y(y_s)^\top \zeta_s^i] \circ dw_s^i,\ 0 \le t \le T, \tag{39}$$

where $F_y(\cdot)$ *denotes the gradient of* F *w.r.t.* y*.*

Proposition 5 basically says that the Stratonovich integral obeys the ordinary chain rule.

Based on the definition of Stratonovich integral and Remark 2, we generalize SDE (36) to the Stratonovich sense (referred to as S-SDE) by letting $y_t = \sigma^i(t,x_t)$. Then (36) is equivalent to

$$x_t = x + \int_0^t \bar{b}(t,x_t,u_t)dt + \sum_{i=1}^m \int_0^t \sigma^i(t,x_t) \circ dw_t^i,\quad 0 \le t \le T, \tag{40}$$

where $\sigma^i : [0,T]\times\mathbb{R}^n \to \mathbb{R}^n$ is the i-th column of σ, $i = 1, \cdots, m$, and $\bar{b}(t,x,u) = b(t,x,u) - \frac{1}{2}\sum_{i=1}^m \sigma_x^i \sigma^i(t,x)$. Here $\sigma_x^i\sigma^i(t,x)$ denotes an $n\times 1$ vector with $\sum_{j=1}^n \frac{\partial \sigma^{ki}}{\partial x_j}(t,x)\sigma^{ji}(t,x)$ being its k-th entry and $\sigma^{ki}(\cdot)$ is the k-th component of $\sigma^i(\cdot)$. Since the stochastic integral in (40) is in the Stratonovich sense, S-SDE (40) adopts its solution in the space of measurable processes, which may not be adapted to the filtration generated by the Brownian motion. Therefore, we are allowed to consider anticipative policies $\mathbf{u} \in \mathscr{U}(0)$ in (40).

Finally, we need to ensure the existence of a solution to S-SDE (40) if the control strategy $\mathbf{u} \in \mathscr{U}(0)$ is anticipative. Following [20],[18], we have a representation of such a solution using the decomposition technique:

$$x_t = \xi_t(\eta_t), \tag{41}$$

where $\{\xi_t(x)\}_{t\in[0,T]}$ denotes the stochastic flow defined by the adapted equation:

$$\begin{aligned} d\xi_t &= \sum_{i=1}^m \sigma^i(t,\xi_t) \circ dw_t^i, \\ &= \frac{1}{2}\sum_{i=1}^m \sigma_x^i\sigma^i(t,\xi_t)dt + \sigma(t,\xi_t)dw_t,\quad \xi_0 = x, \end{aligned} \tag{42}$$

and $(\eta_t)_{t\in[0,T]}$ solves an ordinary differential equation:

$$\frac{d\eta_t}{dt} = (\frac{\partial \xi_t}{\partial x})^{-1}(\eta_t)\bar{b}(t,\xi_t(\eta_t),u_t), \quad \eta_0 = x, \tag{43}$$

where $\frac{\partial \xi_t}{\partial x}$ denotes the $n\times n$ Jacobian matrix of ξ_t with respect to x. Under some technical conditions (see Section 1 of [20]), the solution (41) is defined almost surely: note that ξ_t does not depend on the control u_t, i.e., it is the solution to a regular SDE in the Ito sense; η_t is not defined by a stochastic integral so it is the solution to an ordinary differential equation parameterized by $\mathbf{w}$ (note that $\frac{\partial \xi_t}{\partial x}$ is well-defined a.s. for $(t,x)\in[0,T]\times\mathbb{R}^n$, because $\xi_t(x)$ is flow of diffeomorphisms a.s..). Hence, $x_t = \xi_t(\eta_t)$ is well-defined regardless of the adaptiveness of $\mathbf{u} = (u_t)_{0\le t\le T}$. To check that $x_t = \xi_t(\eta_t)$ satisfies (36), we need to employ a generalized Ito formula of (39) for Stratonovich integral (see Theorem 4.1 in [18]).

B. Value Function-Based Penalty

The tools we have introduced in the last subsection, especially the Ito formula for Stratonovich integral, enable us to develop the value function-based optimal penalty for the controlled Markov diffusions. This penalty, denoted by $h_v^*(\mathbf{u},\mathbf{w})$, coincides with $\int_0^T V_x(t,x_t)^\top \sigma(t,x_t)\,dw_t$, when $\mathbf{u}\in\mathscr{U}_{\mathbb{F}}(0)$.

Theorem 5 (Value Function-Based Penalty) *Suppose the value function $V(t,x)$ for the problem(36)(37) (or (40)(37)) satisfies all the assumptions in Theorem 2(b). We also assume that the Ito formula for Stratonovich integral (39) is valid with $F = V(t,x)$ and $\mathbf{y} = (t,x_t)_{t\in[0,T]}$, where $(x_t)_{t\in[0,T]}$ is the solution to (40). Define*

$$\begin{aligned} h_v^*(\boldsymbol{u},\boldsymbol{w}) \triangleq & \sum_{i=1}^m \int_0^T [V_x(t,x_t)^\top \sigma^i(t,x_t)] \circ dw_t^i \\ & - \frac{1}{2}\int_0^T \big[V_x(t,x_t)^\top \big(\sum_{i=1}^m \sigma_x^i \sigma^i(t,x_t)\big) + tr\big(V_{xx}(t,x_t)(\sigma\sigma^\top)(t,x_t)\big)\big]\,dt. \end{aligned} \tag{44}$$

Then

1) *If $\boldsymbol{u}\in\mathscr{U}_{\mathbb{F}}(0)$, (44) reduces to the following form*

$$h_v^*(\boldsymbol{u},\boldsymbol{w}) = \int_0^T V_x(t,x_t)^\top \sigma(t,x_t)\,dw_t,$$

and $h_v^*(\boldsymbol{u},\boldsymbol{w})\in\mathscr{M}_{\mathbb{F}}(0)$.

2) *The strong duality holds in*

$$V(0,x) = \mathbb{E}_{0,x}\big[\sup_{\boldsymbol{u}\in\mathscr{U}(0)}\{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h_v^*(\boldsymbol{u},\boldsymbol{w})\}\big]. \tag{45}$$

Moreover, the following equalities hold almost surely with $x_0 = x$

$$V(0,x) = \sup_{\boldsymbol{u}\in\mathscr{U}(0)}\{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h_v^*(\boldsymbol{u},\boldsymbol{w})\} \tag{46}$$

$$= \Lambda(x_T^*) + \int_0^T g(t,x_t^*,u_t^*)dt - h_v^*(\boldsymbol{u}^*,\boldsymbol{w}), \tag{47}$$

where $(x_t^*)_{t\in[0,T]}$ is the solution of (36) using the optimal control $\boldsymbol{u}^* = (u_t^*)_{t\in[0,T]}$ (defined in Theorem 2(b)) on $[0,t)$ with the initial condition $x_0^* = x$.

Proof: Proof. Suppose $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$ and let $y_t = V_x(t,x_t)^\top \sigma^i(t,x_t)$ in Remark 2 for $i = 1, \cdots, m$. We can immediately obtain

$$h_v^*(\mathbf{u},\mathbf{w}) = \sum_{i=1}^m \int_0^T V_x(t,x_t)^\top \sigma^i(t,x_t)\, dw_t^i = \int_0^T V_x(t,x_t)^\top \sigma(t,x_t)\, dw_t.$$

Note that V_x and σ both satisfy a polynomial growth, since $V(t,x) \in C^{1,2}(Q) \cap C_p(\bar{Q})$ (also see Appendix A). Then we have

$$\mathbb{E}_{0,x}\big[\, \| \int_0^T V_x(t,x_t)^\top \sigma(t,x_t) \, \|^2 \, dt\big] < \infty,$$

and therefore, $\mathbb{E}_{0,x}[h_v^*(\mathbf{u},\mathbf{w})] = 0$ when $\mathbf{u} \in \mathscr{U}_{\mathbb{F}}(0)$. Hence, $h_v^*(\mathbf{u},\mathbf{w}) \in \mathscr{M}_{\mathbb{F}}(0)$.

Then we show the strong duality (45). According to the weak duality (i.e., Proposition 1),

$$V(0,x) \le \mathbb{E}_{0,x}[\sup_{\mathbf{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h_v^*(\mathbf{u},\mathbf{w})\}]. \tag{48}$$

Next we prove the reverse inequality. Note that with $x_0 = x$,

$$\begin{aligned}
&\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h_v^*(\mathbf{u},\mathbf{w}) \\
=&V(0,x) + \int_0^T [V_t(t,x_t) + V_x(t,x_t)^\top \bar{b}(t,x_t,u_t)]dt \\
&+ \sum_{i=1}^m \int_0^T [V_x(t,x_t)^\top \sigma^i(t,x_t)] \circ dw_t^i - h_v^*(\mathbf{u},\mathbf{w}) \\
=&V(0,x) + \int_0^T [g(t,x_t,u_t) + A^{u_t} V(t,x_t)]dt,
\end{aligned}$$

where the first equality is obtained by applying Ito formula for Stratonovich integral (i.e., Proposition 5) on $V(t,x)$ with $V(T,x_T) = \Lambda(x_T)$:

$$\begin{aligned}
V(T,x_T) =&V(0,x_0) + \int_0^T [V_t(t,x_t) + V_x(t,x_t)^\top \bar{b}(t,x_t,u_t)]dt \\
&+ \sum_{i=1}^m \int_0^T [V_x(t,x_t)^\top \sigma^i(t,x_t)] \circ dw_t^i.
\end{aligned}$$

Since we assume the value function satisfies all the assumptions in Theorem 2(b), there exists an optimal control $\mathbf{u}^* = (u_t^*)_{t\in[0,T]}$ with $u_t^* = u^*(t,x_t)$ and it satisfies

$$g(t,x,u^*(t,x)) + A^{u^*(t,x)} V(t,x) = \max_{\mathbf{u}\in\mathscr{U}} \{g(t,x,u) + A^u V(t,x)\} = 0,$$

then we have

$$\begin{aligned}
&\sup_{\mathbf{u}\in\mathscr{U}(0)} \{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h_v^*(\mathbf{u},\mathbf{w})\} \\
= &\sup_{\mathbf{u}\in\mathscr{U}(0)} \{V(0,x) + \int_0^T [g(t,x_t,u_t) + A^{u_t} V(t,x_t)]dt\} \\
\le &V(0,x) + \int_0^T \sup_{u\in\mathscr{U}} \{g(t,x_t,u) + A^u V(t,x_t)\}dt \qquad (49) \\
= &V(0,x) + \int_0^T \big[g(t,x_t^*,u_t^*) + A^{u_t^*} V(t,x_t^*)\big]dt \\
= &V(0,x). \qquad (50)
\end{aligned}$$

Taking the conditional expectation on both sides, we have

$$V(0,x) \geq \mathbb{E}_{0,x}\big[\sup_{\mathbf{u}\in\mathscr{U}(0)}\{\Lambda(x_T)+\int_0^T g(t,x_t,u_t)dt - h_v^*(\mathbf{u},\mathbf{w})\}\big].$$

Together with the weak duality (48) , we reach the equality (45).

Due to the fact of the equality (45)(in expectation sense) and the pathwise inequality (50), we find that the only inequality (49) (that makes (50) an inequality) should be an equality in almost sure sense. So the equality (46) holds immediately in almost sure sense. To achieve the equality in (49), the optimal control $\mathbf{u}^*$ should be applied, which implies the equality (47). ■

By imposing the value function-based optimal penalty the objective value of the dual problem is equal to $V(0,x)$ not only in the expectation sense, but also in the almost sure sense. Therefore, we can view the dual approach as a variance reduction technique. In particular, h_v^* plays the role of control variates. As another obvious fact, $h_v^*(\mathbf{u},\mathbf{w})$ evaluated at $\mathbf{u}=\mathbf{u}^*$ is equal to $h^*(\mathbf{u}^*,\mathbf{w})$ in Proposition 2.

C. Gradient-Based Penalty

In this subsection we review the results in [20], where a Lagrangian term is proposed to penalize the relaxation of the requirement on non-anticipative control strategies. [20] characterizes the properties of this Lagrangian term, which coincides with the complementary slackness condition developed in Theorem 4 if it is regarded as a penalty function. We will show the "gradient-based" flavor of this Lagrangian term by comparing it with the gradient-based penalty proposed in [9] for MDPs.

For simplicity we present the results of [20] in the case that the control set $\mathscr{U}$ is convex in $\mathbb{R}^{u_d}$ and the intermediate reward $g(t,x_t,u_t)=0$ for $t\in[0,T]$. The following Lagrangian term h_g^* (the subscript g refers to "gradient-based") is used to penalize the relaxation of non-anticipative constraints on the control strategies:

$$h_g^*(\mathbf{u},\mathbf{w}) \triangleq \int_0^T \lambda(t,x_t,\mathbf{w})^\top u_t dt.$$

Then we consider the inner optimization problem with h_g^* (parameterized by $\mathbf{w}$):

$$V_g(t,x,\mathbf{w}) = \sup_{\mathbf{u}\in\mathscr{U}(t)}\{\Lambda(x_T) - \int_t^T \lambda(s,x_s,\mathbf{w})^\top u_s ds\},\ x_t = x. \tag{51}$$

Since $x_t=\xi_t(\eta_t)$ and only η_t depends on u_t, we obtain an equivalent problem

$$\Theta(t,\eta,\mathbf{w}) = \sup_{\mathbf{u}\in\mathscr{U}(t)}\{\Lambda\circ\xi_T(\eta_T) - \int_t^T \tilde{\lambda}(s,\eta_s,\mathbf{w})^\top u_s ds\},\ \eta_t=\eta=\xi_t^{-1}(x), \tag{52}$$

where $\tilde{\lambda}(s,\eta_s,\mathbf{w})=\lambda(s,\xi_s(\eta_s),\mathbf{w})$.

Suppose that $\mathbf{u}^*=(u^*(t,x_t))_{0\leq t\leq T}$ is an optimal control to the problem (36)(37)(or (40)(37)). We will present one main result of [20] in Theorem 6 that characterizes $\lambda(t,x_t,\mathbf{w})$ such that $\mathbf{u}^*$ is also optimal to the problem (40)(51) a.s. (in pathwise sense). For the sake of defining a proper $\lambda(t,x_t,\mathbf{w})$, [20] first introduced $\varphi_t(\eta)$, which is the flow of

$$\frac{d\varphi_t}{dt} = \big(\frac{\partial\xi_t}{\partial x}\big)^{-1}(\varphi_t)\big[\bar{b}\big(t,\xi_t(\varphi_t),u^*(t,\xi_t(\varphi_t))\big) - \nabla_u b\big(t,\xi_t(\varphi_t),u^*(t,\xi_t(\varphi_t))\big)u^*(t,\xi_t(\varphi_t))\big] \tag{53}$$

for $t \in [0,T]$ with the terminal condition $\varphi_T = \eta$, where $\nabla_u b$ denotes the $n \times d_u$ Jacobian matrix of b with respect to u. We use φ_t^{-1} to denote the inverse flow of φ_t.

Theorem 6 (Gradient-Based Penalty, Theorem 2.1 in [20]) *Consider the deterministic optimal control problem (43)(52) with the terminal reward $\Theta(T,\eta,\boldsymbol{w}) = (\Lambda \circ \xi_T)(\eta)$ parameterized by $\boldsymbol{w}$, where $\xi_t(\eta)$ is the solution to (42) with $\xi_0 = \eta$. If we define*

$$\begin{aligned} \tilde{\lambda}(t,\eta,\boldsymbol{w}) &\triangleq \frac{\partial[\Lambda \circ \xi_T(\varphi_t^{-1}(\eta))]}{\partial \eta}\big(\frac{\partial \xi_t}{\partial x}\big)^{-1}(\eta)\nabla_u b\big(t,\xi_t(\eta),u^*(t,\xi_t(\eta))\big); \\ \lambda(t,x,\boldsymbol{w}) &\triangleq \tilde{\lambda}(t,\xi_t^{-1}(x),\boldsymbol{w}), \end{aligned} \tag{54}$$

then $u^*(t,\xi_t(\eta))$ is an optimal control for the problem (43)(52) a.s., and hence $u^*(t,x)$ is optimal for the problem (40)(51). We also have

$$\begin{aligned} \Theta(t,\eta,\boldsymbol{w}) &= \Lambda \circ \xi_T(\varphi_t^{-1}(\eta)); \\ V_g(t,x,\boldsymbol{w}) &= \Theta(t,\xi_t^{-1}(x),\boldsymbol{w}); \\ \mathbb{E}_{t,x} V_g(t,x,\boldsymbol{w}) &= V(t,x). \end{aligned}$$

Remark 3

1) *$V_g(t,x,\boldsymbol{w})$ may NOT be equal to $V(t,x)$ almost surely in pathwise sense.*
2) *Based on Theorem 6, we have*

$$\begin{aligned} V(0,x) &= \mathbb{E}_{0,x}[V_g(0,x,\boldsymbol{w})] \\ &= \mathbb{E}_{0,x}[\sup_{\boldsymbol{u}\in\mathscr{U}(0)}\{\Lambda(x_T) - h_g^*(\boldsymbol{u},\boldsymbol{w})\}] \\ &= \mathbb{E}_{0,x}[\Lambda(x_T^*) - h_g^*(\boldsymbol{u}^*,\boldsymbol{w})], \end{aligned}$$

which implies $\mathbb{E}_{0,x}[h_g^*(\boldsymbol{u}^*,\boldsymbol{w})] = 0$. It can be seen that $(\boldsymbol{u}^*, h_g^*)$ satisfies the complementary slackness condition developed in Theorem 4. Therefore, $h_g^*(\boldsymbol{u},\boldsymbol{w})$ behaves exactly the same as an optimal penalty, though we have not shown that $h_g^*(\boldsymbol{u},\boldsymbol{w}) \in \mathscr{M}_{\mathbb{F}}(0)$.

A complete proof of Theorem 6 is in [20]. Here we provide some insight on the design of the Lagrangian multiplier $\tilde{\lambda}(t,\eta,\mathbf{w})$ using the verification argument. If the value function $\Theta(t,\eta,\mathbf{w})$ for the problem (43)(52) is of sufficient regularity, it should satisfy the following HJB equation

$$\frac{\partial \Theta}{\partial t}(t,\eta,\mathbf{w}) + \sup_{u\in\mathscr{U}}\{\frac{\partial \Theta}{\partial \eta}(t,\eta,\mathbf{w})\big(\frac{\partial \xi_t}{\partial x}\big)^{-1}(\eta)\bar{b}(t,\xi_t(\eta),u) - \tilde{\lambda}(t,\eta,\mathbf{w})^\top u\} = 0$$

with the terminal condition $\Theta(T,\eta,\mathbf{w}) = \Lambda \circ \xi_T(\eta)$. If we define

$$\tilde{\lambda}(t,\eta,\mathbf{w}) = \frac{\partial \Theta}{\partial \eta}(t,\eta,\mathbf{w})\big(\frac{\partial \xi_t}{\partial x}\big)^{-1}(\eta)\nabla_u b\big(t,\xi_t(\eta),u^*(t,\xi_t(\eta))\big)$$

as in (54), then the HJB equation becomes

$$\frac{\partial \Theta}{\partial t}(t,\eta,\mathbf{w}) + \sup_{u\in\mathscr{U}}\{\frac{\partial \Theta}{\partial \eta}(t,\eta,\mathbf{w})\big(\frac{\partial \xi_t}{\partial x}\big)^{-1}(\eta)[\bar{b}\big(t,\xi_t(\eta),u\big) - \nabla_u b\big(t,\xi_t(\eta),u^*(t,\xi_t(\eta))\big)u]\} = 0.$$

We define the Hamiltonian $\mathscr{H}(t,\eta,u,\mathbf{w})$ with $\frac{\partial \Theta}{\partial \eta}$ playing the role of the costate:

$$\mathscr{H}(t,\eta,u,\mathbf{w}) = \frac{\partial \Theta}{\partial \eta}(t,\eta,\mathbf{w})\big(\frac{\partial \xi_t}{\partial x}\big)^{-1}(\eta)\big[\bar{b}\big(t,\xi_t(\eta),u\big) - \nabla_u b\big(t,\xi_t(\eta),u^*(t,\xi_t(\eta))\big)u\big].$$

Because $\mathscr{H}(t,\eta,\cdot,\mathbf{w})$ is strictly concave on $\mathscr{U}$ (ensured by some technical conditions) and its gradient $\nabla_u H(t,\eta,\cdot,\mathbf{w})$ is

$$\nabla_u \mathscr{H}(t,\eta,u,\mathbf{w}) = \frac{\partial \Theta}{\partial \eta}(t,\eta,\mathbf{w})\big(\frac{\partial \xi_t}{\partial x}\big)^{-1}(\eta)\big[\nabla_u b\big(t,\xi_t(\eta),u\big) - \nabla_u b\big(t,\xi_t(\eta),u^*(t,\xi_t(\eta))\big)\big]. \tag{55}$$

It can be seen by the first-order condition that

$$\min_{u\in\mathscr{U}} \mathscr{H}(t,\eta,u,\mathbf{w}) = \mathscr{H}(t,\eta,u^*(t,\xi_t(\eta)),\mathbf{w}) \quad a.s..$$

Hence we have shown why $u^*(t,\xi_t(\eta))$ is optimal for the problem (43)(52). [20] used the characteristics method to ensure the existence of a sufficient regular function $\Theta(t,\eta,\mathbf{w})$ as a unique solution to the HJB equation, and a lengthy approximation argument is spent on passing the results in terms of x via the transformation $x_t = \xi_t(\eta_t)$.

The derivation of $h_g^*(\mathbf{u},\mathbf{w})$ for the continuous-time optimal control problem is based on minimizing the Hamiltonian due to the convexity assumption on the control set $\mathscr{U}$ and the first order condition, which is analogous to that of the gradient-based penalty proposed in [9] for MDPs. The latter construction of the optimal penalty is more straightforward, as it only requires some basic knowledge in convex optimization. We briefly review their results in the setting of the introduction section. For simplicity we also assume $g_k(x_k,a_k) = 0$ and $\mathscr{A}$ is convex for $k = 0,\cdots,K-1$, and assume $\Lambda(x_K(\mathbf{a},\mathbf{v}))$ is differentiable and concave in the control strategy sequence $\mathbf{a}$ for every sequence $\mathbf{v}$. Consider the gradient-based penalty of the form

$$M_g^*(\mathbf{a},\mathbf{v}) = \nabla_{\mathbf{a}}\Lambda(x_K(\mathbf{a}^*,\mathbf{v}))^\top(\mathbf{a}-\mathbf{a}^*), \tag{56}$$

where $\mathbf{a}^*$ is the optimal control, and $\nabla_{\mathbf{a}}\Lambda(x_K(\mathbf{a},\mathbf{v}))$ is the gradient of the terminal reward with respect to a feasible strategy $\mathbf{a}\in\mathbb{A}$; the first-order condition for optimizing (2) over control strategy sequences $\mathbf{a}\in\mathbb{A}_{\mathbb{G}}$ implies

$$\mathbb{E}_{0,x}[M_g^*(\mathbf{a},\mathbf{v})] \le 0, \quad \text{for all } \mathbf{a}\in\mathbb{A}_{\mathbb{G}}.$$

With the gradient-based penalty $M_g^*(\mathbf{a},\mathbf{v})$ the inner optimization problem (4) becomes

$$\sup_{\mathbf{a}\in\mathbb{A}}\{\Lambda(x_K(\mathbf{a},\mathbf{v})) - \nabla_{\mathbf{a}}\Lambda(x_K(\mathbf{a}^*,\mathbf{v}))^\top(\mathbf{a}-\mathbf{a}^*)\}. \tag{57}$$

Note that the penalty $M_g^*(\mathbf{a},\mathbf{v})$ is linear in the feasible strategy $\mathbf{a}$, and hence the inner optimization problem is a convex optimization problem in $\mathbf{a}$. As a consequence, the first-order condition is sufficient to guarantee an optimal solution to (57): the gradient of the objective function in (57) is

$$\nabla_{\mathbf{a}}\Lambda(x_K(\mathbf{a},\mathbf{v})) - \nabla_{\mathbf{a}}\Lambda(x_K(\mathbf{a}^*,\mathbf{v})), \tag{58}$$

and it becomes zero if we take $\mathbf{a} = \mathbf{a}^*$. Therefore, $\mathbf{a}^*$ is the optimal solution to the inner optimization (57). Hence, it is straightforward to see that $M_g^*(\mathbf{a}, \mathbf{v})$ is an optimal penalty in the sense that $V_0^*(x) = (\mathscr{L}M_g^*)(x)$ for all $x \in \mathscr{X}$.

It is obvious to see the analogy between using the first-order condition (55) and (58) to derive optimal penalties for the continuous-time problem and discrete-time problem, respectively. This is the reason why h_g^* can be viewed as the "gradient-based penalty" in the dual formulation of controlled Markov diffusions.

APPENDIX E
DETAILS FOR NUMERICAL EXPERIMENTS

The dynamics of the market state and assets returns are the same as those considered in [38]. In particular, $\mu_k^\phi = -\lambda\phi_k$, $\mu_k = \mu_0 + \mu_1\phi_k$, $\sigma_k \equiv \sigma$, $\sigma_k^{\phi,1} \equiv \sigma^{\phi,1}$, and $\sigma_k^{\phi,2} \equiv \sigma^{\phi,2}$. The parameter values are listed in the following tables including r_f, λ, μ_0, μ_1, σ, $\sigma^{\phi,1}$, and $\sigma^{\phi,2}$.

TABLE V
PARAMETER 1

	μ_0	μ_1	σ	r_f
$\log(R)$	$\begin{pmatrix} 0.081 \\ 0.110 \\ 0.130 \end{pmatrix}$	$\begin{pmatrix} 0.034 \\ 0.059 \\ 0.073 \end{pmatrix}$	$\begin{pmatrix} 0.186 & 0.000 & 0.000 \\ 0.228 & 0.083 & 0.000 \\ 0.251 & 0.139 & 0.069 \end{pmatrix}$	0.01
ϕ	λ		$\sigma^{\phi,1}$	$\sigma^{\phi,2}$
	0.336		$\begin{pmatrix} -0.741 & -0.037 & -0.060 \end{pmatrix}$	0.284

TABLE VI
PARAMETER 2

	μ_0	μ_1	σ	r_f
$\log(R)$	$\begin{pmatrix} 0.081 \\ 0.110 \\ 0.130 \end{pmatrix}$	$\begin{pmatrix} 0.034 \\ 0.059 \\ 0.073 \end{pmatrix}$	$\begin{pmatrix} 0.186 & 0.000 & 0.000 \\ 0.228 & 0.083 & 0.000 \\ 0.251 & 0.139 & 0.069 \end{pmatrix}$	0.01
ϕ	λ		$\sigma^{\phi,1}$	$\sigma^{\phi,2}$
	1.671		$\begin{pmatrix} -0.017 & 0.149 & 0.058 \end{pmatrix}$	1.725

TABLE VII

PARAMETER 3

	μ_0	μ_1	σ	r_f
$\log(R)$	$\begin{pmatrix} 0.142 \\ 0.109 \\ 0.089 \end{pmatrix}$	$\begin{pmatrix} 0.065 \\ 0.049 \\ 0.049 \end{pmatrix}$	$\begin{pmatrix} 0.256 & 0.000 & 0.000 \\ 0.217 & 0.054 & 0.000 \\ 0.207 & 0.062 & 0.062 \end{pmatrix}$	0.01
ϕ	λ		$\sigma^{\phi,1}$	$\sigma^{\phi,2}$
	0.336		$\begin{pmatrix} -0.741 & -0.040 & -0.034 \end{pmatrix}$	0.288

TABLE VIII

PARAMETER 4

	μ_0	μ_1	σ	r_f
$\log(R)$	$\begin{pmatrix} 0.142 \\ 0.109 \\ 0.089 \end{pmatrix}$	$\begin{pmatrix} 0.061 \\ 0.060 \\ 0.067 \end{pmatrix}$	$\begin{pmatrix} 0.256 & 0.000 & 0.000 \\ 0.217 & 0.054 & 0.000 \\ 0.206 & 0.062 & 0.062 \end{pmatrix}$	0.01
ϕ	λ		$\sigma^{\phi,1}$	$\sigma^{\phi,2}$
	1.671		$\begin{pmatrix} -0.017 & 0.212 & 0.096 \end{pmatrix}$	1.716