Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Dual Formulation of Controlled Markov Diffusions and Its Application ⋆

Fan Ye * **Enlu Zhou** *

* *H. Milton Stewart School of Industrial & Systems Engineering (e-mail: fye8@gatech.edu, enlu.zhou@isye.gatech.edu).*

Abstract: Information relaxation and duality in Markov decision processes have been studied recently to derive upper bounds on the maximal expected reward (or lower bounds on the minimal expected cost). The idea is to relax the non-anticipativity constraint on the controls and impose a penalty to punish such a violation. In this paper we generalize this dual approach to controlled Markov diffusions. We develop the weak duality and strong duality results, and explore the structure of the optimal penalty. We demonstrate the use of this dual formulation by computing upper bounds on the optimal expected utility in a dynamic portfolio choice problem.

1. INTRODUCTION

Markov decision processes (MDPs) play a central role in modeling discrete-time sequential decision making problems under uncertainty. MDPs can be solved, in principle, via dynamic programming; however, the exact computation of dynamic programming suffers from the "curse of dimensionality", i.e., the size of the state space increases exponentially with the number of the state variables. To overcome this difficulty, approximate dynamic programming techniques have been developed including Chang et al. [2007], Bertsekas [2007], Powell [2011], de Farias and van Roy [2003]. These methods often generate sub-optimal policies, which lead to lower bounds on the optimal expected reward by simulating the dynamic system under the aforementioned policies. However, the accuracy of the sub-optimal policies is generally unknown. Motivated by the lack of performance guarantee on sub-optimal policies, a dual representation of MDPs was recently developed by Rogers [2007], Brown et al. [2010] to provide an upper bound on the optimal expected reward. If the difference between the lower and upper bounds is small, it may be concluded that the quality of the existing policy is acceptable. The main idea of this dual approach is to allow the decision maker to foresee the future uncertainty but impose a penalty on getting access to the information in advance; particularly, the dual bound can recover the optimal reward by applying proper penalties. Therefore, various approximation methods based on different types of optimal penalties have been proposed in order to derive tight dual bounds, such as Brown et al. [2010], Brown and Smith [2011], Desai et al. [2011], Ye and Zhou [2012].

The goal of this paper is to extend the information relaxation-based dual representation of MDPs to controlled Markov diffusions, which are typical sequential decision making problems in continuous-time setting. The Hamlton-Jacobi-Bellman (HJB) equation, a standard approach solving controlled Markov diffusions, rarely allows a closed-form solution, especially when the state space is of high dimension or there are constraints imposed on the control space. There are many numerical methods based on different approximation schemes: Kushner and Dupuis [2001] considered the Markov chain approximation method by discretizing the HJB equation; Han and van Roy [2011] extended the approximate linear programming method to controlled Markov diffusions. Another numerical approach is to discretize the time space, which reduces the original continuous-time problems to MDPs and the technique of approximate dynamic programming can be applied.

In this paper we intend to answer the following questions. First, is there a dual formulation of controlled Markov diffusions based on information relaxation as that of MDPs? If yes, what is the form of the optimal penalty? Second, is there any practical use of this dual formulation? To answer the first question, we establish the weak duality and strong duality results that parallel those in the dual formulation of MDPs; moreover, we investigate a class of optimal penalties, the so-called "value function-based penalty", which can be written compactly as an Ito stochastic integral under the natural filtration generated by the Brownian motion. We then consider the application of this dual formulation in a dynamic portfolio choice problem: based on the aforementioned value function-based penalty we propose an effective class of penalties, which are easy to evaluate and can be used to derive tight dual bounds on the optimal expected value.

We note that Davis and Burstein [1992, 1991] have pioneered a series of related work for controlled Markov diffusions under the name "anticipative stochastic control". They also adopted the recipe of relaxing the future information and then penalizing, prior to the dual framework of MDPs. The main difference of their work from ours is that we propose a more general framework that may incorporate their approach as a special case; in addition, their proposed Lagrangian approach for penalization differs from our value function-based penalty. A discussion of the connection between their work and our work together with Rogers [2007], Brown et al. [2010], Brown and Smith [2011] is described in the Appendix of Ye and Zhou [2013].

⋆ This work was supported by National Science Foundation under Grant CMMI-1130273, and by the Air Force Office of Scientific Research under YIP Grant FA-9550-12-1-0250.

The rest of the paper is organized as follows. In section 2, we review the dual formulation of MDPs and develop the dual formulation of controlled Markov diffusions. We then illustrate the dual approach in a dynamic portfolio choice problem in Section 3. Finally, we conclude in Section 4.

2. CONTROLLED MARKOV DIFFUSIONS AND ITS DUAL FORMULATION

2.1 Review of Duality in Markov Decision Processes

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

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

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

Denote by $\mathbb{A}$ the set of all control strategies $\mathbf{a} \triangleq (a_1, \cdots, a_{K-1})$, i.e., $a_k \in \mathcal{A}$ for every $k \in \mathcal{K}$. Let $\mathbb{A}_{\mathbb{G}}$ be the set of control strategies that are adapted to the filtration $\mathbb{G}$, i.e., a_k is $\mathcal{G}_k$-adapted for every k. We also call $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$ a *non-anticipative* policy. The objective is to maximize the expected sum of intermediate rewards $\{g_k\}_{k=0}^{K-1}$ and final reward Λ by selecting a non-anticipative policy $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$:

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

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:

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

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 future uncertainty.

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

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

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

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

Note that the supremum in (3) is over the set $\mathbb{A}$ not the set $\mathbb{A}_{\mathbb{G}}$. The optimization problem inside the expectation in (3) will be 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, which leads to an unbiased estimator of $(\mathcal{L}M)(x)$.

Theorem 1 below establishes a strong duality in the sense that for all $x \in \mathcal{X}_0$,

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

In particular, Theorem 1(a) suggests that $\mathcal{L}M(x_0)$ can be used to derive an upper bound on the value function $V_0(x_0)$ given any $M \in \mathcal{M}_{\mathbb{G}}(0)$; Theorem 1(b) states that the duality gap vanishes if the dual problem is solved by choosing M in the form of (4).

Theorem 1. (Theorem 2.1 in Brown et al. [2010]).

(a) (Weak Duality) For all $M \in \mathcal{M}_{\mathbb{G}}(0)$ and all $x \in \mathcal{X}$, $V_0(x) \leq \big(\mathcal{L}M\big)(x)$.
(b) (Strong Duality) For all $x \in \mathcal{X}$, $V_0(x) = \big(\mathcal{L}M^*\big)(x)$, where

$$M^*(\mathbf{a}, \mathbf{v}) = \sum_{k=0}^{K-1} \big(V_{k+1}(x_{k+1}) - \mathbb{E}[V_{k+1}(x_{k+1}) | x_k, a_k]\big). \qquad (4)$$

Remark 1.

(1) Note that the right hand side of (4) is a function of $(\mathbf{a},\mathbf{v})$, since $\{x_k\}$ depends on $(\mathbf{a}, \mathbf{v})$ through the state equation (1).
(2) Since $\mathbb{E}_{0,x}[M^*(\mathbf{a}, \mathbf{v})] = 0$ for all $x \in \mathcal{X}$ and $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$, we know $M^* \in \mathcal{M}_{\mathbb{G}}(0)$. The penalty M^* will be refereed to as the *value function-based penalty* for the Markov decision problem (1)-(2), as M^* depends on the value functions $\{V_k\}_{k=1}^K$.

The optimal penalty (4) that achieves the strong duality involves the value functions $\{V_k\}_{k=1}^K$, 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 approximate value functions $\{\hat{V}_k\}_{k=1}^K$. However, this heuristic approach suffers at least two difficulties. The first difficulty 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; second, even if $\mathbb{E}[\hat{V}_{k+1}(x_{k+1}) | x_k, a_k]$ can be computed analytically, the inner optimization problem may still be difficult to solve since no convex structure can be guaranteed (even assuming that all the rewards $\{g_1, \cdots, g_{K-1}, \Lambda\}$ are concave functions).

2.2 Controlled Markov Diffusions and HJB Equation

This subsection is concerned with the control of Markov diffusions. Consider an $\mathbb{R}^n$-valued controlled Markov diffusion process $(x_t)_{0 \leq t \leq T}$ governed by the stochastic differential equation (SDE) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

$$dx_t = b(t, x_t, u_t)dt + \sigma(t, x_t)dw_t, \ \ 0 \leq t \leq T, \qquad (5)$$

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

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

(1) $\mathbf{u} = (u_s)_{s \in [t,T]}$ is an $\mathbb{F}$-progressively measurable process taking value in $\mathcal{U}$ (i.e., $\mathbf{u}$ 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 $\mathcal{U}_{\mathbb{F}}(t)$.

With the following standard assumptions imposed on b and σ, SDE (5) admits a unique pathwise solution when $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$, i.e., $(x_t)_{t\in[0,T]}$ is $\mathbb{F}$-progressively measurable and has continuous sample paths almost surely.

Assumption 1. b and σ are continuous on their domains, respectively, and for some constants C_1, C_2, and $C_\sigma > 0$,

(1) $\| b(t,x,u) \| + \| \sigma(t,x,u) \| \leq C_1(1+ \| x \| + \| u \|)$ for all $(t,x,u) \in \bar{Q} \times \mathcal{U}$;
(2) $\| b(t,x,u) - b(s,y,u) \| + \| \sigma(t,x,u) - \sigma(s,y,u) \| \leq C_2(|t-s| + \| x-y \|)$ for all $(t,x),(s,y) \in \bar{Q}$.
(3) $\xi^\top(\sigma\sigma^\top)(t,x)\xi \geq C_\sigma \| \xi \|^2$ for all $(t,x) \in [0,T] \times Q$ and $\xi \in \mathbb{R}^n$.

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 $g : \bar{Q} \times \mathcal{U} \to \mathbb{R}$ as the final reward and intermediate reward, respectively. Assume the rewards Λ and g satisfy the polynomial growth conditions.

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

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

Then we introduce the reward functional

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

Given an initial condition $(t,x) \in Q$, the objective is to maximize $J(t,x,u)$ over all the controls $\mathbf{u}$ in $\mathcal{U}_{\mathbb{F}}(t)$:

$$V(t,x) = \sup_{\mathbf{u}\in\mathcal{U}_{\mathbb{F}}(t)} J(t,x;\mathbf{u}). \tag{6}$$

Here we abuse the notations 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

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

where L_t, L_x, and L_{xx} denote the t-partial derivative, the gradient and the Hessian with respect to x respectively, and $(\sigma\sigma^\top)(t,x) \triangleq \sigma(t,x)\sigma^\top(t,x)$. 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 satisfying a polynomial growth condition in x, i.e.,

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

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

Theorem 2. (Theorem 4.3.1 in Fleming and Soner [2006]). Suppose Assumptions 1 and 2 hold, and $\bar{V} \in C^{1,2}(Q) \cap C_p(\bar{Q})$ satisfies $\bar{V}(T,x) = \Lambda(x)$ and

$$\sup_{u\in\mathcal{U}}\{g(t,x,u) + A^u\bar{V}(t,x)\} = 0, \quad (t,x) \in Q. \tag{7}$$

Then
(a) $J(t,x;\mathbf{u}) \leq \bar{V}(t,x)$ for any $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(t)$ and $(t,x) \in \bar{Q}$.
(b) If there exists a function $u^* : \bar{Q} \to \mathcal{U}$ such that

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

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

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

Equation (7) is the well-known HJB equation associated with the stochastic optimal control problem (5)-(6).

2.3 Dual Formulation of Controlled Markov Diffusions

In this subsection we present the dual formulation of controlled Markov diffusions based on 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 $\mathcal{F} = \mathcal{F}_T$. To expand the set of feasible controls, we use $\mathcal{U}(t)$ to denote the set of measurable $\mathcal{U}$-valued control strategies at time t, i.e., $\mathbf{u} = (u_s)_{s\in[t,T]} \in \mathcal{U}(t)$ if $\mathbf{u}$ is $\mathcal{B}([t,T]) \times \mathcal{F}$-measurable and u_s takes value in $\mathcal{U}$, where $\mathcal{B}([t,T])$ is the Borel σ-algebra on $[t,T]$. In particular, $\mathcal{U}(0)$ can be viewed as the counterpart of $\mathbb{A}$ introduced in Section 2.1 for MDPs.

A technical problem we have to face with is to define a solution of (5) with an anticipative control $\mathbf{u} \in \mathcal{U}(0)$. Since it involves the concept of "anticipating stochastic calculus", we refer the readers to Appendix B in Ye and Zhou [2013], where the decomposition technique is used to define the solution of an anticipating SDE following Davis and Burstein [1992], Ocone and Pardoux [1989].

Right now we only need to assume that given a control strategy $\mathbf{u} \in \mathcal{U}(0)$ there exists a unique solution $(x_t)_{t\in[0,T]}$ to (5) that is $\mathcal{B}([0,T]) \times \mathcal{F}$-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 $\mathbf{u} \in \mathcal{U}(0)$ and a sample path of Brownian motion $\mathbf{w} \triangleq (w_t)_{t\in[0,T]}$. We define the set $\mathcal{M}_{\mathbb{F}}(0)$ of *dual feasible penalties* h that do not penalize non-anticipative policies in expectation, i.e.,

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

With an arbitrary choice of $h \in \mathcal{M}_{\mathbb{F}}(0)$, we can determine an upper bound on (6) at $t = 0$ by relaxing the non-anticipativity constraint on control strategies.

Proposition 3. (Weak Duality). If $h \in \mathcal{M}_{\mathbb{F}}(0)$, then for all $x \in \mathbb{R}^n$,

$$\begin{aligned}&\sup_{\mathbf{u}\in\mathcal{U}_{\mathbb{F}}(0)} J(0,x;\mathbf{u})\\ \leq&\mathbb{E}_{0,x}\left[\sup_{\mathbf{u}\in\mathcal{U}(0)}\left\{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h(\mathbf{u},\mathbf{w})\right\}\right].\end{aligned} \tag{8}$$

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

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

Then inequality (8) can be obtained by taking the supremum over $\bar{\mathbf{u}} \in \mathcal{U}_{\mathbb{F}}(0)$ on the left hand side of the last inequality.

The optimization problem inside the conditional expectation in (8) is the *inner optimization problem* in the context of controlled Markov diffusions (cf. (3)): an entire path of $\mathbf{w}$ is known beforehand (i.e., *perfect information relaxation*), and the objective function depends on the trajectory of $\mathbf{w}$. The expectation term on the right hand side of (8) is a *dual bound* on the value function $V(0,x)$. An interesting case is

$$h^*(\mathbf{u},\mathbf{w}) = \Lambda(x_T)+\int_0^T g(t,x_t,u_t)dt - V(0,x). \qquad (9)$$

Note that $h^* \in \mathcal{M}_{\mathbb{F}}(0)$, since for all $x \in \mathbb{R}^n$ and $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$,

$$\mathbb{E}_{0,x}\left[\Lambda(x_T)+\int_0^T g(s,x_s,u_s)ds\right] \leq V(0,x),$$

by the definition of $V(0,x)$. We also note that by plugging $h = h^*$ in the inner optimization problem in (8), 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 4. (Strong Duality). For all $x \in \mathbb{R}^n$,

$$\begin{aligned} &\sup_{\mathbf{u}\in\mathcal{U}_{\mathbb{F}}(0)} J(0,x;\mathbf{u}) \\ &= \inf_{h\in\mathcal{M}_{\mathbb{F}}(0)}\left\{\mathbb{E}_{0,x}\left[\sup_{\mathbf{u}\in\mathcal{U}(0)}\left\{\Lambda(x_T)+\int_0^T g(t,x_t,u_t)dt - h(\mathbf{u},\mathbf{w})\right\}\right]\right\}. \end{aligned} \qquad (10)$$

The minimum of the right hand side of (10) can always be achieved by choosing an $h \in \mathcal{M}_{\mathbb{F}}(0)$ in the form of (9).

Due to the strong duality result, the left hand side problem of (10) is referred to as the *primal problem* and the right hand side problem of (10) is referred to as the *dual problem*. Since the relaxation of the nonanticipativity requirement on admissible strategies is compensated by a proper function in $\mathcal{M}_{\mathbb{F}}(0)$, therefore, we call $h \in \mathcal{M}_{\mathbb{F}}(0)$ a dual feasible penalty. If h is a dual feasible penalty that achieves the infimum on the right side of (10), we call it an optimal solution to the dual problem.

We note that the optimal penalty in (9) is intractable in practice as it requires knowing the exact value of $V(0,x)$. Proposition 5 below presents an optimal penalty that can be viewed as the continuous-time analogue of M^* in (4), which also guides the numerical studies in Section 3. We fully develop the relevant results in Appendix B of Ye and Zhou [2013].

Proposition 5. (Value Function-based Penalty) Suppose the value function $V(t,x)$ defined in (6) satisfies all the assumptions in Theorem 2(b). Then under the technical conditions specified in Theorem 5 of Ye and Zhou [2013], there is an optimal solution to the dual problem, i.e., an optimal penalty $h_v^*(\mathbf{u},\mathbf{w}) \in \mathcal{M}_{\mathbb{F}}(0)$ in the form of

$$h_v^*(\mathbf{u},\mathbf{w}) = \int_0^T V_x(t,x_t)^\top \sigma(t,x_t)dw_t \ \text{ for } \mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0), \ (11)$$

where x_t is the solution of (5) using the control $\mathbf{u} = (u_t)_{t\in[0,T]}$ on $[0,t)$ with the initial condition $x_0 = x$.

Since the value functions $\{V(t,x), 0 \leq t \leq T\}$ are unknown in real applications, (11) implies that if an approximate value function $\{\hat{V}(t,x), 0 \leq t \leq T\}$ of sufficient regularity is given, h_v^* 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 $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$. 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 $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$. As a result, $\hat{h}_v(\mathbf{u},\mathbf{w}) \in \mathcal{M}_{\mathbb{F}}(0)$, i.e., $\hat{h}_v$ is a dual feasible penalty and can be used to derive an upper bound on the value function $V(0,x)$ through (8). Therefore, in terms of the approximation scheme implied by the form of the optimal penalty, h_v^* is a *value function-based penalty* for controlled Markov diffusions (5)-(6).

3. DYNAMIC PORTFOLIO CHOICE PROBLEM

The purpose of this section is to illustrate how the value function-based penalty in Proposition 5 helps to solve a *discrete-time* dynamic portfolio choice problem with predictable returns and intermediate consumptions. Since most dynamic portfolio problems can only be solved numerically with suboptimal policies (see, e.g, Cvitanic et al. [2003], Tauchen and Hussey [1991], Brandt et al. [2005], Han and van Roy [2011]), it is often hard to tell how far these policies are from the optimal one. By generating a tight upper bound on the value function, the duality gap indicates the performance of suboptimal polices (see, e.g., Haugh et al. [2006], Brown and Smith [2011]). Under the dual formulation of MDPs in Section 2.1, we compute upper bounds on the optimal expected utility using a new class of penalties that avoid evaluating any conditional expectation and keep the inner optimization problem easy to solve.

3.1 The Dynamic Portfolio Choice Model

We first consider a continuous-time financial market with finite horizon $[0,T]$, which is built on a probability space $(\Omega,\mathcal{F},\mathbb{P})$. There are one risk-free asset(cash) and n risky assets the investor can choose among. The instantaneously risk-free rate of return is denoted by r_f. An m-dimensional market state variable ϕ_t follows a diffusion process

$$d\phi_t = \mu_t^\phi dt + \sigma_t^{\phi,1} dz_t + \sigma_t^{\phi,2} d\tilde{z}_t, \qquad (12)$$

where $\mathbf{z} \triangleq (z_t)_{0\leq t\leq T}$ and $\tilde{\mathbf{z}} \triangleq (\tilde{z}_t)_{0\leq t\leq T}$ are two independent standard Brownian motions of dimensions n and d, respectively; $\mu_t^\phi \triangleq \mu^\phi(t,\phi_t)$, $\sigma_t^{\phi,1} \triangleq \sigma^{\phi,1}(t,\phi_t)$ and $\sigma_t^{\phi,2} \triangleq \sigma^{\phi,2}(t,\phi_t)$ are deterministic functions of ϕ_t and are of dimension m, $m\times n$, and $m\times d$, respectively. Denote the filtration by $\mathbb{F} = \{\mathcal{F}_t, 0 \leq t \leq T\}$, where $\mathcal{F}_t$ is generated by $\{(z_s,\tilde{z}_s), 0 \leq s \leq t\}$.

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

$$\begin{aligned} dW_t =& W_t\big[\pi_t^\top(\mu_t dt + \sigma_t dz_t) + r_f(1-\pi_t^\top \mathbf{1}_n)dt - \tilde{c}_t dt\big] \\ =& W_t\big(\pi_t^\top(\mu_t - r_f\mathbf{1}_n) + r_f - \tilde{c}_t\big)dt + W_t\pi_t^\top\sigma_t dz_t, \end{aligned} \tag{13}$$

where the drift $\mu_t = \mu(t,\phi_t)$ and the volatility $\sigma_t = \sigma(t,\phi_t)$ of the risky assets are deterministic functions of ϕ_t and are of dimension n and $n\times n$, respectively; the covariance matrix $\sigma_t\sigma_t^\top$ is denoted by Σ_t. We use $\mathbf{1}_n$ to denote the n-dimensional all-ones vector. The control $\mathbf{u} \triangleq (u_t)_{0\le t\le T}$ with $u_t \triangleq (\pi_t, \tilde{c}_t)$ is assumed to be an $\mathcal{U}$-valued admissible strategy (see Definition 1), where the control space $\mathcal{U}$ will be specified later. We still use $\mathcal{U}_{\mathbb{F}}(t)$ to denote the set of admissible strategies at time t.

Define $U(x) \triangleq \frac{1}{1-\gamma}x^{1-\gamma}$, an utility function of constant relative risk aversion (CRRA) type with coefficient $\gamma > 0$, which is widely used in economics and finance. The investor's objective is to maximize the weighted sum of the expected utility of the intermediate consumption and the final wealth:

$$V(t,\phi_t,W_t) = \sup_{\mathbf{u}\in\mathcal{U}_{\mathbb{F}}(t)} \mathbb{E}_t\left[\int_t^T \alpha U(\tilde{c}_s W_s)ds + (1-\alpha)U(W_T)\right], \tag{14}$$

where α implies the relative importance of the intermediate consumption, and $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot|\phi_t, W_t]$.

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 (12)-(14) by discretizing its time space. Suppose the decision takes place at equally spaced time $\{0 = t_0, t_1 \cdots, t_K\}$ 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 $\{0, 1, \cdots, K\}$ and discretize (12) and (13) as follows:

$$\phi_{k+1} = \phi_k + \mu_k^\phi\delta + \sigma_k^{\phi,1}\sqrt{\delta}Z_{k+1} + \sigma_k^{\phi,2}\sqrt{\delta}\tilde{Z}_{k+1}, \tag{15a}$$

$$\log(R_{k+1}) = (\mu_k - \frac{1}{2}\sigma_k^2)\delta + \sigma_k\sqrt{\delta}Z_{k+1}, \tag{15b}$$

$$\begin{aligned} W_{k+1} &= W_k(\pi_k^\top R_{k+1}) + W_k(1-\pi_k^\top\mathbf{1}_n)R_f - W_k c_k, \\ &= W_k(R_f + (R_{k+1} - R_f\mathbf{1}_n)^\top\pi_k - c_k), \end{aligned} \tag{15c}$$

where $\{(Z_k, \tilde{Z}_k), k = 1, \cdots, K\}$ is a sequence of identically and independently distributed standard Gaussian random vectors, (15b) follows from the fact that $d\log(R_t) = (\mu_t - \frac{1}{2}\sigma_t^2)dt + \sigma_t dz_t$ is equivalent to $dR_t/R_t = \mu_t dt + \sigma_t dz_t$, where σ_t^2 denotes the vector that consists of the diagonal of Σ_t. In particular, we use $R_f \triangleq 1 + r_f\delta$ and c_k to approximate $e^{r_f\delta}$ and $\tilde{c}_k\delta$ due to the time-discretization.

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 subscripts $t \in [0,T]$ and $k = 0, \cdots, K$ are used in the continuous-time model and the discrete-time model, respectively.

Denote the filtration of the process (15) by $\mathbb{G} = \{\mathcal{G}_0, \cdots, \mathcal{G}_K\}$, where $\mathcal{G}_k$ 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

$$\mathcal{A} \triangleq \{(\pi, c) \in \mathbb{R}^{n+1} | \pi \ge 0, c \ge 0, c \le R_f(1 - \mathbf{1}_n^\top\pi)\}. \tag{16}$$

which corresponds to a control set $\mathcal{U}$ for the continuous-time model that is defined as

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

Let $\mathbb{A}_{\mathbb{G}}$ again denote the set of $\mathcal{A}$-valued control strategies $\mathbf{a} \triangleq (a_1, \cdots, a_{K-1})$ that are adapted to the filtration $\mathbb{G}$. Then the value function to the discrete-time problem that is the discretization of (14) is

$$H_0(\phi_0, W_0) = \sup_{\mathbf{a}\in\mathbb{A}_{\mathbb{G}}} \mathbb{E}_0[\sum_{k=0}^{K-1} \alpha U(c_k W_k)\delta + (1-\alpha)U(W_K)], \tag{17}$$

which can be solved via dynamic programming:

$$\begin{aligned} H_K(\phi_K, W_K) &= (1-\alpha)U(W_K); \\ H_k(\phi_k, W_k) &= \sup_{a_k\in\mathcal{A}} \{\alpha U(c_k W_k)\delta + \mathbb{E}_k[H_{k+1}(\phi_{k+1}, W_{k+1})]\}. \end{aligned} \tag{18}$$

The rest of this section is devoted to compute the lower and upper bounds on H_0. Particularly, since the utility function is of CRRA type, both value functions (14) and (17) have simplified structures

$$V(t,\phi_t,W_t) = W_t^{1-\gamma}\tilde{J}(t,\phi_t), \tag{19}$$

$$H_k(\phi_k, W_k) = W_k^{1-\gamma}J_k(\phi_k), \tag{20}$$

where $\tilde{J}(t,\phi) = V(t,\phi_t,1)$ and $J_k(\phi_k) = H_k(\phi_k, 1)$. In particular, J_k is defined recursively as $J_K(\phi_K) = (1-\alpha)/(1-\gamma)$ and for $k = K-1, \cdots, 0$,

$$\begin{aligned} J_k(\phi_k) = \sup_{(\pi_k,c_k)\in\mathcal{A}} \big\{ & \frac{\alpha}{1-\gamma}c_k^{1-\gamma}\delta \\ & + \mathbb{E}\big[\big(R_f + (R_{k+1} - R_f)^\top\pi_k - c_k\big)^{1-\gamma} J_{k+1}(\phi_{k+1})|\phi_k\big]\big\}. \end{aligned} \tag{21}$$

3.2 Dual Bounds and Penalties

In this subsection we focus on generating an upper bound on H_0 based on the dual formulation of MDPs (see Section 2.1). We first show that the value function-based penalty M^* in Theorem 1 does not directly suggest a tractable feasible penalty for the problem (15)-(17). As a remedy we will introduce a heuristic and tractable penalty by discretizing the value function-based penalty of the continuous-time problem, assuming that an approximate function of $J_k(\phi)$, say $\hat{J}_k(\phi)$ (therefore, $\hat{H}_k(\phi_k, W_k) \triangleq W_k^{1-\gamma}\hat{J}_k(\phi_k)$ is an approximation of H_k), 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.

To derive an upper bound on H_0, we consider the perfect information relaxation that assumes the investor can foresee the future uncertainty $\mathbf{Z} = (Z_1, \cdots, Z_K)$ and $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \cdots, \tilde{Z}_K)$. A function $M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})$ is a *dual feasible penalty* in the setting of dynamic portfolio choice problem (15)-(17) if for any (ϕ_0, W_0),

$$\mathbb{E}[M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})|\phi_0, W_0] \le 0 \quad \text{for all} \quad \mathbf{a} \in \mathbb{A}_{\mathbb{G}}. \tag{22}$$

Let $\mathcal{M}_{\mathbb{G}}(0)$ denote the set of all dual feasible penalties. For $M \in \mathcal{M}_{\mathbb{G}}(0)$ we define $(\mathcal{L}M)(\phi_0, W_0)$ as (cf. (3))

$$\mathbb{E}\Big[\sup_{\mathbf{a}\in\mathbb{A}}\Big\{\sum_{k=0}^{K-1} \alpha U(c_k W_k)\delta + (1-\alpha)U(W_K) - M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})\Big\}\Big|\phi_0, W_0\Big]. \tag{23}$$

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

To ease the inner optimization problem in the conditional expectation in (23), we introduce 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 admissible strategy, though in terms of (Π_k, C_k) now. Then we can rewrite the inner optimization problem in $\mathcal{L}M$ as follows:

$$\max_{\{\Pi_k, C_k\}} \{\sum_{k=0}^{K-1} \alpha U(C_k)\delta + (1-\alpha)U(W_K) - M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})\} \quad (24a)$$

$$\text{s.t.} \quad \phi_{k+1} = \phi_k + \mu_k^{\phi}\delta + \sigma_k^{\phi,1}\sqrt{\delta}Z_{k+1} + \sigma_k^{\phi,2}\sqrt{\delta}\tilde{Z}_{k+1}, \quad (24b)$$

$$\log(R_{k+1}) = (\mu_k - \frac{1}{2}\sigma_k^2)\delta + \sigma_k\sqrt{\delta}Z_{k+1}, \quad (24c)$$

$$W_{k+1} = W_k R_f + (R_{k+1} - R_f \mathbf{1}_n)^\top \Pi_k - C_k, \quad (24d)$$

$$\Pi_k \geq 0, \quad C_k \geq 0, \quad (24e)$$

$$C_k \leq R_f(W_k - 1_n^\top \Pi_k), \text{ for } k = 0, \cdots, K-1. \quad (24f)$$

Note that (24b)-(24d) is equivalent to (15a)-(15c), and(24e)-(24f) are equivalent to (16). The advantage of this reformulation is that the inner optimization problem (24) has linear constraints. Therefore, we may find the global maximizer of (24) as long as the objective function in (24a) is jointly concave in $\mathbf{a}$.

The challenge is to design effective penalties that lead to tight upper bound and also keep the inner optimization problem easy to solve. We first investigate the optimal penalty for the problem (15)-(17) according to (4):

$$\begin{aligned} M^*(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) &= \sum_{k=0}^{K-1} \Delta H_{k+1}(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \\ &= \sum_{k=0}^{K-1} \Big(H_{k+1}(\phi_{k+1}, W_{k+1}) - \mathbb{E}[H_{k+1}(\phi_{k+1}, W_{k+1})|\phi_k, W_k, a_k] \Big) \end{aligned} \quad (25)$$

We may approximate H_k by $\hat{H}_k = W_k^{1-\gamma}\hat{J}_k$, however, it does not mean that $\Delta\hat{H}_{k+1}$ can be easily computed, since an intractable conditional expectation $\mathbb{E}_k[H_{k+1}]$, i.e., an integral over $(n+d)$-dimensional space is involved in (25). Another difficulty is that M^* enters into (24a) with possibly positive or negative signs for different realizations of $(\mathbf{Z}, \tilde{\mathbf{Z}})$, making the objective of (24) nonconcave, even if U_1 and U_2 are concave functions. Therefore, it can be extremely hard to locate the global maximizer of (24).

As an alternative approximation scheme, we exploit the value function-based penalty h_v^* for the continuous-time problem (12)-(14) based on Proposition 5, assuming that all the technical conditions hold. Note that by selecting $x_t = (\phi_t, W_t)$ and $V_x = (V_\phi, V_W)$, we can formally write

$$\begin{aligned} h_v^*(\mathbf{u}, \mathbf{z}, \tilde{\mathbf{z}}) &= \int_0^T \begin{pmatrix} V_\phi(t, \phi_t, W_t) \\ V_W(t, \phi_t, W_t) \end{pmatrix}^\top \begin{pmatrix} \sigma_t^{\phi,1} & \sigma_t^{\phi,2} \\ W_t\pi_t\sigma_t & 0 \end{pmatrix} \begin{pmatrix} dz_t \\ d\tilde{z}_t \end{pmatrix} \\ &= \sum_{k=0}^{K-1} \int_{k\delta}^{(k+1)\delta} \big[V_\phi(t, \phi_t, W_t)^\top \sigma_t^{\phi,1} dz_t \\ &\quad + V_\phi(t, \phi_t, W_t)^\top \sigma_t^{\phi,2} d\tilde{z}_t + V_W(t, \phi_t, W_t) W_t \pi_t \sigma_t dz_t \big], \\ &= \sum_{k=0}^{K-1} \int_{k\delta}^{(k+1)\delta} \big[W_t^{1-\gamma} \nabla_\phi \tilde{J}(t, \phi_t)^\top \sigma_t^{\phi,1} dz_t \\ &\quad + W_t^{1-\gamma} \nabla_\phi \tilde{J}(t, \phi_t)^\top \sigma_t^{\phi,2} d\tilde{z}_t + (1-\gamma) W_t^{1-\gamma} \tilde{J}(t, \phi_t) \pi_t \sigma_t dz_t \big], \end{aligned} \quad (26)$$

for $\mathbf{u} = (\pi_t, \tilde{c}_t)_{0 \leq t \leq T} \in \mathcal{U}_{\mathbb{F}}(0)$, where the last equality holds due to structure of the value function (19). 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 to approximate (25), i.e.,

$$\begin{aligned} &M_1(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \\ &\triangleq \sum_{k=0}^{K-1} \big[\Psi_{k+1}^1(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) + \Psi_{k+1}^2(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) + \Psi_{k+1}^3(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \big], \end{aligned} \quad (27)$$

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

Here $\bar{\phi}_k$, $\bar{\sigma}_k^{\phi,1}$,$\bar{\sigma}_k^{\phi,2}$ are the realization of ϕ_k, $\sigma_k^{\phi,1}$,$\sigma_k^{\phi,2}$ based on the realization of $(\mathbf{Z}, \tilde{\mathbf{Z}})$, and $\bar{W}_k$ is the realization of W_k under the strategy $\hat{\mathbf{a}} = (\hat{a}_0, \cdots, \hat{a}_{K-1})$; $\hat{J}_k(\cdot)$ and $\nabla_\phi \hat{J}_k(\cdot)$ are approximations of J_k and its (sub)gradient with respect to ϕ.

Note that for a fixed realization of $(\mathbf{Z}, \tilde{\mathbf{Z}})$, $M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})$ in (27) is linear in Π_k (hence, in $\mathbf{a}$), therefore, the objective function (24a) is jointly concave in $\mathbf{a}$. As a result, the inner optimization problem (24) remains a convex optimization problem and can be efficiently solved.

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

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

where $\bar{R}_k$ is the realization of R_k based on the realization $(\mathbf{Z}, \tilde{\mathbf{Z}})$, while $\bar{\Pi}_{k-1}$ and $\bar{C}_{k-1}$ are the realized decisions made according to the strategy $\hat{a}_{k-1}$. Then $\breve{\Psi}_{k+1}^1$ is affine in Π_{k-1} and C_{k-1}. We can also obtain a variant of Ψ_{k+1}^2, say $\breve{\Psi}_{k+1}^2$, in exactly the same way. Since Ψ_{k+1}^3 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 induced by (24) with $M = M_2$, where

$$\begin{aligned} &M_2(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \\ &\triangleq \sum_{k=0}^{K-1} \big[\breve{\Psi}_{k+1}^1(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) + \breve{\Psi}_{k+1}^2(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) + \Psi_{k+1}^3(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \big]. \end{aligned} \quad (28)$$

Finally, we can easily justify that M_1 and M_2 are dual feasible penalties in the sense of (22).

Proposition 1. The functions M_1 in (27) and M_2 in (28) are dual feasible penalties, i.e., $M_1, M_2 \in \mathcal{M}_{\mathbb{G}}(0)$. Hence, $\mathcal{L}M_1$ and $\mathcal{L}M_2$ are upper bounds on H_0.

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 $\mathcal{G}_k$-adapted for $k = 0, \cdots, K-1$. We also note that Π_k is $\mathcal{G}_k$-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 $\mathcal{G}_k$ and (ϕ_0, W_0), we have for any (ϕ_0, W_0),

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

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 $M_1 \in \mathcal{M}_{\mathbb{G}}(0)$. Since the same argument can apply on $\breve{\Psi}_{k+1}^i(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})$ for $i = 1, 2$, it can be concluded that $M_2 \in \mathcal{M}_{\mathbb{G}}(0)$.

The penalties in the forms of (27) and (28) 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 (24) computationally tractable.

3.3 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 (17).

- Problem parameters: we consider a model with three risky assets ($n = 3$) and one market state variable ($m = 1$); $T = 1$ year and $\delta = 0.1$ year; the weight of the intermediate utility function: $\alpha = 0.5$; the initial condition: $\phi_0 = 0$ and $W_0 = 1$; control space: $\mathcal{A}$ in (16); $\mu_k^\phi = -\lambda\phi_k$, $\mu_k = \mu_0 + \mu_1\phi_k$, $\sigma_k \equiv \sigma$, $\sigma_k^{\phi,1} \equiv \sigma^{\phi,1}$, and $\sigma_k^{\phi,2} \equiv \sigma^{\phi,2}$. The values of r_f, λ, μ_0, μ_1, σ, $\sigma^{\phi,1}$, and $\sigma^{\phi,2}$ are listed in Table 1 and 2.

Table 1. Parameter Set 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 2. Parameter Set 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

We first derive $\hat{J}_k(\cdot)$, $\nabla_\phi \hat{J}_k(\cdot)$, and a suboptimal policy $\hat{\mathbf{a}}$ using the state-space discretization method: we approximate each ϕ_k with 21 evenly spaced grids from -2 to 2, and the random variables Z_k and $\tilde{Z}_k$ are approximated by Gaussian quadrature method with 3 points for each dimension(see, e.g., Judd [1998]). To compute the conditional expectation in (21), we simply ignore the correlation between ϕ_{k+1} and R_{k+1} and assume they are independent conditional on ϕ_k. For the optimization problem in (21) we use CVX, a package to solve convex optimization problems in MATLAB, to determine the optimal 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 $k = 0, \cdots, K$. Since the market state variable ϕ_k is one dimensional, the value function and the policy can be naturally defined on 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 $\hat{J}_k(\phi_k)$ and $\hat{\mathbf{a}}$; 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 apply the aforementioned policy $\hat{\mathbf{a}}$ to get an estimate of the lower bound on H_0 by generating 100 random sequences of $(\mathbf{Z}, \tilde{\mathbf{Z}})$, which is referred to as "Lower Bound" in Table 3; based on each random sequence we can solve the inner optimization problem (24) with the penalty M_1 in (27) and M_2 in (28), respectively, which leads to an estimate of upper bounds on H_0; these two bounds are referred to as "Dual Bound 1" and "Dual Bound 2". 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". We present our numerical results in Table 3 and 4: these bounds on H_0 are reported in the sub-column "Value", where each entry shows the sample average and the standard error (in parentheses) of 10 independent runs of the above procedures; in the sub-column "CE" we report the certainty equivalent of the expected utility (also reported in Cvitanic et al. [2003]), i.e., the equivalent wealth at time $T = 1$, where "CE" is defined through $U(\text{CE}) = \text{Value}$. For ease of comparison, in the column "Duality Gap" we report the smaller difference (in relative sense) between "Lower bound" and two "Dual Bounds" on the expected utility and its certainty equivalent.

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 the 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_{k+1}^3 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 H_0 are generally smaller when γ is small, implying that both the approximate policy and penalties are near optimal. As γ increases, the duality gaps of the expected utility generally become larger; however, the duality gaps in terms of "CE" are kept at a relatively constant range for different γ, which implies that the sub-optimal policies derived for different γ are good enough considering the the certainty equivalent wealth. There are several possible reasons for the enlarged duality gaps 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 (24a) is more sensitive to the solution of the inner optimization problem (24), and hence the quality of the penalty functions. In other words, even a small deviation from the optimal penalty will lead to significant deterioration on the quality 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 farther away from optimal for the discrete-time problem when γ increases. 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.

4. CONCLUSION

In this paper we study the dual formulation of controlled Markov diffusions based on the information relaxation

Table 3. Results with Parameter Set 1

	Lower Bound		Dual Bound 1		Dual Bound 2		Zero Penalty		Duality Gap	
γ	Value	CE	Value	CE	Value	CE	Value	CE	Value	CE
1.5	−5.480	0.1332	−5.391	0.1376	−5.392	0.1376	-4.861	0.1693	1.61%	3.30%
	(0.003)	(0.0001)	(0.008)	(0.0004)	(0.007)	(0.0004)	(0.012)	(0.0008)		
3.0	−42.887	0.1080	−39.227	0.1129	−39.873	0.1120	-27.562	0.1347	7.53%	3.70%
	(0.036)	(0.0001)	(0.164)	(0.0002)	(0.317)	(0.0004)	(0.252)	(0.0006)		
5.0	−2445.9	0.1005	−2066.5	0.1049	−2025.5	0.1054	-1105.7	0.1226	15.51%	4.38%
	(1.635)	(0.0001)	(22.019)	(0.0003)	(17.833)	(0.0002)	(16.438)	(0.0004)		

Table 4. Results with Parameter Set 2

	Lower Bound		Dual Bound 1		Dual Bound 2		Zero Penalty		Duality Gap	
γ	Value	CE	Value	CE	Value	CE	Value	CE	Value	CE
1.5	−5.466	0.1339	−5.380	0.1382	−5.381	0.1381	-4.864	0.1691	1.56%	3.14%
	(0.005)	(0.0001)	(0.011)	(0.0006)	(0.015)	(0.0008)	(0.020)	(0.0008)		
3.0	−42.585	0.1084	−39.645	0.1123	−39.690	1.1122	-27.708	0.1343	6.80%	3.51%
	(0.081)	(0.0001)	(0.229)	(0.0003)	(0.155)	(0.0002)	(0.209)	(0.0005)		
5.0	−2431.6	0.1007	−2043.8	0.1052	−2040.7	0.1052	-1122.1	0.1222	15.95%	4.47%
	(7.510)	(0.0001)	(11.881)	(0.0002)	(19.882)	(0.0003)	(9.842)	(0.0004)		

duality approach. This dual formulation can be used to derive a dual bound on the value function associated with the controlled diffusion. In particular, we explore the structure of the value function-based optimal penalty, which is the underpinning of developing near-optimal penalties that lead to tight dual bounds. We illustrate the use of this dual formulation in a dynamic portfolio choice problem that is discretized from a continuous-time model: we proposed a class of penalties that can be viewed as discretizing the value function-based penalty for the continuous-time problem, and these new penalties makes the inner optimization problem computationally tractable. These numerical studies complement the existing examples of applying the dual approach to continuous-state MDPs.

REFERENCES

D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 3rd edition, 2007.

M.W. Brandt, A. Goyal, P. Santa-Clara, and J.R. Stroud. A simulation approach to dynamic portfolio choice with an application to learning about return predictability. *Review of Financial Studies*, 18(3):831–873, 2005.

D.B. Brown and J.E. Smith. Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds. *Management Science*, 57(10):1752–1770, 2011.

D.B. Brown, J.E. Smith, and P. Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4):758 – 801, 2010.

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

J. Cvitanic, L. Goukasian, and F. Zapatero. Monte Carlo computation of optimal portfolios in complete markets. *Journal of Economic Dynamics and Control*, 27(6):971–986, 2003.

M.H.A. Davis and G. Burstein. Anticipative stochastic control. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pages 1830–1835, 1991.

M.H.A. 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*, 40(3-4):203–256, 1992.

D.P. de Farias and B. van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, pages 850–865, 2003.

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

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

J. Han and B. van Roy. Control of diffusions via linear programming. *Stochastic Programming*, pages 329–353, 2011.

M.B. Haugh, L. Kogan, and J. Wang. Evaluating portfolio policies: A duality approach. *Operations Research*, 54 (3):405–418, 2006.

K.L. Judd. *Numerical Methods in Economics*. The MIT press, 1998.

H.J. Kushner and P. Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24. Springer Verlag, 2001.

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*, 25(1):39–71, 1989.

W.B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley and Sons, 2nd edition, 2011.

L. C. G. Rogers. Pathwise stochastic optimal control. *SIAM J.Control Optimization*, 46(3):1116 – 1132, 2007.

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

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*, pages 870–876, 2012.

F. Ye and E. Zhou. Information relaxation and dual formulation of controlled Markov diffusions. *arXiv preprint arXiv:1303.2388*, 2013.