INFORMATION RELAXATION IN STOCHASTIC OPTIMAL CONTROL

A Thesis
Presented to
The Academic Faculty

by

Fan Ye

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2015

INFORMATION RELAXATION IN STOCHASTIC OPTIMAL CONTROL

Approved by:

Professor Enlu Zhou, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Shabbir Ahmed
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Alexander Shapiro
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Chelsea C. White III
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Fumin Zhang
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 22 April 2015

To my parents,

Pinhua Ye and Yunying Fang.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Professor Enlu Zhou, for her endless support and the wonderful opportunities she has provided throughout my doctoral study. She has guided me the direction of research, triggered valuable discussion, and passed on her knowledge in writing and presentations. Her insights inspired my deep thinking, and led me to overcome obstacles in research. Without her guidance, trust, and encouragement, I would not have been able to complete this thesis. Enlu was not only a mentor in my research, but also cared about my professional development and displayed a genuine concern for my well being.

I would also like to express gratitude to my other thesis committee members, Professor Shabbir Ahmed, Professor Alexander Shapiro, Professor Chelsea White, and Professor Fumin Zhang for their time, efforts, and valuable comments.

I have spent the first three years of my doctoral study in the University of Illinois at Urbana-Champaign and the last two years in Georgia Institute of Technology. I cherish these precious experiences very much. It was my great fortune to attend the lectures given by the distinguished faculty at both schools including Professor P. R. Kumar, Professor Sean P. Meyn, Professor Shabbir Ahmed, Professor Xin Chen, Professor Tamer Basar, Professor Pierre Moulin, Professor Richard Sowers, Professor Uday V. Shanbhag, and many others. Their inspiring lectures broadened my vision and enriched my knowledge, which underpined my thesis research. I greatly appreciate Professor Jong-Shi Pang, Professor Shabbir Ahmed, and Professor Chelsea White for their generosity to be my referees on many occasions.

I am also indebted to the great staff members such as Pamela Morrison, Judith

Norback, and Jennifer Harris at Georgia Tech, and Holly Michelle Kizer, Amy Summers, and Barbara Bohlen at the University of Illinois. I sincerely thank their time and help that have made my school life much easier.

The life at graduate school could have been very different without my good friends. I was fortunate to have the fellow group members Helin Zhu, Hale Joshua, Changhan Rhee, Xi Chen, and Di Wu. Helin has been my long-time collaborator and I have benefited a lot from our interactions. There are a number of close friends without whose help I would not have made my accomplishment so far. I would like to thank Tao Zhu for his help in my job search, and the road trip with him in the West has been a great memory of mine. I also missed the travel to Italy with Jingnan Chen, whom I have always enjoyed the conversation with. Qiushi Chen has amazed me by his musical talent and broad interests, and I greatly appreciate his optimistic attitude towards life. I am grateful to Shun Yao for his care and constant encouragement thorough times of my self-doubt. Especially, I owe a debt of gratitude to Hanlin Ouyang for his companionship and support during the past five years. I would also like to thank Can Zhang, Wenjia Wang, Zihao Li, Linwei Xin, Niao He, Weijun Xie, Jikai Zou, Fang Cao, and Chengliang Zhang at Georgia Tech, and Rui Yang, Ziqiao Ding, Yunwen Xu, Hui Xu, Zhenyu Hu, Xiangyu Gao, Chengcheng Guo, Tao Yang, Hao Jiang, Peng Hu and many others at the University of Illinois for their friendship and good time spent together.

Finally, I would like to thank my parents, Pinhua Ye and Yunying Fang, for raising me with their unconditional love. I appreciate their high expectation and unwavering belief in me. I owe more to them than I can ever express.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Dynamic programming is a principal method for analyzing stochastic optimal control problems. However, the exact computation of dynamic programming can be intractable in large-scale problems due to the "curse of dimensionality". Various approximate dynamic programming methods have been proposed to address this issue and they can often generate sub-optimal policies. However, it is generally difficult to tell how far these suboptimal policies are from optimal. To this end, this thesis concerns with studying the stochastic control problems from a duality perspective and generating upper bounds on maximal rewards (or lower bounds on minimal costs), which complements lower bounds on maximal rewards (or upper bounds on minimal costs) that can be derived by simulation under heuristic policies. If the gap between the lower and upper bounds is small, it implies that the heuristic policy must be close to optimal.

The approach considered in this thesis is called "information relaxation" (see [20, 77]), that is, it relaxes the non-anticipativity constraint that requires the decisions to depend only on the information available to the decision maker and impose a penalty that punishes such a violation. This methodology has been applied successfully in finite-horizon stochastic dynamic programs with various applications. This thesis further explores theories of information relaxation and computational methods in several stochastic optimal control problems.

First, we study the interaction of Lagrangian relaxation and information relaxation in weakly coupled dynamic program. A commonly studied approach builds on the property that this high-dimensional problem can be decoupled by dualizing the resource constraints via Lagrangian relaxation. We generalize the information relaxation approach, by generating penalties based on the Lagrangian relaxation bound, to improve upon the Lagrangian bound and also develop a computational method to tackle large-scale problems. We implement the algorithm on two examples and effectively reduce the duality gap between the performance of heuristic policies and Lagrangian bounds.

Second, we formulate the information relaxation-based duality in an important class of continuous-time decision-making models – controlled Markov diffusion, which is widely used in risk management and portfolio optimization. We find that this continuous-time model admits an optimal penalty in compact expression – an Ito stochastic integral, which enables us to construct approximate penalties in simple forms and achieve tight dual bounds, and to facilitate the computation of dual bounds significantly compared with that of the discrete-time model. We demonstrate its use in a dynamic portfolio choice problem subject to position and consumption constraints.

Third, we consider the problem of optimal stopping of discrete-time continuous-state partially observable Markov processes. We develop a filtering-based dual approach, which relies on the martingale duality formulation of the optimal stopping problem and the particle filtering technique. We carry out error analysis and illustrate the effectiveness of our method in an example of pricing American options under partial observation of stochastic volatility.

CHAPTER I

INTRODUCTION

1.1 Stochastic Optimal Control

Stochastic optimal control studies the sequential decision making problems in the presence of uncertainty, where the decision can earn profits or cost resources, and can also have an impact on the future by influencing the probabilistic dynamics. The goal of stochastic optimal control is to design control strategy such that it performs the desired task with maximal reward or minimum cost. In many situations, decisions that are chosen with the largest immediate profits may not be good in view of future events.

Stochastic optimal control has attracted the attention of researchers for decades because they are important from both the practical and intellectual point of view (see, e.g., [73, 9]). Its wide applications have been seen in different fields such as supply chain management [80], financial engineering [70], planning in robotic control [54], network models [65], and medical treatment decisions [2]. Among different types of the stochastic optimal control problems, Markov decision processes(MDPs) provide a powerful paradigm for modeling optimal decision making under uncertainty in the discrete-time setting. Though the optimal policies for such problems are generally known to exist and to satisfy the Bellman's principle of dynamic programming [6], the exact computation of optimal policies suffers from the "curse of dimensionality". Many approximate dynamic programming methods have been proposed to combat this curse of dimensionality such as [9, 10, 22, 72, 30, 85]. It is worth noting that these approximate dynamic programming methods often generate suboptimal policies, and simulation under a suboptimal policy leads to a lower bound on the optimal expected reward (or an upper bound on the optimal expected cost), though it is generally difficult to tell how far they are from the optimal ones.

The lack of performance guarantee on a sub-optimal policy can be potentially addressed by providing a dual bound, i.e., an upper bound (or lower bound) on the optimal expected reward (or cost). With such a complementary dual bound on the optimal value, the decision maker can easily evaluate the quality of heuristic policies and justify the need of improvement. It is worth noting that the problem-specific dual bound may be derived from relaxations of two intrinsic constraints in general stochastic optimal control problems. The first constraint is "resource constraint" or the feasibility of the control, which means the decision or control should take values in a feasible region. Another constraint is the "information constraint" or non-anticipativity of the control policy, that is, the decision should depend on the information up to the time that the decision is made. The relaxation of these constraints may lead to a simpler dynamic optimization problem: the resource constraint that exists universally in mathematical programs can be tackled by the commonly known Lagrangian relaxation (see, e.g., [8]), which relaxes the feasibility of the decisions and results in a less complicated unconstrained stochastic dynamic program; the information constraint can be approached by a recently developed technique – "*information relaxation*", which is proposed by [20] and [77]. The main idea of this approach is to relax the non-anticipativity constraint on decisions but impose a penalty for such a violation. In particular, a perfect information relaxation assumes that the decision maker can acquire all the system randomness in advance and allows her to make decisions based on the extra information; therefore, decisions are determined according to a scenario-based optimization problem that may be easier to solve than the original stochastic dynamic program, making this relaxation useful to evaluate the quality of sub-optimal policies in complex stochastic systems. There also exist other relaxation methods. For example, the LP-based approximate dynamic programming

(ALP) method proposed by [81] and [30] employs a parameterized class of functions to approximate the optimal value based on the linear programming formulation of the Bellman optimality equation; moreover, the ALP method not only provide approximate values that can be used to generate heuristic policies, but also upper bounds (or lower bounds) on the optimal expected rewards (or expected costs).

Recent years have seen growing research interests and numerous attempts to derive valid and tight information relaxation bounds based on the dual representation of MDPs or more general stochastic dynamic programs. This relaxation method has also found increasing applications including but not limited to natural gas storage valuation [57], dynamic portfolio optimization and execution [17, 66, 48], optimization of commodity procurement, processing and trade operations [34], and inventory management [20, 19].

1.2 Motivation of Thesis Research

Despite the growing interests in applying the information relaxation approach in various problems, there are many unanswered questions in both theoretical and computational aspects.

- First, prior research mainly focus on discrete-time and finite-horizon sequential decision making problems. A natural question is whether the idea of information relaxation can be extended to the continuous-time/infinite-horizon setting; more importantly, we aim to develop a tractable computational method that can be used to generate dual bounds in these settings.

- Second, many dynamic programs involve constraints on the controls or decision variables, which implies that the Bellman equation is a constrained stochastic optimization problem. To tackle the feasibility constraints on the controls, Lagrangian relaxation is a commonly used relaxation method that may simplify the original constrained dynamic program. On the other hand, information relaxation that relaxes non-anticipativity constraints of the controls, at first glance, contrasts with the Lagrangian relaxation that relaxes the feasibility of the controls. We expect to use a unifying framework to interpret both relaxations; furthermore, we are also interested in the interaction between the two relaxations.

- Third, the information relaxation approach can be naturally applied in the optimal stopping problems under imperfect state information, since a well-known technique can be used to transform a partially observable MDP to a fully observable one with belief state. However, this belief state could be infinite-dimensional in general and hence become intractable to represent, making it less straightforward to apply the information relaxation approach. Moreover, it is generally difficult to characterize in theory the gap between the optimal values of the optimal stopping problems under perfect and partial observations. Therefore, computing a tight dual bound on the optimal value of the partially observable problem has the potential to numerically capture this difference.

1.3 Contributions

This thesis attempts to develop theories of information relaxation and computational methods in several stochastic optimal control problems to address the aforementioned questions.

- We consider weakly coupled dynamic program, which describes a broad class of stochastic optimization problems in which multiple controlled stochastic processes evolve independently but subject to a set of linking constraints imposed on the controls. For example, a supplier needs to dynamically allocate the limited capacity among different customers with stochastic demand, in order to maximize the expected profits. One feature of the weakly coupled dynamic program is that it decouples into lower-dimensional dynamic programs by dualizing the linking constraint via the Lagrangian relaxation, which also yields a bound on the optimal value of the original dynamic program. Together with the Lagrangian relaxation, we generalize the information relaxation approach to obtain a guaranteed tighter dual bound than the Lagrangian bound. To tackle the large-scale problems, we further propose a computationally tractable method by relaxing the weakly-coupled inner optimization problem. Preliminary results of this work appear in our paper [94].

- Though the information relaxation approach can be generalized to the infinite-horizon MDPs in a straightforward way, the main difficulty is that the inner optimization problem involves infinite number of decision variables due to the future information of infinite length. To adapt this dual approach to the infinite-horizon problem, we consider a randomization idea to reformulate the original problem such that the inner optimization problem within the information relaxation approach is of finite-time horizon.

- We develop the information relaxation-based dual formulation of an important class of the continuous-time stochastic optimal control problems – the controlled Markov diffusions. Based on the technical machinery "anticipating stochastic calculus" (see, e.g., [68, 67]), we establish the weak duality, strong duality and complementary slackness results in parallel as those in the dual formulation of MDPs. We investigate one type of optimal penalties, i.e., the so-called "value function-based penalty", which admits a stochastic integral form under the natural filtration generated by the Brownian motion. This compact expression potentially enables us to design sub-optimal penalties in simple forms and also facilitates the computation of the dual bound. An application is illustrated by a 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 compute is proposed to derive dual bounds on the optimal value of the discrete-time model. The development of this framework and technical details appear in our papers [95, 98].

- We propose a filtering-based duality approach for partially observable optimal stopping problem, in order to complement the suboptimal policy with an asymptotic upper bound on the value function. This method relies on the martingale duality formulation of the optimal stopping problem. Our work focuses on employing the particle filtering technique, which is used to approximate general filtering distribution, to generate penalty (i.e., the martingale term) in the dual formulation via Monte Carlo simulation. We apply our approach to price American put options in stochastic volatility models, under more realistic assumption that the volatility cannot be directly observed but can be inferred from the asset prices. The numerical results confirm a higher price of the option if we alternatively assume that the volatility is directly observable. The price difference becomes more significant when the effect of volatility is high, indicating the importance of taking the partial observability into account. The development and analysis of the approach appear in our papers [97, 96].

1.4 Thesis Outline

The rest of the dissertation is organized as follows.

Chapter 2 provides the background and literature review on information relaxation and duality in stochastic dynamic programs and martingale dual representation of optimal stopping problems.

Chapter 3 discusses the interaction of Lagrangian relaxation and information relaxation in weakly coupled dynamic program that is formulated as a discounted infinite horizon MDP. We develop a computational method that involves the idea of time randomization and relaxing the inner optimization problem. We implement the algorithm on a restless bandit problem and a linear quadratic control problem under non-convex linking constraints. Our method effectively reduces the duality gap between the heuristic policy and Lagrangian bound.

Chapter 4 characterizes the dual formulation of controlled Markov diffusions. We establish the duality results in a parallel way as those in Markov decision processes. We further explore the structure of the optimal penalties and expose the connection between the optimal penalties for Markov decision processes and controlled Markov diffusions. We demonstrate the use of this dual representation in a classic dynamic portfolio choice problem through a new class of penalties, which require little extra computation and produce small duality gap on the optimal value.

Chapter 5 presents a filtering-based duality approach to solve the discrete-time continuous-state partially observable optimal stopping problem. This method relies on the martingale duality formulation of the optimal stopping problem and the particle filtering technique. We show that this approach complements an asymptotic lower bound derived from a suboptimal stopping time with an asymptotic upper bound on the value function. We carry out error analysis and illustrate the effectiveness of our method on an example of pricing American options under partial observation of stochastic volatility.

Chapter 6 concludes the dissertation and outlines some future research.

CHAPTER II

INFORMATION RELAXATION

In this section we review the information relaxation-based dual formulation of Markov decision processes(MDPs) and the martingale duality formulation of optimal stopping problems.

2.1 Duality in Markov Decision Process

Consider a finite-horizon MDP on the probability space $(\Omega, \mathcal{G}, \mathbb{P})$, where Ω is the set of possible outcomes or scenarios ω, $\mathcal{G}$ is an σ-algebra containing the events in Ω, and P is a probability measure. 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

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

where $a_k \in \mathcal{A}_k \subset \mathcal{A}$ is the control or decision variable chosen at time k, and $\{v_k\}_{k=0}^{K-1}$ are independent random variables taking values in the set $\mathcal{V}$ with known distributions. The natural filtration associated with this MDP is denoted by $\mathbb{G} = \{\mathcal{G}_0, \cdots, \mathcal{G}_K\}$, where $\mathcal{G}_k$ is the σ-algebra generated by $\{x_0, a_0, v_1, a_2, v_2, \cdots, a_k, v_k\}$; in particular, $\mathcal{G}_0 = \sigma\{x_0\}$. Therefore, $\mathcal{G}_k$ contains the information that is known to the decision maker at the beginning of time k.

Note that a scenario $\omega \in \Omega$ refers to a realization of $\mathbf{v} = \{v_1, \cdots, v_K\}$. Given a scenario $\omega \in \Omega$, the decision maker chooses a sequence of controls $\mathbf{a} = (a_0, \cdots, a_{K-1})$ with $a_k \in \mathcal{A}_k$. Such a selection is called a control policy, i.e., $\alpha : \Omega \to \mathcal{A}_0 \times \cdots \times \mathcal{A}_{K-1}$. We denote the set of such control policies as $\mathbb{A}$.

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 $\mathcal{G}_k$-adapted. We also call $\alpha \in \mathbb{A}_{\mathbb{G}}$ a *non-anticipative* policy. Given the initial

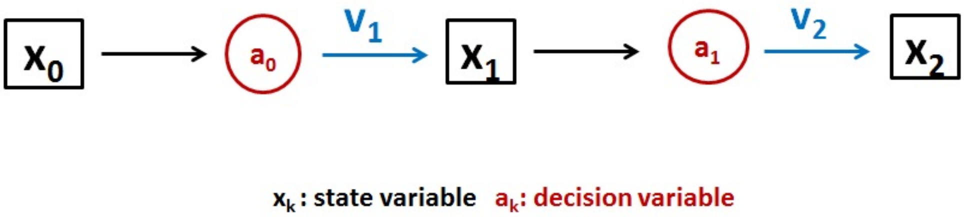

Figure 1: Markov Decision Process

condition $x_0 \in \mathcal{X}$, the objective is to maximize the expected sum of intermediate rewards $\{g_k(x_k, a_k)\}_{k=0}^{K-1}$ (that depend on the state and the control) and final reward $\Lambda(x_K)$ (that depends only on the final state) by selecting a non-anticipative policy $\alpha \in \mathbb{A}_{\mathbb{G}}$:

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

where a_k is selected by α depending on the scenario ω, and the expectation in (2) is taken over ω or all possible realizations of $\mathbf{v}$. The value function V_0 is a solution to the following dynamic programming recursion:

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

Next we describe the information relaxation-based dual formulation of Markov decision process. Here we only consider the *perfect information relaxation*, i.e., we have full knowledge of the future randomness.

Define $\mathbb{E}_{k,x}[\cdot] \triangleq \mathbb{E}[\cdot|x_k = x]$. 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 $\mathcal{L} : \mathcal{M}_{\mathbb{G}}(0) \to \mathcal{D}$:

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

Note that the supremum in (3) is over the set $\mathbb{A}$ not the set $\mathbb{A}_{\mathbb{G}}$, i.e., the control or decision a_k can be selected based on the 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 following inner optimization problem:

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

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

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

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

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

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

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

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 $M \in \mathcal{M}_{\mathbb{G}}(0)$, i.e., $I(x_0, M, \mathbf{v})$ is a high-biased estimator of $V_0(x_0)$ for all $x_0 \in \mathcal{X}$; Theorem 1(b) claims 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 [20])**.**

(a) (Weak Duality) For all $M \in \mathcal{M}_{\mathbb{G}}(0)$ and all $x_0 \in \mathcal{X}$, $V_0(x_0) \leq (\mathcal{L}M)(x)$.

(b) (Strong Duality) For all $x_0 \in \mathcal{X}$, $V_0(x_0) = (\mathcal{L}M^*)(x)$, where

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

In addition, the following equality holds almost surely, i.e.,

$$V_0(x_0) = \sup_{\boldsymbol{a}\in\mathbb{A}} \left\{ \sum_{k=0}^{K-1} g_k(x_k, a_k) + \Lambda(x_K) - M^*(\boldsymbol{a}, \boldsymbol{v}) \right\} \quad a.s..$$

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 equation (1).*

2. *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, $M^*(\boldsymbol{a}, \boldsymbol{v}) \in \mathcal{M}_{\mathbb{G}}(0)$. Since M^* depends on the value function $\{V_k\}$, it is referred to as the* value function-based penalty*.*

2.2 Duality in Optimal Stopping

Optimal stopping is concerned with the problem of choosing a time to take a particular action based on sequentially observed random variables whose joint distribution is known, in order to maximize an expected reward or to minimize an expected cost. References [76, 45] use the martingale duality approach to compute upper bounds on the prices of American options, which is essentially an optimal stopping problem. This martingale-based dual approach can be viewed as a case of the perfect information relaxation.

Consider a finite-horizon Markov process $\{x_k\}_{k\in\mathcal{K}}$ on the probability space $(\Omega, \mathcal{G}, \mathbb{P})$, where time is indexed by $\mathcal{K} = \{0, 1, \cdots, K\}$ and the transition probability $P_k(x_{k+1}|x_k)$ is known. The filtration generated by the processes $\{x_k\}_{k\in\mathcal{K}}$ is denoted by $\mathbb{G} = \{\mathcal{G}_0, \cdots, \mathcal{G}_K\}$ with $\mathcal{G}_k \triangleq \sigma\{x_0, \cdots, x_k\}$.

A random variable $\tau : \Omega \to \mathcal{K}$ is a $\mathcal{G}_k$-stopping time if $\{\tau \leq k\} \in \mathcal{G}_k$ for every $k \in \mathcal{K}$. We define $\mathbb{S}_{\mathbb{G}}$ as the set of $\mathcal{G}_k$-stopping times that take values in $\mathcal{K}$. We consider the finite-horizon optimal stopping problem assuming that the initial x_0 is known:

$$V_0(x_0) = \sup_{\tau \in \mathbb{S}_{\mathbb{G}}} \mathbb{E}\left[g(\tau, x_\tau) \middle| x_0 \right], \tag{6}$$

where $g(k, \cdot)$ is the reward at time k that only depends on the state x_k.

The optimal stopping problem is a special case of the Markov decision process. The only differences are (i) the state dynamic is an uncontrolled process and (ii) the only decision to be made at each stage is to "stop" or "continue" the process, that is, to compare the immediate profit based on the current state information and the expected reward considering all the future outcome. Therefore, the value function V_0 can also be solved by a dynamic program:

$$\begin{aligned} &V_K(x_K) \triangleq g(K, x_K); \\ &V_k(x_k) \triangleq \max\{g(k, x_k), \mathbb{E}[V_{k+1}(x_{k+1})|x_k]\}, \ k = K-1, \cdots, 0. \end{aligned}$$

Theorem 2 (Theorem 2.1 in [76]). *Let $\mathcal{M}$ represent the space of $\mathcal{G}_k$-adapted martingales $\{M_k\}$ with $M_0 = 0$ and $\sup_{k \in \mathcal{K}} \mathbb{E}[|M_k|] < \infty$. Then*

$$V_0(x_0) = \min_{M \in \mathcal{M}} \left\{ \mathbb{E}\left[\max_{k \in \mathcal{K}} \{g(k, x_k) - M_k\} | X_0 = x_0 \right] \right\}. \tag{7}$$

The optimal martingale $\{M_k^*\}$ that achieves the minimum on the right hand side of (7) is of the form

$$M_k^* = \sum_{j=1}^{k} (V_j(x_j) - \mathbb{E}[V_j(x_j)|x_{j-1}]), \tag{8}$$

In addition, the following equality holds in the almost sure sense, i.e.,

$$V_0(x_0) = \max_{k \in \mathcal{K}} (g(k, x_k) - M_k^*) \quad a.s..$$

Theoretically, the strong duality results hold in both Markov decision processes and optimal stopping problems. However, the optimal penalty (5) and the optimal martingale (8) that achieve the strong duality involve the value function $\{V_k\}$, and hence are intractable in practical problems. The hope is that we can construct a good approximation of the optimal penalties based on some approximate value functions $\{\hat{V}_k\}$ or sub-optimal policy $\hat{\mathbf{a}}$ (or τ), which may result in a tight dual bound by incorporating the approximate penalties in the dual approach. Methods based on these ideas have been successfully implemented in the American option pricing problems [76, 45, 3], and also in various stochastic dynamic programs [20, 57, 34].

2.3 Literature Review

Information relaxation-based duality for general stochastic dynamic program [20, 77] builds on the research work of American option pricing (or optimal stopping) and stochastic programming. [76, 45] propose a general algorithm for constructing an upper bound on American option price based on its dual formulation (see section 2.2), which complements the lower bound derived from the approximate dynamic programming method [60]. The same duality technique was also developed in the earlier work [28]. In particular, [76] uses the approximate value function to generate penalties (or "dual martingale" in their terminology), while [3] develops an alternative computational algorithm by using the approximate policies. To improve the quality of dual bounds, [23] proposes an iterative approach to construct a sequence of the dual martingales, and [33] considers the idea of parameterized martingales and uses a convex optimization procedure to produce upper and lower bounds. In the special cases that the asset process is modeled as a diffusion process or a jump-diffusion process, the structure of the optimal martingale (i.e., the optimal penalty) is investigated by [7, 91, 101], which leads to practical algorithms for fast computation of tight upper bounds on the American option prices. A nice overview of American option pricing can be found in [40].

The idea of relaxing the non-anticipativity has also been studied in stochastic programming literature [83, 75]. The stochastic programming formulation requires that the cost functions and the set of feasible action (or control set) to be convex and penalties to be linear in actions. In contrast, the information relaxation approach does not impose any assumption on the convexity of the cost functions. In case that the reward function is convex in the feasible actions, [17] develops a gradient-based optimal penalty, which is equivalent to the Lagrangian term that dualizes the non-anticipativity constraints in stochastic programming (see Chap 3.2.4 of [83]).

The relaxation of the non-anticipativity constraint on the control policies in MDPs also has a long history, dating back at least to [29]. Due to the work [20, 77], the information relaxation technique has attracted researchers' attention in both theoretical and practical aspects during recent years. It is worth noting that the optimal penalty is not unique: for general problems we have the value function-based penalty derived in [77] and [20]; for problems with convex structure there is a class of alternative optimal penalties [19], which extends the aforementioned gradient-based penalty in [17]. It is shown in [47] that the value function-based penalty and the gradient-based penalty are different in linear quadratic control problems. In order to derive tight dual bounds, various algorithms based on different approximation schemes [20, 17, 19] and the idea of parameterized penalties [32, 93] have been proposed. In addition, [55] studies a robust model of the multi-armed bandits using the information relaxation approach. [44] extends this dual approach to the zero-sum game. Information relaxation has found various applications such as natural gas storage valuation [57], dynamic portfolio optimization and execution [17, 66, 48], optimization of commodity procurement, processing and trade operations [34], and inventory management [20, 19].

CHAPTER III

LAGRANGIAN AND INFORMATION RELAXATIONS IN WEAKLY COUPLED DYNAMIC PROGRAM

We study the interactions between Lagrangian relaxation and information relaxation in the weakly coupled dynamic program(WCDP), which consists of multiple subproblems that are independent of each other except for a set of budget or linking constraints on the controls (see, e.g., [49, 1]). The WCDPs have many interesting and practical applications including multi-armed and restless bandits ([52, 39, 92, 12]), resources allocation ([41]), network revenue management ([86, 89]), and optimal learning ([11, 21, 38]). Unfortunately, the exact solution to WCDPs quickly becomes intractable as the number of subproblems increases. Therefore, we resort to heuristic policies as well as a good performance bound in high-dimensional problems. [49] decomposes the original problem by dualizing the linking constraints, which leads to a dual bound on the optimal value. Despite the computational advantage of Lagrangi5an relaxation, we cannot expect the bound to be tight, because the general WCDPs may lack the convex structure. Some recent literatures suggest two main approaches for improving the Lagrangian relaxation bound. The first approach in [1] shows that the ALP method can be used to obtain a tighter upper bound compared with the Lagrangian bound, and it has been successfully implemented in large-scale bandit-like problems. The second approach in [19] studies how to improve the Lagrangian bound using the information relaxation approach, and develops a gradient penalty for computing the bound in convex WCDPs. However, these approaches have their own limitations. For example, the gradient penalty approach in [19] is not suitable for nonconvex problems, and the efficiency of the ALP method may deteriorate quickly with increasing dimensions of linking constraints due to the implementation of the column generation. Another issue of the ALP method is that the column generation procedure may require designing problem-specific sampling technique to achieve the optimal solution quickly, when the stochastic decision model has continuous states or decisions.

In this paper, we consider an alternative approach that utilizes the information relaxation technique to generate upper bounds on the optimal value of a general class of WCDPs. Our approach does not require convexity assumption and can apply on infinite-horizon problems with discounted rewards. Our work is partly motivated by the work of [19] that considers finite-horizon WCDPs with convex structure, and the work of [32] that studies the ALP and the information relaxation methods. To apply the state-of-art of the information relaxation technique in an infinite-horizon discounted WCDP, there are several challenges. First, a perfect information relaxation means that the system randomness of infinite length is revealed beforehand, which implies that the associated scenario-based inner problem has infinite number of decision variables and can be difficult to solve. To address this problem, we use a standard technique in simulation - a geometric distributed randomized time - to convert the discounted infinite-horizon inner problem to a finite (but random) horizon problem (see, .e.g., [37]). This reformulation makes it possible to solve the scenario-based inner problem with finite computational costs, though the costs depend on the length of the random horizon that is affected by the discount factor. By coupling the randomized time with penalties derived from approximate value functions, we can adapt the weak and strong duality to the discounted infinite-horizon problem, which parallels the results in finite-horizon stochastic dynamic programs (see, e.g., [20]). We also observe in principle the information relaxation approach can always generate a tighter bound than the approximate value function, as long as it is a supersolution to the Bellman's equation. In particular, both Lagrangian bound and

ALP bound qualify as supersolutions. We then compare the respective sufficient and necessary conditions such that Lagrangian relaxation and information relaxation give tight bounds on the optimal value. We provide an example in which the Lagrangian bound can be arbitrarily loose, whereas the information relaxation bound that builds upon it is always tight.

Despite the finite-horizon scenario-based inner problem, solving its optimal solution effectively remains another main challenge, especially when the number of subproblems is large: with a fixed scenario, the state transition is purely determined by the decision subject to linking constraints; therefore, finding the optimal decisions suffers from the curse of dimensionality (in terms of the number of subproblems) and becomes more difficult in the possible long-horizon problem. Instead of computing its optimal value exactly, we solve a relaxed problem by dualizing linking constraints for the purpose of decomposition, which leads to an upper bound on the optimal value that is computationally tractable for each scenario. With this relaxation we generally have a weaker bound (referred to as the "practical information relaxation bound") compared with the exact information relaxation bound, but we can show it is still superior to the Lagrangian relaxation bound; therefore, this relaxation lies intermediately between the Lagrangian relaxation and the exact information relaxation. We also provide theoretical analysis on the relative gap between the exact and practical information relaxation bounds: with certain conditions on the linking constraints, this relative gap will vanish as the number of subproblems goes to infinity.

Overall, we can compute various dual bounds via relaxations with different performance guarantees and computational complexities. Due to the trade-off between the quality versus complexity, we may start with one relaxation that requires the least computational cost, and based on its bound performance we may decide how much more we should invest to derive better policies or/and tighter dual bounds.

3.0.1 Literature Review

The weakly coupled dynamic program was first systematically studied in [49], which employs Lagrangian relaxation to develop heuristic policies in high-dimensional problems. The work [1] shows that the ALP method can guarantee a tighter dual bound than the Lagrangian bound, though Lagrangian relaxation is much easier to compute. The WCDPs have many interesting and practical applications including multi-armed and restless bandits in [52, 39, 92, 12], resources allocation ([41]), network revenue management ([86, 89]), and optimal learning ([11, 21, 38]).

Recently, information relaxation was developed in [20] to study performance bounds in general dynamic programs. [77] independently proposes a dual formulation of Markov decision process that can be interpreted as perfect information relaxation. Information relaxation was further explored in different settings such as convex dynamic programs in [17, 19], continuous-time stochastic control in [95], and zero-sum stochastic games in [44].

There are several works related to information relaxation and the subject of our study. The work [19] develops a gradient-based penalty method to compute dual bounds on the revenue in an airline network problem, which is a case of WCDPs; their method can also be applied in general convex stochastic dynamic programs. [32] explores the theoretical formulation of the information relaxation bound in the infinite-horizon discounted MDP and compared it to the ALP bound. A recent independent work by [16] studies the infinite-horizon discounted MDP using a change-of-measure technique called "weak formulation"; they also consider the "strong formulation" (i.e., the randomized time) and solve the relaxed inner problem, but they do not characterize the relative gap between the exact and practical information relaxation bounds. The work [55] studies a robust model of the multi-armed bandit using the information relaxation approach. [18] applies the information relaxation and decouple the inner problem in an optimal sequential exploration problem (a generalized multi-armed bandit with dependent arms) without using penalty. The work [17] and [82] use information relaxation to derive tighter dual bounds from the optimal value or policy of a simplified model in a dynamic portfolio optimization problem and a commodity storage problem, respectively. In stochastic programming literature, [31] studies variants of Lagrangian relaxations and the associated decomposition scheme with duality gaps in nonconvex stochastic optimization problems. The work [13] shows the structure of stochastic integer programs that leads to a vanishing Lagrangian duality gap as the number of scenarios increases.

The rest of this chapter is organized as follows. In Section 3.1, we review the formulation of the weakly coupled dynamic program and its decomposition using the Lagrangian relaxation approach. In Section 3.2, we present the information relaxation-based dual bounds for the infinite-horizon problem, and compare it to the Lagrangian relaxation and ALP method. In Section 3.3, we address the computational issue of the information relaxation bound in the large-scale setting. We present our numerical studies in Section 3.4, and provide the concluding remarks in Section 3.5.

3.1 Formulation of the Weakly Coupled Dynamic Program

In this section, we present the general framework of the weakly coupled stochastic dynamic program and the Lagrangian relaxation approach.

3.1.1 Problem Formulation

Consider a collection of N projects or subproblems labeled by $n = 1, \cdots, N$. The state of each project or subproblem transits independently according to a homogenous transition law and yields a reward that is dependent only on the individual state and control. However, at each time period there are constraints imposed on the controls of these projects, which are referred to as the "linking constraints" or "budget constraints". The underlying probability space is described by $(\Omega, \mathcal{F}, P)$, where Ω is the set of possible outcomes or scenarios ω, $\mathcal{F}$ is a σ-algebra containing the events in Ω, and P is a probability measure.

We use the following notations to describe the mathematical formulation of the weakly coupled stochastic optimization problem.

1. Time is indexed by $t = 0, 1, 2 \cdots$.

2. $\mathbf{x}_t = (x_t^1, \cdots, x_t^N)$ is the joint state of the N projects, and it takes value in the state space $\mathcal{X} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^N$.

3. $\mathbf{a}_t = (a_t^1, \cdots, a_t^N)$ is the control (or decision variable) that takes value in the control (or action) space $\mathcal{A} = \prod_{n=1}^N \mathcal{A}^n$.

4. The state of N-project transits in a Markovian fashion; in particular, it evolves as N independent Markov decision processes according to a known homogenous transition law
$$P(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t) = \prod_{n=1}^N P_n(x_{t+1}^n | x_t^n, a_t^n),$$
where $\{P_n\}_{n=1}^N$ denotes the controlled transition probability of the individual project. Note that each state $\mathbf{x}_{t+1}$ depends on the prior control sequence $\mathbf{a}(t) \triangleq (\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_t)$ and the scenario ω, i.e., $\mathbf{x}_0 = \mathbf{x}_0(\omega)$ and $\mathbf{x}_{t+1} = \mathbf{x}_{t+1}(\mathbf{a}(t), \omega)$ for $t \geq 0$, where ω represents the underlying uncertainty.

5. At period t the control $\mathbf{a}_t$ is chosen by the decision maker subject to a set of L time-invariant linking constraints $\sum_{n=1}^N \mathbf{B}^n(x_t^n, a_t^n) \leq \mathbf{b}$, where $\mathbf{b} \in \mathbb{R}^L$. Denote the feasible control space at time t by
$$\bar{\mathcal{A}}_t = \{\mathbf{a}_t \in \mathcal{A} : \ \mathbf{B}(\mathbf{x}_t, \mathbf{a}_t) \triangleq \sum_{n=1}^N \mathbf{B}^n(x_t^n, a_t^n) \leq \mathbf{b}\}. \tag{9}$$
Here, the dependence of $\bar{\mathcal{A}}_t$ on the state $\mathbf{x}_t$ is omitted for convenience.

6. At period t the n-th project or subproblem receives a reward of $R^n(x_t^n, a_t^n)$. The total reward received at time t is of the additive form
$$R(\mathbf{x}_t, \mathbf{a}_t) \triangleq \sum_{n=1}^{N} R^n(x_t^n, a_t^n).$$

7. Given a scenario ω, the decision maker chooses a sequence of controls $\mathbf{a} = (\mathbf{a}_0, \mathbf{a}_1, \cdots)$, where each $\mathbf{a}_t$ takes value in $\bar{\mathcal{A}}_t$. Such a selection is called a control policy, i.e., $\alpha : \Omega \to \bar{\mathcal{A}}_0 \times \bar{\mathcal{A}}_1 \times \cdots$. We denote the set of such control policies as $\bar{\mathbb{A}}$.

8. The filtration $\mathbb{F} = \{\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \cdots\}$ describes the evolution of the state information, where $\mathcal{F}_0 \triangleq \sigma\{\mathbf{x}_0\}$ and $\mathcal{F}_t \triangleq \sigma\{\mathbf{x}_0, \cdots, \mathbf{x}_t, \mathbf{a}_0, \cdots, \mathbf{a}_{t-1}\}$ for $t \geq 1$. Since the decision maker determines $\mathbf{a}_t$ based only on the information known up to period t, each $\mathbf{a}_t$ is then $\mathcal{F}_t$-measurable; we call such a control policy α to be non-anticipative and denote the set of non-anticipative policies by
$$\bar{\mathbb{A}}_{\mathbb{F}} = \{\alpha \in \bar{\mathbb{A}} |\ \alpha \text{ is non-anticipative}\}.$$

9. The expected discounted infinite-horizon reward induced by a control policy α is
$$V(\mathbf{x}_0; \alpha) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R\left(\mathbf{x}_t, \mathbf{a}_t\right) \middle| \mathbf{x}_0\right], \tag{10}$$
where $\beta \in (0, 1)$ is a discount factor, and $\mathbf{a}_t$ is selected by α depending on the scenario ω. The objective of the decision maker is to maximize the expected infinite-horizon reward over all non-anticipative policies, given the initial condition $\mathbf{x}_0 \in \mathcal{X}$:
$$V(\mathbf{x}_0) = \sup_{\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}} V(\mathbf{x}_0; \alpha). \tag{11}$$

To avoid technical complication, we assume that $\{R^n\}_{n=1}^N$ are uniformly bounded on their respective domain (therefore, V is also bounded), and the supremum in (11) can be achieved (this is the case, for example, when $\mathcal{X}$ and $\mathcal{A}$ are finite). So V is well-defined for all $\mathbf{x}_0 \in \mathcal{X}$. Thus, the exact solution to (11) can be obtained by solving the following Bellman optimality equation:

$$V(\mathbf{x}_0) = \max_{\mathbf{a}_0 \in \bar{\mathcal{A}}_0} \left\{ R(\mathbf{x}_0, \mathbf{a}_0) + \beta \mathbb{E}\left[V(\mathbf{x}_1) | \mathbf{x}_0, \mathbf{a}_0 \right] \right\}. \tag{12}$$

We assume (12) have an optimal stationary and Markov policy α^*, where $\alpha^* : \mathcal{X} \to \mathcal{A}$ satisfies

$$\alpha^*(\mathbf{x}_0) \in \arg\max_{\mathbf{a}_0 \in \bar{\mathcal{A}}_0} \{ R(\mathbf{x}_0, \mathbf{a}_0) + \beta \mathbb{E}\left[V(\mathbf{x}_1) | \mathbf{x}_0, \mathbf{a}_0 \right] \}.$$

The standard value iteration or policy iteration algorithm that can be used to solve (12) becomes intractable as N increases, since the size of its state space is $|\mathcal{X}| = \prod_{n=1}^{N} |\mathcal{X}^n|$.

3.1.2 Lagrangian Relaxation

In this subsection we consider the Lagrangian dual of (12) that relaxes the linking constraints on the controls. The motivation of relaxing the linking constraint is to decompose the original high-dimensional problem to several low-dimensional subproblems.

Denote by $\mathbb{A} \triangleq \{\alpha : \Omega \to \mathcal{A} \times \mathcal{A} \times \cdots\}$, which contains $\bar{\mathbb{A}}$ as a subset. By dualizing the linking constraint with the Lagrangian multiplier $\boldsymbol{\lambda} \in \mathbb{R}_+^L$, we define $J^{\boldsymbol{\lambda}}(\mathbf{x}_0)$ for $\mathbf{x}_0 \in \mathcal{X}$:

$$J^{\boldsymbol{\lambda}}(\mathbf{x}_0) \triangleq \max_{\alpha \in \mathbb{A}_{\mathbb{F}}} J^{\boldsymbol{\lambda}}(\mathbf{x}_0; \alpha), \tag{13}$$

where

$$J^{\boldsymbol{\lambda}}(\mathbf{x}_0; \alpha) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \left(R(\mathbf{x}_t, \mathbf{a}_t) + \boldsymbol{\lambda}^\top \left[\mathbf{b} - \mathbf{B}(\mathbf{x}_t, \mathbf{a}_t) \right] \right) \middle| \mathbf{x}_0 \right],$$

and $\mathbb{A}_{\mathbb{F}} \triangleq \{\alpha \in \mathbb{A} |\ \alpha \text{ is non-anticipative}\}$.

We list some properties of $J^{\boldsymbol{\lambda}}$ in Lemma 1; in particular, $J^{\boldsymbol{\lambda}}$ is an upper bound on V given any $\boldsymbol{\lambda} \geq 0$, which will be referred to as the "Lagrangian bound" in the following.

Lemma 1 (Properties of $J^{\boldsymbol{\lambda}}$). *1. For any $\boldsymbol{\lambda} \geq 0$, $J^{\boldsymbol{\lambda}}(\boldsymbol{x}) \geq V(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$.*

2. $J^{\boldsymbol{\lambda}}(\boldsymbol{x})$ is convex and piecewise linear in $\boldsymbol{\lambda} \geq 0$.

3. For all $\boldsymbol{x} \in \mathcal{X}$, $J^{\boldsymbol{\lambda}}(\boldsymbol{x})$ can be written as

$$J^{\boldsymbol{\lambda}}(\boldsymbol{x}) = \frac{\boldsymbol{\lambda}^\top \boldsymbol{b}}{1-\beta} + \sum_{n=1}^{N} J^{\boldsymbol{\lambda},n}(x^n), \tag{14}$$

where $H^{\boldsymbol{\lambda},n}(x_0^n)$ is the solution to the following Bellman optimality equation for each $n = 1, \cdots, N$:

$$J^{\boldsymbol{\lambda},n}(x_0^n) = \max_{a_0^n \in \mathcal{A}^n} \left\{ R^n(x_0^n, a_0^n) - \boldsymbol{\lambda}^\top \boldsymbol{B}^n(x_0^n, a_0^n) + \beta \mathbb{E}\left[J^{\boldsymbol{\lambda},n}(x_1^n) | x_0^n, a_0^n \right] \right\}. \tag{15}$$

The proof of these results can be found in Theorem 1 and Theorem 2 of Section 2 in [49], or Proposition 1 and Proposition 2 in [1].

In the case that $\mathcal{X}$ and $\mathcal{A}$ are finite, we may compute the tightest Lagrangian bound over $\boldsymbol{\lambda} \geq 0$ via a linear program. To be more specific, suppose $\{\upsilon(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ is a probability distribution on $\mathcal{X}$, which can be viewed as the initial distribution of $\mathbf{x}_0$. Let $\upsilon_n(\cdot)$ denote the marginal distribution of υ with respect to the n-th project, i.e, $\upsilon_n(x_0^n) = \sum_{\{\mathbf{x}=(x^1,\cdots,x^n) \in \mathcal{X}: x^n = x_0^n\}} \upsilon(\mathbf{x})$. From (14) we define the Lagrangian bound based on the initial distribution υ as the weighted sum

$$\sum_{\mathbf{x} \in \mathcal{X}} \upsilon(\mathbf{x}) \cdot J^{\boldsymbol{\lambda}}(\mathbf{x}) = \frac{\boldsymbol{\lambda}^\top \mathbf{b}}{1-\beta} + \sum_{n=1}^{N} \sum_{x^n \in \mathcal{X}_n} \upsilon_n(x^n) J^{\boldsymbol{\lambda},n}(x^n).$$

The optimal $\boldsymbol{\lambda}^* = \arg\min_{\{\boldsymbol{\lambda} \geq 0\}} \{\sum_{\mathbf{x} \in \mathcal{X}} \upsilon(\mathbf{x}) \cdot J^{\boldsymbol{\lambda}}(\mathbf{x})\}$ and the corresponding $\{J^{\boldsymbol{\lambda}^*,n}(\cdot)\}_{n=1}^N$ can be determined by the following linear program (with variables $\boldsymbol{\lambda}$ and $\{H^n(\cdot)\}_{n=1}^N$).

$$\min_{\boldsymbol{\lambda}, H^n(\cdot)} \quad \frac{\boldsymbol{\lambda}^\top \mathbf{b}}{1-\beta} + \sum_{n=1}^{N} \sum_{x^n \in \mathcal{X}_n} \upsilon_n(x^n) H^n(x^n) \tag{16}$$

$$s.t. \quad \boldsymbol{\lambda} \geq 0,$$

$$H^n(x_0^n) \geq R^n(x_0^n, a_0^n) - \boldsymbol{\lambda}^\top \mathbf{B}^n(x_0^n, a_0^n) + \beta \sum_{x_1^n \in \mathcal{X}^n} P_n(x_1^n | x_0^n, a_0^n) H^n(x_1^n),$$

$$\text{for all } (x_0^n, a_0^n) \text{ with } a_0^n \in \mathcal{A}^n(x_0^n).$$

In the continuous-state or continuous-action case, noting that $J^{\boldsymbol{\lambda}}(v)$ is convex in $\boldsymbol{\lambda}$ with a fixed probability distribution v, the Lagrangian bound $\boldsymbol{J}^{\boldsymbol{\lambda}^*}$ may be solved using the stochastic subgradient method (see, e.g., Section 2.2.1 of [49]). We also review the ALP method to derive an upper bound H^{LP} on V and compare its bound performance with the Lagrangian bound in Appendix A.1.

3.2 Information Relaxation-based Dual Bound

Information relaxation has been used to compute a dual bound on the optimal value of finite-horizon stochastic dynamic programs. In this section, we propose a computational method based on a randomization idea to extend information relaxation to the discounted infinite-horizon problem. This computational approach will be used to improve the quality of the Lagrangian bound; we show in one example that the improvement can be significant. We also analyze the conditions that the two bounds equal the optimal value.

We will use the following notations. Given $T \in \mathbb{N}$, we denote by $\mathcal{A}(T) \triangleq \mathcal{A}_0 \times \cdots \times \mathcal{A}_T$, where each $\mathcal{A}_t = \mathcal{A}$. Respectively, we define $\bar{\mathcal{A}}(T) \triangleq \bar{\mathcal{A}}_0 \times \cdots \times \bar{\mathcal{A}}_T$.

3.2.1 Information Relaxation-based Bounds for Discounted Infinite-Horizon Problem

The Lagrangian relaxation approach in Section 3.1.2 relaxes the feasible set of the controls, where the term $\sum_{t=0}^{\infty} \beta^t \boldsymbol{\lambda}^\top \left([\mathbf{b} - \mathbf{B}(\mathbf{x}_t, \mathbf{a}_t)] \right)$ plays the role of a penalty when the decision takes value outside the feasible region. As an alternative relaxation technique, the "information relaxation" relaxes the non-anticipativity constraint on the control policy and impose a class of penalties that penalize this violation.

We will construct a penalty from a function H defined on the state space $\mathcal{X}$. This penalty is the discounted sum of martingale difference sequence under any policy $\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}$, implying that charging this penalty does not influence the expected rewards under such a policy.

To begin with, we first define a partial discounted sum M_k based on a measurable and bounded function $H : \mathcal{X} \to \mathbb{R}$, that is,

$$M_k(\mathbf{a}, \omega) \triangleq \sum_{t=0}^{k} \beta^{t+1} \Delta H(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{a}_t), \quad k = 0, 1, \cdots, \tag{17}$$

where $\Delta H(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{a}_t) = H(\mathbf{x}_{t+1}) - \mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t, \mathbf{a}_t]$ and . Note that given a control policy $\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}$, $\{M_k(\alpha(\omega), \omega)\}_{k=0}^{\infty}$ is an $\mathbb{F}$-martingale, since $\{\Delta H(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{a}_t)\}_{t=0}^{\infty}$ is an $\mathbb{F}$-martingale difference sequence. In particular, $\mathbb{E}[M_k(\alpha(\omega), \omega)|\mathbf{x}_0] = 0$ for any $\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}$.

We then consider the discounted infinite sum of ΔH, that is,

$$M(\mathbf{a}, \omega) \triangleq \sum_{t=0}^{\infty} \beta^{t+1} \Delta H(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{a}_t).$$

Define $\mathcal{D} \triangleq \{H : \mathcal{X} \to \mathbb{R} | H \text{ is measurable and bounded}\}$. It is straightforward to verify that M is well-defined given any function $H \in \mathcal{D}$ and M has expectation zero under a non-anticipative policy.

Lemma 2. *Suppose $H \in \mathcal{D}$. Then $M(\boldsymbol{a}, \omega)$ is well defined given any control sequence $\boldsymbol{a}$ and scenario ω; moreover, $\mathbb{E}[M(\alpha(\omega), \omega)|\boldsymbol{x}_0] = 0$ for all $\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}$.*

Proof. We can show that $M(\mathbf{a}, \omega)$ is well defined for any $\mathbf{a}$ and ω given $H \in \mathcal{D}$, i.e., $|H(\cdot)| < \Lambda$ for some $\Lambda > 0$; the sequence $\{M_k\}_{k=0}^{\infty}$ is then uniformly bounded for all $k \geq 0$, since for $t = 0, 1, \cdots, k$,

$$|M_k(\mathbf{a}, \omega)| \leq \sum_{t=0}^{k} \beta^{t+1} |\Delta H(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{a}_t)| \leq \frac{2\Lambda}{1-\beta} \text{ for all } \omega \in \Omega \text{ and } \mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t),$$

Therefore, $M(\mathbf{a}, \omega) \triangleq \lim_{k\to\infty} M_k(\mathbf{a}, \omega)$ is well-defined for every $\mathbf{a}$ and ω. In particular, $\mathbb{E}[M(\alpha(\omega), \omega)|\mathbf{x}_0] = \lim_{k\to\infty} \mathbb{E}[M_k(\alpha(\omega), \omega)|\mathbf{x}_0] = 0$ for $\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}$ due to the dominated convergence theorem, noting that $\mathbb{E}[M_k(\alpha(\omega), \omega)|\mathbf{x}_0] = 0$ for all k. □

Since $\mathbb{E}[M(\alpha(\omega),\omega)|\mathbf{x}_0]=0$ for any $\alpha\in\bar{\mathbb{A}}_{\mathbb{F}}$, then

$$\begin{aligned}V(\mathbf{x}_0;\alpha)=&\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t R(\mathbf{x}_t,\mathbf{a}_t)\middle|\mathbf{x}_0\right]-\mathbb{E}\left[M(\alpha(\omega),\omega)|\mathbf{x}_0\right]\\ =&\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t\big(R(\mathbf{x}_t,\mathbf{a}_t)-\beta(H(\mathbf{x}_{t+1})-\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t,\mathbf{a}_t])\big)\middle|\mathbf{x}_0\right]\\ =&H(\mathbf{x}_0)+\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t\big(R(\mathbf{x}_t,\mathbf{a}_t)+\beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t,\mathbf{a}_t]-H(\mathbf{x}_t)\big)\middle|\mathbf{x}_0\right].\end{aligned}\tag{18}$$

The second equality holds due to the definition of ΔH, and the last equality holds since $\sum_{t=0}^{\infty}\beta^t R(\mathbf{x}_t,\mathbf{a}_t)$, $\sum_{t=0}^{\infty}\beta^{t+1}\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t,\mathbf{a}_t]$, and $\sum_{t=0}^{\infty}\beta^{t+1}H(\mathbf{x}_t)$ are absolutely convergent for all $\omega\in\Omega$ and $\mathbf{a}\in\mathbb{A}$.

To develop a computational method that reduces the infinite sum inside the conditional expectation in (18) to a finite sum, we consider a random time τ (see, .e.g., [37]) that is independent of $\{\mathcal{F}_t,\ t=0,1,\cdots\}$, and τ is of geometric distribution with parameter β, i.e.,

$$P(\tau=t)=(1-\beta)\beta^t,\quad t=0,1,\cdots.$$

A complete definition of τ is in Appendix A.2.1.

Lemma 3. *Suppose τ is a random time of geometric distribution with parameter β and it is independent of $\{\mathcal{F}_t,\ t=0,1,\cdots\}$. Then for all $\alpha\in\bar{\mathbb{A}}_{\mathbb{F}}$ and $H\in\mathcal{D}$,*

$$V(\boldsymbol{x}_0;\alpha)=H(\boldsymbol{x}_0)+\mathbb{E}\left[I_H(\alpha(\omega),\omega,\tau)|\boldsymbol{x}_0\right],\tag{19}$$

where

$$I_H(\boldsymbol{a},\omega,\tau)\triangleq\sum_{t=0}^{\tau}\big(R(\boldsymbol{x}_t,\boldsymbol{a}_t)+\beta\mathbb{E}[H(\boldsymbol{x}_{t+1})|\boldsymbol{x}_t,\boldsymbol{a}_t]-H(\boldsymbol{x}_t)\big).\tag{20}$$

Proof. Noting that $P(t\le\tau)=\mathbb{E}\left[\mathbb{1}_{\{t\le\tau\}}\right]=\beta^t$, we can rewrite the second term in

(18) as

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{\infty}\mathbb{E}\left[\mathbb{1}_{\{t\leq\tau\}}\right]\cdot\left(R(\mathbf{x}_t,\mathbf{a}_t)+\beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t,\mathbf{a}_t]-H(\mathbf{x}_t)\right)\middle|\mathbf{x}_0\right]\\
=&\mathbb{E}\left[\sum_{t=0}^{\infty}\mathbb{1}_{\{t\leq\tau\}}\cdot\left(R(\mathbf{x}_t,\mathbf{a}_t)+\beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t,\mathbf{a}_t]-H(\mathbf{x}_t)\right)\middle|\mathbf{x}_0\right]\\
=&\mathbb{E}\left[\sum_{t=0}^{\tau}\left(R(\mathbf{x}_t,\mathbf{a}_t)+\beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t,\mathbf{a}_t]-H(\mathbf{x}_t)\right)\middle|\mathbf{x}_0\right],
\end{aligned}
$$

where the first equality holds due to the Fubini's theorem, noting that the boundedness of R and H implies the integrability of the integrand in $\mathbb{E}_0[\cdot]$. □

The conditional expectation in (20) is now taken with respect to both the random outcome ω and the random time τ. We can better interpret this conditional expectation via Monte Carlo simulation: in each trial of simulation, we first generate a realization of the random horizon τ (that is finite) and scenario ω, i.e., the underlying uncertainty that affects the evolution of $\{\mathbf{x}_t\}_{t=0}^{\tau}$; we then apply the policy α on the scenario ω up to time τ to evaluate the value of $I_H(\alpha(\omega),\omega,\tau)$. According to (19), $H(\mathbf{x}_0)+I_H(\alpha(\omega),\omega,\tau)$ is an unbiased estimator of $V(\mathbf{x}_0;\alpha)$.

To obtain an upper bound on the optimal value V that complements the lower bound $V(\mathbf{x}_0;\alpha)$ in (19) induced by the policy α and function H, we introduce the operator $\mathcal{L}:\mathcal{D}\to\mathcal{D}$

$$
\mathcal{L}H(\mathbf{x}_0)\triangleq H(\mathbf{x}_0)+\mathbb{E}\left[\max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)}\{I_H(\mathbf{a},\omega,\tau)\}\middle|\mathbf{x}_0\right], \tag{21}
$$

Since the dependence of I_H on $\mathbf{a}$ is only through the first $\tau+1$ actions, namely, $\mathbf{a}(\tau)$. Thus, $\max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)}\{I_H(\mathbf{a},\omega,\tau)\}$ is short for $\max_{\mathbf{a}(\tau)\in\bar{\mathcal{A}}(\tau)}\{I_H(\mathbf{a}(\tau),\omega,\tau)\}$, which is referred to as the *inner optimization problem.* In each trial of simulation, we maximize $I_H(\mathbf{a},\omega,\tau)$ subject to $\mathbf{a}\in\bar{\mathcal{A}}(\tau)$ given a realization of the random horizon τ and scenario ω. We show that the estimator $\max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)}\{I_H(\mathbf{a},\omega,\tau)\}$ has finite mean and variance in Appendix A.2.2.

We next show for any $H \in \mathcal{D}$, the optimal value V is upper bounded by $\mathcal{L}H$, which will be referred to as the "information relaxation bound". The relaxed information is reflected in the scenario-based inner optimization problem, while $M(\mathbf{a}, \omega) = \sum_{t=0}^{\infty} \beta^{t+1} \Delta H(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{a}_t)$ induced by the function H plays the role of a penalty: if H is chosen to be V, then the upper bound $\mathcal{L}H$ is tight, i.e, $\mathcal{L}H = V$.

Theorem 3 (Information Relaxation Bound)**.** *Let τ be a random time of geometric distribution with parameter β and it is independent of $\{\mathcal{F}_t,\ t = 0, 1, \cdots\}$. Then*

(a) (Weak Duality) For any $H \in \mathcal{D}$, $V(\boldsymbol{x}) \leq \mathcal{L}H(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{X}$.

(b) (Tighter Bound) For any $H \in \mathcal{D}^*(\varepsilon)$ with $\varepsilon \geq 0$, where

$$\mathcal{D}^*(\varepsilon) \triangleq \{H \in \mathcal{D} : R(\boldsymbol{x}_0, \boldsymbol{a}_0) + \beta \mathbb{E}[H(\boldsymbol{x}_1) | \boldsymbol{x}_0, \boldsymbol{a}_0] \leq H(\boldsymbol{x}_0) - \varepsilon \text{ for } \boldsymbol{x}_0 \in \mathcal{X} \text{ and } \boldsymbol{a}_0 \in \bar{\mathcal{A}}(\boldsymbol{x}_0)\},$$

then $\max_{\boldsymbol{a} \in \bar{\mathcal{A}}(\tau)} \{I_H(\boldsymbol{a}, \omega, \tau)\} \leq -\varepsilon(\tau + 1)$ for every $\omega \in \Omega$ and $\tau \in \mathbb{N}$; therefore, for $\boldsymbol{x} \in \mathcal{X}$,

$$\mathcal{L}H(\boldsymbol{x}) \leq H(\boldsymbol{x}) - \frac{\varepsilon}{1 - \beta}.$$

(c) (Strong Duality) $V(\boldsymbol{x}) = \mathcal{L}V(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{X}$. Moreover, $\max_{\boldsymbol{a} \in \bar{\mathcal{A}}(\tau)} \{I_V(\boldsymbol{a}, \omega, \tau)\} = 0$ for every $\omega \in \Omega$ and $\tau \in \mathbb{N}$.

Proof. (a) For $\mathbf{x}_0 \in \mathcal{X}_0$ and $\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}$,

$$V(\mathbf{x}_0; \alpha) = H(\mathbf{x}_0) + \mathbb{E}_0 \left[I_H(\alpha(\omega), \omega, \tau) \right] \leq H(\mathbf{x}_0) + \mathbb{E}_0 \left[\max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} \{I_H(\mathbf{a}, \omega, \tau)\} \right],$$

where $\mathbb{E}_0[\ \cdot\ ] = \mathbb{E}[\ \cdot\ | \mathbf{x}_0]$. By maximizing $V(\mathbf{x}_0; \alpha)$ over $\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}$, the weak duality $V(\mathbf{x}_0) \leq \mathcal{L}H(\mathbf{x}_0)$ holds.

(b) Note that given any $H \in \mathcal{D}^*$ and $\mathbf{x}_t \in \mathcal{X}$, $R(\mathbf{x}_t, \mathbf{a}_t) + \beta \mathbb{E}[H(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - H(\mathbf{x}_t) \leq -\varepsilon$ for all $\mathbf{a}_t \in \bar{\mathcal{A}}(\mathbf{x}_t)$. It is straightforward to see that for any $\tau \in \mathbb{N}$ and $\omega \in \Omega$,

$$I_H(\mathbf{a}, \omega, \tau) = \sum_{t=0}^{\tau} \big(R(\mathbf{x}_t, \mathbf{a}_t) + \beta \mathbb{E}[H(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - H(\mathbf{x}_t) \big) \leq -(\tau + 1)\varepsilon$$

for any $\mathbf{a}_t \in \bar{\mathcal{A}}(\mathbf{x}_t)$, $t = 0, 1, \cdots, \tau$. Therefore, for all $\mathbf{x}_0 \in \mathcal{X}$ we have

$$\mathcal{L}H(\mathbf{x}_0) \leq H(\mathbf{x}_0) + \mathbb{E}_0[-(\tau+1)\varepsilon] = H(\mathbf{x}_0) - \frac{\varepsilon}{1-\beta}.$$

Together with the weak duality, we have shown that $V(\mathbf{x}_0) \leq \mathcal{L}H(\mathbf{x}_0) \leq H(\mathbf{x}_0) - \frac{\varepsilon}{1-\beta}$.

(c) The strong duality follows from the weak duality and the results in (b) by choosing $H = V$ noting that $V \in \mathcal{D}^*(0)$. Moreover, since $V(\mathbf{x}_0) = \max_{\mathbf{a}\in\bar{\mathcal{A}}}\{R(\mathbf{x}_0, \mathbf{a}_0) + \beta\mathbb{E}[V(\mathbf{x}_1)|\mathbf{x}_0, \mathbf{a}_0]\}$ for every $\mathbf{x}_0 \in \mathcal{X}$, we can use the dynamic program to show that $\max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)}\{I_V(\mathbf{a}, \omega, \tau)\} = 0$ for every $\omega \in \Omega$ and $\tau \in \mathbb{N}$.

□

The function $H \in \mathcal{D}^*(0)$ is sometimes referred to as a "supersolution" to the problem (11), and it is a standard result that the optimal value V is upper bounded by a supersolution H (see, e.g., [9]). Theorem 3(b) indicates that the scenario-dependent inner optimization problem of an arbitrary time horizon τ is upper bounded by zero provided $H \in \mathcal{D}^*(0)$; therefore, $\mathcal{L}H$ improves the quality of the supersolution H as an upper bound on V. The strong duality implies that we may obtain a tight dual bound, given some approximate function of V that induces a good approximation of $\sum_{t=0}^{\infty}\beta^{t+1}\Delta V(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{a}_t)$. In addition, Theorem 3 is true not only for weakly coupled dynamic program, but also for general discounted infinite-horizon stochastic dynamic program due to the applicable randomization technique.

As a corollary of Theorem 3, we present the information relaxation-based dual representation of the Lagrangian bound $J^{\boldsymbol{\lambda}}$. To this end, we dualize the linking constraints at each period in $\max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)}\{I_H(\mathbf{a}, \omega, \tau)\}$ up to time τ, and introduce the Lagrangian function $I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu})$ for $\boldsymbol{\mu} \triangleq (\boldsymbol{\mu}_0, \cdots, \boldsymbol{\mu}_\tau)$ with each $\boldsymbol{\mu}_t \in \mathbb{R}^L_+$:

$$I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu}) \triangleq \sum_{t=0}^{\tau} \left(\mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \boldsymbol{\mu}_t^\top[\mathbf{b} - \mathbf{B}(\mathbf{x}_t, \mathbf{a}_t)] + \beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t, \mathbf{a}_t] - H(\mathbf{x}_t)\right). \tag{22}$$

In particular, $I_H(\mathbf{a}, \omega, \tau) = I_H(\mathbf{a}, \omega, \tau; \mathbf{0})$.

Corollary 4. *Suppose* $\boldsymbol{\lambda} \in \mathbb{R}_+^L$ *and* $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\lambda}, \cdots, \boldsymbol{\lambda})$. *Suppose* τ *is of geometric distribution with parameter* β *and it is independent of* $\{\mathcal{F}_t,\ t = 0, 1, \cdots\}$. *Then*

(a) (Weak Duality) For any $H \in \mathcal{D}$,

$$J^{\boldsymbol{\lambda}}(\boldsymbol{x}_0) \leq H(\boldsymbol{x}_0) + \mathbb{E}\left[\max_{\boldsymbol{a} \in \mathcal{A}(\tau)} \{I_H(\boldsymbol{a}, \omega, \tau; \tilde{\boldsymbol{\mu}})\} \middle| \boldsymbol{x}_0\right].$$

(b) (Strong Duality) $J^{\boldsymbol{\lambda}}(\boldsymbol{x}_0) = J^{\boldsymbol{\lambda}}(\boldsymbol{x}_0) + \mathbb{E}\left[\max_{\boldsymbol{a} \in \mathcal{A}(\tau)} \{I_{J^{\boldsymbol{\lambda}}}(\boldsymbol{a}, \omega, \tau; \tilde{\boldsymbol{\mu}})\} \middle| \boldsymbol{x}_0\right]$. *Moreover,* $\max_{\boldsymbol{a} \in \mathcal{A}(\tau)}\{I_{J^{\boldsymbol{\lambda}}}(\boldsymbol{a}, \omega, \tau; \boldsymbol{\mu})\} = 0$ *for every* $\omega \in \Omega$ *and* $\tau \in \mathbb{N}$.

Proof. Note that $J^{\boldsymbol{\lambda}}$ is the optimal value to the discounted infinite-horizon MDP with one-period reward $\mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \boldsymbol{\lambda}^\top[\mathbf{b} - \mathbf{B}(\mathbf{x}_t, \mathbf{a}_t)]$ and control set $\mathcal{A}(\tau)$. Following the proof of Theorem 3, it is straightforward to verify the weak duality and strong duality results. □

3.2.2 Comparing Lagrangian Relaxation Bound

In weakly coupled stochastic dynamic program, a natural candidate of the approximate value function is the Lagrangian bound $J^{\boldsymbol{\lambda}}$. It can be shown that the information relaxation approach can be used to improve the performance of the Lagrangian bound.

Corollary 5. *For any* $\boldsymbol{\lambda} \geq 0$, $\mathcal{L}J^{\boldsymbol{\lambda}}(\boldsymbol{x}) \leq J^{\boldsymbol{\lambda}}(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \mathcal{X}$.

Proof. This is an immediate corollary of Theorem 3(c) since $J^{\boldsymbol{\lambda}} \in \mathcal{D}^*(0)$ (see Lemma 4(b) in Appendix A.1). Here we consider an alternative proof based on the dual representation of $J^{\boldsymbol{\lambda}}$. Let $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\lambda}, \cdots, \boldsymbol{\lambda})$. Note that for each scenario ω and $\tau \in \mathbb{N}$,

$$0 = \max_{\mathbf{a} \in \mathcal{A}(\tau)} \{I_H(\mathbf{a}, \omega, \tau; \tilde{\boldsymbol{\mu}})\} \geq \max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} \{I_H(\mathbf{a}, \omega, \tau; \tilde{\boldsymbol{\mu}})\} \geq \max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} \{I_H(\mathbf{a}, \omega, \tau)\}, \tag{23}$$

where the equality follows Corollary 4(b), the first inequality holds because $\mathcal{A}(\tau) \supset \bar{\mathcal{A}}(\tau)$, and the second inequality holds since $\boldsymbol{\lambda} \geq 0$ and each $\mathbf{b} - \mathbf{B}(\mathbf{x}_t, \mathbf{a}_t) \geq 0$ for $\mathbf{a}_t \in \bar{\mathcal{A}}$. Hence, $J^{\boldsymbol{\lambda}}(\mathbf{x}) \geq \mathcal{L}J^{\boldsymbol{\lambda}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

□

Corollary 5 generalizes the result in [19] from finite-horizon to discounted infinite-horizon setting. It is worth noting that the information relaxation approach can improve the ALP bound H^{LP}, since it is also a supersolution to (11) (see Lemma 6 in Appendix A.1).

A natural question is whether the improvement of the information relaxation bound over the Lagrangian bound can be significant. In Appendix A.3, we provide an affirmative answer by investigating the example proposed in [1], where the Lagrangian bound can be arbitrarily poor compared with the optimal value; as opposed to the performance of the Lagrangian bound, we show that the optimal value can be recovered based on it using the information relaxation approach.

A significant difference of the information relaxation and Lagrangian relaxation in the weakly coupled dynamic program is that the strong duality exists in the former relaxation (at least theoretically), while such a result does not hold in general for the latter approach. The following theorem characterizes the sufficient and necessary conditions such that $V(\mathbf{x}; \alpha') = \mathcal{L}H(\mathbf{x}_0)$, where α' is a stationary Markov policy and $H \in \mathcal{D}$. This result resembles Theorem 2.2 in [20] for the finite-horizon problem.

Theorem 6. *Suppose $H \in \mathcal{D}$ and $\alpha' : \mathcal{X} \to \mathcal{A}$ is a stationary Markov policy such that $\alpha'(\boldsymbol{x}) \in \bar{\mathcal{A}}(\boldsymbol{x})$. A necessary and sufficient condition for $V(\boldsymbol{x}_0; \alpha') = \mathcal{L}H(\boldsymbol{x}_0)$ for all $\boldsymbol{x}_0 \in \mathcal{X}$ is that*

$$\begin{aligned} &\max_{\boldsymbol{a} \in \bar{\mathcal{A}}(T)} \left\{ \sum_{t=0}^{T} \left(R(\boldsymbol{x}_t, \boldsymbol{a}_t) + \beta \mathbb{E}\left[H(\boldsymbol{x}_{t+1}) | \boldsymbol{x}_t, \boldsymbol{a}_t \right] - H(\boldsymbol{x}_t) \right) \right\} \\ =& \sum_{t=0}^{T} \left(R(\boldsymbol{x}_t, \alpha'(\boldsymbol{x}_t)) + \beta \mathbb{E}[H(\boldsymbol{x}_{t+1}) | \boldsymbol{x}_t, \alpha'(\boldsymbol{x}_t)] - H(\boldsymbol{x}_t) \right) \end{aligned} \tag{24}$$

for $\omega \in \Omega$ *almost surely,* $T = 0, 1, 2, \cdots$ *. In particular, by considering the case* $T = 0$*,*

$$\alpha'(\boldsymbol{x}_0) \in \arg \max_{\boldsymbol{a}_0 \in \bar{\mathcal{A}}(\boldsymbol{x}_0)} \{R(\boldsymbol{x}_0, \boldsymbol{a}_0) + \beta \mathbb{E}[H(\boldsymbol{x}_1)|\boldsymbol{x}_0, \boldsymbol{a}_0]\}$$

for all $\boldsymbol{x}_0 \in \mathcal{X}$*.*

Proof. Given $\alpha' \in \bar{\mathbb{A}}_{\mathbb{F}}$ and $\mathbf{x}_0 \in \mathcal{X}$,

$$\begin{aligned} V(\mathbf{x}_0; \alpha') =& H(\mathbf{x}_0) + \mathbb{E}_0 \left[I_H(\alpha', \omega, \tau) \right] \\ \leq& H(\mathbf{x}_0) + \mathbb{E}_0 \left[\max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} \{I_H(\mathbf{a}, \omega, \tau)\} \right], \qquad (25) \\ =& H(\mathbf{x}_0) + \sum_{T=0}^{\infty} P(\tau = T) \cdot \mathbb{E}_0 \left[\max_{\mathbf{a} \in \bar{\mathcal{A}}(T)} \{I_H(\mathbf{a}, \omega, T)\} \right] \\ =& \mathcal{L}H(\mathbf{x}_0). \end{aligned}$$

To show necessity, $V(\mathbf{x}; \alpha') = \mathcal{L}H(\mathbf{x})$ means that the inequality (25) is an equality; we also note that $I_H(\mathbf{a}, \omega, \tau) \leq \max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} \{I_H(\mathbf{a}, \omega, \tau)\}$ for each $\tau = 0, 1, 2, \cdots$ and $\omega \in \Omega$, which implies that for every $T \in \mathbb{N}$, the equality (24) holds for $\omega \in \Omega$ almost surely, observing that $P(\tau = T) > 0$ for each $T \in \mathbb{N}$.

The sufficiency is straightforward, since the condition (24) holds for $\omega \in \Omega$ almost surely and $T \in \mathbb{N}$ implies that (25) is an equality. □

Theorem 6 characterizes the optimality conditions of a policy α' to (11) and value approximation H in (21) as a pair: the optimal policy to the inner optimization problem of any horizon T induced by the approximate value function is non-anticipative and also stationary, though these decisions can be chosen to be anticipative and non-stationary. In particular, the policy α' should be the greedy policy induced by the approximate value function H.

As a special case, if we choose Lagrangian bound as the approximate value, the analogous optimality conditions developed in Theorem 2 of [1] can be recovered using the information relaxation argument. We review the sufficient and necessary conditions therein and present them in parallel with the statement of Theorem 6.

Lemma 4. *Suppose $\boldsymbol{\lambda}' \geq 0$ and $\alpha' : \mathcal{X} \to \mathcal{A}$ is a stationary Markov policy such that $\alpha'(\boldsymbol{x}) \in \bar{\mathcal{A}}(\boldsymbol{x})$. A necessary and sufficient condition for $V(\boldsymbol{x}_0; \alpha') = J^{\boldsymbol{\lambda}'}(\boldsymbol{x}_0)$ for all $\boldsymbol{x}_0 \in \mathcal{X}$ is that for all $\boldsymbol{x}_0 \in \mathcal{X}$, $\boldsymbol{\lambda}'^{\top} [\boldsymbol{b} - \boldsymbol{B}(\boldsymbol{x}_0, \alpha'(\boldsymbol{x}_0))] = 0$ and*

$$\alpha'(\boldsymbol{x}_0) \in \arg \max_{\boldsymbol{a}_0 \in \mathcal{A}(\boldsymbol{x}_0)} \left\{ \boldsymbol{R}(\boldsymbol{x}_0, \boldsymbol{a}_0) + \boldsymbol{\lambda}'^{\top} [\boldsymbol{b} - \boldsymbol{B}(\boldsymbol{x}_0, \boldsymbol{a}_0)] + \beta \mathbb{E}[J^{\boldsymbol{\lambda}'}(\boldsymbol{x}_1) | \boldsymbol{x}_0, \boldsymbol{a}_0] \right\}. \quad (26)$$

We can use information relaxation to obtain the optimality conditions in Lemma 4. Note that for any $\boldsymbol{\lambda} \geq 0$ (let $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\lambda}, \cdots, \boldsymbol{\lambda})$) and stationary Markov policy α' (in $\bar{\mathbb{A}}_{\mathbb{F}}$), we have

$$I_{J^{\boldsymbol{\lambda}}}(\alpha', \omega, \tau) \leq I_{J^{\boldsymbol{\lambda}}}(\alpha', \omega, \tau; \tilde{\boldsymbol{\mu}}) \leq \max_{\mathbf{a} \in \mathcal{A}(\tau)} \{ I_{J^{\boldsymbol{\lambda}}}(\mathbf{a}, \omega, \tau; \tilde{\boldsymbol{\mu}}) \}$$

for every ω and τ; therefore, $V(\mathbf{x}_0; \alpha') \leq J^{\boldsymbol{\lambda}}(\mathbf{x}_0)$ for all $\mathbf{x}_0 \in \mathcal{X}$, according to Lemma 3 and Corollary 4(b).

If $V(\mathbf{x}_0; \alpha') = J^{\boldsymbol{\lambda}}(\mathbf{x}_0)$ for some $\alpha' \in \bar{\mathbb{A}}_{\mathbb{F}}$ and $\boldsymbol{\lambda} \geq 0$, it implies

$$I_{J^{\boldsymbol{\lambda}}}(\alpha', \omega, \tau) = I_{J^{\boldsymbol{\lambda}}}(\alpha', \omega, \tau; \tilde{\boldsymbol{\mu}}) = \max_{\mathbf{a} \in \mathcal{A}(\tau)} \{ I_{J^{\boldsymbol{\lambda}}}(\mathbf{a}, \omega, \tau; \tilde{\boldsymbol{\mu}}) \}.$$

The above equality renders the conditions in Lemma 4 more stringent than those in Theorem 6. Consider the special case $\tau = 0$ and recall that $\boldsymbol{\lambda}^{\top} [\mathbf{b} - \mathbf{B}(\mathbf{x}_0, \alpha(\mathbf{x}_0))] \geq 0$. Then the first equality implies $\boldsymbol{\lambda}^{\top} [\mathbf{b} - \mathbf{B}(\mathbf{x}_0, \alpha(\mathbf{x}_0))] = 0$ and the second equality implies (26).

3.3 Practical Information Relaxation Bound for Large-scale Problems

The information relaxation approach has the desirable property that it generates tighter upper bound based on the Lagrangian bound; however, computing the information relaxation bound can be challenging in large-scale weakly coupled dynamic program due to the intractable inner optimization problem. To be specific, the size of this scenario-dependent optimization problem increases exponentially with respect to the number of the projects or subproblems N, and also increases at least linearly in the horizon τ. Instead of computing the optimal value of the inner optimization problem, we discuss how to derive its upper bound that is computationally tractable. Therefore, this sub-optimal method still leads to a valid upper bound on the value function, which is referred to as the "practical information relaxation bound". We will show its performance guarantee under certain conditions.

Throughout this section we assume that the approximate value function is of the additively separable form $H(\mathbf{x}) = \theta + \sum_{n=1}^{N} H^n(x^n)$, where θ is a constant and $H^n : \mathcal{X}^n \to \mathbb{R}$ for $n = 1, \cdots, N$. We denote by $\mathcal{D}^\circ$ the space of additively separable functions. By substituting $H(\cdot)$ in (20) by $\theta + \sum_{n=1}^{N} H^n(\cdot)$, we can rewrite I_H as

$$I_H(\mathbf{a}, \omega, \tau) = \sum_{n=1}^{N}\left[\sum_{t=0}^{\tau}\big(R^n(x_t^n, a_t^n) + \beta\mathbb{E}[H^n(x_{t+1}^n)|x_t^n, a_t^n] - H^n(x_t^n)\big)\right] - (\tau+1)(1-\beta)\theta. \tag{27}$$

3.3.1 Relaxation of the Inner Optimization Problem

Note that the scenario-dependent primal problem $\max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)}\{I_H(\mathbf{a}, \omega, \tau)\}$ is also *weakly coupled* due to the additively separable structure of (27) and the feasible control set $\bar{\mathcal{A}}(\tau)$. To obtain an upper bound on its optimal value, we consider its Lagrangian dual $\max_{\mathbf{a}\in\mathcal{A}(\tau)}\{I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu})\}$, where

$$\begin{aligned} I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu}) =& \sum_{t=0}^{\tau}\big(\mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \boldsymbol{\mu}_t^\top[\mathbf{b} - \mathbf{B}(\mathbf{x}_t, \mathbf{a}_t)] + \beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t, \mathbf{a}_t] - H(\mathbf{x}_t)\big) \\ =& \sum_{t=0}^{\tau}\left[\sum_{n=1}^{N}\big(R^n(x_t^n, a_t^n) + \beta\mathbb{E}[H^n(x_{t+1}^n)|x_t^n, a_t^n] - H^n(x_t^n) - \boldsymbol{\mu}_t^\top\mathbf{B}^n(x_t^n, a_t^n)\big)\right. \\ & \left. - (1-\beta)\theta\right] + \sum_{t=0}^{\tau}\boldsymbol{\mu}_t^\top\ \mathbf{b} \\ =& \sum_{n=1}^{N} I_{H^n}^n(\mathbf{a}^n, \omega, \tau; \boldsymbol{\mu}) - (\tau+1)(1-\beta)\theta + \sum_{t=0}^{\tau}\boldsymbol{\mu}_t^\top\mathbf{b}, \end{aligned} \tag{28}$$

where $I_{H^n}^n$ in (28) is defined as

$$I_{H^n}^n(\mathbf{a}^n, \omega, \tau; \boldsymbol{\mu}) \triangleq \sum_{t=0}^{\tau}\big(R^n(x_t^n, a_t^n) + \beta\mathbb{E}[H^n(x_{t+1}^n)|x_t^n, a_t^n] - H^n(x_t^n) - \boldsymbol{\mu}_t^\top\mathbf{B}^n(x_t^n, a_t^n)\big)$$

with $\mathbf{a}^n \triangleq (a_0^n, \cdots, a_\tau^n)$.

Given any $\boldsymbol{\mu} \geq 0$, it is straightforward to see

$$\max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} I_H(\mathbf{a}, \omega, \tau) \leq \max_{\mathbf{a} \in \mathcal{A}(\tau)} I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu}).$$

According to (28), the Lagrangian dual function can be decomposed as

$$\max_{\mathbf{a} \in \mathcal{A}(\tau)} \{I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu})\} = \sum_{n=1}^{N} \max_{\mathbf{a}^n \in \mathcal{A}^n(\tau)} \{I_{H^n}^n(\mathbf{a}^n, \omega, \tau; \boldsymbol{\mu})\} - (\tau+1)(1-\beta)\theta + \sum_{t=0}^{\tau} \boldsymbol{\mu}_t^\top \mathbf{b}, \tag{29}$$

where $\mathcal{A}^n(\tau) \triangleq \mathcal{A}_0^n \times \cdots \times \mathcal{A}_\tau^n$ with each $\mathcal{A}_t^n = \mathcal{A}^n$. The equality (29) implies that the computational cost on solving $\max_{\mathbf{a} \in \mathcal{A}(\tau)}\{I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu})\}$ is linear rather than exponential in the number of the subproblems N. Therefore, the Lagrangian relaxation significantly reduces the computational complexity, and hence solving (29) to optimality becomes potentially tractable.

It remains to find the optimal $\boldsymbol{\mu}^*$ that achieves the minimum of $I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu})$ over $\boldsymbol{\mu} \geq 0$. To this end, we list some properties of $\max_{\mathbf{a} \in \mathbb{A}(\tau)} I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu})$ as a function of $\boldsymbol{\mu}$, based on properties of Lagrangian relaxation.

Lemma 5. *Given $I_H(\boldsymbol{a}, \omega, \tau; \boldsymbol{\mu})$ defined in (28), where $\omega \in \Omega$ and $\tau \in \mathbb{N}$. Then*

(a) $\max_{\boldsymbol{a} \in \mathcal{A}(\tau)} I_H(\boldsymbol{a}, \omega, \tau; \boldsymbol{\mu})$ *is convex in* $\boldsymbol{\mu}$.

(b) Let $\boldsymbol{a}^\circ = (\boldsymbol{a}_0^\circ, \cdots, \boldsymbol{a}_\tau^\circ) \in \arg\max_{\boldsymbol{a} \in \mathcal{A}(\tau)} I_H(\boldsymbol{a}, \omega, \tau; \boldsymbol{\mu})$ *for a fixed* $\boldsymbol{\mu} \geq 0$. *Then*

$$[\boldsymbol{b} - \boldsymbol{B}(\boldsymbol{x}_0^\circ, \boldsymbol{a}_0^\circ), \cdots, \boldsymbol{b} - \boldsymbol{B}(\boldsymbol{x}_\tau^\circ, \boldsymbol{a}_\tau^\circ)] \in \partial I_H(\boldsymbol{a}^\circ, \omega, \tau; \boldsymbol{\mu}),$$

where $\{\boldsymbol{x}_t^\circ\}_{t=0}^\tau$ *is the state trajectory under* $\boldsymbol{a}^\circ$ *and* ω, *and* $\partial I_H(\boldsymbol{a}^\circ, \omega, \tau; \boldsymbol{\mu})$ *is the subdifferential of* $I_H(\boldsymbol{a}, \omega, \tau; \boldsymbol{\mu})$ *with respect to* $\boldsymbol{\mu}$ *at* $\boldsymbol{a} = \boldsymbol{a}^\circ$.

(c) $\max_{\boldsymbol{a} \in \bar{\mathcal{A}}(\tau)} I_H(\boldsymbol{a}, \omega, \tau) \leq \min_{\boldsymbol{\mu} \geq 0} \max_{\boldsymbol{a} \in \mathcal{A}(\tau)} I_H(\boldsymbol{a}, \omega, \tau; \boldsymbol{\mu})$.

The duality gap in Lemma 5(c) is zero if the primal problem is convex and the strong duality holds. Since the primal problem may lack the convex structure, we cannot expect zero duality gap in general. To find the optimal solution $\boldsymbol{\mu}^*$ to the dual problem, Lemma 5 indicates that it is is convex in $\boldsymbol{\mu}$ and has explicit subgradient at every $\boldsymbol{\mu}$; therefore, we can employ the standard subgradient method or its variant to locate the optimal solution efficiently. Due to Lemma 5(c), we refer to $\min_{\boldsymbol{\mu}\geq 0}\max_{\mathbf{a}\in\mathcal{A}(\tau)} I_H(\mathbf{a},\omega,\tau;\boldsymbol{\mu})$ as the "relaxed inner optimization problem".

Based on the relaxed inner optimization problem we define a new operator $\mathcal{L}^\circ$ that can be viewed as a "relaxed" version of $\mathcal{L}$ on the additively separable function space $\mathcal{D}^\circ$:

$$\mathcal{L}^\circ H(\mathbf{x}) \triangleq H(\mathbf{x}) + \mathbb{E}_0\left[\min_{\boldsymbol{\mu}\geq 0}\max_{\mathbf{a}\in\mathcal{A}(\tau)} I_H(\mathbf{a},\omega,\tau;\boldsymbol{\mu})\right]. \tag{30}$$

Due to the computational tractability of $\mathcal{L}^\circ H(\mathbf{x})$, it will be referred to as "practical information relaxation bound". In the next theorem we formalize the bound performance of $\mathcal{L}^\circ H(\mathbf{x})$, which naturally places an upper bound on the information relaxation bound $\mathcal{L}H$; moreover, the performance of $\mathcal{L}^\circ J^{\boldsymbol{\lambda}}(\mathbf{x})$ is no worse than the Lagrangian bound $J^{\boldsymbol{\lambda}}(\mathbf{x})$.

Theorem 7. *Suppose $H\in\mathcal{D}^\circ$. Then*

(a) $\mathcal{L}H(\boldsymbol{x}) \leq \mathcal{L}^\circ H(\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathcal{X}$.

(b) Suppose $H = J^{\boldsymbol{\lambda}}$ is a Lagrangian bound for some $\boldsymbol{\lambda}\geq 0$. Then for every $\omega\in\Omega$ and $\tau\in\mathbb{N}$,

$$\min_{\boldsymbol{\mu}\geq 0}\max_{\boldsymbol{a}\in\mathcal{A}(\tau)}\{I_{J^{\boldsymbol{\lambda}}}(\boldsymbol{a},\omega,\tau;\boldsymbol{\mu})\} \leq 0.$$

Consequently, $\mathcal{L}^\circ J^{\boldsymbol{\lambda}}(\boldsymbol{x}) \leq J^{\boldsymbol{\lambda}}(\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathcal{X}$.

Proof. (a) This is because for every $\omega\in\Omega$ and $\tau\in\mathbb{N}$,

$$\max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)}\{I_H(\mathbf{a},\omega,\tau)\} \leq \min_{\boldsymbol{\mu}\geq 0}\max_{\mathbf{a}\in\mathcal{A}(\tau)}\{I_H(\mathbf{a},\omega,\tau;\boldsymbol{\mu})\}.$$

(b) Note that for any $\boldsymbol{\lambda} \geq 0$, $J^{\boldsymbol{\lambda}}(\mathbf{x}_0) = J^{\boldsymbol{\lambda}}(\mathbf{x}_0) + \mathbb{E}_0\left[\max_{\mathbf{a}\in\mathcal{A}(\tau)}\{I_{J^{\boldsymbol{\lambda}}}(\mathbf{a},\omega,\tau;\tilde{\boldsymbol{\mu}})\}\right]$, where $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\lambda},\boldsymbol{\lambda},\cdots,\boldsymbol{\lambda})$. According to Corollary 4, we have

$$0 = \max_{\mathbf{a}\in\mathcal{A}(\tau)}\{I_{J^{\boldsymbol{\lambda}}}(\mathbf{a},\omega,\tau;\tilde{\boldsymbol{\mu}})\} \geq \min_{\boldsymbol{\mu}\geq 0}\max_{\mathbf{a}\in\mathcal{A}(\tau)}\{I_{J^{\boldsymbol{\lambda}}}(\mathbf{a},\omega,\tau;\boldsymbol{\mu})\}, \tag{31}$$

for every $\omega \in \Omega$ and $\tau \in \mathbb{N}$. Therefore,

$$J^{\boldsymbol{\lambda}}(\mathbf{x}_0) \geq J^{\boldsymbol{\lambda}}(\mathbf{x}_0) + \mathbb{E}_0\left[\min_{\boldsymbol{\mu}\geq 0}\max_{\mathbf{a}\in\mathcal{A}(\tau)}\{I_{J^{\boldsymbol{\lambda}}}(\mathbf{a},\omega,\tau;\boldsymbol{\mu})\}\right] = \mathcal{L}^{\circ}J^{\boldsymbol{\lambda}}(\mathbf{x}_0).$$

□

The inequality (31) highlights the comparison between two scenario-based inner optimization problems: the right term of the inequality in (31) allows $\boldsymbol{\mu} = (\boldsymbol{\mu}_0,\cdots,\boldsymbol{\mu}_\tau)$ $\big($contained in $\sum_{t=0}^{\tau}\boldsymbol{\mu}_t^{\top}[\mathbf{b}-\mathbf{B}(\mathbf{x}_t,\mathbf{a}_t)]\big)$ to be different across periods; on the other hand, the left term forces $\boldsymbol{\mu}_t$ to be constant (equal to $\boldsymbol{\lambda}$) over time. Therefore, $\mathcal{L}^{\circ}J^{\boldsymbol{\lambda}}$ can be viewed as an intermediate relaxation between the "exact" information relaxation $\mathcal{L}J^{\boldsymbol{\lambda}}$ and the Lagrangian relaxation $J^{\boldsymbol{\lambda}}$. Another useful observation is that $\boldsymbol{\mu} = (\boldsymbol{\lambda},\cdots,\boldsymbol{\lambda})$ can naturally serve as the initial point to solve $\min_{\boldsymbol{\mu}\geq 0}\max_{\mathbf{a}\in\mathcal{A}(\tau)}\{I_{J^{\boldsymbol{\lambda}}}(\mathbf{a},\omega,\tau;\boldsymbol{\mu})\}$ via the subgradient method.

Note that the computational complexity of the inner optimization problem also depends on the time horizon τ. In case of drawing a sample of τ that is a large number (often occurs when β that is close to 1), we propose a simple remedy to ease computation, i.e., to truncate the random horizon of the relaxed inner optimization problem up to some deterministic time $\mathcal{T} \in \mathbb{N}$ that is sufficiently large. This operation reduces the computational cost in some extreme cases. The next result shows the complexity versus quality trade-off in choosing an appropriate $\mathcal{T}$: a greater truncated horizon $\mathcal{T}$ implies a more difficult inner optimization problem but guarantees better bound.

Corollary 8. *Suppose* $\mathcal{T} \in \mathbb{N}$. *Define*

$$\mathcal{L}^{\circ}_{\mathcal{T}}J^{\boldsymbol{\lambda}}(\boldsymbol{x}) \triangleq J^{\boldsymbol{\lambda}}(\boldsymbol{x}) + \mathbb{E}_0\left[\min_{\boldsymbol{\mu}\geq 0}\max_{\boldsymbol{a}\in\mathcal{A}(\tau)} I_{J^{\boldsymbol{\lambda}}}(\boldsymbol{a},\omega,\tau\wedge\mathcal{T};\boldsymbol{\mu})\right],$$

where $\tau \wedge \mathcal{T} = \min\{\tau, \mathcal{T}\}$*. Then*

(a) $\mathcal{L}^\circ J^{\boldsymbol{\lambda}}(\boldsymbol{x}) \leq \mathcal{L}^\circ_{\mathcal{T}+1} J^{\boldsymbol{\lambda}}(\boldsymbol{x}) \leq \mathcal{L}^\circ_{\mathcal{T}} J^{\boldsymbol{\lambda}}(\boldsymbol{x}) \leq J^{\boldsymbol{\lambda}}(\boldsymbol{x})$.

(b) $\lim_{\mathcal{T}\to\infty} \mathcal{L}^\circ_{\mathcal{T}} J^{\boldsymbol{\lambda}}(\boldsymbol{x}) = \mathcal{L}^\circ J^{\boldsymbol{\lambda}}(\boldsymbol{x})$.

Proof. Proof of Corollary 8 Note that by fixing $\omega \in \Omega$ and $\tau \in \mathbb{N}$, the following inequality holds for any $\mathcal{T} \in \mathbb{N}$:

$$\min_{\boldsymbol{\mu}\geq 0} \max_{\mathbf{a}\in\mathcal{A}(\tau)} I_{J^{\boldsymbol{\lambda}}}(\mathbf{a}, \omega, \tau; \boldsymbol{\mu}) \leq \min_{\boldsymbol{\mu}\geq 0} \max_{\mathbf{a}\in\mathcal{A}(\tau)} I_{J^{\boldsymbol{\lambda}}}(\mathbf{a}, \omega, \tau \wedge (\mathcal{T}+1); \boldsymbol{\mu}) \leq \min_{\boldsymbol{\mu}\geq 0} \max_{\mathbf{a}\in\mathcal{A}(\tau)} I_{J^{\boldsymbol{\lambda}}}(\mathbf{a}, \omega, \tau \wedge \mathcal{T}; \boldsymbol{\mu}) \leq 0.$$

Therefore, the inequality in (a) follows from the above inequality immediately, and the equality in (b) is true due to the monotone convergence theorem. □

3.3.2 The Gap between Two Information Relaxation Bounds

The practical information relaxation bound $\mathcal{L}^\circ H(x)$ effectively reduces the computational cost compared to deriving the exact information relaxation bound $\mathcal{L}H(x)$, though yields a less tight bound. In this subsection we investigate the gap $\mathcal{L}^\circ H(x) - \mathcal{L}H(x)$, which is the average difference between the optimal values of the exact and relaxed inner optimization problems, i.e.,

$$\min_{\boldsymbol{\mu}\geq 0} \max_{\mathbf{a}\in\mathcal{A}(\tau)} I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu}) - \max_{\mathbf{a}\in\bar{\mathcal{A}}(\tau)} I_H(\mathbf{a}, \omega, \tau). \tag{32}$$

[8] established the sufficient conditions such that the Lagrangian duality gap of the weakly coupled deterministic optimization problem is uniformly bounded regardless of the number of the subproblems (see Appendix A.4). We will show a similar result for $\mathcal{L}^\circ H(x) - \mathcal{L}H(x)$ assuming that H is additively separable.

We begin with an intuitive interpretation on the duality gap (32) by looking at two equivalent linear program formulations of (29). We fix $\omega \in \Omega$ and $\tau \in \mathbb{N}$, and assume that the control space $\mathcal{A}$ is finite. For each project $n = 1, \cdots, N$, we can then enumerate all state trajectories of $(x^n_1, \cdots, x^n_\tau)$ (denoted by $(x^{n,n_k}_1, \cdots, x^{n,n_k}_\tau)$ with index n_k) associated with the control sequence $(a^n_1, \cdots, a^n_\tau) \in \mathcal{A}^n(\tau)$ (denoted by

$\mathbf{a}^{n,n_k}$). Noting that $I_H^n(\mathbf{a}^{n,n_k},\omega,\tau;0)-\sum_{t=0}^{\tau}\boldsymbol{\mu}_t^\top\ \mathbf{B}_t^n(x^{n,n_k},a^{n,n_k})=I_H^n(\mathbf{a}^{n,n_k},\omega,\tau;\boldsymbol{\mu})$. Then (29) can be equivalently written as the following linear program,

$$\min_{\{y_n,\boldsymbol{\mu}_t\}}\sum_{n=1}^{N}y_n+\sum_{t=0}^{\tau}\boldsymbol{\mu}_t^\top\mathbf{b}-(\tau+1)(1-\beta)\theta$$

$$\text{s.t. } y_n\geq I_H^n(\mathbf{a}^{n,n_k},\omega,\tau;0)-\sum_{t=0}^{\tau}\boldsymbol{\mu}_t^\top\mathbf{B}^n(x^{n,n_k},a^{n,n_k})\ \text{ for all } n_k,\ n=1,\cdots,N; \tag{33}$$

$$\boldsymbol{\mu}_t\geq 0,\quad t=0,\cdots,\tau.$$

We use p^{n,n_k} to denote the dual variable associated with (33), so the dual linear program is

$$\max_{\{p^{n,n_k}\}}\sum_{n=1}^{N}\sum_{n_k}p^{n,n_k}I_H^n(\mathbf{a}^{n,n_k},\omega,\tau;0)-(\tau+1)(1-\beta)\theta$$

$$\text{s.t. }\sum_{n=1}^{N}\sum_{n_k}p^{n,n_k}\mathbf{B}^n(x_t^{n,n_k},a_t^{n,n_k})\leq\mathbf{b},\quad t=1,\cdots,\tau;$$

$$\sum_{n_k}p^{n,n_k}=1,\quad n=1,\cdots,N;$$

$$p^{n,n_k}\geq 0 \text{ for all } n_k \textit{ and } n=1,\cdots,N,$$

where p^{n,n_k} can be interpreted as the probability assigned to the n_k-th scenario associated with project n. Comparing the above linear program to (27), it can be seen that the feasible control set $\bar{\mathcal{A}}(\tau)$ is enlarged to include all the randomized controls subject to the linking constraint. Therefore, the relaxed inner optimization problem can be viewed as the convexification of the exact inner optimization problem. In addition, the optimal solution to the above linear programs also provides benchmark result on (29), which may help to adjust the parameters used in the subgradient method.

To characterize the gap $\mathcal{L}^\circ H(x)-\mathcal{L}H(x)$, we list some technical assumptions based on Lemma 2 in Appendix A.4. In particular, we denote $\mathbf{B}^n(x_t^n,a_t^n)$ equivalently as $\mathbf{B}_t^n(\mathbf{a}^n,\omega)$, as x_t^n depends on $\mathbf{a}^n$ and ω.

Assumption 1. *For every state* $\boldsymbol{x}\in\mathcal{X}$, $\bar{\mathcal{A}}(\boldsymbol{x})\neq\phi$.

Assumption 2. *Given $\omega \in \Omega$ and $T \in \mathbb{N}$, the sets*

$$S_n \triangleq \{(\boldsymbol{a}^n, \boldsymbol{B}_0^n(\boldsymbol{a}^n, \omega), \cdots, \boldsymbol{B}_T^n(\boldsymbol{a}^n, \omega), I_H^n(\boldsymbol{a}^n, \omega, T)) | \boldsymbol{a}^n \in \mathcal{A}^n(T)\}$$

are non-empty and compact for $n = 1, \cdots, N$.

This assumption is automatically true if each $\mathcal{A}^n$ is finite, or $\mathcal{A}^n(T)$ is compact and each $\mathbf{B}_t^n(\mathbf{a}^n, \omega)$ and $I_H^n(\mathbf{a}^n, \omega, T)$ are continuous functions on $\mathcal{A}^n(T)$.

Assumption 3. *Given $\omega \in \Omega$ and $T \in \mathbb{N}$. For every $n = 1, \cdots, N$, we assume that for any $\tilde{\boldsymbol{a}}^n \in conv(\mathcal{A}^n(T))$, there exists $\boldsymbol{a}^n \in \mathcal{A}^n(T)$ such that*

$$\boldsymbol{B}_t^n(\boldsymbol{a}^n, \omega) \leq (\check{cl}\ \boldsymbol{B}_t^n)(\tilde{\boldsymbol{a}}^n, \omega),\ t = 0, \cdots, T, \tag{34}$$

where $\check{cl}\ \boldsymbol{B}_t^n$ is the function whose component is the convex closure of the corresponding component of $\boldsymbol{B}_t^n$, i.e.,

$$\check{cl}\ \boldsymbol{B}_t^n(\tilde{\boldsymbol{a}}^n, \omega) \triangleq \inf \left\{ \sum_{n_k} p^{n,n_k} \boldsymbol{B}_t^n(\boldsymbol{a}^{n,n_k}, \omega) \middle| \tilde{\boldsymbol{a}}^n = \sum_{n_k} p^{n,n_k} \boldsymbol{a}^{n,n_k},\ \boldsymbol{a}^{n,n_k} \in \mathcal{A}^n(T); \right.$$
$$\left. \sum_{n_k} p^{n,n_k} = 1,\ p^{n,n_k} \geq 0 \right\}.$$

Remark 2. *All the sums in the definition of $\check{cl}\ \boldsymbol{B}_t^n(\tilde{\boldsymbol{a}}^n, \cdot)$ are finite sums.*

This assumption is not trivially satisfied, as (34) can be a vector inequality. However, we can directly verify Assumption 3 is true in several cases.

Case 1. Each $|\mathcal{A}^n|$ is finite, the number of the linking constraints $L = 1$ (therefore, each inequality in (34) is a scalar inequality), and each $\mathbf{B}_t^n(\mathbf{a}^n, \omega)$ (i.e., $\mathbf{B}^n(x_t^n, a_t^n)$) only depends on a_t^n. A typical example is the restless bandit problem, in which the linking constraint is $\sum_{n=1}^N \mathbf{B}^n(x_t^n, a_t^n) = \sum_{n=1}^N a_t^n = 1$ with $a_t^n \in \{0, 1\}$.

Case 2. If $\mathcal{A}^n(T)$ is convex, and the components of each $\mathbf{B}_t^n(\mathbf{a}^n, \omega)$ are convex over $\mathcal{A}^n(T)$ for $t = 0, \cdots, T$. Then $conv(\mathcal{A}^n(T)) = \mathcal{A}^n(T)$, and $(\check{\text{cl}}\ \mathbf{B}_t^n)(\tilde{\mathbf{a}}^n, \omega) = \mathbf{B}_t^n(\tilde{\mathbf{a}}^n, \omega)$.

We present our main result on the gap $\mathcal{L}^\circ H(x) - \mathcal{L}H(x)$.

Theorem 9. *Suppose that H is of the additively separable form $H(\boldsymbol{x}) = \theta + \sum_{n=1}^{N} H^n(x^n)$, and Assumptions 1-3 hold for every $\omega \in \Omega$ and $T \in \mathbb{N}$. Then for all $\boldsymbol{x} \in \mathcal{X}$,*

$$\mathcal{L}^\circ H(\boldsymbol{x}) - \mathcal{L}H(\boldsymbol{x}) \leq \frac{(L-1)\beta + L + 1}{(1-\beta)^2} \max_{n=1,\cdots,N} \Gamma^n, \tag{35}$$

where

$$\begin{aligned}\Gamma^n = & \sup_{x_0^n \in \mathcal{X}^n, a_0^n \in \mathcal{A}^n} \{R^n(x_0^n, a_0^n) + \beta \mathbb{E}[H^n(x_1^n)|x_0^n, a_0^n] - H^n(x_0^n)\} \\ & - \inf_{x_0^n \in \mathcal{X}^n, a_0^n \in \mathcal{A}^n} \{R^n(x_0^n, a_0^n) + \beta \mathbb{E}[H^n(x_1^n)|x_0^n, a_0^n] - H^n(x_0^n)\}.\end{aligned}$$

The proof of Theorem 9 is in Appendix A.4. Theorem 9 not only characterizes the gap between $\mathcal{L}^\circ H(x)$ and $\mathcal{L}H(x)$, but also allows controlling this gap by restricting the feasible region of $\{H^n(\cdot)\}_{n=1}^N$. To be specific, we can add to the linear program (16) or (94) the following constraints on the *Bellman error* of each subproblem (i.e., $R^n(x_0^n, a_0^n) + \beta \mathbb{E}[H^n(x_1^n)|x_0^n, a_0^n] - H^n(x_0^n)$):

$$\Gamma^{n,2} \ \geq R^n(x_0^n, a_0^n) + \beta \mathbb{E}[H^n(x_1^n)|x_0^n, a_0^n] - H^n(x_0^n) \geq -\Gamma^{n,1},$$

for all (x_0^n, a_0^n) with $a_0^n \in \mathcal{A}^n(x_0^n)$, where $\Gamma^{n,1}$ and $\Gamma^{n,2}$ are two positive numbers for $n = 1, \cdots, N$. Suppose that there is a feasible solution to the linear program (16) or (94), then $\mathcal{L}^\circ H(\mathbf{x}) - \mathcal{L}H(\mathbf{x})$ can be bounded by $\frac{(L-1)\beta+L+1}{(1-\beta)^2} \max_{n=1,\cdots,N}\{\Gamma^{n,1} + \Gamma^{n,2}\}$. Note that the greater $\Gamma^{n,1}$ and $\Gamma^{n,2}$ are, the larger the feasible region of $\{H^n(\cdot)\}_{n=1}^N$ is, which implies a better bound $J^{\boldsymbol{\lambda}}(\mathbf{x})$ or $H^{LP}(\mathbf{x})$; they may be used to generate tighter bounds $\mathcal{L}J^{\boldsymbol{\lambda}}(\mathbf{x})$ or $\mathcal{L}H^{LP}(\mathbf{x})$ according to Theorem 5. As a trade-off, the gap between the practical information relaxation bound $\mathcal{L}^\circ H(\mathbf{x})$ and the exact $\mathcal{L}H(\mathbf{x})$ may be enlarged.

As a corollary, Theorem 9 indicates that the gap $\mathcal{L}^\circ H(x) - \mathcal{L}H(x)$ has a uniform bound in N, if the Bellman errors of individual subproblems (and hence Γ^n) are uniformly bounded for all state-action pairs $\{(x_0^n, a_0^n)\}$. Therefore, the relative gap $\frac{\mathcal{L}^\circ H(\mathbf{x}) - \mathcal{L}H(\mathbf{x})}{N}$ vanishes as N goes to infinity. We provide an instance in which $\{\Gamma^n\}_{n=1}^N$ are uniformly bounded with mild conditions on rewards and linking constraints.

Corollary 10.

(a) If $\{\Gamma^n\}_{n=1}^N$ *are uniformly bounded for all subproblems, then* $\mathcal{L}^\circ H(\boldsymbol{x}) - \mathcal{L}H(\boldsymbol{x})$ *is also uniformly bounded with respect to the number of subproblems* N*.*

(b) Let $H(\boldsymbol{x}) = J^{\boldsymbol{\lambda}}(\boldsymbol{x})$ *for some* $\boldsymbol{\lambda} \geq 0$*. Suppose there exists a constant* $C > 0$ *such that* $\{|R^n|, |R^n - \boldsymbol{\lambda}^\top \boldsymbol{B}^n|\}_{n=1}^N$ *are uniformly bounded by* C*. Then* $\{\Gamma^n\}_{n=1}^N$ *are uniformly bounded by* $\frac{4C}{1-\beta}$*.*

Proof. (a) The result directly follows from Theorem 9.

(b) Since $|R^n - \boldsymbol{\lambda}^\top \mathbf{B}^n| \leq C$, it can be seen from (15) that $\{J^{n,\lambda}\}_{n=1}^N$ are uniformly bounded by $\frac{C}{1-\beta}$. Therefore, for all (x_0^n, a_0^n) with $a_0^n \in \mathcal{A}^n$ and $n = 1, \cdots, N$,

$$\frac{2C}{1-\beta} \geq R^n(x_0^n, a_0^n) + \beta \mathbb{E}[J^{\boldsymbol{\lambda},n}(x_1^n)|x_0^n, a_0^n] - J^{\boldsymbol{\lambda},n}(x_0^n) \geq -\frac{2C}{1-\beta},$$

i.e., $\{\Gamma_n\}_{n=1}^N$ are uniformly bounded by $\frac{4C}{1-\beta}$.

□

In other words, if the optimal value is proportional to the number of the subproblems, i.e., $NC_1 \leq V \leq NC_2$ for some $C_1, C_2 > 0$ (e.g., $C_1(1-\beta) \leq |R^n| \leq C_2(1-\beta)$ for all $n = 1 \cdots, N$), then the relative gap $\frac{\mathcal{L}^\circ H(\mathbf{x}) - V(\mathbf{x})}{V(\mathbf{x})}$ converges to the relative gap $\frac{\mathcal{L}H(\mathbf{x}) - V(\mathbf{x})}{V(\mathbf{x})}$ as the number of subproblems N increases.

Remark 3. *All results presented in Section 3.3 have counterparts in the finite-horizon setting; we refer the readers to Appendix A.5 for details.*

3.4 Numerical Examples

To investigate the performance of the information relaxation bounds, we test our method in both discrete-state and continuous-state WCDPs. We compare some heuristic policies with both the Lagrangian bound and the practical information relaxation bound.

3.4.1 Dynamic Product Promotion

We consider dynamic promotion management of perishable items in retail stores or supermarkets following [51]. By dynamically allocating products of different categories to a limited promotion space, these products are more likely to attract customers and bring in more revenues to the retailers. The limited promotion space may refer to the promotion counters, the shelves close to the cashier, or the space available on the advertisement of weekly specials and sales. A perishable item is a product unit that worsens in quality over time and can no longer be sold at a deadline (e.g., the "best by" date). A profit is obtained by the retailer if an item is sold before its deadline; otherwise, a loss is received.

Since perishable products must reach consumers in a timely manner, at each time period the retailer considers selecting a collection of products to the promotion space, which changes the probability the chosen product is sold. Such a selection is subject to the capacity of the promotion space with the goal of maximizing the expected profits in the long run. This problem can be formulated as a weakly coupled dynamic problem with knapsack constraints, and can be generalized to a variety of dynamic resource allocation problems.

3.4.1.1 MDP Model

Our model generalizes the model in [51] in that we assume the products will be restocked and the selection of products is under multi-dimensional knapsack constraints. Suppose There are N items. The n-th item has the deadline S^n. The state space of this item is described by $\mathcal{X}^n = \{0\} \cup \mathcal{S}^n$, where state $x^n \in \mathcal{S}^n = \{1, 2, \cdots, S^n\}$ means that there are x^n remaining periods to deadline (i.e., the item does not perished) and it is not sold, while state 0 means this product needs to be reordered either because it has perished or has been sold. One feature of our model is that we assume the retailer will replenish one item when it is sold or becomes perished, while there is no act of reordering in the model of [51].

At each period, the retailer decides whether to include the n-th item in the promotion space ($a^n = 1$) or not ($a^n = 0$). Therefore, the action space for item n is $\mathcal{A}^n(x^n) = \{0, 1\}$ if $x^n \in \mathcal{S}^n$; otherwise, $\mathcal{A}^n(0) = \{0\}$.

The retailer's decision results in a different probability $\xi^n_{a^n}$ that the n-th item can be sold during this period.

$$P_n(x^n_{t+1} = s - 1 | x^n_t = s, a^n_t) = 1 - \xi^n_{a^n_t}, \quad P_n(x^n_{t+1} = 0 | x^n_t = s, a^n_t) = \xi^n_{a^n_t}, \quad \text{if} \;\; s \in \mathcal{S}^n;$$
$$P_n(x^n_{t+1} = S^n | x^n_t = 0, a^n_t) = 1.$$

In particular, the transition from state 1 to state 0 is not influenced by the action a^n_1 (though the expected revenue is influenced as explained later). When the n-th item is sold or becomes perished, the retailer reorders this item immediately and the new products will arrive the next day in state S^n.

If the item n is sold before the deadline, it yields a profit margin $r^n > 0$. Otherwise, a loss $\varphi^n r^n$ with $\varphi^n \leq 0$ is obtained. Therefore, the expected one-period revenue is $R^n(x^n, a^n) = r^n \xi^n_{a^n}$ for $x^n \in \mathcal{S}^n/\{1\}$, $R^n(1, a^n) = r^n(\xi^n_{a^n} + \varphi^n(1 - \xi^n_{a^n}))$, and $R^n(0, 0) = 0$.

Suppose that the promotion space is available with capacity of $W_0 \geq 1$, and each item n occupies w^n units of promotion space. So the retailer's decision is subject to the constraint $\sum_{n=1}^{N} w^n a^n \leq W_0$. In practice, the retailer may promote at most a certain number of products among the same category or brand per period by allowing a limited capacity of the promotion space. To this end, we can impose extra linking constraints such as $\sum_{n \in \mathcal{N}_k} w^n a^n \leq W_k$ with $\mathcal{N}_k \subseteq \mathcal{N}_0 = \{1, 2, \cdots, N\}$, where $W_k \in [0, W_0]$ for $k = 1, \cdots, K$. Therefore, the resource constraints can be represented as

$$\bar{\mathcal{A}}_t = \left\{ \mathbf{a}_t \in \{0, 1\}^N \middle| a^n_t \in \mathcal{A}^n(x^n_t), \; \sum_{n \in \mathcal{N}_k} w^n a^n_t \leq W_k \text{ for } k = 0, 1, \cdots, K \right\}.$$

Under the multiple capacity constraints, the objective of the retailer is to sequentially select certain products at each period in order to maximize the discounted

Table 1: Parameter Set

	Category 1 $1 \leq n \leq \frac{N}{3}$	Category 2 $\frac{N}{3} + 1 \leq n \leq \frac{2N}{3}$	Category 3 $\frac{2N}{3} + 1 \leq n \leq N$
S^n	8	10	12
w^n	1	2	4
r^n	$[2.5, 3.5]$	$[4.5, 5.5]$	$[7.5, 8.5]$
ϕ^n	-1	-2	-4
ξ_0^n	$[0.10, 0.80]$	$[0.10, 0.80]$	$[0.10, 0.80]$
ξ_1^n	$[\xi_0^n + 0.05, 0.90]$	$[\xi_0^n + 0.05, 0.90]$	$[\xi_0^n + 0.05, 0.90]$

expected reward:

$$V(\mathbf{x}_0) = \max_{\mathbf{a}_0 \in \bar{\mathcal{A}}_0} \left\{ R(\mathbf{x}_0, \mathbf{a}_0) + \beta \mathbb{E}\left[V(\mathbf{x}_1) | \mathbf{x}_0, \mathbf{a}_0 \right] \right\}, \tag{36}$$

where $R(\mathbf{x}_0, \mathbf{a}_0) = \sum_{n=1}^{N} R^n(x_0, a_0)$.

3.4.1.2 Heuristics and Bounds

We consider a 4-dimensional knapsack constraints ($K = 3$) by divide all N perishable items into three categories and the items in the same category share similar property. The parameter values are listed in Table 1 including the deadline S^n, the size w^n, the profit margin r^n, the loss ϕ^n, and the transition probabilities ξ_0^n and ξ_1^n. For instance, the item in the first category has $S^n = 8$, $w^n = 1$, $\phi^n = -1$; r^n is sampled from the uniform distribution on $[2.5, 3.5]$, and ξ_0^n and ξ_1^n are sampled from the uniform distribution $[0.10, 0.80]$ and $[\xi_0^n + 0.05, 0.90]$ (to satisfy $\xi_1^n > \xi_0^n$), respectively.

The set of capacities on the promotion space are chosen as $(W_0, W_1, W_2, W_3) = (10, 4, 6, 4)$: there is a total capacity of 10 units for all the items, while the capacity of each of the category is upper bounded by 4 units, 6 units, and 4 units, respectively, i.e., at most 4 items from category one, 3 items from category two, and 1 item from category three can be promoted. We then solve the problem with the initial condition $\mathbf{x}_0 = (S^1, S^2, \cdots, S^N)^\top$ under different discount factors $\beta = 0.9, 0.95$, and 0.99, and different number of subproblems $N = 12$ and 24.

We first solve the Lagrangian bound $J^{\boldsymbol{\lambda}^*}(\mathbf{x}_0)$ via the linear program (16), where $\upsilon(\cdot)$ has all probability mass on the initial $\mathbf{x}_0$. In all cases, we observe that the only non-zero Lagrangian multipliers are associated with the total capacity constraint (i.e., W_0) and the capacity constraint on the third category (i.e., W_3) of the promotion space. As we increase W_3 from 4 to 8 and keep other W_k ($k = 0, 1, 2$) unchanged, the only non-zero Lagrangian multiplier is associated with the total capacity constraint W_0, meaning that the total capacity is the main constraint on the promotion space. Therefore, by considering different capacity constraints $(W_0) = (10)$ and $(W_0, W_1, W_2, W_3) = (10, 4, 6, 4)$, respectively, we can derive two upper bounds "Lag. Bound 1" and "Lag. Bound 2" on the optimal value V_0; "Lag. Bound 1" can be viewed as the Lagrangian relaxation bound that ignores the respective capacity constraints on three categories, i.e, $(W_1, W_2, W_3) = (4, 6, 8)$, which is implied by the values of the Lagrangian multipliers. In particular, both "Lag. Bound 1" and "Lag. Bound 2" are supersolutions to (36) under the capacity constraints parameters $(W_0, W_1, W_2, W_3) = (10, 4, 6, 4)$. Given these two different approximate values, we compute lower bounds from the one-step greedy policies, as well as upper bounds from the practical information relaxation approach on the optimal value V_0:

- Lag. Policy 1/Lag. Policy 2: By generating 400 random horizons τ and scenarios ω, we estimate $V(\mathbf{x}_0; \alpha)$ in (19) by applying the one-step greedy policy α,

$$\alpha(\mathbf{x}_t) \in \arg\max_{\mathbf{a}_t \in \bar{\mathcal{A}}_t} \left\{ R(\mathbf{x}_t, \mathbf{a}_t) + \beta \mathbb{E}\left[J^{\boldsymbol{\lambda}^*}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t \right] \right\},$$

  where $J^{\boldsymbol{\lambda}^*}$ is "Lag. Bound 1"/"Lag. Bound 2". The average of the sample rewards provides a lower bound on V_0.

- Info. Bound 1/Info. Bound 2: We compute the practical information relaxation bound $\mathcal{L}^{\circ} J^{\boldsymbol{\lambda}^*}(\mathbf{x}_0)$ in (30) based on the same 400 random horizons τ and scenarios ω and solve the associated relaxed inner optimization problems using the subgradient method, where $J^{\boldsymbol{\lambda}^*}$ is "Lag. Bound 1"/"Lag. Bound 2". The average of these optimal values provides an upper bound on V_0.

Table 2: Bound computation times (in seconds)

		LB	UB	UB	LB	UB	UB
N	β	Lag. Bound 1	Lag. Bound 2	Lag. Policy 1	Lag. Policy 2	Info. Bound 1	Info . Bound 2
12	0.90	0.8	2.1	382.1	385.3	1447.2	1482.3
12	0.95	0.8	2.0	840.6	838.3	3793.6	3821.3
12	0.99	0.8	2.1	426.3	424.7	15190.2	15810.4
24	0.90	1.0	2.9	960.5	962.3	2912.7	2925.6
24	0.95	1.0	2.5	165.3	165.7	5527.2	5549.3
24	0.99	1.0	2.5	605.4	603.2	29362.5	29490.1

3.4.1.3 Numerical Results

In Table 2 we list the running time of the lower and upper bounds on a laptop with 1.70GHz Intel Core(TM)i5 with 4GB RAM using Matlab2013b. The running time (in seconds) of solving the Lagrangian relaxation bound by CVX (see [42]) is reported; the total running time (in seconds) of other bounds is calculated over 400 scenarios . It can be observed that more running time is needed in problems with a larger discount factor β, since a larger β implies a longer horizon with higher probability. To compute the information relaxation bound, we use the subgradient method to solve the relaxed inner optimization problem with at most 2000 iterations, or until the norm of the subgradient is exactly zero. To save computational time, we truncate the random time horizon τ up to $\mathcal{T} = 120$. In practice, The actual number of iterations mainly depends on the realization of τ: the greater τ is, generally more iterations are needed to attain convergence in the subgradient method. We observe that the actual computational time of two lower bounds and upper bounds are roughly proportional to $1/(1-\beta)$, which is the average number of horizons under the discount factor β.

In Table 3 we list the numerical results and the corresponding parameters including the discount factor β and N. The estimated bounds are reported with standard errors in parentheses. To facilitate the comparison, we also report the gaps between two upper bounds("UB") and one lower bound("LB"). The relative gaps are also computed as the percentage of the Lagrangian bound and reported in parentheses following the associated gaps.

Table 3: Bounds on dynamic product promotion

		LB	UB	UB	LB	UB	UB
N	β	Lag. Policy 1	Lag. Bound 1	Info. Bound 1	Lag. Policy 2	Lag. Bound 2	Info . Bound 2
12	0.90	220.1(0.41)	232.5	222.9 (0.32)	220.3(0.13)	224.0	221.2(0.10)
Gap	(%)		12.4 (5.3%)	1.8 (0.8%)		3.7 (1.7%)	0.9 (0.4%)
12	0.95	429.6(1.25)	456.4	435.2(0.98)	431.1(0.36)	437.6	432.6(0.22)
Gap	(%)		26.8 (5.9%)	5.6 (1.2%)		6.5 (1.5%)	1.5 (0.3%)
12	0.99	2125.5(3.26)	2247.9	2151.3 (2.63)	2121.7(0.97)	2144.4	2129.4(0.68)
Gap	(%)		122.4 (5.4%)	25.8 (1.1%)		22.7 (1.1%)	7.7 (0.4%)
24	0.90	366.0(0.19)	378.9	368.6(0.16)	366.5(0.06)	369.6	367.9(0.03)
Gap	(%)		12.9 (3.5%)	2.6 (0.7%)		3.1 (0.8%)	1.4 (0.4%)
24	0.95	721.7(0.40)	745.5	727.1(0.22)	720.9(0.08)	725.7	722.9(0.05)
Gap	(%)		23.8 (3.2%)	5.4 (0.7%)		4.8 (0.7%)	2.0 (0.2%)
24	0.99	3559.3(2.92)	3676.0	3575.5(2.43)	3554.4(0.64)	3572.9	3564.5(0.32)
Gap	(%)		116.7 (3.2%)	16.2 (0.4%)		18.5 (0.5%)	10.1 (0.3%)

In all the cases “Lag. Bound 2” are superior to “Lag. Bound 1” as an upper bound on the optimal value V_0, since “Lag. Bound 1” corresponds to the Lagrangian relaxation with only a total capacity constraint on the promotion space. We use these two approximate values to derive respective one-step greedy policies and generate lower bounds based on the same set of scenarios. The relative gaps between “Lag. Bound 1” and “Lag. Policy 1” are comparatively larger (ranging from 3.2% to 5.9%), while the relative gaps between “Lag. Bound 2” and “Lag. Policy 2” are greatly reduced (ranging from 0.5% to 1.7%). It is expected that “Lag. Policy 2” has an advantage over “Lag. Policy 1” in terms of the standard errors, since “Lag. Bound 2” is a better approximate value than “Lag. Bound 1”; therefore, we can obtain an accurate lower bound with a relatively smaller number of scenarios using the approximate value “Lag. Bound 2”.

The practical information relaxations bounds “Info. Bound 1” and “Info. Bound 2” improve the quality of the upper bounds “Lag. Bound 1” and “Lag. Bound 2”, respectively. We observe that in all scenarios the optimal value of the inner optimization problem is no greater than zero; the optimal value generally becomes farther away from zero as the random horizon τ increases. “Info. Bound 1” are quite good upper bounds in terms of the relative gaps (ranging from 0.4% to 1.2%) considering that it is derived from the less satisfying approximate value “Lag. Bound 1” (with relative gaps ranging from 3.2% to 5.9%). This great improvement is because all four capacity constraints are incorporated in the relaxed inner optimization problems. Comparatively, "Info. Bound 2" has a moderate improvement over "Lag. Bound 2" (e.g., the relative reduced gap is reduced from around 1.1% to 0.4% when $N = 12$), as "Lag. Bound 2" is already a good upper bound. In problems with larger discount factors (e.g., $\beta = 0.99$), "Lag. Bound 2" can be even better than "Info. Bound 1"; this may be because the truncated horizon has stronger effects in problems with large discount factor, and a longer horizon τ worsens the performance of practical information relaxation bounds, as the relaxation of the inner optimization problem tends to be weaker with increasing horizons.

In all cases, "Info. Bound 2" that derived upon the better approximate value "Lag. Bound 2" are tighter than "Info. Bound 1". Another advantage of having a good approximate value is reflected in the standard errors of its induced information relaxation bounds: "Info. Bound 2" always has a smaller standard error than "Info. Bound 1", since there is not much space for "Info. Bound 2" to improve upon "Lag. Bound 2". This observation is consistent to the comparison of standard errors of two lower bounds. To conclude, information relaxation approach strengthens the upper bound performance and shows that the Lagrangian relaxation-based greedy policy is very close to optimal. On the other hand, the choice of the approximate values can be critical in the information relaxation approach to generate a tight and accurate dual bound.

3.4.2 Linear Quadratic Control with Nonconvex linking constraint

We next consider a finite horizon linear quadratic control (LQC) problem with a nonconvex linking constraint. We refer the readers to [47] on the information relaxation approach in (unconstrained) finite horizon LQC. Let $\mathbf{x}_t \in \mathcal{X}_t = \mathbb{R}^N$ and $\mathbf{a}_t \in \mathcal{A}_t = \mathbb{R}^N$ denote the state and the action at time t, respectively. The state equation is described by

$$\mathbf{x}_{t+1} = A_t\mathbf{x}_t + B_t\mathbf{a}_t + \mathbf{w}_{t+1}, \quad t = 0, \cdots, T-1, \tag{37}$$

where A_t, B_t are diagonal matrices for $t = 0, \cdots, T-1$, and $\mathbf{w}_t'$s are N-dimensional zero-mean random vectors with finite second moments. In particular, $cov(\mathbf{w}_t) = \Sigma_t$ is a diagonal matrix for $t = 1, \cdots, T$. We denote by $\mathbb{F}$ the natural filtration generated by $\{\mathbf{w}_0, \cdots, \mathbf{w}_{T-1}\}$.

The objective is to minimize the expected cost

$$U_0(\mathbf{x}_0) = \min_{\alpha \in \bar{\mathbb{A}}_{\mathbb{F}}(T)} \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{a}_t^\top \tilde{R}_t \mathbf{a}_t + \mathbf{x}_T^\top Q_T \mathbf{x}_T \middle| \mathbf{x}_0\right], \tag{38}$$

where each $\tilde{R}_t$ and Q_T are diagonal positive definite matrices, and $\bar{\mathbb{A}}_{\mathbb{F}}(T)$ is the set of non-anticipative policies α, where α selects $\mathbf{a} = (\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_{T-1})$ over time such that $\mathbf{a}_t \in \bar{\mathcal{A}}_t = \{\mathbf{a}_t \in \mathbb{R}^N | \tilde{B}(\mathbf{a}_t) \triangleq \sum_{n=1}^N (a_t^n)^2 \geq b\}$ with $b \in \mathbb{R}_+$ for each $t = 0, 1, \cdots, T-1$. The system (37)-(38) is weakly-coupled, since $A_t,\ B_t,\ \Sigma_t,\ \tilde{R}_t$, and Q_T are all diagonal matrices and the linking constraint at time t is $\tilde{B}(\mathbf{a}_t) \geq b$. It is simple to verify that the value function U_0 is well defined for all $b \geq 0$.

Note that the control set $\bar{\mathcal{A}}_t$ is nonconvex, so the optimal policy for (38) cannot be solved to optimality. Instead we consider a simple heuristic. At each period t we compute the one-step greedy policy induced by the value function to the unconstrained problem: we apply such an action if it is already feasible subject to the linking constraint; otherwise, we project it onto the sphere $\partial\bar{\mathcal{A}}_t \triangleq \{\mathbf{a} \in \mathbb{R}^N | \tilde{B}(\mathbf{a}) = b\}$, and use the projection as the action at time t. We call this heuristic "projection policy". The performance of this policy provides an upper bound on (38) (since it is a minimization problem), which will be referred to as "Projection Policy" in Table 4.

To derive a lower bound on U_0 we first consider the Lagrangian relaxation of (38), which turns out to be an unconstrained LQC problem:

$$\begin{aligned} J_0^\lambda(\mathbf{x}_0) &\triangleq \min_{\alpha\in\mathbb{A}_{\mathbb{F}}(T)} \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{a}_t^\top \tilde{R}_t \mathbf{a}_t + \mathbf{x}_T^\top Q_T \mathbf{x}_T - \sum_{t=0}^{T-1} \lambda_t \cdot \left[\tilde{B}(\mathbf{a}_t) - b\right] \middle| \mathbf{x}_0\right] \\ &= \min_{\alpha\in\mathbb{A}_{\mathbb{F}}(T)} \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{a}_t^\top \left(\tilde{R}_t - \lambda_t \cdot \mathbf{I}_N\right) \mathbf{a}_t + \mathbf{x}_T^\top Q_T \mathbf{x}_T \middle| \mathbf{x}_0\right] + \sum_{t=0}^{T-1} \lambda_t^\top b, \end{aligned}$$

where each λ_t is a scalar and $\lambda = (\lambda_0, \cdots, \lambda_{T-1}) \geq 0$, and $\mathbf{I}_N$ is the N-dimensional identity matrix. Noting that $J_0^\lambda(\mathbf{x}_0)$ admits a closed form solution that is quadratic in $\mathbf{x}_0$, provided that every $\tilde{R}_t - \lambda_t \cdot \mathbf{I}_N$ is positive definite:

$$J_t^\lambda(\mathbf{x}_0) = \mathbf{x}_t^\top K_t \mathbf{x}_t + \sum_{s=t}^{T-1} \mathrm{trace}(K_{s+1}\Sigma_{s+1}) + \sum_{s=t}^{T-1} \lambda_s \cdot b, \quad t = 0, \cdots, T.$$

where K_0 is obtained by the Riccati equation $K_T = Q_T$, and

$$K_t = A_t'\left(K_{t+1} - K_{t+1}B_t\left(B_t'K_{t+1}B_t + (\tilde{R}_t - \lambda_t \cdot \mathbf{I}_N)\right)^{-1} B_t'K_{t+1}\right)A_t, \; t = T{-}1, \cdots, 0.$$

We can use stochastic subgradient method to derive a tightest Lagrangian bound on the domain $\mathcal{S} \triangleq \{\lambda \geq 0 | \tilde{R}_t - \lambda_t \cdot \mathbf{I}_N \succ \mathbf{0}, t = 0, \cdots, T-1\}$. Due to the restricted range, the Lagrangian multiplier λ may not be optimal, but J_0^λ is still a valid lower bound on U_0.

Based on the Lagrangian bounds $\{J_t^\lambda\}_{t=1}^T$ we can derive the information relaxation bound through (10) in Appendix A.5 by choosing $H_t = J_t^\lambda(\mathbf{x}_t)$, that is,

$$\begin{aligned} \mathbb{E}_0\Bigg[\max_{\boldsymbol{\mu}\geq 0} \min_{\mathbf{a}\in\mathcal{A}(T)} \Big\{ &\mathbf{x}_T^\top Q_T \mathbf{x}_T + \sum_{t=0}^{T-1} \mathbf{a}_t^\top \tilde{R}_t \mathbf{a}_t + \mu_t \cdot (b_t - B_t(\mathbf{x}_t, \mathbf{a}_t)) \\ &+ \mathbb{E}[J_{t+1}^\lambda(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - J_{t+1}^\lambda(\mathbf{x}_{t+1}) \Big\} \Bigg], \end{aligned} \tag{39}$$

where $\mu = (\mu_0, \cdots, \mu_{T-1})$, and

$$\begin{aligned} &\mathbb{E}[J_{t+1}^\lambda(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - J_{t+1}^\lambda(\mathbf{x}_{t+1}) \\ =& -2(A_t\mathbf{x}_t + B_t\mathbf{a}_t)^\top K_{t+1}\mathbf{w}_{t+1} - \mathbf{w}_{t+1}' K_{t+1}\mathbf{w}_{t+1} + \mathrm{trace}(K_{t+1}\Sigma_{t+1}). \end{aligned}$$

Restricting μ in $\mathcal{S}$, the optimization problem inside the conditional expectation in (39) is

$$\max_{\mu \in \mathcal{S}} \min_{\mathbf{a} \in \mathcal{A}(T)} \Bigg\{ \mathbf{x}_T^\top Q_T \mathbf{x}_T + \sum_{t=0}^{T-1} \mathbf{a}_t^\top (\tilde{R}_t - \mu_t \cdot \mathbf{I}_N)\mathbf{a}_t - 2(A_t\mathbf{x}_t + B_t\mathbf{a}_t)^\top K_{t+1}\mathbf{w}_{t+1} \\ - \mathbf{w}'_{t+1} K_{t+1} \mathbf{w}_{t+1} + \mathrm{trace}(K_{t+1}\Sigma_{t+1}) \Bigg\} \qquad (40)$$

subject to the state dynamics (37). Then the minimization problem in (40) remains a standard deterministic LQ problem, and can be solved efficiently.

Table 4: LQ problem with Nonconvex linking constraint

			Proj. Policy		Unconstrained	Lag. Bound	Info. Relaxation		Duality Gap	
N	b	T	Value	S.E.	Value	Value	Value	S.E.	1	2
10	5	10	61.4693	0.211	34.7883	59.8636	60.0606	0.0028	2.29%	2.01%
20	5	10	88.0457	0.242	71.2836	87.2886	87.6371	0.0085	0.46%	0.25%
50	5	10	189.1857	0.099	182.7984	188.7715	189.0481	0.0039	0.07%	0.02%
100	5	10	364.2132	0.023	361.7224	364.1160	364.1729	0.0004	0.01%	0.00%
10	10	10	104.6974	0.306	34.7883	103.6067	103.7460	0.0026	0.91%	0.79%
20	10	10	123.4789	0.444	71.2836	120.7735	121.5403	0.0090	1.57%	1.13%
50	10	10	209.3848	0.099	182.7984	208.8579	209.1757	0.0046	0.10%	0.04%
100	10	10	374.4066	0.193	361.7224	373.7121	374.2227	0.0118	0.05%	0.01%

In our numerical experiments we set $A_t = B_t = \tilde{R}_t = \mathbf{I}_N$ for $t = 0, \cdots, T-1$, and each diagonal entry of Q_T is sampled from the uniform distribution on $[1, 2]$. We set the initial point $\mathbf{x}_0 = (1, 1, \cdots, 1)^\top$. Here is the procedure to get the bounds in Table 4:

- "Proj. Policy": We generate 10000 sample paths $\mathbf{w} \triangleq (\mathbf{w}_0, \cdots, \mathbf{w}_{T-1})$ and apply the projection policy to compute the sample cost. To reduce the variance, we use the unconstrained problem as a control variate. The average of the adjusted sample costs provides an upper bound on U_0.

- "Unconstrained": The value function to the problem (38) without the linking constraint, i.e., $\bar{\mathcal{A}}_t = \mathbb{R}^N$. It can be seen that the "Unconstrained" is equal to J^0, which is a lower bound on U_0.

- "Lag. Bound": we use (stochastic) subgradient method and run 500 iterations to compute the tightest Lagrangian bound $J_0^{\lambda^*}(\mathbf{x}_0)$. We restrict λ in the range $\mathcal{S}' = \{\lambda \geq 0 | \tilde{R}_t - \lambda_t \cdot \mathbf{I}_N \succeq 0.001 \cdot \mathbf{I}_N, t = 0, \cdots, T-1\} \subseteq \mathcal{S}$ (therefore, $\tilde{R}_t - \lambda_t \cdot \mathbf{I}_N$ is positive definite) to ease the optimization. In our numerical experiments the stochastic gradient with respect to λ is very close to zero, which implies that our Lagrangian bound is already near optimal.

- "Info. Relaxation": We generate another 100 sample paths of $\mathbf{w}$. Based on these sample paths and the Lagrangian bound $J_0^{\lambda^*}$, we compute the relaxed inner optimization problem (40) (also replace $\mathcal{S}$ by $\mathcal{S}'$) using subgradient method that runs at most 80 iterations or until the norm of the subgradient is under the tolerance level (we set it to be 0.001). For most scenarios, this relaxed inner optimization problem can be solved to optimality after around 40 iterations.

- "Duality Gap": We report the relative gaps between the upper bounds and the lower bound as the percentage of the Lagrangian bound. "Duality Gap 1" is the relative gap between "Lag. Bound" and "Proj. Policy", and "Duality Gap 2" is the relative gap between "Info. Relaxation" and "Proj. Policy".

Observing the small gaps between "Proj. Policy" and "Lag. Bound", it is a little surprising to see the excellent performance of the simple projection policy. We also note that this simple policy is not trivial by comparing "Proj. Policy" to "Unconstrained": the weak lower bound of "Unconstrained" indicates that the "projection" should occur in some scenarios if not many. The "Info. Relaxation" improves the quality of the "Lag. Bound", where the duality gaps also behave quite consistently as those in the dynamic production promotion problem. The "Info. Relaxation" bound shows that the projection policy becomes closer to optimal as N increases. In this example, the linking constraint is a non-decreasing function in the number of subproblems. Therefore, the linking constraint becomes weaker as N increases, i.e., the action derived from the unconstrained problem becomes unlikely to violate the linking constraint. As we observe, the optimal value to the constrained problem gets closer to the unconstrained one with increasing N.

3.5 Conclusion

Lagrangian relaxation and information relaxation are developed to tackle the budget and non-anticipativity constraints that exist universally in general stochastic dynamic programs. The attraction of studying the interaction of these relaxations particularly in the setting of weakly coupled dynamic programs is due to the decomposed structure of the Lagrangian bound, as well as the theoretical strong duality guaranteed by the information relaxation. We show that a tighter dual bound, compared with the Lagrangian bound, can be derived by incorporating it into the information relaxation approach. For large-scale problem, we further develop a computational method to obtain the practical information relaxation bound, which implies an intermediate relaxation between the Lagrangian and exact information relaxations. The computation of the practical information relaxation bound is easy to implement, and requires little structure of the linking constraints. We may apply this computational method to the case in which both "easy" and "complicated" linking constraints exist: to balance the complexity and quality of the dual bound, we may choose to dualize the "complicated" constraints in Lagrangian relaxation and incorporating the "easy" constraints in computing the information relaxation bounds.

CHAPTER IV

DUAL FORMULATION OF CONTROLLED MARKOV DIFFUSIONS

The goal of this chapter 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 Hamilton-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 variable. There are several numerical methods based on different approximation schemes: [56] considered the Markov chain approximation method by discretizing the HJB equation; [43] 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 chapter we intend to answer the following questions.

- 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?
- If certain optimal penalty exists, does its structure imply any computational advantage in deriving dual bounds on the optimal value of practical problems?

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 present the information relaxation-based dual formulation of controlled Markov diffusions based on the technical machinery "anticipating stochastic calculus" (see, e.g., [68, 67]). We establish the weak duality, strong duality and complementary slackness results in a parallel way as those in the dual formulation of MDPs. We investigate one type of optimal penalties, i.e., the so-called "value function-based penalty", to answer the second 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. Then we emphasize on the computational aspect using the value function-based optimal penalty so as to answer the third question. 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 compute is proposed to derive dual bounds on the optimal value of the discrete-time model.

It turns out that [28, 27, 26] have pioneered a series of related work for controlled Markov diffusions. They also adopted the approach of relaxing the future information and penalizing. In particular, [28] proposed a Lagrangian approach for penalization, where the Lagrangian term plays essentially the same role as a penalty in our dual framework; in addition, this Lagrangian term has a similar flavor as the gradient-based penalty proposed by [17] for MDPs in terms of their linear forms in actions. The main difference of the work [28] from ours is that we propose a general framework that may incorporate their Lagrangian approach as a special case; the optimal penalty we develop in this chapter is value function-based, which differs from their proposed Lagrangian approach. In addition, their work is purely theoretical and does not suggest any computational method. In contrast, we provide an example to demonstrate the practical use of the value function-based penalty.

Another closely-related literature focuses on the dual representation of the American option pricing problem (that is essentially an optimal stopping problem) [76, 45, 3]. In particular, the structure of the optimal martingale (i.e., the optimal penalty) under the diffusion process is investigated by [7, 91], which leads to practical algorithms for fast computation of tight upper bounds on the American option prices. The form of the optimal martingale also reflects its inherent relationship with the value function-based optimal penalty in the controlled diffusion setting.

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 our numerical experiments the upper bounds on the optimal value 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 this chapter is organized as follows. In Section 4.1, we derive the dual formulation of controlled Markov diffusions. In Section 4.2, we illustrate the dual approach and carry out numerical studies in a dynamic portfolio choice problem. Finally, we conclude with future directions in Section 4.3. We put some of the proofs and discussion of the connection between [7, 91] and our work in Appendix.

4.1 Controlled Markov Diffusion and Its Dual Representation

We begin with a basic setup of the controlled Markov diffusion and its associated Hamilton-Jacobi-Bellman equation in Section 4.1.1. We then develop the dual representation of controlled Markov diffusions and present the main results in Section 4.1.2.

4.1.1 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 [35].

Let us consider a $\mathbb{R}^n$-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 $(\Omega, \mathcal{F}, \mathbb{P})$, following the stochastic differential equation (SDE):

$$dx_t = b(t, x_t, u_t)dt + \sigma(t, x_t)dw_t, \quad 0 \le t \le T, \tag{41}$$

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

A control strategy $\mathbf{u}$ at time t is defined as a stochastic process $\mathbf{u} : [t, T] \times \Omega \to \mathcal{U}$. Given an outcome in Ω (i.e., a realization of $\mathbf{w} \triangleq (w_s)_{t\le s\le T}$), the decision maker chooses the control $u_s \in \mathcal{U}$ at time $s \in [t, T]$.

Definition 1. *A control strategy $\boldsymbol{u} : [t, T] \times \Omega \to \mathcal{U}$ is called an admissible strategy at time t if $\boldsymbol{u}$ is $\mathbb{F}$-progressively measurable (therefore, $\boldsymbol{u}(s, \cdot)$ is $\mathcal{F}_s$-adapted for $s \in [t, T]$), and satisfying $\mathbb{E}\left[\int_t^T ||\boldsymbol{u}(s, \cdot)||^2 ds\right] < \infty$.*

The set of admissible strategies at time t is denoted by $\mathcal{U}_{\mathbb{F}}(t)$.

Let $Q = [0, T) \times \mathbb{R}^n$ and $\bar{Q} = [0, T] \times \mathbb{R}^n$. With the following standard technical conditions imposed on b and σ, the SDE (41) admits a unique pathwise solution when $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$, i.e., $(x_t)_{0 \leq t \leq T}$ is $\mathbb{F}$-progressively measurable and has continuous sample paths almost surely given $x_0 = x \in \mathbb{R}^n$.

Assumption 4. *b and σ are continuous on their domains, respectively, and for some constants $C_1, C_2,$ and $C_\sigma > 0$,*

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

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 that Λ and g satisfy the following polynomial growth conditions.

Assumption 5. *For some constants $C_\Lambda, c_\Lambda, C_g, c_g > 0$,*

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

Given an initial condition $(t, x) \in Q$, the objective is to maximize the expected sum of intermediate rewards and final reward by selecting an admissible strategy $\mathbf{u}$ in $\mathcal{U}_{\mathbb{F}}(t)$:

$$V(t, x) = \sup_{\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(t)} J(t, x; \mathbf{u}), \tag{42}$$
$$\text{where} \quad J(t, x; \mathbf{u}) = \mathbb{E}_{t,x}\left[\Lambda(x_T) + \int_t^T g(s, x_s, u_s)ds\right].$$

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 continuously differentiable in (i.e., C^1) in t and twice continuously differentiable (i.e., C^2) in x on Q. For $L \in C^{1,2}(Q)$, define a partial differential operator A^u by

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

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

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

for some constants $C_L > 0$ and $c_L \geq 0$. The following well-known verification theorem under Assumptions 1 and 2 provides a sufficient condition for the value function and an optimal control strategy using Bellman's principle of dynamic programming.

Theorem 11 (Verification Theorem, Theorem 4.3.1 in [35]). *Suppose Assumptions 1 and 2 hold, and $\bar{V} \in C^{1,2}(Q) \cap C_p(\bar{Q})$ satisfies*

$$\sup_{u\in\mathcal{U}}\{g(t,x,u) + A^u\bar{V}(t,x)\} = 0 \text{ for } (t,x) \in Q, \tag{43}$$

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

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

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

$$g(t,x,u^*(t,x)) + A^{u^*(t,x)}\bar{V}(t,x) = \max_{u\in\mathcal{U}}\{g(t,x,u) + A^u\bar{V}(t,x)\} = 0 \tag{44}$$

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

1. $\bar{V}(t,x) = V(t,x) = \sup_{\boldsymbol{u}\in\mathcal{U}_{\mathbb{F}}(t)} J(t,x;\boldsymbol{u})$*. 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 (43) is the well-known HJB equation associated with the problem (41)-(42).

4.1.2 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 past information and impose a penalty to punish the access to 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 4.2.

We consider the information relaxation that the decision maker 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 the feasible controls, we use $\mathcal{U}(t)$ to denote the set of measurable $\mathcal{U}$-valued control strategies at time t, i.e., $\mathbf{u} \in \mathcal{U}(t)$ if $\mathbf{u}$ is $\mathcal{B}([t,T]) \times \mathcal{F}$-measurable and $\mathbf{u}(s,\cdot)$ takes value in $\mathcal{U}$ for $s \in [t,T]$, 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.

Unlike the case of MDPs, the first technical problem we have to face is to define a solution of (41) with an anticipative control $\mathbf{u} \in \mathcal{U}(0)$. Since it involves the concept of "anticipating stochastic calculus" and Stratonovich integral, we postpone the technical details to Appendix B.1, where we use the decomposition technique to define the solution of an anticipating SDE following [28], [68].

Right now we assume that given a control strategy $\mathbf{u} \in \mathcal{U}(0)$ there exists a unique solution $(x_t)_{t\in[0,T]}$ to (41) and it 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 penalty that is a function of a control strategy $\mathbf{u}\in\mathcal{U}(0)$ and the Brownian motion $\mathbf{w}=(w)_{t\in[0,T]}$. Denote by $\mathcal{M}_{\mathbb{F}}(0)$ the set of *dual feasible penalties*, which are penalties that do not penalize non-anticipative policies in expectation, i.e.,

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

We will show in the dual formulation of controlled Markov diffusions, the set $\mathcal{M}_{\mathbb{F}}(0)$ parallels the role of $\mathcal{M}_{\mathbb{G}}(0)$ in the dual formulation of MDPs.

With an arbitrary choice of $h\in\mathcal{M}_{\mathbb{F}}(0)$, we can determine an upper bound on (42) with $t=0$ by relaxing the constraint on the adaptiveness of control strategies.

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

$$\sup_{\boldsymbol{u}\in\mathcal{U}_{\mathbb{F}}(0)} J(0,x;\boldsymbol{u})\le\mathbb{E}_{0,x}\left[\sup_{\boldsymbol{u}\in\mathcal{U}(0)}\left\{\Lambda(x_T)+\int_0^T g(t,x_t,u_t)dt-h(\boldsymbol{u},\boldsymbol{w})\right\}\right]. \qquad (45)$$

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

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

Then inequality (45) 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 (45) is the counterpart of that in (3) in the context of controlled Markov diffusions: an entire realization of $\mathbf{w}$ is known beforehand, and the objective function depends on this specific realization. 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 side of (45) is a *dual bound* on the value function $V(0,x)$.

References [28, 26, 27] have conducted a series of research on this problem under the name of “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; [28] characterized the value of relaxed information in this case. We would expect that the dual bound associated with the zero penalty can be loose as that in MDPs.

An interesting case is when we choose

$$h^*(\mathbf{u}, \mathbf{w}) = \Lambda(x_T) + \int_0^T g(t, x_t, u_t)dt - V(0, x). \tag{46}$$

Note that $h^* \in \mathcal{M}_{\mathbb{F}}(0)$, since by the definition of $V(0, x)$,

$$\mathbb{E}_{0,x}\left[\Lambda(x_T) + \int_0^T g(t, x_t, u_t)ds\right] \le V(0, x)$$

for all $x \in \mathbb{R}^n$ and $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$.

We also note that by plugging $h = h^*$ in the inner optimization problem in (45), 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 12 (Strong Duality)**.** *For all $x \in \mathbb{R}^n$,*

$$\sup_{\boldsymbol{u}\in\mathcal{U}_{\mathbb{F}}(0)} J(0, x; \boldsymbol{u}) = \inf_{h\in\mathcal{M}_{\mathbb{F}}(0)} \left\{ \mathbb{E}_{0,x}\left[\sup_{\boldsymbol{u}\in\mathcal{U}(0)} \left\{\Lambda(x_T) + \int_0^T g(t, x_t, u_t)dt - h(\boldsymbol{u}, \boldsymbol{w})\right\}\right]\right\}. \tag{47}$$

The minimum of the right hand side of (47) can always be achieved by choosing an $h \in \mathcal{M}_{\mathbb{F}}(0)$ in the form of (46).

Proof. According to the weak duality, the left side of (47) should be less than or equal to the right side of (47); the equality is achieved by choosing $h = h^*$ in (46). □

Theorem12 is the counterpart of Theorem 2.1 in [20] that is developed for the discrete-time problem. Due to the strong duality result, the left side of (47) is referred to as the *primal problem* and the right side of (47) is referred to as the *dual problem*. If $\mathbf{u}^\star$ is a control strategy that achieves the supremum in the primal problem, and $h^\star$ is a dual feasible penalty that achieves the infimum in the dual problem, then they are optimal solutions to the primal and dual problems, respectively. The "complementary slackness condition" in the next theorem, which parallels the result in the discrete-time problem (Theorem 2.2 in [20]), characterizes such a pair $(\mathbf{u}^\star, h^\star)$.

Theorem 13 (Complementary Slackness)**.** *Given $\boldsymbol{u}^\star \in \mathcal{U}_{\mathbb{F}}(0)$ and $h^\star \in \mathcal{M}_{\mathbb{F}}(0)$, 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

$$\begin{aligned}&\mathbb{E}_{0,x}\left[\Lambda(x_T^\star) + \int_t^T g(s, x_s^\star, u_s^\star)ds - h^\star(\boldsymbol{u}^\star, \boldsymbol{w})\right]\\ =&\mathbb{E}_{0,x}\left[\sup_{\boldsymbol{u}\in\mathcal{U}(0)}\left\{\Lambda(x_T) + \int_0^T g(s, x_s, u_s)ds - h^\star(\boldsymbol{u}, \boldsymbol{w})\right\}\right], \qquad (48)\end{aligned}$$

where $x_t^\star$ is the solution of (41) using the control strategy $\boldsymbol{u}^\star = (u_t^\star)_{t\in[0,T]}$ on $[0, t)$ with the initial condition $x_0^\star = x$.

Proof. We first consider sufficiency. Let $\mathbf{u}^\star \in \mathcal{U}_{\mathbb{F}}(0)$ and $h^\star \in \mathcal{M}_{\mathbb{F}}(0)$. We assume $\mathbb{E}_{0,x}[h^\star(\mathbf{u}^\star, \mathbf{w})] = 0$ and (48) holds. Then by the weak duality, $\mathbf{u}^\star$ and $h^\star$ should be optimal to the primal and dual problem, respectively.

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

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

The last inequality holds due to $h^\star \in \mathcal{M}_{\mathbb{F}}(0)$. Since we know $\mathbf{u}^\star$ and $h^\star$ are optimal to the primal and dual problem respectively, then the strong duality result (the equality (47)) implies all the inequalities above are equalities. Therefore, we know $\mathbb{E}_{0,x}[h^{\star}(\mathbf{u}^{\star},\mathbf{w})] = 0$ and (48) holds. □

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 (41)-(42) does exist (see, e.g., Theorem 11(b)), then $\mathbf{u}^*$ and $h^*(\mathbf{u},\mathbf{w})$ defined in (46) is a pair of the optimal solutions to the primal and dual problems. However, we note that the optimal penalty in the form of (46) has no practical use, as it requires knowing the value of $V(0,x)$. Theorem 14 characterizes the form of another optimal penalty, which motivates the numerical approximation scheme that will be illustrated in Section 4.2. The proof of Theorem 14 is in Appendix B.2.

Theorem 14 (Value Function-Based Penalty)**.** *Suppose that the value function $V(t,x)$ for the problem (41)-(42) satisfies the assumptions in Theorem11(b), and $\boldsymbol{y} = (t,x_t)_{t\in[0,T]}$ satisfies the conditions in Proposition 4 in Appendix B.1 (i.e., the Ito formula for Stratonovich integral (103) is valid for $F = V(t,x)$ and $\boldsymbol{y} = (t,x_t)_{t\in[0,T]}$), where $(x_t)_{t\in[0,T]}$ is the solution to (41) with $\boldsymbol{u}\in\mathcal{U}(0)$. For $\boldsymbol{u}\in\mathcal{U}(0)$, define*

$$h_v^*(\boldsymbol{u},\boldsymbol{w}) \triangleq \sum_{i=1}^{m}\int_0^T \left[V_x^\top(t,x_t)\sigma^i(t,x_t)\right]\circ dw_t^i - \frac{1}{2}\int_0^T \left[V_x^\top(t,x_t)\left(\sum_{i=1}^{m}\sigma_x^i\sigma^i(t,x_t)\right)\right.$$
$$\left. + \, tr\left(V_{xx}(t,x_t)(\sigma\sigma^\top)(t,x_t)\right)\right]dt, \qquad (49)$$

where $\boldsymbol{w} = (w_t^1,\cdots,w_t^m)_{0\le t\le T}$, σ^i is the i-th column of σ, σ^{ki} is the (k,i)-th entry of σ, and $\sigma_x^i\sigma^i$ denotes an $n\times 1$ vector with $\sum_{j=1}^n \frac{\partial\sigma^{ki}}{\partial x_j}\sigma^{ji}$ being its k-th entry. Then

1. *If $\boldsymbol{u}\in\mathcal{U}_{\mathbb{F}}(0)$, (49) reduces to the form*

$$h_v^*(\boldsymbol{u},\boldsymbol{w}) = \int_0^T V_x^\top(t,x_t)\sigma(t,x_t)\,dw_t, \qquad (50)$$

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

2. *The strong duality holds in* $V(0,x) =$

$$\mathbb{E}_{0,x}\left[\sup_{\boldsymbol{u}\in\mathcal{U}(0)}\left\{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h_v^*(\boldsymbol{u},\boldsymbol{w})\right\}\right].$$

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

$$V(0,x) = \sup_{\boldsymbol{u}\in\mathcal{U}(0)}\left\{\Lambda(x_T) + \int_0^T g(t,x_t,u_t)dt - h_v^*(\boldsymbol{u},\boldsymbol{w})\right\} \tag{51}$$

$$=\Lambda(x_T^*) + \int_0^T g(t,x_t^*,u_t^*)dt - h_v^*(\boldsymbol{u}^*,\boldsymbol{w}), \tag{52}$$

where $(x_t^*)_{t\in[0,T]}$ *is the solution of (41) using the optimal control* $\boldsymbol{u}^* = (u^*(t,x_t))_{t\in[0,T]}$ *(defined in Theorem11(b)) on* $[0,t)$ *with the initial condition* $x_0^* = x$*.*

Although the value functions $\{V(t,x), 0 \le t \le T\}$ are unknown in real applications, (50) implies that if an approximate value function $\{\hat{V}(t,x), 0 \le t \le T\}$ is differentiable with respect to x, then heuristically, h_v^* can be approximated by $\hat{h}_v(\mathbf{u},\mathbf{w}) \triangleq \int_0^T \hat{V}_x^\top(t,x_t)\sigma(t,x_t)dw_t$ at least for $\mathbf{u}\in\mathcal{U}_{\mathbb{F}}(0)$. Noting that $\{\int_0^t \hat{V}_x^\top(s,x_s)\sigma(s,x_s)dw_s\}_{0\le t\le T}$ is an $\mathbb{F}$-martingale if $\mathbf{u}\in\mathcal{U}_{\mathbb{F}}(0)$ (assuming that $\hat{V}_x^\top(t,x)\sigma(t,x)$ satisfies the polynomial growth condition in x); therefore, $\mathbb{E}_{0,x}[\hat{h}_v(\mathbf{u},\mathbf{w})] = 0$ 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 dual feasible, which means that $\hat{h}_v$ can be used to derive an upper bound on the value function $V(0,x)$ through (45). Hence, in terms of the approximation scheme implied by the form of the optimal penalty, Theorem14 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 [20]) and continuous-time (Theorem13) 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 Theorem14 in Appendix B.2 for controlled Markov diffusions. In Appendix B.3 we briefly review the dual representation of the optimal stopping problem, where an analogous result of Theorem14 exists provided that the evolution of the state is modelled as a diffusion process.

4.2 Dynamic Portfolio Choice Problem

We illustrate the practical use of the dual formulation of controlled Markov diffusions, especially the value function-based optimal penalty developed in Theorem14, in a classic dynamic portfolio choice problem with predictable returns and intermediate consumptions (see, e.g., [79, 63, 64]). Since most portfolio choice problems of practical interest cannot be solved analytically, various numerical methods have been developed including the martingale approach [24, 50], state-space discretization methods [87, 5], and approximate dynamic programming methods [14, 43]. These methods all produce sub-optimal policies, and it is not difficult to obtain lower bounds on the optimal expected utility by Monte Carlo simulation under these policies; on the other hand, an upper bound is constructed by [46] and [17] respectively based on the work by [25] and [20]. The gap between the lower bound and the upper bound can be used to justify the performance of a candidate policy.

In this section we solve a *discrete-time* dynamic portfolio choice problem that is discretized from a continuous-time model (see, e.g., [25, 59]). We consider the time-discretization as it is a common approach to numerically solve the continuous-time problem, and the decisions of investment only occur at discrete-time points. We focus on generating upper bounds on the optimal expected utility of the discrete-time problem using the information relaxation dual approach. In particular, we propose *a new class of penalties for the discrete-time problem* by discretizing the value function-based optimal penalties of the continuous-time problem. These penalties make the inner optimization problem much easier to solve compared with the penalties that directly approximates the optimal penalty of the discrete-time model. We demonstrate the effectiveness of our method in computing dual bounds through numerical experiments. We also discuss more general settings (other than the portfolio choice problem) in which our method can be successfully applied.

4.2.1 The Portfolio Choice Model

We first consider a continuous-time financial market with finite horizon $[0, T]$, which is built on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. There are one risk-free asset and n risky assets that the investor can invest on. The prices of the risk-free asset and risky assets are denoted by S_t^0 and $S_t = (S_t^1, \cdots, S_t^n)^\top$, respectively, and the instantaneous asset returns depend on the m-dimensional state variable ϕ_t:

$$\begin{aligned} dS_t^0 &= r_f S_t^0 dt \\ dS_t &= S_t \bullet (\mu_t dt + \sigma_t dz_t), && (53) \\ d\phi_t &= \mu_t^\phi dt + \sigma_t^{\phi,1} dz_t + \sigma_t^{\phi,2} d\tilde{z}_t, && (54) \end{aligned}$$

where r_f is the instantaneous risk-free rate of return, and $\mathbf{z} \triangleq (z_t)_{0 \leq t \leq T}$ and $\tilde{\mathbf{z}} \triangleq (\tilde{z}_t)_{0 \leq t \leq T}$ are two independent standard Brownian motions that are of dimension n and d, respectively; the drift vector $\mu_t = \mu(t, \phi_t)$ and the diffusion matrix $\sigma_t = \sigma(t, \phi_t)$ in (53) are of dimension n and $n \times n$, where the symbol $\bullet$ denotes the component-wise multiplication of two vectors; the terms $\mu_t^\phi = \mu^\phi(t, \phi_t)$, $\sigma_t^{\phi,1} = \sigma^{\phi,1}(t, \phi_t)$, $\sigma_t^{\phi,2} = \sigma^{\phi,2}(t, \phi_t)$ in (54) are of dimension m, $m \times n$, and $m \times d$, respectively.

We denote the filtration by $\mathbb{F} = \{\mathcal{F}_t, 0 \leq t \leq T\}$, where $\mathcal{F}_t$ is generated by the Brownian motions $\{(z_s, \tilde{z}_s), 0 \leq s \leq t\}$.

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

$$\begin{aligned} dW_t =& W_t \left[\pi_t^\top (\mu_t dt + \sigma_t dz_t) + r_f \left(1 - \pi_t^\top \mathbf{1}_n\right) dt - \tilde{c}_t dt\right] \\ =& W_t \left(\pi_t^\top (\mu_t - r_f \mathbf{1}_n) + r_f - \tilde{c}_t\right) dt + W_t \pi_t^\top \sigma_t dz_t, \end{aligned} \tag{55}$$

where $\mathbf{1}_n$ is the n-dimensional all-ones vector. A control strategy $\mathbf{u}$ with $u_t \triangleq (\pi_t, \tilde{c}_t)$ is an admissible strategy in the sense that

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

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

3. $u_t \in \mathcal{U}$, where $\mathcal{U}$ is a closed convex set in $\mathbb{R}^{n+1}$.

We still use $\mathcal{U}_{\mathbb{F}}(t)$ to denote the set of admissible strategies at time t and we will specify the control space $\mathcal{U}$ later. Suppose that U is a strictly increasing and concave utility function (see, e.g., [62]). The investor's objective is to maximize the weighted sum of the expected utility of the intermediate consumption and the final wealth:

$$V(t, \phi_t, W_t) = \sup_{\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(t)} \mathbb{E}\left[\int_t^T \alpha \beta^s U\left(\tilde{c}_s W_s\right) ds + (1-\alpha)\beta^T U(W_T) \middle| \phi_t, W_t\right], \tag{56}$$

where $\beta \in [0, 1)$ is the discount factor, and $\alpha \in [0, 1]$ indicates the relative importance of the intermediate consumption.

The value function (56) sometimes admits an analytic solution, for example, under the assumption that μ_t is a constant vector and σ_t is a constant matrix in (53), and there is no constraint on $u_t = (\pi_t, \tilde{c}_t)$. A recent progress on the analytic tractability of (56) can be found in [59]. However, (56) 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 discretize the continuous-time problem (54)-(56). 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 (53) is equivalent to

$$d\log(S_t) = \left(\mu_t - \frac{1}{2}\cdot \text{Pdiag}\left(\Sigma_t\right)\right) dt + \sigma_t dz_t,$$

where $\text{Pdiag}(\Sigma_t)$ denotes an n-dimensional vector that is the principal diagonal of $\Sigma_t = \sigma_t\sigma_t^\top$, the covariance matrix of the instantaneous return. That is to say, $S_{k+1} = R_{k+1} \bullet S_k$ with distribution $\log(R_{k+1}) \sim N(\int_{k\delta}^{(k+1)\delta}(\mu_s - \frac{1}{2}\sigma_s^2)ds, \int_{k\delta}^{(k+1)\delta}\Sigma_s ds)$. Hence, we can discretize (54),(53), and (55) as follows:

$$\phi_{k+1} = \phi_k + \mu_k^\phi \delta + \sigma_k^{\phi,1}\sqrt{\delta}Z_{k+1} + \sigma_k^{\phi,2}\sqrt{\delta}\tilde{Z}_{k+1}, \tag{57a}$$

$$\log(R_{k+1}) = \left(\mu_k - \frac{1}{2}\sigma_k^2\right)\delta + \sigma_k\sqrt{\delta}Z_{k+1}, \tag{57b}$$

$$\begin{aligned} W_{k+1} &= W_k\left(R_{k+1}^\top \pi_k\right) + W_k\left(1 - \mathbf{1}_n^\top \pi_k\right)R_f - W_k c_k, \\ &= W_k\left(R_f + (R_{k+1} - R_f\mathbf{1}_n)^\top \pi_k - c_k\right), \end{aligned} \tag{57c}$$

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 (57) 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 risk-free asset. Then the constraint, on the control $a_k \triangleq (\pi_k, c_k)$ for the discrete-time problem, can be defined as

$$\mathcal{A} \triangleq \{(\pi, c) \in \mathbb{R}^{n+1} | \pi \geq 0, c \geq 0, c \leq R_f(1 - \mathbf{1}_n^\top \pi)\}. \tag{58}$$

Since c_k is used to approximate $\tilde{c}_k \delta$, (58) corresponds to a control set for the continuous-time model, which is defined as

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

Let $\mathbb{A}_\mathbb{G}$ again denote the set of $\mathcal{A}$-valued non-anticipative control strategies $\mathbf{a}$, which selects the decisions $(a_0, \cdots, a_{K-1})$ that are adapted to the filtration $\mathbb{G}$. The discretization of (56) serves as the value function to the discrete-time problem:

$$H_0(\phi_0, W_0) = \sup_{\mathbf{a} \in \mathbb{A}_\mathbb{G}} \mathbb{E}_0 \left[\sum_{k=0}^{K-1} \alpha \beta^{k\delta} U(c_k W_k) \delta + (1 - \alpha) \beta^{K\delta} U(W_K) \right], \tag{59}$$

which can be solved via dynamic programming:

$$\begin{aligned} H_K(\phi_K, W_K) &= (1 - \alpha) \beta^{K\delta} U(W_K); \\ H_k(\phi_k, W_k) &= \sup_{a_k \in \mathcal{A}} \left\{ \alpha \beta^{k\delta} U(c_k W_k) \delta + \mathbb{E}\left[H_{k+1} | \phi_k, W_k, a_k\right] \right\}. \end{aligned} \tag{60}$$

We will focus on *solving the discrete-time model* (57)-(59), which is discretized from the continuous-time model (54)-(56). Though our methods proposed later can be applied 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*, $U(x) = \frac{1}{1-\gamma} x^{1-\gamma}$, which are widely used in economics and finance. Since the utility functions are of CRRA type, both value functions (56) and (59) have simplified structures. To be specific, the value function to the continuous-time problem can be written as the factorization (see, e.g., [59])

$$V(t, \phi_t, W_t) = \beta^t W_t^{1-\gamma} \tilde{J}(t, \phi_t), \tag{61}$$

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

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

and the value function to the discrete-time problem, due to the factorization scheme, can be written as

$$H_k(\phi_k, W_k) = \beta^{k\delta} W_k^{1-\gamma} J_k(\phi_k), \tag{62}$$

where J_k, the discrete-time reward functional, is defined recursively as $J_K(\phi_K) = (1-\alpha)/(1-\gamma)$ and

$$J_k(\phi_k) = \sup_{(\pi_k, c_k)\in\mathcal{A}} \left\{ \frac{\alpha c_k^{1-\gamma}\delta}{1-\gamma} + \beta^\delta \mathbb{E}\left[\left(R_f + (R_{k+1} - R_f)^\top \pi_k - c_k\right)^{1-\gamma} J_{k+1}(\phi_{k+1}) | \phi_k \right] \right\}. \tag{63}$$

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 are independent of the wealth W_t given ϕ_t (respectively, the optimal (π_k, c_k) in discrete-time model are independent of the wealth W_k given ϕ_k). So the dimension of the state space in (60) 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 (63) including the state-space discretization approach [87, 5], and a simulation-based method [14].

4.2.2 Penalties and Dual Bounds

In this subsection, we compute upper bounds on the optimal value H_0 of the discrete-time (and continuous-state) model (57)-(59) based on the dual approach for MDPs in Theorem 1. We illustrate how to generate two dual feasible penalties for the discrete-time problem: one directly approximates the value function-based penalty of the discrete-time problem, while the other one is derived by discretizing the value function-based penalty of the continuous-time problem (54)-(56). We discuss why the latter approach is more desirable to generate upper bounds on H_0 in terms of computational tractability of the inner optimization problem.

Throughout this subsection we assume 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 from $\hat{J}_k(\phi)$ or 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, assuming 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),

$$\mathbb{E}\left[M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})|\phi_0, W_0\right] \leq 0 \quad \text{for all} \quad \mathbf{a} \in \mathbb{A}_{\mathbb{G}}. \tag{64}$$

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

$$\begin{aligned}&(\mathcal{L}M)(\phi_0, W_0)\\ =&\mathbb{E}\left[\sup_{\mathbf{a}\in\mathbb{A}}\{\sum_{k=0}^{K-1}\alpha\beta^{k\delta}U(c_kW_k)\delta + (1-\alpha)\beta^{K\delta}U(W_K) - M(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})\}\middle|\phi_0, W_0\right].\end{aligned} \tag{65}$$

Based on Theorem1(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, we introduce equivalent decision variables $\Pi_k = W_k\pi_k$ and $C_k = W_kc_k$, 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 inside the conditional expectation in

(65) as follows:

$$\max_{\Pi,C,W} \sum_{k=0}^{K-1} \alpha\beta^{k\delta}U(C_k)\delta + (1-\alpha)\beta^{K\delta}U(W_K) - M(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) \tag{66a}$$

$$\text{s.t. } \phi_{k+1} = \phi_k + \mu_k^{\phi}\delta + \sigma_k^{\phi,1}\sqrt{\delta}Z_{k+1} + \sigma_k^{\phi,2}\sqrt{\delta}\tilde{Z}_{k+1}, \tag{66b}$$

$$\log(R_{k+1}) = (\mu_k - \frac{1}{2}\sigma_k^2)\delta + \sigma_k\sqrt{\delta}Z_{k+1}, \tag{66c}$$

$$W_{k+1} = W_kR_f + (R_{k+1} - R_f\mathbf{1}_n)^\top\Pi_k - C_k, \tag{66d}$$

$$\Pi_k \geq 0, \quad C_k \geq 0, \tag{66e}$$

$$C_k \leq R_f(W_k - 1_n^\top\Pi_k), \text{ for } k = 0, \cdots, K-1. \tag{66f}$$

Note that (66b)-(66d) are equivalent to (57a)-(57c), and(66e)-(66f) are equivalent to (58). The advantage of this reformulation is that the inner optimization problem (66) has linear constraints. Therefore, we may find the global maximizer of (66) as long as the objective function in (66a) 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 investigate the optimal penalty M^* for the discrete-time problem according to (5):

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

where ΔH_{k+1} is the deviation in H_{k+1} from the conditional mean. In practice we can 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 the conditional expectation (that is, $\mathbb{E}_k[\hat{H}_{k+1}]$) over $(n+d)$-dimensional space is involved. We may use sample average estimation to obtain its accurate approximation, though, at the expense of substantial computational efforts. Another difficulty is that $M^* = \sum_{k=0}^{K-1}\Delta H_{k+1}$ enters into (66a) with possibly positive or negative signs for different realizations of $(\mathbf{Z},\tilde{\mathbf{Z}})$, making the objective function of (66) nonconcave, even if U is a concave function. Therefore, it might be extremely hard to locate the global maximizer of (66).

To address these problems, we propose *another dual feasible penalty for the discrete-time problem*, and describe how to efficiently compute this penalty. This new class of penalties are derived by exploiting the value function-based optimal penalty h_v^* for the continuous-time problem (54)-(56), recalling that our discrete-time problem is discretized from the continuous-time model. We assume that all the technical conditions in Theorem14 hold, and then we can apply the result (50) by selecting $x_t = (\phi_t, W_t)$, $V(t, x_t) = V(t, \phi_t, W_t)$, $\sigma(t, x_t) = \begin{pmatrix} \sigma_t^{\phi,1} & \sigma_t^{\phi,2} \\ W_t\pi_t\sigma_t & 0 \end{pmatrix}$, and $dw_t = \begin{pmatrix} dz_t \\ d\tilde{z}_t \end{pmatrix}$ such that

$$\begin{aligned} h_v^*(\mathbf{u}, \mathbf{z}, \tilde{\mathbf{z}}) =& \int_0^T \begin{pmatrix} V_\phi(t, \phi_t, W_t) \\ V_W(t, \phi_t, W_t) \end{pmatrix}^\top \begin{pmatrix} \sigma_t^{\phi,1} & \sigma_t^{\phi,2} \\ W_t\pi_t\sigma_t & 0 \end{pmatrix} \begin{pmatrix} dz_t \\ d\tilde{z}_t \end{pmatrix} \\ =& \sum_{k=0}^{K-1} \int_{k\delta}^{(k+1)\delta} \Big[V_\phi^\top(t, \phi_t, W_t)\sigma_t^{\phi,1} dz_t \\ &+ V_\phi^\top(t, \phi_t, W_t)\sigma_t^{\phi,2} d\tilde{z}_t + V_W(t, \phi_t, W_t)W_t\pi_t\sigma_t dz_t \Big] \\ =& \sum_{k=0}^{K-1} \int_{k\delta}^{(k+1)\delta} \beta^t \Big[W_t^{1-\gamma}\nabla_\phi \tilde{J}^\top(t, \phi_t)\sigma_t^{\phi,1} dz_t + W_t^{1-\gamma}\nabla_\phi \tilde{J}^\top(t, \phi_t)\sigma_t^{\phi,2} d\tilde{z}_t \\ &+ (1-\gamma)W_t^{1-\gamma}\tilde{J}(t, \phi_t)\pi_t\sigma_t dz_t \Big], \end{aligned} \tag{67}$$

for $\mathbf{u} = (\pi_t, \tilde{c}_t)_{0 \le t \le T} \in \mathcal{U}_{\mathbb{F}}(0)$, and the last equality holds due to the structure of the value function (61). In particular, we use $\nabla_\phi \tilde{J}$ to denote the gradient of the function $\tilde{J}$ with respect to ϕ. By discretizing the Ito stochastic integrals in (67), we propose a heuristic – using the $(k+1)$-th term in the summation – to approximate ΔH_{k+1} in M^*, that is,

$$\begin{aligned} \Delta H_{k+1} \approx& \beta^{k\delta}\big[W_k^{1-\gamma}\nabla_\phi J_k^\top(\phi_k)\sigma_k^{\phi,1}\sqrt{\delta}Z_{k+1} + W_k^{1-\gamma}\nabla_\phi J_k^\top(\phi_k)\sigma_k^{\phi,2}\sqrt{\delta}\tilde{Z}_{k+1} \\ &+ (1-\gamma)W_k^{-\gamma}J_k(\phi_k)\Pi_k^\top\sigma_k\sqrt{\delta}Z_{k+1} \big], \end{aligned} \tag{68}$$

where we use $J_k(\phi)$ to approximate $\tilde{J}(k\delta, \phi)$ and also use the substitution $\Pi_k = W_k\pi_k$.

We then describe a procedure to numerically approximate ΔH_{k+1} using simulation based on (68). Given a realization of $(\mathbf{Z}, \tilde{\mathbf{Z}})$, we can obtain the realized terms of $\bar{\phi}_k \triangleq$

$\phi_k(\phi_0, \mathbf{Z}, \tilde{\mathbf{Z}})$, $\bar{\sigma}_k \triangleq \sigma(\bar{\phi}_k)$, $\bar{\sigma}_k^{\phi,1} \triangleq \sigma^{\phi,1}(k, \bar{\phi}_k)$, $\bar{\sigma}_k^{\phi,2} \triangleq \sigma^{\phi,2}(k, \bar{\phi}_k)$; with an admissible strategy $\hat{\mathbf{a}}$, we can also obtain $\bar{W}_k \triangleq W_k(W_0, \hat{\mathbf{a}}(\phi_0, W_0, \mathbf{Z}, \tilde{\mathbf{Z}}), \mathbf{Z}, \tilde{\mathbf{Z}})$ via (57c) as an approximation to the wealth under the optimal policy. Then we can approximate ΔH_{k+1} by $\Psi_k^1\left(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}\right) Z_{k+1} + \Psi_k^2(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})\tilde{Z}_{k+1}$, where

$$\begin{aligned}\Psi_k^1(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) =&\beta^{k\delta}\left[\bar{W}_k^{1-\gamma}\Xi_k^{2\top}\left(\bar{\phi}_k\right)\bar{\sigma}_k^{\phi,1}\sqrt{\delta} + (1-\gamma)\bar{W}_k^{-\gamma}\Xi_k^1(\bar{\phi}_k)\Pi_k^\top\bar{\sigma}_k\sqrt{\delta}\right], \qquad (69)\\ \Psi_k^2(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) =&\beta^{k\delta}\bar{W}_k^{1-\gamma}\Xi_k^{2\top}(\bar{\phi}_k)\bar{\sigma}_k^{\phi,2}\sqrt{\delta},\end{aligned}$$

and where $\Xi_k^1(\cdot)$ is a scalar function of ϕ, whereas $\Xi_k^2(\cdot)$ is an m-dimensional function of ϕ. Therefore, we can further approximate $M^* = \sum_{k=0}^{K-1}\Delta H_{k+1}$ by

$$M_1(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \triangleq \sum_{k=0}^{K-1}\left(\Psi_k^1\left(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}\right) Z_{k+1} + \Psi_k^2(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})\tilde{Z}_{k+1}\right), \qquad (70)$$

We will verify in Proposition 2 below that M_1 is dual feasible in the sense of (64) given any functions Ξ_k^1 and Ξ_k^2, and hence $\mathcal{L}M_1$ is an upper bound on H_0. To derive a tight upper bound, it is suggested by (68) that $\Xi_k^1(\cdot)$ and $\Xi_k^2(\cdot)$ are preferably chosen as $\hat{J}_k(\cdot)$ – an approximation of $J_k(\cdot)$, and $\nabla_\phi\hat{J}_k(\phi_k)$ – an approximation of $\nabla_\phi J_k(\phi_k)$, respectively. It is worth noting that the differentiability of $\hat{J}_k(\phi)$ is not required to validate the dual feasibility of the penalty M_1. In the case that $\hat{J}_k(\phi)$ is not differentiable in ϕ, we may apply the finite difference method on $\hat{J}_k(\phi_k)$ to obtain the difference quotient as $\Xi_k^2(\cdot)$ (i.e., a nominal approximation of $\nabla_\phi\hat{J}_k(\phi_k)$).

It remains to show why the forms of Ψ_k^1 and Ψ_k^2 make the inner optimization problem (66) easy to solve. This is because both functions are *affine* in $\mathbf{a}$, regardless of the realizations of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. To be specific, when a realization of $(\mathbf{Z}, \tilde{\mathbf{Z}})$ is fixed, Ψ_k^2 is a constant with respect to $\mathbf{a}$, while Ψ_k^1 is affine in Π_k (hence, in $\mathbf{a}$). Therefore, together with the concave property of $U(\cdot)$, the inner optimization problem (66) is guaranteed to be convex with $M = M_1$. To find some variants of the penalties while still keeping the convexity of the inner optimization problem, we also generate $\breve{\Psi}_{k+1}^1$ based on a first-order Taylor expansion of Ψ_{k+1}^1 in (69) around the strategy $\hat{a}_{k-1}$,

$k = 1, \cdots, K$ (we only expand the first term, since the second term is already linear in Π_k):

$$\begin{aligned}\breve{\Psi}^1_{k+1}(\mathbf{a},\mathbf{Z},\tilde{\mathbf{Z}}) = \beta^{k\delta}\big[&\bar{W}_k^{1-\gamma} + (1-\gamma)\bar{W}_k^{-\gamma}\big((\bar{R}_k - R_f\mathbf{1}_n)^\top \\ &(\Pi_{k-1}-\bar{\Pi}_{k-1}) - (C_{k-1}-\bar{C}_{k-1})\big)\big]\cdot \Xi_k^{2\top}(\bar{\phi}_k)\bar{\sigma}_k^{\phi,1}\sqrt{\delta} \\ &+ \beta^{k\delta}(1-\gamma)\bar{W}_k^{1-\gamma}\Xi_k^1(\bar{\phi}_k)\Pi_k^\top\bar{\sigma}_k\sqrt{\delta},\end{aligned}$$

where $\bar{R}_k \triangleq R_k(\phi_0, \mathbf{Z}, \tilde{\mathbf{Z}})$, $(\bar{\Pi}_k, \bar{C}_k) \triangleq \hat{a}_k(\phi_0, W_0, \mathbf{Z}, \tilde{\mathbf{Z}})$. Then $\breve{\Psi}^1_{k+1}$ is affine in Π_{k-1} and C_{k-1}. We can also obtain a variant of Ψ^2_{k+1} that is is affine in Π_{k-1} and C_{k-1}, say $\breve{\Psi}^2_{k+1}$, in exactly the same way. In our numerical examples we will consider dual bounds generated by M_1 as well as M_2, where

$$M_2(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}}) \triangleq \sum_{k=0}^{K-1} \left(\breve{\Psi}^1_k(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})Z_{k+1} + \breve{\Psi}^2_k(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})\tilde{Z}_{k+1}\right). \tag{71}$$

To go further, we can also generate a penalty function by linearizing Ψ^1_{k+1} around $(\hat{a}_0, \cdots, \hat{a}_{k-1})$. We show $M_2 \in \mathcal{M}_{\mathbb{G}}(0)$ in Proposition 2 as well.

Proposition 2. *Both M_1 and M_2 are dual feasible in the sense of (64), i.e., $M_1, M_2 \in \mathcal{M}_{\mathbb{G}}(0)$. Hence, both $\mathcal{L}M_1$ and $\mathcal{L}M_2$ are upper bounds on H_0.*

Proof. First, we show that $\Psi^i_k(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})$ is $\mathcal{G}_k$-adapted given any $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$ for $i = 1, 2$. Noting that $\bar{\phi}_k$, $\Xi^1_k(\bar{\phi}_k)$, $\Xi^2_k(\bar{\phi}_k)$, $\bar{\sigma}_k$, $\bar{\sigma}^{\phi,j}_k$ $(j = 1, 2)$, and $\bar{W}_k$ are naturally $\mathcal{G}_k$-adapted under a fixed non-anticipative policy $\hat{\mathbf{a}} \in \mathbb{A}_{\mathbb{G}}$. Therefore, $\Psi^2_{k+1}(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})$ is $\mathcal{G}_k$-adapted. We also observe that Π_k is $\mathcal{G}_k$-adapted as $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$; therefore, $\Psi^1_k(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})$ is $\mathcal{G}_k$-adapted for any $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$.

Second, since Z_{k+1} and $\tilde{Z}_{k+1}$ have zero means and are independent of $\mathcal{G}_k$ and (ϕ_0, W_0), along with the linearity of Ψ^1_k (resp., Ψ^2_k) in Z_{k+1} (resp., $\tilde{Z}_{k+1}$), we have for $k = 0, \cdots, K-1$,

$$\begin{aligned}\mathbb{E}\left[\Psi^1_k \cdot Z_{k+1} \middle| \phi_0, W_0\right] &= \mathbb{E}_0\left[\Psi^1_k \cdot \mathbb{E}_k[Z_{k+1}]\right] = 0 \text{ for all } \mathbf{a} \in \mathbb{A}_{\mathbb{G}}; \\ \mathbb{E}\left[\Psi^2_k \cdot \tilde{Z}_{k+1} \middle| \phi_0, W_0\right] &= \mathbb{E}_0\left[\Psi^2_k \cdot \mathbb{E}_k[\tilde{Z}_{k+1}]\right] = 0 \text{ for all } \mathbf{a} \in \mathbb{A}_{\mathbb{G}}.\end{aligned}$$

Therefore, $\mathbb{E}[M_1(\mathbf{a}, \mathbf{Z}, \tilde{\mathbf{Z}})|\phi_0, W_0] = 0$ for all $\mathbf{a} \in \mathbb{A}_{\mathbb{G}}$, and hence $M \in \mathcal{M}_{\mathbb{G}}(0)$. The same argument can also apply on M_2. Therefore, $M_2 \in \mathcal{M}_{\mathbb{G}}(0)$. □

4.2.3 Discussion on Penalties

In this subsection we compare our penalty-generating method with some available approaches that are designed for stochastic dynamic programs in the literature. We also discuss a broader class of controlled diffusion problems in which our proposed penalties can be applied.

The first approach of constructing penalties of the discrete-time problem is proposed in [20], which suggests directly approximating the optimal values in the value function-based penalty (e.g, the first approach discussed in Section 4.2.2). The dual feasibility of this class of penalties is ensured by computing the conditional expectation term accurately, which may involve a substantial amount of computational work. Later, [17] proposes a gradient-based penalty (the gradient is taken with respect to a policy) that requires to solve a stochastic decision-making problem, which is generally simpler than the problem of interest but with similar problem structure. The dual feasibility of these gradient-based penalties relies on the computational tractability of the optimal policy to the simpler problem. In the setting of convex stochastic dynamic programs, the recent work [19] develops a new class of gradient-based penalties, which can be viewed as the combination of the previous two classes of penalties. This new penalty circumvents the requirement of deriving an optimal policy, though it involves conditional expectations over the subgradients of an approximate value with respect to a suboptimal policy.

In contrast, relying on the settings of Markov diffusions, our proposed penalties (70) and (71) do not involve any conditional expectation, while the only extra computational work comes from estimating the difference quotient (or gradient) of the approximate value function with respect to the state variable. Therefore, this new class of penalties can be evaluated very efficiently. Furthermore, the design of our proposed penalties is quite flexible: we can use any suboptimal policy to obtain a dual feasible penalty, and linearize around this policy if necessary, which guarantees the convexity of the inner optimization problem (66).

We will provide some insights on how to generalize our penalty-generating method to more controlled diffusion problems other than the dynamic portfolio choice problem. Recall that the purpose of our penalty-generating methods is to make the inner optimization problem a *convex program*. So our proposed penalty can be applied in problems with the following two features.

(i) The discrete-time state dynamic (derived by discretizing the continuous-time dynamic (41)) is linear in its decision variables, which may be done by reformulation or introducing extra decision variables.

(ii) The reward function (resp., cost function) to be maximized (resp., minimized) is concave (resp., convex) in the decision variables.

Since our proposed penalty can be linearized with respect to the decision variables, the inner optimization problem remains convex with the linearized penalty, provided the above two assumptions hold. To illustrate our points, we provide two examples below and they are in scalar cases for simplicity.

1. Suppose $b(t, x, u) = A^1 x + A^2 u$ and $\sigma(t, x, u) = A^3 x + A^4 u$, where all A^i are constants. The state equation after discretization is

   $$x_{k+1} = x_k + (A^1 x_k + A^2 u_k)\delta + (A^3 x_k + A^4 u_k)\sqrt{\delta} \cdot Z_{k+1},$$

   where the index k denotes the time $k\delta$, and Z_{k+1} is a standard Gaussian random variable for $k = 0, \cdots, K-1$. In addition, we require the reward function $\sum_{k=0}^{K-1} R_k(x_k, u_k) + \Lambda(x_T)$ is jointly concave in $\{x_k\}_{k=1}^{K}$ and $\{u_k\}_{k=0}^{K-1}$. The linear convex (e.g., quadratic) control problem with convex constraints lies in this category.

2. Suppose $b(t,x,u) = A^1 x + A^2 xu$ and $\sigma(t,x,u) = A^3 x + A^4 xu$, where all A^i are constants. The state equation after discretization is

$$\begin{aligned} x_{k+1} =& x_k + (A^1 x_k + A^2 x_k u_k)\delta + (A^3 x_k + A^4 x_k u_k)\sqrt{\delta} \cdot Z_{k+1}, \\ =& x_k + (A^1 x_k + A^2 U_k)\delta + (A^3 x_k + A^4 U_k)\sqrt{\delta} \cdot Z_{k+1}, \end{aligned}$$

where $U_k = x_k u_k$, and Z_{k+1} is a standard Gaussian random variable for $k = 0, \cdots, K-1$. In addition, we require the reward function can be reformulated as $\sum_{k=0}^{K-1} R_k(x_k, U_k) + \Lambda(x_T)$, which is jointly concave in $\{x_k\}_{k=1}^{K}$ and $\{U_k\}_{k=0}^{K-1}$. The dynamic portfolio choice problem and many financial decision-making problems lie in this category.

4.2.4 Numerical Examples

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 (59). We consider a model with three risky assets ($n = 3$) and one market state variable ($m = 1$). The dynamics (53)-(54) of the market state and assets returns are the same as those considered in [46]. In particular, let $\mu_k^{\phi} = -\lambda\phi_k$, $\mu_k = \mu_0 + \mu_1\phi_k$, $\sigma_k \equiv \sigma$, $\sigma_k^{\phi,1} \equiv \sigma^{\phi,1}$, and $\sigma_k^{\phi,2} \equiv \sigma^{\phi,2}$, in (57a)-(57b). The parameter values are listed in the following tables including r_f, λ, μ_0, μ_1, σ, $\sigma^{\phi,1}$, and $\sigma^{\phi,2}$. Note from (54) that the market state ϕ follows a mean-reverting Ornstein-Uhlenbeck process: it has relatively small mean reversion rate and volatility in the parameter set 1, while it has relatively large mean reversion rate and volatility in the parameter set 2. 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. We assume $\phi_0 = 0$ and $W_0 = 1$ as the initial condition and impose the constraint (58) on the control space $\mathcal{A}$ in the following numerical tests.

Table 5: 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 6: 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

For each parameter set, we first solve the the recursion (63) assuming that ϕ_{k+1} and R_{k+1} are independent conditioned on ϕ_k. We will use the numerical solution to this simplified continuous-state problem as $\hat{J}_k(\phi)$ and $\hat{\mathbf{a}}$, which are presumed to be available in Section 4.2.2. The numerical method we employ is the discrete state-space approximation method. To be specific, we approximate the market state variable ϕ_k using a grid with 21 equally spaced grids from -2 to 2, and the transition between these grid points is determined by (57a) 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., [53]). 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 (63), i.e., a finite weighted sum. For the optimization problem in (63) 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 and the corresponding policy on this grid at each time $k = 0, \cdots, K$. We then extend these value functions and policies on the real line (noting that the market state variable ϕ_k is one dimensional) by piecewise linear interpolation. These extended functions are regarded as the numerical solution to the recursion (63) with the assumption that ϕ_{k+1} and R_{k+1} are conditionally independent.

In our numerical implementation these piecewise linear value function and policy function play the roles of $\hat{J}_k(\phi)$ (i.e., $\Xi_k^1(\phi)$) and approximate policy $\hat{\mathbf{a}}$ to the continuous-state problem (57)-(59) with the assumption that ϕ_{k+1} and R_{k+1} are conditionally dependent. We take the slope of the piecewise linear function as $\nabla\hat{J}_k(\phi)$ (i.e., $\Xi_k^2(\phi)$), if ϕ is between the grid points; otherwise, we can use the average slope of two consecutive lines as $\Xi_k^2(\phi)$, which is equivalent to computing the difference quotient of $\hat{J}_k(\phi)$ via central difference method.

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 (57)-(59) (assuming that ϕ_{k+1} and R_{k+1} are conditionally dependent). 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 (66) with penalty M_1 in (70) or M_2 in (71), 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 [40] 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.

Table 7: 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 8: 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	0.1122	-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)		

We also compute the certainty equivalent of the expected utility in the sub-column "CE", i.e., the equivalent wealth left 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 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 the second term in (69) 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.

There are several possible reasons for 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 (66a) is more sensitive to the solution of the inner optimization problem (66), 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 the same approximation scheme of the recursion (63) 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 sub-optimal 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.

4.3 Conclusion

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 the 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.

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, we investigate the sensitivity of the quality of both lower and upper bounds in terms of duality gaps with respect to different parameters. These numerical studies complement the existing examples of applying the dual approach to continuous-state MDPs.

This dual formulation also offers a straightforward extension to the jump-diffusion models. By relaxing the non-anticipativity constraints on the admissible control strategies, we expect to derive the value function-based penalty also in compact form (under natural filtration) as that in the setting of controlled Markov diffusions. The recent work [101] has exploited the martingale structure of this penalty in optimal stopping problems and found its application in pricing financial derivatives.

CHAPTER V

OPTIMAL STOPPING OF PARTIALLY OBSERVABLE MARKOV PROCESSES: A FILTERING-BASED DUALITY APPROACH

Optimal stopping of a partially observable Markov process (POMP) is a sequential decision making problem under partial observation of the underlying state. This type of problems arise in a number of applications, including change point detection in a production line, launching of a new technology under incomplete information of the market, and selling of an asset or a financial derivative. Optimal stopping of a POMP is more challenging than its counterpart of a fully observable process, since the inference of the hidden state and the choice of an optimal action should be accomplished at the same time. As a special class of the partially observable Markov decision processes (POMDPs), optimal stopping of a POMP can be transformed to a fully observable optimal stopping problem by introducing a new state variable, often referred to as the filtering distribution. However, this concise representation does not reduce the complexity of the problem, because the filtering distribution is usually infinite dimensional when the unobserved state takes values in a continuous space. Some recent work proposed to solve continuous-state POMDP include [15],[71],[78],[88] and [99], most of which can be viewed as a combination of dimension reduction on the filtering distribution and the approximate dynamic programming. These methods can also be adapted to solving OSPO with some modification.

To the best of our knowledge, only [36], [61], [74], [69] and [100] have studied numerical methods in the specific context of OSPO. In particular, they all interpreted the problem in the setting of American option pricing under partial observation of the stochastic volatility. [36] proposed a multinomial tree method that combines with particle filtering; [61] and [74] also utilize the particle filtering technique, and incorporate it into the regression-based approximate dynamic programming approach; whereas [69] used a grid-based method to approximate the filtering distribution. On the other hand, [100] proposed the method of approximate value iteration that avoids filtering step. It is worth noting that the first three methods all use particle filtering, as it is so far the most successful and versatile numerical method to solve nonlinear filtering problems. All of the above four approaches provide approximate solutions, and some are proven to converge asymptotically to the true option price. However, in practice with a finite computational power, the difference between their approximate solutions and the true option price is not known. Some other intrinsic problems of these methods also prevent their wide use in practice: for example, the computation of the multinomial tree method grows exponentially in the number of the exercise opportunities; the choice of basis functions is always problem-specific for the regression-based methods, which usually provides a lower bound on the option price.

In view of the lack of performance guarantee and computational complexity of the aforementioned methods, in this chapter we focus on developing a lower-and-upper-bound approach with moderate computational cost. We propose a filtering-based duality approach that complements a suboptimal stopping time (hence an asymptotic lower bound) with an asymptotic upper bound on the value function. Since our approach does not tie to a particular model and only involves Monte Carlo simulation, it can be generalized to any POMP as long as the particle filtering technique can be applied. Our method relies on the martingale duality formulation of the fully observable optimal stopping problem, which is proposed by [76] and [45] in the setting of pricing American options under constant volatility.

From the perspective of modeling fidelity versus computational complexity, it is not trivial to compare optimal stopping of POMPs with its counterpart in fully observable Markov processes. In particular, the difference of their value functions cannot be quantified in general and is problem dependent, so we are also interested in learning the features that influence this difference in the underlying probabilistic model. Indeed, as an example, our numerical experiments on pricing American options under partially observable stochastic volatility show that our asymptotic upper bound is strictly less than the option price of the model where the volatility is treated directly observable, and the difference is especially obvious when the effect of the volatility is dominant. This in turn shows that our method provides a better criterion to evaluate the performance of a suboptimal policy in the partially observable model.

The rest of this chapter is organized as follows. In Section 5.1, we describe the general problem formulation of optimal stopping of POMPs and the transformation to an equivalent fully observable optimal stopping problem. In Section 5.2, we develop the filtering-based duality approach, and its error analysis and convergence result are presented in Section 5.3. We present some numerical examples in Section 5.4, and finally conclude in Section 5.5.

5.1 Problem Formulation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider a hidden Markov model $\{(X_t, Y_t), t = 0, 1, \cdots, T\}$ satisfying the following equations

$$X_{t+1} = f(X_t, Z^1_{t+1}), \quad t = 0, 1, \cdots, T-1; \tag{72a}$$

$$Y_0 = h_0(X_0, Z^2_0); \tag{72b}$$

$$Y_{t+1} = h(X_{t+1}, Y_t, Z^2_{t+1}), \quad t = 0, 1, \cdots, T-1; \tag{72c}$$

where the unobserved state X_t is in a continuous state space $\mathcal{X} \subseteq \mathbb{R}^{n_x}$, the observation Y_t is in a continuous observation space $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$. The noises $\{(Z^1_t, Z^2_t), t = 1, \cdots, T\}$, which are independent of the initial state X_0 and the initial observation Y_0, are independent random vectors with known distributions, but the components of each vector can be correlated. Equations (72a) and (72b)-(72c) are often referred to as the state equation and the observation equation, respectively. Note that $\{(X_t, Y_t)\}$ is a bivariate Markov process adapted to the filtration $\Big\{\mathcal{F}_t \triangleq \sigma\{(X_i, Y_i); i = 0, \ldots, t\}\Big\}$.

Let $\mathcal{J} \triangleq \{1, \cdots, T\}$. Denote by $\Big\{\mathcal{F}_t^Y \triangleq \sigma\{Y_0, \ldots, Y_t\}\Big\}$ the filtration generated by the processes (72b)-(72c). A random variable $\tau : \Omega \to \mathcal{J}$ is an $\mathcal{F}_t^Y$-stopping time if $\{\tau \leq t\} \in \mathcal{F}_t^Y$ for every $t \in \mathcal{J}$. We define $\mathcal{T}^Y$ as the set of $\mathcal{F}_t^Y$-stopping times that take values in $\mathcal{J}$. Assume that the initial Y_0 is a known constant, and the initial X_0 follows a known distribution π_0, which is derived from the historical data (including Y_0). We consider the finite-horizon partially observable optimal stopping problem

$$V_0(\pi_0, y_0) = \sup_{\tau \in \mathcal{T}^Y} \mathbb{E}[g(\tau, X_\tau, Y_\tau) | X_0 \sim \pi_0, Y_0 = y_0], \tag{73}$$

where $g : \mathcal{J} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the reward function. In this setting the decision maker has access to only state Y_t so that her decision at time t is made purely depending on the observation history up to time t, i.e.,$\{Y_0, \cdots, Y_t\}$. For convenience, in the following we use $g(X_t, Y_t)$ and $g(X_\tau, Y_\tau)$ in short for $g(t, X_t, Y_t)$ and $g(\tau, X_\tau, Y_\tau)$ respectively.

The optimal stopping problem of a POMP can be transformed to an equivalent fully observable optimal stopping problem by introducing a new state variable Π_t, often referred to as the *filtering distribution*, which is the conditional distribution of X_t given the observations $Y_{0:t} \triangleq \{Y_0, \ldots, Y_t\}$. More specifically, given a set A in the Borel σ-algebra over $\mathcal{X}$, define

$$\Pi_t(A) \triangleq \text{Prob}(X_t \in A | Y_0, \ldots, Y_t), \quad t = 0, \ldots, T.$$

Given a realization of the observations $y_{0:t} \triangleq \{y_0, \ldots, y_t\}$, the probability density π_t of the filtering distribution Π_t evolves as follows:

$$\pi_t(x_t) = \frac{\int_{\mathcal{X}} p(x_t, y_t | x_{t-1}, y_{t-1}) \pi_{t-1}(x_{t-1})\, dx_{t-1}}{\int_{\mathcal{X}} p(y_t | x_{t-1}, y_{t-1}) \pi_{t-1}(x_{t-1})\, dx_{t-1}}, \ t = 1, \ldots, T, \tag{74}$$

where the conditional probability density functions $p(x_t, y_t | x_{t-1}, y_{t-1})$ and $p(y_t | x_{t-1}, y_{t-1})$ are induced by (72a), (72c), and the distributions of Z_t^1 and Z_t^2. Noticing that π_t

only depends on π_{t-1}, y_{t-1}, and y_t, and letting the realization $y_{0:t}$ be replaced by the random variables $Y_{0:t}$, we can abstractly rewrite the filtering recursion (74) as

$$\Pi_t = \Phi(\Pi_{t-1}, Y_{t-1}, Y_t), \quad t = 1, 2, \ldots, T.$$

Then problem (73) can be transformed to an equivalent optimal stopping problem (see, e.g., Chapter 5 in [9]) with fully observable state (Π_t, Y_t):

$$V_0(\pi_0, y_0) = \sup_{\tau \in \mathcal{T}^Y} \mathbb{E}[\tilde{g}(\Pi_\tau, Y_\tau) | X_0 \sim \pi_0, Y_0 = y_0],$$

where

$$\tilde{g}(\Pi_t, Y_t) \triangleq \mathbb{E}[g(X_t, Y_t) | \mathcal{F}_t^Y] = \int g(x_t, Y_t) \Pi_t(x_t)\, dx_t.$$

Theoretically, we can solve (73) following the dynamic programming recursion:

$$V_t(\Pi_t, Y_t) = \max\left(\tilde{g}(\Pi_t, Y_t), C_t(\Pi_t, Y_t)\right), \quad t = T, \ldots, 1, \tag{75}$$

where $C_t(\Pi_t, Y_t)$ is the *continuation value* at time t defined as

$$\begin{aligned} C_T(\Pi_T, Y_T) &\triangleq \tilde{g}(\Pi_T, Y_T); \\ C_t(\Pi_t, Y_t) &\triangleq \mathbb{E}[V_{t+1}(\Pi_{t+1}, Y_{t+1}) | \Pi_t, Y_t], \quad t = T-1, \ldots, 0. \end{aligned}$$

Here $\mathbb{E}[\cdot | \Pi_t, Y_t]$ is interpreted as $\mathbb{E}[\cdot | X_t \sim \Pi_t, Y_t]$. Then $V_0 = C_0$ and the optimal stopping time is

$$\tau^* = \min\left\{t \in \mathcal{J} \mid \tilde{g}(\Pi_t, Y_t) \geq C_t(\Pi_t, Y_t)\right\}.$$

We also define its associated t-indexed stopping time τ_t^* for each $t \in \mathcal{J}$:

$$\tau_t^* \triangleq \min\left\{i \in \mathcal{J}_t \mid \tilde{g}(\Pi_i, Y_i) \geq C_i(\Pi_i, Y_i)\right\} \tag{76}$$

with $\mathcal{J}_t \triangleq \{t, t+1, \ldots, T\}$. The above recursion also shows that (Π_t, Y_t) are the sufficient statistics that determine the optimal stopping time. The process $\{V_t \triangleq V_t(\Pi_t, Y_t)\}$ defined in (75) is called the Snell envelope process (see, e.g., Chapter 2 in [58]) of the process $\{\tilde{g}(\Pi_t, Y_t)\}$, which is the smallest $\mathcal{F}_t^Y$-supermartingale that dominates $\tilde{g}$ in the sense that $V_t(\Pi_t, Y_t) \geq \tilde{g}(\Pi_t, Y_t)$. In particular, by shifting the time index in (73) we can interpret V_t as

$$\begin{aligned} V_t(\pi_t, y_t) &= \sup_{\tau \in \mathcal{T}^Y,\ t \leq \tau \leq T} \mathbb{E}[g(X_\tau, Y_\tau) | X_t \sim \pi_t, Y_t = y_t] \\ &= \mathbb{E}[g(X_{\tau_t^*}, Y_{\tau_t^*}) | X_t \sim \pi_t, Y_t = y_t], \quad t = 1, \ldots, T. \end{aligned} \tag{77}$$

However, it is often impossible to solve the problem exactly following (75) due to two main difficulties. One is that in general the filtering distribution Π_t is infinite dimensional and the filtering recursion (74) cannot be computed exactly. The other difficulty lies in the accurate estimation of the continuation value $C_t(\Pi_t, Y_t)$ that leads to the optimal stopping time τ^*. So we develop an approximation method in the next section.

5.2 Filtering-Based Martingale Duality Approach

In this section, we construct a dual problem to the original optimal stopping of POMPs, and develop a numerical method that yields an asymptotic upper bound on the value function. Our dual formulation is a straightforward extension of the dual formulation for the optimal stopping problem proposed in [76], [45], and [3], by replacing the filtration with $\mathcal{F}_t^Y$.

Theorem 15 (c.f. (5) in [3])**.** *Let $\mathcal{M}$ represent the space of $\mathcal{F}_t^Y$-adapted martingales $\{M_t\}$ with $M_0 = 0$ and $\sup_{t \in \mathcal{J}} \mathbb{E}|M_t| < \infty$. Then*

$$V_0(\pi_0, y_0) = \min_{M \in \mathcal{M}} \left\{ \mathbb{E}[\max_{t \in \mathcal{J}} \{\tilde{g}(\Pi_t, Y_t) - M_t\} | X_0 \sim \pi_0, Y_0 = y_0] \right\}. \tag{78}$$

The optimal martingale $\{M_t^*\}$ that achieves the minimum on the right hand side of (78) is of the form

$$M_t^* = \sum_{i=1}^{t} \Delta_i^*, \tag{79}$$

where$\{\Delta_t^*\}$ is the martingale difference sequence defined as

$$\Delta_t^* \triangleq \mathbb{E}[V_t | \mathcal{F}_t^Y] - \mathbb{E}[V_t | \mathcal{F}_{t-1}^Y],\ t \in \mathcal{J}. \tag{80}$$

In addition, the following equality holds pathwisely in the almost sure sense, i.e.,

$$V_0(\pi_0, y_0) = \max_{t\in\mathcal{J}}(\tilde{g}(\Pi_t, Y_t) - M_t^*) \quad a.s..$$

The proof of Theorem 15 follows the same line in [3] and hence is omitted here. Theorem 15 characterizes a strong duality relation between the primal problem (73) and its dual problem on the right side of (78); the equality (78) suggests that any $\mathcal{F}_t^Y$-adapted martingale $\{M_t\}$ can lead to an upper bound on $V_0(\pi_0, y_0)$ and that the optimal martingale (79) is derived from the Doob-Meyer decomposition of the supermartingale $\{V_t\}$. In particular, we can rewrite (80) as

$$\Delta_t^* = \mathbb{E}[V_t|\Pi_t, Y_t] - \mathbb{E}[V_t|\Pi_{t-1}, Y_{t-1}] \tag{81a}$$

$$= \mathbb{E}[g(X_{\tau_t^*}, Y_{\tau_t^*})|\Pi_t, Y_t] - \mathbb{E}[g(X_{\tau_t^*}, Y_{\tau_t^*})|\Pi_{t-1}, Y_{t-1}]. \tag{81b}$$

Note that it is impossible to compute the optimal martingale $\{M_t^*\}$, since the martingale difference term (81a) (or (81b)) involves the intractable filtering distribution Π_t and the Snell envelop process $\{V_t\}$ (or the optimal stopping time τ_t^*). Therefore, we need to introduce approximation schemes to address both aspects. On the one hand, the intractable filtering distribution Π_t can be approximated by a discrete distribution using particle filtering, which will be stated in Section 5.2.1. On the other hand, (81a) and (81b) suggest that we approximate Δ_t^* using either approximate value functions of V_t or suboptimal $\mathcal{F}_t^Y$-stopping times that approximate τ_t^*. In addition, some other heuristic constructions can be considered. For example, we can take $\Delta_t = \mathbb{E}[U_t(X_t, Y_t)|\mathcal{F}_t^Y] - \mathbb{E}[U_t(X_t, Y_t)|\mathcal{F}_{t-1}^Y]$, where $U_t(X_t, Y_t)$ is the value function to the corresponding optimal stopping problem with fully observable state (X_t, Y_t):

$$U_t(x_t, y_t) = \sup_{\kappa\in\mathcal{T}_t} \mathbb{E}[g(X_\kappa, Y_\kappa)|X_t = x_t, Y_t = y_t], \tag{82}$$

where $\mathcal{T}_t$ is the set of $\mathcal{F}_t$-stopping times κ that take values in $\mathcal{J}_t$; or equivalently we can take $\Delta_t = \mathbb{E}[g(X_{\kappa_t^*}, Y_{\kappa_t^*})|\Pi_t, Y_t] - \mathbb{E}[g(X_{\kappa_t^*}, Y_{\kappa_t^*})|\Pi_{t-1}, Y_{t-1}]$, where κ_t^* is the optimal $\mathcal{F}_t$-stopping time to problem (82). Even if the explicit forms of U_t and κ_t^*

are not known, their approximations can be used in Δ_t and its martingale difference property can still be preserved. The advantage of approximating U_t or κ_t^* is their simple structure as functions of only (X_t, Y_t), whereas either V_t or τ_t^* is a function of $(Y_0, \cdots, Y_t)$. Thus, it may be easier to generate martingale difference terms based on approximate U_t or κ_t^*, even though they may yield less optimal values.

In the rest of this section we focus on approximating Δ_t^* in (81b) by the following Δ_t^m based on a fixed stopping time τ (see, e.g., (87) in Section 5.2.2), which is either $\mathcal{F}_t^Y$ or $\mathcal{F}_t$-adapted:

$$\Delta_t^m \triangleq \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\Pi_t^m, Y_t] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\Pi_{t-1}^m, Y_{t-1}], \tag{83}$$

where τ_t is the t-indexed stopping time associated with τ, and Π_t^m (see details in Section 5.2.1) is the approximate filtering distribution at time t obtained by particle filtering (the superscript m in Π_t^m denotes the number of particles), which will be elaborated in the next section. A lower-case notation π_t^m denotes the corresponding approximate filtering distribution based on a realization of the observations $y_{0:t}$. Then we define $\{M_t^m\}$ as

$$M_0^m = 0; \quad M_t^m = \Delta_1^m + \ldots + \Delta_t^m, \quad t \in \mathcal{J}. \tag{84}$$

Incorporating the above ideas, we propose the following algorithm that yields an asymptotic upper bound on V_0.

Algorithm 1. *Filtering-Based Martingale Duality Approach*

Step 1. For $k = 1, 2, \ldots, N$, **do**

- Generate a path of observations $y_{1:T}^{(k)}$ according to the processes (72a)-(72c) with initial condition $Y_0 = y_0$ and $X_0 \sim \pi_0$, and then follow Algorithm 2 (particle filtering) to generate the approximate filtering distribution $\{\pi_1^{m(k)}, \ldots, \pi_T^{m(k)}\}$.
- For $t = 1, \ldots, T$, use Algorithm 3 to compute $\tilde{\Delta}_t^{m(k)}$, which is an approximation for

$$\Delta_t^{m(k)} = \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t^{m(k)}, y_t^{(k)}] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}^{m(k)}, y_{t-1}^{(k)}]. \tag{85}$$

- Sum the approximate martingale differences to obtain

$$\tilde{M}_t^{m(k)} = \tilde{\Delta}_1^{m(k)} + \ldots + \tilde{\Delta}_t^{m(k)},\ t = 1, \ldots, T.$$

- Evaluate $V^{(k)} = \max_{t \in \mathcal{J}} \left(\tilde{g}(\pi_t^{m(k)}, y_t^{(k)}) - \tilde{M}_t^{m(k)} \right)$. **end**

Step 2. Set $V_N^\tau = \frac{1}{N} \sum_{k=1}^N V^{(k)}$. V_N^τ is an asymptotic upper bound on the value function $V_0(\pi_0, y_0)$.

In the next two subsections, we will discuss how to generate approximate filtering distribution using particle filtering via Algorithm 2 and how to compute the approximate martingale difference via Algorithm 3.

5.2.1 Particle Filtering

We approximate π_t using particle filtering, which is a successful and versatile numerical method for solving nonlinear filtering problems. A good introduction on particle filtering can be found in the book [4]. The particle filtering method approximates π_t by a finite number (say m) of particles $\{x_t^{(1)}, \ldots, x_t^{(m)}\}$, i.e., a discrete distribution π_t^m written as follows

$$\pi_t^m = \frac{1}{m} \sum_{i=1}^m \delta_{x_t^{(i)}}, \tag{86}$$

where δ is the Dirac measure. As the number of particles m goes to infinity, it can be ensured that π_t^m converges to π_t in certain sense.

Algorithm 2. *Particle Filtering*

Input: $X_0 \sim \pi_0$ and a sequence of observations $y_{0:T}$.

Output: The approximate filtering distribution $\pi_0^m, \ldots, \pi_T^m$.

Step 1. Initialization: Set $t = 0$. Draw m i.i.d. samples $\{x_0^{(1)}, \ldots, x_0^{(m)}\}$ from the distribution π_0. Set $\pi_0^m = \frac{1}{m} \sum_{i=1}^m \delta_{x_0^{(i)}}$.

Step 2. For $t = 1, \ldots, T$, **do**

$-$ Prediction: For each $i = 1, \ldots, m$, draw one sample $\bar{x}_t^{(i)}$ from $P(X_t | X_{t-1} = x_{t-1}^{(i)})$.

$-$ Bayes' Updating: Compute $w_t^{(i)} = \frac{p(y_t | \bar{x}_t^{(i)}, y_{t-1})}{\sum_{i=1}^m p(y_t | \bar{x}_t^{(i)}, y_{t-1})}, \quad i = 1, \ldots, m.$

$-$ Resampling: Draw i.i.d. samples $\{x_t^{(1)}, \ldots, x_t^{(m)}\}$ from the discrete distribution $\{\text{Prob}(\bar{x}_t^{(i)}) = w_t^{(i)},\ i = 1, \ldots, m\}$. Set $\pi_t^m = \frac{1}{m} \sum_{i=1}^m \delta_{x_t^{(i)}}$. **end**

5.2.2 Approximate Martingale Difference

The remaining issue is how to compute the martingale difference (85). Throughout this subsection we assume a suboptimal stopping time τ of the form,

$$\tau = \min\{t \in \mathcal{J} | g(X_t, Y_t) \geq \tilde{C}_t(X_t, Y_t)\}, \tag{87}$$

where $\{\tilde{C}_t, t \in \mathcal{J}\}$ is a sequence of approximate continuation functions of U_t. The approximate continuation functions $\tilde{C}_t$ can be derived, for example, by regression on some basis functions as suggested by [60] and [90]. We choose an $\mathcal{F}_t$-stopping time τ of the form (87) only for ease of exposition, though Algorithm 3 can be adjusted using any other $\mathcal{F}_t$(or $\mathcal{F}_t^Y$)-stopping time with the same principle.

Given a realization of observations $y_{0:T}$, we employ nested simulation to estimate Δ_t^m in (85). Note that π_t^m in Algorithm 1 is of the form (86). Therefore,

$$\begin{aligned} \Delta_t^m =& \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t}) | X_t = x_t^{(i)}, Y_t = y_t] \\ &- \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t}) | X_{t-1} = x_{t-1}^{(i)}, Y_{t-1} = y_{t-1}], \end{aligned}$$

where τ_t is the t-indexed stopping time associated with τ defined as

$$\tau_t = \min\{i \in \mathcal{J}_t | g(X_i, Y_i) \geq \tilde{C}_i(X_i, Y_i)\}.$$

To estimate $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t}) | x_t^{(i)}, y_t]$ (resp., $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t}) | x_{t-1}^{(i)}, y_{t-1}]$), we generate l subpaths that are stopped according to τ_t with the initial condition $X_t = x_t^{(i)}, Y_t = y_t$ (resp., $X_{t-1} = x_{t-1}^{(i)}, Y_{t-1} = y_{t-1}$) for each i and t, and we average $g(X_{\tau_t}, Y_{\tau_t})$ over these subpaths. So there are a total number of $m \cdot l$ subpaths generated to estimate each expectation term in (85). The details of the nested simulation are presented below.

Algorithm 3. *Estimation of Δ_t^m Using Nested Simulation*

Input: y_{t-1}, y_t, $\pi_{t-1}^m = \frac{1}{m}\sum_{i=1}^m \delta_{x_{t-1}^{(i)}}$ and $\pi_t^m = \frac{1}{m}\sum_{i=1}^m \delta_{x_t^{(i)}}$ from Algorithm 1 and Algorithm 2.

(Step 1 - Step 2 are used to estimate $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}^m, y_{t-1}]$.)

Step 1. For $i = 1, \ldots, m$, **do**

- Simulate $\{(x_t^{(ij)}, y_t^{(ij)}), \ldots, (x_T^{(ij)}, y_T^{(ij)})\}_{j=1}^l$ from the processes (72a)-(72c) with the initial condition $X_{t-1} = x_{t-1}^{(i)}$ and $Y_{t-1} = y_{t-1}$.

- To apply τ_t on these sample paths, find

$$t_{ij} = \min\left\{k \in \mathcal{J}_t : g(x_k^{(ij)}, y_k^{(ij)}) \geq \tilde{C}_k(x_k^{(ij)}, y_k^{(ij)})\right\}.$$

- Set $b_i = \frac{1}{l}\sum_{j=1}^l g(x_{t_{ij}}^{(ij)}, y_{t_{ij}}^{(ij)})$. **end**

Step 2. Set $G_{t-1,t}^{m,l} \triangleq \frac{1}{m}\sum_{i=1}^m b_i$, which is an unbiased estimator of $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}^m, y_{t-1}]$.

(Step 3 - Step 4 are used to estimate $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t^m, y_t]$.)

Step 3. For $i = 1, \ldots, m$, **do**

If $g(x_t^{(i)}, y_t) \geq \tilde{C}_t(x_t^{(i)}, y_t)$, i.e., $(x_t^{(i)}, y_t)$ is in the stopping region, set $\tilde{b}_i = g(x_t^{(i)}, y_t)$. Otherwise, repeat Step 1 with the initial condition $X_t = x_t^{(i)}$ and $Y_t = y_t$ to obtain $\tilde{b}_i$. **end**

Step 4. Set $G_{t,t}^{m,l} \triangleq \frac{1}{m}\sum_{i=1}^m \tilde{b}_i$, which is an unbiased estimator of $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t^m, y_t]$.

Step 5. Set $\tilde{\Delta}_t^m = G_{t,t}^{m,l} - G_{t-1,t}^{m,l}$.

5.3 Error Analysis

In this section, we analyze the error bound and asymptotic convergence of our algorithm. To lighten the notations, we use $\mathbb{E}_0[\cdot]$ to denote $\mathbb{E}[\cdot|X_0 \sim \pi_0, Y_0 = y_0]$ in the rest of note. The following assumption is used throughout our analysis.

Assumption 6.

i. $\| g \|_\infty \triangleq \max_{t\in\mathcal{J}} \| g(t, \cdot, \cdot) \|_\infty < \infty$.

ii. For any observation sequence $y_{0:T}$,

$$\sup_{x_t\in\mathcal{X}} p(y_t|x_t, y_{t-1}) < \infty, \quad \forall t \in \mathcal{J}.$$

We first introduce an $\mathcal{F}_t^Y$-adapted martingale difference sequence $\{\Delta_t^\tau\}$ and martingale $\{M_t^\tau\}$ induced by an $\mathcal{F}_t$(or $\mathcal{F}_t^Y$)-stopping time τ:

$$\Delta_t^\tau = \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\Pi_t, Y_t] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\Pi_{t-1}, Y_{t-1}],$$
$$M_0^\tau \triangleq 0; \quad M_t^\tau \triangleq \Delta_1^\tau + \ldots + \Delta_t^\tau, \quad t \in \mathcal{J}.$$

Since M_t^τ is an $\mathcal{F}_t^Y$-adapted martingale, then $\mathbb{E}_0[\max_{t\in\mathcal{J}}(\tilde{g}(\Pi_t, Y_t) - M_t^\tau)]$ is an upper bound on $V_0(\pi_0, y_0)$ by Theorem 15.

Recall that the approximate martingale difference Δ_t^m based on a realization of observations $y_{0:t}$ is

$$\Delta_t^m = \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t^m, y_t] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}^m, y_{t-1}].$$

In Algorithm 3 the empirical estimates of $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t^m, y_t]$ and $\mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}^m, y_{t-1}]$ are denoted by $G_{t,t}^{m,l}$ and $G_{t-1,t}^{m,l}$, respectively. Therefore, we use

$$\tilde{\Delta}_t^m = G_{t,t}^{m,l} - G_{t-1,t}^{m,l} \text{ and } \tilde{M}_t^m = \sum_{i=1}^{t} \tilde{\Delta}_i^m$$

to approximate Δ_t^m and M_t^m. Instead of obtaining $\max_{t\in\mathcal{J}}\{\tilde{g}(\pi_t, y_t) - M_t^\tau\}$ exactly along each path of the observations $y_{0:T}$, we compute $\max_{t\in\mathcal{J}}\{\tilde{g}(\pi_t^m, y_t) - \tilde{M}_t^m\}$. Note that conditional on a fixed observation sequence, the former term is a constant, while the latter one is a random term due to sampling. The difference between these two terms is caused by two sources of errors: one from the difference between the deterministic density π_t and the random measure π_t^m, and this gap will go to zero (in expectation) by increasing the number of particles m under Assumption 6; another difference is from the variability of the nested (Monte Carlo) simulation, which can be eliminated by increasing the number of sample paths $m \cdot l$.

We will show in the next theorem that $\mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t^m, Y_t) - \tilde{M}_t^m\}]$ converges to $\mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\}]$ when the particle number m increases to infinity. Hence, $\mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t^m, Y_t) - \tilde{M}_t^m\}]$ is an asymptotic (as $m \to \infty$) upper bound on $V_0(\pi_0, y_0)$. Moreover, the gap between $\mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\}]$ and $V_0(\pi_0, y_0)$ is purely due to the suboptimal stopping time τ.

Theorem 16. *Suppose τ is an $\mathcal{F}_t$(or $\mathcal{F}_t^Y$)-stopping time. Then*

$$\lim_{m\to\infty} \mathbb{E}_0\left[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t^m, Y_t) - \tilde{M}_t^m\}\right] = \mathbb{E}_0\left[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\}\right]. \tag{88}$$

Moreover, we have the following inequalities:

$$\begin{aligned}&\mathbb{E}_0\left[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\}\right] - V_0(\pi_0, y_0)\\ \leq&2\sqrt{\sum_{t=1}^{T}\mathbb{E}_0\left[(\Delta_t^* - \Delta_t^\tau)^2\right]}\\ \leq&2\sqrt{\sum_{t=1}^{T}\mathbb{E}_0\left[\left(\mathbb{E}[g(X_{\tau_t^*}, Y_{\tau_t^*})|\Pi_t, Y_t,] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\Pi_t, Y_t]\right)^2\right]}.\end{aligned} \tag{89}$$

From (88), the output V_N^τ in Algorithm 1 is an asymptotic (as the sample path number $N \to \infty$ and the particle number $m \to \infty$) upper bound on the true value function V_0. According to (89), a large m will lead to a tight upper bound provided that the martingale $\{M_t^\tau\}$ induced by the stopping time τ does not differ too much from the optimal $\{M_t^*\}$, or more intuitively, the suboptimal stopping time τ_t does not differ too much from the optimal τ_t^*.

Proof. We need the following proposition for the proof of the theorem.

Proposition 3 (Corollary 10.28, [4])**.** *Let $\{\pi_0^m, \ldots, \pi_T^m\}$ be the random measure generated by Algorithm 2 for the observation sequence $y_{0:T}$. Suppose that the following assumption holds:*

$$\parallel f \parallel_\infty < \infty \quad \textit{and} \quad \sup_{x_t} p(y_t|x_t, y_{t-1}) < \infty, \quad t = 1, \ldots, T.$$

Then

$$\mathbb{E}\left[\left(\int_{\mathcal{X}} f(x_t)\pi_t(x_t)dx_t - \int_{\mathcal{X}} f(x_t)\pi_t^m(x_t)dx_t\right)^2\right] \leq \frac{k_t^2 \parallel f \parallel_\infty^2}{m}, \quad t = 0, \ldots, T,$$

where the constant k_t does not depend on m (but it does depend on t and $y_{0:t}$). In particular, $k_0 = 1$.

We first prove (88). Given a sample path of the observations $\{y_0, \ldots, y_T\}$, the difference between $\tilde{g}(\pi_t, y_t)$ and $\tilde{g}(\pi_t^m, y_t)$ is

$$\vartheta_t^m \triangleq \int_{\mathcal{X}} g(x_t, y_t)\pi_t(x_t)dx_t - \int_{\mathcal{X}} g(x_t, y_t)\pi_t^m(x_t)dx_t.$$

Guaranteed by Proposition 3, $\mathbb{E}[|\vartheta_t^m|] \leq \sqrt{\mathbb{E}[(\vartheta_t^m)^2]} \leq \frac{k_t \|g\|_\infty}{\sqrt{m}}$ for some constant k_t. The difference between M_t^τ and $\tilde{M}_t^m$ is the sum of the differences between Δ_t^τ and $\tilde{\Delta}_t^m$:

$$\Delta_t^\tau - \tilde{\Delta}_t^m = \chi_{t,t}^m - \chi_{t-1,t}^m + \epsilon_{t,t}^{m,l} - \epsilon_{t-1,t}^{m,l},$$

where

$$\begin{aligned}
\chi_{t,t}^m &\triangleq \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t, y_t] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t^m, y_t], \\
\chi_{t-1,t}^m &\triangleq \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}, y_{t-1}] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}^m, y_{t-1}], \\
\epsilon_{t,t}^{m,l} &\triangleq \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_t^m, y_t] - G_{t,t}^{m,l}, \\
\epsilon_{t-1,t}^{m,l} &\triangleq \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\pi_{t-1}^m, y_{t-1}] - G_{t-1,t}^{m,l}.
\end{aligned}$$

The first two errors are filtering errors, since we can rewrite $\chi_{t,t}^m$ as

$$\begin{aligned}
\chi_{t,t}^m &= \mathbb{E}\left[\sum_{j=t}^T g(X_j, Y_j)1_{\{\tau_t = j\}}\big|\pi_t, y_t\right] - \mathbb{E}\left[\sum_{j=t}^T g(X_j, Y_j)1_{\{\tau_t = j\}}\big|\pi_t^m, y_t\right] \\
&= \int_{\mathcal{X}} I_t(x_t, y_t)\pi_t(x_t)dx_t - \int_{\mathcal{X}} I_t(x_t, y_t)\pi_t^m(x_t)dx_t. \qquad (90)
\end{aligned}$$

$I_t(x_t, y_t)$ is defined as the integrand of $\mathbb{E}[\sum_{j=t}^T g(X_j, Y_j)1_{\{\tau_t=j\}}|\pi_t, y_t]$, i.e.,

$$I_t(x_t, y_t) \triangleq g(x_t, y_t)1_{\{\tau_t = t\}} + \sum_{j=t+1}^T \int g(x_j, y_j)1_{\{\tau_t = j\}}p(dx_{t+1}dy_{t+1}\ldots dx_j dy_j|x_t, y_t),$$

where $p(dx_{t+1}dy_{t+1}\ldots dx_j dy_j|x_t, y_t)$ denotes the joint probability distribution of $(x_{t+1}, y_{t+1}, \ldots, x_j, y_j)$ conditional on (x_t, y_t). As $\{\tau_t = j\}$ are disjoint sets for each $t \leq j \leq T$, it implies $\| I_t \|_\infty \leq \| g \|_\infty$. Based on (90) and using Proposition 3 with $f = I_t$, it is ensured that $\mathbb{E}[|\chi^m_{t,t}|] \leq \frac{k^{'}_t \|g\|_\infty}{\sqrt{m}}$ for some constant $k^{'}_t$. Similarly, $\mathbb{E}[|\chi^m_{t-1,t}|] \leq \frac{b^{'}_{t-1}\|g\|_\infty}{\sqrt{m}}$ for some constant $b^{'}_{t-1}$. The latter two errors are from the sampling variability of Monte Carlo simulation; the error bounds are guaranteed by Proposition 3 with $t = 0$ (here π^m_t or π^m_{t-1} plays the role of π_0), i.e., $\mathbb{E}[|\epsilon^{m,l}_{t,t}|] \leq \frac{\|g\|_\infty}{\sqrt{ml}}$ and $\mathbb{E}[|\epsilon^{m,l}_{t-1,t}|] \leq \frac{\|g\|_\infty}{\sqrt{ml}}$.

So given a sample path of the observations $y_{0:t}$ we have for each $t \in \mathcal{J}$,

$$\begin{aligned}&\lim_{m\to\infty} \mathbb{E}[|(\tilde{g}(\pi_t, y_t) - M^\tau_t) - (\tilde{g}(\pi^m_t, y_t) - \tilde{M}^m_t)|]\\ =&\lim_{m\to\infty} \mathbb{E}[|\vartheta^m_t + (\sum_{i=1}^{t}(\tilde{\Delta}^m_i - \Delta^\tau_i))|] = 0. \end{aligned} \tag{91}$$

Since

$$\begin{aligned}&|\max_{t\in\mathcal{J}}\{\tilde{g}(\pi_t, y_t) - M^\tau_t\} - \max_{t\in\mathcal{J}}\{\tilde{g}(\pi^m_t, y_t) - \tilde{M}^m_t\}|\\ \leq&\max_{t\in\mathcal{J}}\{|(\tilde{g}(\pi_t, y_t) - M^\tau_t) - (\tilde{g}(\pi^m_t, y_t) - \tilde{M}^m_t)|\}\\ \leq&\sum_{t=1}^{T}|(\tilde{g}(\pi_t, y_t) - M^\tau_t) - (\tilde{g}(\pi^m_t, y_t) - \tilde{M}^m_t)|,\end{aligned}$$

by taking expectation and letting m go to infinity we have

$$\lim_{m\to\infty} \mathbb{E}[|\max_{t\in\mathcal{J}}\{\tilde{g}(\pi^m_t, y_t) - \tilde{M}^m_t\} - \max_{t\in\mathcal{J}}\{\tilde{g}(\pi_t, y_t) - M^\tau_t\}|] = 0.$$

Note that $\tilde{\Delta}^m_t$ is bounded by $2 \parallel g \parallel_\infty$ for each $t \in \mathcal{J}$, and therefore, $\tilde{g}(\Pi^m_t, Y_t) - \tilde{M}^m_t$ is bounded by $(2t+1)\cdot \parallel g \parallel_\infty$ and $\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi^m_t, Y_t) - \tilde{M}^m_t\}$ is bounded by $(2T+1)\cdot \parallel g \parallel_\infty$. The same conclusions are also valid for Δ^τ_t, $\tilde{g}(\Pi_t, Y_t) - M^\tau_t$ and $\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M^\tau_t\}$. Then

$$\begin{aligned}&\lim_{m\to\infty} \mathbb{E}_0\big[|\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi^m_t, Y_t) - \tilde{M}^m_t\} - \max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M^\tau_t\}|\big]\\ =&\lim_{m\to\infty} \mathbb{E}_0\big[\mathbb{E}\big[|\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi^m_t, Y_t) - \tilde{M}^m_t\} - \max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M^\tau_t\}|\big|\mathcal{F}^Y_T\big]\big]\\ =&\mathbb{E}_0\big[\lim_{m\to\infty} \mathbb{E}\big[|\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi^m_t, Y_t) - \tilde{M}^m_t\} - \max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M^\tau_t\}|\big|\mathcal{F}^Y_T\big]\big]\\ =&0,\end{aligned}$$

where the second equality follows from the boundedness of the integrand and the dominated convergence theorem. Hence,

$$\lim_{m\to\infty} \mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t^m, Y_t) - \tilde{M}_t^m\}] = \mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\}].$$

Now we prove (89). First we have

$$\begin{aligned} &\mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\}] - V_0 \\ =&\mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\}] - \mathbb{E}_0[\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^*\}] \\ \leq&\mathbb{E}_0[\max_{t\in\mathcal{J}}\{M_t^* - M_t^\tau\}], \end{aligned}$$

following the fact that

$$\max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^\tau\} - \max_{t\in\mathcal{J}}\{\tilde{g}(\Pi_t, Y_t) - M_t^*\} \leq \max_{t\in\mathcal{J}}\{M_t^* - M_t^\tau\}.$$

Then (89) follows from

$$\begin{aligned} &\mathbb{E}_0[\max_{t\in\mathcal{J}}\{M_t^* - M_t^\tau\}] \\ \leq&2\sqrt{\mathbb{E}_0[(M_T^* - M_T^\tau)^2]} \\ =&2\sqrt{\sum_{t=1}^T \mathbb{E}_0\left[\left((M_t^* - M_t^\tau) - (M_{t-1}^* - M_{t-1}^\tau)\right)^2\right]} \\ =&2\sqrt{\sum_{t=1}^T \mathbb{E}_0[(\Delta_t^* - \Delta_t^\tau)^2]} \\ \leq&2\sqrt{\sum_{t=1}^T \mathbb{E}_0\left[\left(\mathbb{E}[g(X_{\tau_t^*}, Y_{\tau_t^*})|\Pi_t, Y_t] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\Pi_t, Y_t]\right)^2\right]}, \end{aligned}$$

where the first inequality follows from the fact that $M_t^* - M_t^\tau$ is a martingale and applying Doob's martingale inequality, and the first equality uses the orthogonality property of martingale difference (see p.331 in [84]). To show the last inequality, recall that

$$\begin{aligned} \Delta_t^* - \Delta_t^\tau =&(\mathbb{E}[g(X_{\tau_t^*}, Y_{\tau_t^*})|\mathcal{F}_t^Y] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\mathcal{F}_t^Y]) \\ &- (\mathbb{E}[g(X_{\tau_t^*}, Y_{\tau_t^*})|\mathcal{F}_{t-1}^Y] - \mathbb{E}[g(X_{\tau_t}, Y_{\tau_t})|\mathcal{F}_{t-1}^Y]); \end{aligned}$$

then the last inequality can be shown by simple algebra and iterated expectation on $\mathcal{F}^Y_{t-1}$. □

5.4 Numerical Examples

We apply our method to price American put options under stochastic volatility. Following the model in [74] we considered a d_S-dimensional process of asset price $\{S_t, t = 0 : T\}$:

$$S^i_{t+1} = S^i_t \exp\left\{ \left(r - \frac{(\sigma^i_{t+1})^2}{2} \right) \delta + \sigma^i_{t+1} \sqrt{\delta} Z^{i,1}_{t+1} \right\}, \ i = 1, \ldots, d_S, \tag{92}$$

where r is the constant interest rate, δ is the time period between the equally-spaced time points, $\{Z^{i,1}_t, t = 1 : T\}, i = 1, \ldots, d_S$ are independent sequences of Gaussian random variables with $Z^{i,1}_t \sim \mathcal{N}(0,1)$, and the volatility $\sigma^i_t \triangleq \exp(X^i_t)$ is a deterministic function of a $d_X (= d_S)$-dimensional process $\{X_t, t = 0 : T\}$ that evolves as a discretized Ornstein-Uhlenbeck process:

$$X^i_{t+1} = X^i_t e^{-\lambda_i \delta} + \theta_i (1 - e^{-\lambda_i \delta}) + \gamma_i \sqrt{\frac{1 - e^{-2\lambda_i \delta}}{2\lambda_i}} Z^{i,2}_{t+1}, \ i = 1, \ldots, d_X, \tag{93}$$

where the positive constant θ_i is the mean reversion value, the constant λ_i is the mean reversion rate, the constant γ_i is a measure of the process volatility, and $\{Z^{i,2}_t, t = 1 : T\}, i = 1, \ldots, d_X$ are independent sequences of Gaussian random variables with $Z^{i,2}_t \sim \mathcal{N}(0, \mu_i^2)$, which are also independent of $\{Z^{i,1}_t\}$. Here μ_i is used to control the observation noise. For simplicity, in our numerical experiments we use $\lambda_i = \lambda$, $\theta_i = \theta$, $\gamma_i = \gamma$, $\mu_i = \mu$ for all $i = 1, \ldots, d_X$. Assume that only the asset price is observed, and exercise opportunities take place at $t = 1, \ldots, T$. We consider the put option on the minimum of d_S assets, i.e., the payoff function is of the form

$$g(t, S_t) = \max\left\{ e^{-r\delta t} \left(K - \min\{S^1_t, \ldots, S^{d_S}_t\} \right), 0 \right\}.$$

In the rest of this section, “exercise policy” simply means “stopping time” in the general optimal stopping problem.

Remark 4. *In this example, the conditional probability density function is*

$$p(S_t|X_t, S_{t-1}) = \prod_{i=1}^{d_X} p(S_t^i|X_t^i, S_{t-1}^i),$$

where

$$p(S_t^i|X_t^i, S_{t-1}^i) = \frac{\exp\left\{-\frac{\left(\ln(S_t^i/S_{t-1}^i)-(r-\exp^2(X_t^i)/2)\delta\right)^2}{2\exp^2(X_t^i)\delta\mu^2}\right\}}{S_t^i\sqrt{2\pi\exp^2(X_t^i)\delta\mu^2}}.$$

It can be shown that $p(S_t|X_t, S_{t-1})$ satisfies Assumption 6(ii) and that Assumption 6(i) is also trivially satisfied.

Since the stochastic volatility cannot be directly observed in reality but can be "partially observable" through the inference from the observed asset price, pricing American option under the above model (92)-(93) falls into the framework of optimal stopping of POMPs. We illustrate our algorithm through a series of numerical experiments with $d_S = 1$ (one asset) and $d_S = 2$ (two assets). In particular, we are interested in how the variance of the volatility (corresponding to the parameters $(\theta, \lambda, \gamma)$) and observation noise (corresponding to the parameter μ) influence the price difference due to the difference between the fully observable and partially observable volatilities. We list the parameter sets in Table 9. To compute option prices under both full and partial observations, we implement our algorithm as well as the Least-Squares Monte Carlo (LSMC) method of [60], which provides suboptimal exercise policies, and the primal-dual (PD) method of [3], which parallels our method in the fully observable models. The numerical results of the option prices under different parameter sets are listed in Table 10 (for one asset) and Table 11 (for two assets), where "*LB*" represents the lower bound obtained by the LSMC method for the fully/partially observable model with the following two sets of basis functions for the one-asset and two-asset problems respectively:

$$H_1 =\{L_0(S_t^1), L_0^2(S_t^1), L_1(S_t^1), L_1^2(S_t^1), L_0(S_t^1)L_1(S_t^1), 1\},$$

$$H_2 =\{L_0(S_t^1), L_0^2(S_t^1), L_0(S_t^2), L_0^2(S_t^2), L_0(S_t^1)L_0(S_t^2), L_2(S_t^1, S_t^2), L_2^2(S_t^1, S_t^2), 1\},$$

where $L_0(x) = x$, $L_1(x) = \max\{K - x, 0\}$ and $L_2(x, y) = \max\{K - \min\{x, y\}, 0\}$. Please note that the basis functions only depend on the asset price S_t not the volatility $\exp(X_t)$, so the suboptimal policy is $\mathcal{F}_t^Y$-adapted and the results are guaranteed to be lower bounds for the partially observable model. In the tables, "UB" represents the corresponding upper bound yielded by our filtering-based duality method for the partially observable model, and "$Full.\widetilde{UB}$" represents the corresponding upper bound yielded by the PD method for the fully observable model. It is clear that we can improve the exercise policy for the fully observable model by employing more basis functions that use the information of the volatility $\exp(X_t)$: "$Full.LB$" and "$Full.UB$" are the lower bound and upper bound for the fully observable model, still obtained by the LSMC method and PD method with additional basis functions for each problem:

$$H_1^{add} = \{L_0(e^{X_t^1}), L_0(e^{X_t^1})L_1(S_t^1)\}$$

$$H_2^{add} = \{L_0(e^{X_t^1}), L_0^2(e^{X_t^1}), L_0(e^{X_t^2}), L_0^2(e^{X_t^2}), L_0(e^{X_t^1})L_2(S_t^1, S_t^2), L_0(e^{X_t^2})L_2(S_t^1, S_t^2)\}.$$

Each entry in Table 10 and Table 11 shows the sample average and the standard error (in parentheses) of the numerical results of 20 independent runs using the following procedure: we implement the LSMC method with 50000 sample paths to obtain a suboptimal policy τ, and then apply this policy on another independent set of 50000 paths to get the lower bound LB; the dual upper bound UB is obtained by implementing Algorithm 1 using the suboptimal policy τ with the number of sample paths $N = 500$, number of particles $m = 500$, and number of subpaths $l = 10$; to investigate the option prices under the fully observable stochastic volatility, we use the PD method with 500 sample paths and 5000 subpaths in nested simulation (which is equal to $m \cdot l$) to obtain an upper bound $Full.\widetilde{UB}$, since the policy τ obtained before is also a suboptimal policy for the fully observable model. Except the new sets of basis functions, the LSMC and PD methods are implemented exactly the same way as before to generate another set of lower bound $Full.LB$ and upper bound $Full.UB$ for the fully observable model. In practice we often use the average of LB and UB, and the average of $Full.LB$ and $Full.UB$ as estimates of the option prices to the partially observable and fully observable problems, respectively.

Table 9: Parameter Sets

#	$(\theta, \lambda, \gamma)$	μ
1	$(\log(0.1), 1.0, 1.0)$	0.3
2	$(\log(0.1), 1.0, 1.0)$	1.0
3	$(\log(0.2), 0.5, 1.0)$	0.3
4	$(\log(0.2), 0.5, 1.0)$	1.0
5	$(\log(0.2), 1.5, 1.0)$	0.3
6	$(\log(0.2), 1.5, 1.0)$	1.0
7	$(\log(0.2), 1.0, 0.5)$	0.3
8	$(\log(0.2), 1.0, 0.5)$	1.0
9	$(\log(0.3), 2.0, 0.3)$	0.3
10	$(\log(0.3), 2.0, 0.3)$	1.0

Table 10: American Put Option Prices on One Asset ($r = 0.05$, $K = 40$, $\delta = 0.1$, $T = 10$, $S_0 = 36$, $X_0 = \theta$)

	Volatility not observable		Volatility directly observable		
#	LB	UB	$Full.\widetilde{UB}$	$Full.LB$	$Full.UB$
1	3.820(0.000)	3.820(0.000)	3.825(0.001)	3.820(0.000)	3.821(0.000)
2	3.853(0.001)	3.887(0.001)	3.954(0.003)	3.905(0.002)	3.912(0.001)
3	3.892(0.001)	4.019(0.003)	4.321(0.005)	4.197(0.003)	4.209(0.001)
4	5.009(0.006)	5.216(0.005)	5.368(0.009)	5.297(0.005)	5.328(0.001)
5	3.881(0.001)	3.898(0.001)	3.995(0.004)	3.928(0.002)	3.938(0.001)
6	4.842(0.003)	4.935(0.002)	5.028(0.003)	4.973(0.004)	4.997(0.001)
7	3.869(0.001)	3.870(0.000)	3.876(0.001)	3.871(0.001)	3.872(0.000)
8	4.632(0.002)	4.653(0.001)	4.704(0.002)	4.679(0.003)	4.689(0.001)
9	4.010(0.001)	4.022(0.001)	4.049(0.001)	4.030(0.001)	4.044(0.001)
10	5.881(0.003)	5.902(0.001)	5.907(0.001)	5.896(0.005)	5.904(0.001)

The numerical results are divided into two categories: the first six rows report the numerical results under the dominant volatility effects, i.e., γ is comparatively large and λ is comparatively small; the last four rows report the results under moderate/weak volatility effects. It can be seen from the tables that $[Full.LB, Full.UB]$

Table 11: American Put Option Prices on the Minimum of Two Assets ($r = 0.05$, $K = 40$, $\delta = 0.1$, $T = 10$, $S_0 = (36, 36)^\top$, $X_0 = (\theta, \theta)^\top$)

	Volatility not observable		Volatility directly observable		
#	LB	UB	$Full.\widetilde{UB}$	$Full.LB$	$Full.UB$
1	4.027(0.002)	4.032(0.001)	4.068(0.002)	4.039(0.001)	4.043(0.001)
2	5.004(0.006)	5.147(0.004)	5.256(0.006)	5.143(0.005)	5.222(0.003)
3	5.274(0.005)	5.378(0.002)	5.565(0.004)	5.467(0.004)	5.489(0.001)
4	8.045(0.006)	8.171(0.004)	8.289(0.006)	8.188(0.010)	8.268(0.003)
5	4.641(0.002)	4.782(0.001)	4.918(0.005)	4.833(0.006)	4.870(0.001)
6	7.531(0.006)	7.638(0.002)	7.723(0.007)	7.606(0.007)	7.704(0.002)
7	4.429(0.002)	4.456(0.001)	4.514(0.001)	4.477(0.002)	4.500(0.001)
8	6.984(0.004)	7.042(0.003)	7.074(0.004)	6.997(0.007)	7.080(0.001)
9	5.417(0.002)	5.428(0.001)	5.449(0.001)	5.431(0.003)	5.447(0.001)
10	9.084(0.006)	9.130(0.002)	9.138(0.002)	9.071(0.009)	9.133(0.002)

is usually a tighter interval than $[LB, Full.\widetilde{UB}]$ for the fully observable option price, since more information is used to determine a better exercise policy. To differentiate the option prices under full and partial observations of stochastic volatility, [74] pointed out that the partial observation of stochastic volatility has an impact especially when the effect of the volatility (*i.e.*, $\frac{\gamma^2}{2\lambda}$) is high. Our numerical results also support their viewpoints in terms of the differences between UB and $Full.\widetilde{UB}$, which demonstrate the effectiveness of introducing the filtering step. In particular, it can be observed that we can reduce relatively more overpricing for problems with dominant volatility (i.e., the first category). Considering the differences between LB and $Full.UB$, partially observable and fully observable option prices have relatively small gaps under moderate/weak volatility effects compared with the gaps in the first category. Larger observation noise μ challenges the performance of suboptimal exercise policy and also deteriorates the performance of particle filtering, so it generally increases the gap between $Full.LB$ and $Full.UB$ and the gap between LB and UB. Compared with [74] and [61], whose approaches provide asymptotic lower bounds on the option prices, our main contribution is to provide an asymptotic upper bound on the option price, which is less than or similar to the lower bound ($Full.LB$) of the corresponding fully observable option price in the first category. Hence, our method provides a better criterion to evaluate the performance of LB: the smaller the gap between UB and LB, the better the bounds. If the gap between UB and LB is small enough, they can be both regarded as approximate option prices under partial observation. Otherwise, improvement on the exercise policy should be considered.

5.5 Conclusion

We propose a numerical approach to solve for the value function of the partially observable optimal stopping problem. We represent the value function as a solution of a dual minimization problem, based on which we develop an algorithm that complements a suboptimal stopping time with an asymptotic upper bound on the value function. Our approach provides a practical way to judge whether more computational effort is needed to improve the quality of the approximate solution. We apply our approach to price American put options in stochastic volatility models, with the realistic assumption that the volatility cannot be directly observed but can be inferred from the asset prices. The numerical results confirm a higher price of the option if we alternatively assume that the volatility is directly observable. The price difference is more significant when the effect of volatility is high, indicating the importance of taking the partial observability into account.

CHAPTER VI

CONCLUSIONS AND FUTURE RESEARCH

6.1 Conclusions

This thesis has developed new theories and computational methods that extend the scope of information relaxation in three important dynamic decision making problems.

The first part of the thesis studies the interactions of Lagrangian relaxation and information relaxations in weakly coupled dynamic programs. We generalize the information relaxation approach to obtain a tighter dual bound than the Lagrangian relaxation bound in discounted infinite-horizon problems. To develop this approach, we first employ a geometric distributed randomized time to convert the discounted infinite-horizon inner problem to a finite (but random) horizon problem. Next, we propose a computationally tractable method that relaxes the inner problem to tackle large-scale problems. We provide insightful interpretation and theoretical analysis on the relative gap between the exact and practical information relaxation bounds.

The second part of the thesis is devoted to establish the information relaxation-based dual representation of controlled Markov diffusion. We derive the weak an strong duality as well as the complementary slackness conditions in parallel with the results in MDPs. In particular, we explore the structure of the value function-based optimal penalty and show that it takes the compact form of a stochastic integral under the natural filtration generated by the Brownian motion. We discuss the connection between the dual formulations of MDPs and controlled Markov diffusions. An application is illustrated by a 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 and easy-to-compute penalty is proposed to derive dual bounds on the optimal value of the discrete-time model and produces small duality gaps.

The third part of the thesis focuses on discrete-time continuous-state partially observable optimal stopping problem. We develop a simulation-based method to provide an approximate solution to the optimal value. By treating the filtering distribution as a state, the partially observable problem is transformed to an equivalent fully observable optimal stopping problem. We extend the martingale duality to this formulation, based on which we apply the particle filtering technique to develop a numerical method and show that it is an asymptotic upper bound on the optimal value. We use this approach to price American put options on one and two assets respectively, and compare with the option prices from the models assuming fully observable volatility. The numerical results indicate that different assumptions on the observability of stochastic volatility have an impact on the option price, and show that our method effectively reduces overpricing of the option.

6.2 Future Research

My future research will study the impact of future information on the performance of decision strategies in dynamic decision-making problems including applications in resources allocation and revenue management. The current information relaxation technique is mainly used to place a dual bound on the performance of non-anticipative decision strategy. In many real-world problems, the non-anticipativity constraint may be relieved thanks to the modern forecasting technology and business strategy. For instance,

- The sensor technology allows monitoring and collecting traffic information in office buildings or hotels, which can be used to predict future requests for elevator service.
- Some airline companies offer the option of locking in the current fare for a

specified period of time to the potential passengers, which helps the company to forecast the future demands within this time window.

Under the circumstances that the decision maker has partial access to the future information, I plan to (i) develop new algorithm that can effectively utilize the inexact future information; (ii) study the conditions under which the new algorithm can gain significant improvement over the conventional "non-anticipative" algorithm; (iii) analyze the limitation of the anticipative algorithm, especially on its sensitivity to the inaccurate forecast. As a concrete example, incorporating advance passenger information into the scheduling of group elevator system may help reduce the average waiting time of passengers during peak times.

APPENDIX A

WEAKLY COUPLED DYNAMIC PROGRAM

A.1 Approximate Linear Programming Approach

The approximate linear programming (ALP) method aims to find a good approximation of V within a parameterized class of functions with a lower-dimensional representation [30]. In the setting of weakly coupled stochastic dynamic program, we can set $H(\mathbf{x}) = \theta + \sum_{n=1}^{N} H^n(x^n)$, where θ is a constant and $H^n(\cdot)$ only depends on x^n for $n = 1, \cdots, N$. This approximation scheme is probably motivated by the additive form of Lagrangian function.

Note that each $H^n(\cdot)$ is a mapping from $\mathcal{X}^n$ to $\mathbb{R}$ determined by $|\mathcal{X}^n|$ values, which implies that $H(\mathbf{x})$ can be represented with $1 + \sum_{i=1}^{n} |\mathcal{X}^n|$ variables. To determine appropriate parameter values, we are seeking a best feasible and additively separable solution from the following linear program (with variables θ and $\{H^n(\cdot)\}_{n=1}^{N}$):

$$H^{LP}(\upsilon) \triangleq \min_{\{\theta, H^n(\cdot)\}} \theta + \sum_{n=1}^{N} \sum_{x^n \in \mathcal{X}_n} \upsilon_n(x^n) H^n(x^n) \tag{94}$$

$$\text{s.t.} \;\; \theta(1-\beta) + \sum_{n=1}^{N} H^n(x_0^n) \geq \sum_{n=1}^{N} R^n(x_0^n, a_0^n) + \beta \sum_{n=1}^{N} \sum_{x_1^n \in \times \mathcal{X}_n} P_n(x_1^n | x_0^n, a_0^n) H^n(x_0^n),$$

$$\text{for all } \;\mathbf{x}_0 \in \mathcal{X} \text{ and } \mathbf{a}_0 \in \bar{\mathcal{A}}(\mathbf{x}_0),$$

where $\upsilon_n(x^n)$ is the marginal distributions of x^n derive from a probability distribution $\upsilon(\cdot)$ on $\mathcal{X}$. This linear programming has $\sum_{n=1}^{N} |\mathcal{X}_n|$ variables and can have as many as $\prod_{n=1}^{N} (\sum_{x^n \in \mathcal{X}_n} |\mathcal{A}_n(x^n)|)$ constraints. We denote by $\{\theta^*, H^{LP,n}(\cdot), \; n = 1, \cdots, N\}$ the optimal solution to (94), and define

$$H^{LP}(\mathbf{x}) \triangleq \theta^* + \sum_{n=1}^{N} H^{LP,n}(x^n).$$

The following lemma shows that the bound derived by the ALP method is tighter than the Lagrangian bound, the proof of which can be found in [1].

Lemma 6. *(a). Let $\{\boldsymbol{\lambda}, H^{\boldsymbol{\lambda},n}(\cdot),\ n = 1, \cdots, N\}$ be a feasible solution to the linear program (16). Then $\{\frac{\boldsymbol{\lambda}^\top \boldsymbol{b}}{1-\beta}, H^{\boldsymbol{\lambda},n}(\cdot),\ n = 1, \cdots, N\}$ is also in the feasible region of the linear program (94), i.e., $J^{\boldsymbol{\lambda}} \in \mathcal{D}^*$ (see the definition of $\mathcal{D}^*$ in Theorem 3(b)).*

(b). $H^{LP}(\upsilon) \leq J^{\boldsymbol{\lambda}}(\upsilon)$ for any $\boldsymbol{\lambda} \in \mathbb{R}^L_+$ and probability distribution υ.

(c). $H^{LP} \in \mathcal{D}^*$, and $V(\boldsymbol{x}) \leq H^{LP}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$ (regardless of the distribution υ).

A.2 Complements to Section 3.2

A.2.1 A formal definition of τ

In this subsection we discuss the augmentation of the probability space $(\Omega, \mathcal{F}, P)$ associated with the original problem (11) due to the introduction of the random time τ. We can assume that the random variable τ is associated with another probability space $(\hat{\Omega}, \hat{\mathcal{G}}, \hat{P})$, where $\tau : \hat{\Omega} \to \mathbb{N}$, $\hat{\mathcal{G}}$ is the σ-algebra generated by τ (i.e., $\sigma(\tau)$), and $\hat{P}(\tau = t) = (1 - \beta)\beta^t$ for $t = 0, 1, 2, \cdots$.

The probability space $(\Omega, \mathcal{F}, P)$ is then augmented to $(\Omega \times \hat{\Omega}, \mathcal{F} \otimes \sigma(\tau), \mathbb{P})$, where $\mathcal{F} \otimes \sigma(\tau)$ is the product σ-algebra of $\mathcal{F}$ and $\sigma(\tau)$, and $\mathbb{P}$ is the product measure of P and $\hat{P}$, i.e., $\mathbb{P}(A \times [t, \infty)) = P(A) \times \hat{P}(\tau \geq t) = P(A) \times \beta^t$ with $A \in \mathcal{F}$. We clarify this (straightforward) augmentation is because we can use the pair (ω, τ) to denote the uncertainty in $\mathbb{E}_0[\cdot]$ in (21) without confusion, though we use P to denote $\mathbb{P}$ to save notations.

A.2.2 $\max_{\mathbf{a} \in \bar{\mathbb{A}}}\{I_H(\mathbf{a}, \omega, \tau)\}$ has finite mean and variance

Let $\mathcal{I}(\omega, \tau) = \max_{\mathbf{a} \in \bar{\mathbb{A}}}\{I_H(\mathbf{a}, \omega, \tau)\}$. Then $\mathcal{L}H(\mathbf{x}_0) \triangleq H(\mathbf{x}_0) + \mathbb{E}_0[\mathcal{I}(\omega, \tau)]$. Since $\mathbf{R}$ and H are both bounded, we can assume for all $(\mathbf{x}_t, \mathbf{a}_t) \in \mathcal{X} \times \mathcal{A}$, $t = 0, 1, 2, \cdots$,

$$|\mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t, \mathbf{a}_t] - H(\mathbf{x}_t)| \leq C$$

for some $C > 0$. Therefore, $|\mathcal{I}(\omega,\tau)| \le (\tau+1)C$ for any $\omega \in \Omega$, which implies

$$|\mathbb{E}_0[\mathcal{I}(\omega,\tau)]| \le \mathbb{E}_0\left[\mathbb{E}\left[|\mathcal{I}(\omega,\tau)||\tau\right]\right] \le \sum_{\tau=0}^{\infty}(1-\beta)\beta^{\tau}(\tau+1)C = \frac{C}{1-\beta} < \infty, \tag{95}$$

and $\mathrm{Var}[\mathcal{I}(\omega,\tau)|\tau] \le \mathbb{E}[\mathcal{I}^2(\omega,\tau)|\tau] \le (\tau+1)^2C^2$. The inequality (95) indicates that $\mathcal{I}(\omega,\tau)$ has finite mean.

We note that $\mathrm{Var}[\mathcal{I}(\omega,\tau)] = \mathbb{E}[\mathrm{Var}[\mathcal{I}(\omega,\tau)|\tau]] + \mathrm{Var}[\mathbb{E}[\mathcal{I}(\omega,\tau)|\tau]]$.

It can be seen that $\mathbb{E}[\mathrm{Var}[\mathcal{I}(\omega,\tau)|\tau]] \le \sum_{\tau=0}^{\infty}(1-\beta)\beta^{\tau}(\tau+1)^2C^2 = \frac{1+\beta}{(1-\beta)^2}C^2 < \infty$, and

$$\begin{aligned}\mathrm{Var}[\mathbb{E}[\mathcal{I}(\omega,\tau)|\tau]] \le & \mathbb{E}[(\mathbb{E}[\mathcal{I}(\omega,\tau)|\tau])^2] \le \mathbb{E}[(\tau+1)^2C^2] \\ = & \sum_{\tau=0}^{\infty}(1-\beta)\beta^{\tau}(\tau+1)^2C^2 = \frac{1+\beta}{(1-\beta)^2}C^2 < \infty.\end{aligned}$$

Hence, we conclude that $\mathcal{I}(\omega,\tau)$ has finite variance.

A.3 Whether Information Relaxation Can Improve the Lagrangian Bound

We consider the restless bandit-like problem with $N = 1$ as proposed in Section 3.3 of [1]: the state space contains three states, i.e., $\mathcal{X} = \{0,1,2\}$, and for each state $\mathbf{x} \in \mathcal{X}$ the control space is $\mathcal{A}(\mathbf{x}) = \{0,1\}$. The corresponding reward $\mathbf{R}(\mathbf{x},\mathbf{a})$, weight $\mathbf{B}(\mathbf{x},\mathbf{a})$, and transition probability $P(\mathbf{x}_{t+1}|\mathbf{x}_t,\mathbf{a}_t)$ are listed in Table 12, in which $l > 0$ and $c > 1$ are positive constants. Note that states "1" and "2" are absorbing states regardless of the control applied, however, the state "0" may transit to either "1" or "2" depending on the control chosen. The linking constraint is $\mathbf{B}(\mathbf{x},\mathbf{a}) \le 1$. Therefore, $\bar{\mathcal{A}}(0) = \bar{\mathcal{A}}(2) = \{0,1\}$ and $\bar{\mathcal{A}}(1) = \{0\}$.

Remark 5. *In Table 1 of [1],* $\boldsymbol{D}(2,0) = \epsilon > 0$*. For simplicity we set* $\epsilon = 0$*, and the results therein are still true.*

The exact value function (11) is $V(0) = \frac{c\beta}{1-\beta}$, $V(1) = 0$, and $V(2) = \frac{c}{1-\beta}$. The optimal stationary policy is $\alpha = (\alpha^*, \alpha^*, \cdots)$, where $\alpha^*(0) = \alpha^*(1) = 0$ and $\alpha^*(2) = 1$.

Table 12: One-subproblem with $b = 1$ and $\beta \in (\frac{1}{2}, 1)$

State	Control	Reward	Weight	Transition
0	0	$\mathbf{R}(0,0) = 0$	$\mathbf{B}(0,0) = 0$	$P(2\|0,0) = 1$
0	1	$\mathbf{R}(0,1) = 0$	$\mathbf{B}(0,1) = 0$	$P(1\|0,1) = 1$
1	0	$\mathbf{R}(1,0) = 0$	$\mathbf{B}(1,0) = 0$	$P(1\|1,0) = 1$
1	1	$\mathbf{R}(1,0) = c(2+l)$	$\mathbf{B}(1,1) = 2$	$P(1\|1,1) = 1$
2	0	$\mathbf{R}(2,0) = 0$	$\mathbf{B}(2,0) = 0$	$P(2\|2,0) = 1$
2	1	$\mathbf{R}(2,1) = c$	$\mathbf{B}(2,1) = 0$	$P(2\|2,1) = 1$

The Lagrangian relaxation yields $J^{\lambda}(\mathbf{x}) = \frac{\lambda}{1-\beta} + H^{\lambda}(\mathbf{x})$ for $\mathbf{x} = 0, 1, 2$, where $H^{\lambda}(\cdot)$ are the solution to (15). According to [1], the optimal Lagrangian multiplier is

$$\lambda^* = \arg\min_{\lambda \geq 0} H^{\lambda}(v) = c + cl/2,$$

which implies $H^{\lambda^*}(0) = 0$, $H^{\lambda^*}(1) = 0$, and $H^{\lambda^*}(2) = 0$. Therefore,

$$J^{\lambda^*}(0) = \frac{\lambda^*}{1-\beta},\ J^{\lambda^*}(1) = \frac{\lambda^*}{1-\beta},\ \text{and } J^{\lambda^*}(2) = \frac{\lambda^*}{1-\beta}.$$

Note that $J^{\lambda^*}(\cdot)$ are unbounded on $\mathcal{X}$ as $l \to \infty$, though the exact values $V(\cdot)$ are constant with respect to l.

By applying the information relaxation approach with $H = J^{\lambda^*}$,

$$\begin{aligned}
&\mathcal{L}J^{\lambda^*}(\mathbf{x}_0) \\
=&H(\mathbf{x}_0) + \mathbb{E}_0\left[\max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} \left\{\sum_{t=0}^{\tau} \big(\mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \beta\mathbb{E}[H(\mathbf{x}_{t+1})|\mathbf{x}_t, \mathbf{a}_t] - H(\mathbf{x}_t)\big)\right\}\right] \\
=&H(\mathbf{x}_0) + \sum_{T=0}^{\infty}(1-\beta)\beta^T \cdot \mathbb{E}_0\left[\max_{\mathbf{a} \in \bar{\mathbb{A}}} \left\{\sum_{t=0}^{T} \big(\mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \beta\mathbb{E}[J^{\lambda^*}(\mathbf{x}_{t+1})|\mathbf{x}_t, \mathbf{a}_t] - J^{\lambda^*}(\mathbf{x}_t)\big)\right\}\right].
\end{aligned}$$

Note that $J^{\lambda^*}(\mathbf{x}_0) - (\mathbf{R}(\mathbf{x}_0, \mathbf{a}_0) + \beta\mathbb{E}[J^{\lambda^*}(\mathbf{x}_1)|\mathbf{x}_0, \mathbf{a}]) \geq \lambda^* - c$ for all $\mathbf{x}_0 = 1, 2, 3$ and $\mathbf{a}_0 \in \bar{\mathcal{A}}(\mathbf{x}_0)$. According to Theorem 5(c), $J^{\lambda^*}(\mathbf{x}) - \mathcal{L}J^{\lambda^*}(\mathbf{x}) \geq \frac{\lambda^* - c}{1-\beta}$, which implies $\mathcal{L}J^{\lambda^*}(\mathbf{x}) \leq \frac{c}{1-\beta}$ for $\mathbf{x} = 1, 2, 3$. This bound remains constant with respect to l and it has already been tight as an upper bound on $V(2)$.

We can show that the exact computation of the information relaxation bound also leads to a tight upper bound on $V(0)$. Starting at $\mathbf{x}_0 = 0$ and for each $T \in \mathbb{N}$ and

$\omega \in \Omega$,

$$
\begin{aligned}
&\max_{\mathbf{a}\in\mathcal{A}(T)} \left\{ \sum_{t=0}^{T} \left(\mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \beta \mathbb{E}[J^{\lambda^*}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - J^{\lambda^*}(\mathbf{x}_t) \right) \right\} \\
&= \max \Bigg\{ \mathbf{R}(0,0) + \beta \mathbb{E}[J^{\lambda^*}(\mathbf{x}_1)|0,0] - J^{\lambda^*}(0) \\
&\qquad + \sum_{t=1}^{T} \max_{\mathbf{a}_t \in \bar{\mathcal{A}}(\mathbf{x}_t)} \left\{ \mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \beta \mathbb{E}[J^{\lambda^*}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - J^{\lambda^*}(\mathbf{x}_t) \right\}, \\
&\qquad \mathbf{R}(0,1) + \beta \mathbb{E}[J^{\lambda^*}(x_1)|0,1] - J^{\lambda^*}(0) \\
&\qquad + \sum_{t=1}^{T} \max_{\mathbf{a}_t \in \bar{\mathcal{A}}(\mathbf{x}_t)} \{ \mathbf{R}(\mathbf{x}_t, \mathbf{a}_t) + \beta \mathbb{E}[J^{\lambda^*}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - J^{\lambda^*}(\mathbf{x}_t) \} \Bigg\} \\
&= \max \Bigg\{ \mathbf{R}(0,0) + \beta J^{\lambda^*}(2) - J^{\lambda^*}(0) + \sum_{t=1}^{T} \max_{\mathbf{a}_t \in \bar{\mathcal{A}}(2)} \left\{ \mathbf{R}(2, \mathbf{a}_t) - (1-\beta) J^{\lambda^*}(2) \right\}, \\
&\qquad \mathbf{R}(0,1) + \beta J^{\lambda^*}(1) - J^{\lambda*}(0) + \sum_{t=1}^{T} \max_{\mathbf{a}_t \in \bar{\mathcal{A}}(1)} \left\{ \mathbf{R}(1, \mathbf{a}_t) - (1-\beta) J^{\lambda^*}(1) \right\} \Bigg\} \\
&= \max \left\{ 0 + \beta \frac{\lambda^*}{1-\beta} - \frac{\lambda^*}{1-\beta} + (c - \lambda^*) T,\ 0 + \beta \frac{\lambda^*}{1-\beta} - \frac{\lambda^*}{1-\beta} + (0 - \lambda^*) T \right\} \\
&= -\lambda^* + (c - \lambda^*) T,
\end{aligned}
$$

where the first equality holds since staring at $\mathbf{x}_0 = 0$, the control $\mathbf{a}_0 = 0$ leads to $\mathbf{x}_1 = 2$ (respectively, $\mathbf{a}_0 = 1$ leads to $\mathbf{x}_1 = 1$) with probability 1, and hence determine all the subsequent states $\mathbf{x}_2$, $\mathbf{x}_3$, $\cdots$, since $\mathbf{x} = 1$ and 2 are absorbing states. Consequently, the deterministic dynamic program with time horizon T can be decomposed as the summation of T sub-problems. The last equality holds as the first term dominates the second, meaning that $\mathbf{a}_0 = 0$ and $\mathbf{a}_1 = 1$ for $t \geq 1$ is the solution to the inner optimization problem for all the scenarios $\omega \in \Omega$. Since $J^{\lambda^*}(0) = \frac{\lambda^*}{1-\beta}$, then

$$
\begin{aligned}
\mathcal{L} J^{\lambda^*}(0) = &\frac{\lambda^*}{1-\beta} + \mathbb{E}_0[-\lambda^* + (c - \lambda^*)\tau] \\
= &\frac{\lambda^*}{1-\beta} + \sum_{\tau=0}^{\infty} (1-\beta)\beta^{\tau} [-\lambda^* + (c - \lambda^*)\tau] = \frac{c\beta}{1-\beta}.
\end{aligned}
$$

Hence, $\mathcal{L} J^{\lambda^*}(0) = V(0)$.

A.4 Proof of Theorem 9

To prove Theorem 5, we use the result of Lagrangian duality gap on deterministic separable problem. Consider a separable problem

$$\max_{\mathbf{a}\in\bar{\mathcal{A}}} \sum_{n=1}^{N} f^n(a^n), \tag{96}$$

where $\bar{\mathcal{A}} = \{\mathbf{a} \triangleq (a^1, \cdots, a^N) \in \mathcal{A}^1 \times \cdots \times \mathcal{A}^N |\ \sum_{n=1}^{N} \mathbf{h}^n(a^n) \leq \mathbf{q}\}$ with $\mathbf{q} \in \mathbb{R}^{\tilde{L}}$.

We then define the Lagrangian dual of (96):

$$\min_{\mu\geq 0}\ d(\mu) \triangleq \sum_{n=1}^{N} \max_{a^n\in\mathcal{A}^n} \{f^n(a^n) - \mu^\top \mathbf{h}^n(a^n)\} + \mu^\top \mathbf{q}.$$

Lemma 7 (Proposition 5.26 in [8])**.** *Suppose the following assumptions hold.*

Assumption 1: $\bar{\mathcal{A}} \neq \emptyset$.

Assumption 2: for each $n = 1, \cdots, N$, $\{a^n, \boldsymbol{h}^n(a^n), f^n(a^n) | a^n \in \mathcal{A}^n\}$ *is compact.*

Assumption 3: for each $n = 1, \cdots, N$, *given any vector* $\tilde{a}^n \in conv(\mathcal{A}^n)$, *there exists* $a^n \in \mathcal{A}^n$ *such that*

$$\boldsymbol{h}^n(a^n) \leq (\check{cl}\ \boldsymbol{h}^n)(\tilde{a}^n).$$

Then

$$\min_{\mu\geq 0} d(\mu) - \max_{\boldsymbol{a}\in\bar{\mathcal{A}}} \sum_{n=1}^{N} f^n(a^n) \leq (\tilde{L}+1) \max_{n=1,\cdots,N} \rho_n,$$

where $\rho_n = \sup_{a^n\in conv(\mathcal{A}^n)} \left\{ \widetilde{f^n}(a^n) - (\check{cl}\ f^n)(a^n) \right\}$.

The proof of Theorem 5 uses the following lemma, which is a corollary of Lemma 7.

Lemma 8. *Suppose that* H *is of the additively separable form* $H(\boldsymbol{x}) = \theta + \sum_{n=1}^{N} H^n(x^n)$, *and Assumptions 1-3 in Section 3.2 hold for* $\omega \in \Omega$ *and* $T \in \mathbb{N}$. *Then*

$$\min_{\boldsymbol{\mu}\geq 0} \max_{\boldsymbol{a}\in\mathcal{A}(T)} I_H(\boldsymbol{a}, \omega, T; \boldsymbol{\mu}) - \max_{\boldsymbol{a}\in\bar{\mathcal{A}}(T)} I_H(\boldsymbol{a}, \omega, T) \leq (1 + L(T+1)) \max_{n=1,\cdots,N} \gamma^n,$$

where

$$\gamma^n = \sup_{\tilde{\boldsymbol{a}}^n \in conv(\mathcal{A}^n(T))} \left\{ \widetilde{I^n_{H_n}}(\tilde{\boldsymbol{a}}^n, \omega, T; 0) - (\check{cl}\ I^n_{H_n})(\tilde{\boldsymbol{a}}^n, \omega, T; 0) \right\},$$

$\check{cl}\ I^n_H$ is the convex closure of $I^n_{H_n}$, and $\widetilde{I^n_{H_n}}$ is defined as

$$\widetilde{I^n_{H^n}}(\tilde{\boldsymbol{a}}^n, \omega, T; 0) = \inf_{\boldsymbol{a}^n \in \mathcal{A}^n(T)} \left\{ I^n_{H^n}(\boldsymbol{a}^n, \omega, T; 0) | \boldsymbol{B}^n_t(\boldsymbol{a}^n, \omega) \leq (\check{cl}\ \boldsymbol{B}^n_t)(\tilde{\boldsymbol{a}}^n, \omega),\ t = 0, \cdots, T \right\}.$$

Remark 6. *Note that $\widetilde{I^n_{H^n}}(\tilde{\boldsymbol{a}}^n, \omega, T; 0)$ is well-defined according to Assumption* **??** *in Section* **??**.

Proof. Lemma 8 directly follows from Lemma 7 by setting $f^n = I^n_{H^n}$, $\mathbf{h}^n = (\mathbf{B}^n_0, \cdots, \mathbf{B}^n_T)$, $\mathbf{q} = (\mathbf{b}, \cdots, \mathbf{b}) \in \mathbb{R}^{\tilde{L}}$ with $\tilde{L} = L \times (T+1)$, and the decision variable $a^n = \mathbf{a}^n \in \mathcal{A}^n(T)$. □

A.4.1 Proof of Theorem 9

According to Lemma 8, we have for fixed $\omega \in \Omega$ and $\tau = T$,

$$\min_{\mu \geq 0} \max_{\mathbf{a} \in \mathcal{A}(T)} I_H(\mathbf{a}, \omega, T; \mu) - \max_{\mathbf{a} \in \bar{\mathcal{A}}(T)} I_H(\mathbf{a}, \omega, T) \leq (1 + L(T+1)) \max_{n=1,\cdots,N} \gamma^n,$$

where

$$\begin{aligned}
\gamma^n \leq & \sup_{\mathbf{a}^n \in \mathcal{A}^n(\tau)} \{ I^n_{H^n}(\mathbf{a}^n, \omega, T; 0) \} - \inf_{\mathbf{a}^n \in \mathcal{A}^n(\tau)} \{ I^n_{H^n}(\mathbf{a}^n, \omega, T; 0) \} \\
\leq & (T+1) \sup_{x^n_0 \in \mathcal{X}^n, a^n_0 \in \mathcal{A}^n(x^n_0)} \{ R^n(x^n_0, a^n_0) + \beta \mathbb{E}[H^n(x^n_1) | x^n_0, a^n_0] - H^n(x^n_0) \} \\
& - (T+1) \inf_{x^n_0 \in \mathcal{X}^n, a^n_0 \in \mathcal{A}^n(x^n_0)} \{ R^n(x^n_0, a^n_0) + \beta \mathbb{E}[H^n(x^n_1) | x^n_0, a^n_0] - H^n(x^n_0) \} \\
= & (T+1) \Gamma^n,
\end{aligned}$$

where the first inequality is due to the definitions of $\widetilde{I^n_{H^n}}$ and $\check{\text{cl}}\ I^n_{H^n}$, and the second inequality holds independent of ω. It is straightforward to see

$$\begin{aligned}
\mathcal{L}^\circ H(x) - \mathcal{L} H(x) &= \mathbb{E}\left[\min_{\boldsymbol{\mu} \geq 0} \max_{\mathbf{a} \in \mathcal{A}(\tau)} \{ I_H(\mathbf{a}, \omega, \tau; \boldsymbol{\mu}) \} - \max_{\mathbf{a} \in \bar{\mathcal{A}}(\tau)} \{ I_H(\mathbf{a}, \omega, \tau) \} \right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[(1 + L(\tau+1))(\tau+1) \max_{n=1,\cdots,N} \Gamma^n \middle| \tau \right] \right].
\end{aligned}$$

Then we can obtain (35), since

$$\mathbb{E}\left[\mathbb{E}\left[(1+L(\tau+1))(\tau+1)\max_{n=1,\cdots,N}\Gamma^n\middle|\tau\right]\right] = \max_{n=1,\cdots,N}\Gamma^n\cdot\mathbb{E}\left[(1+L(\tau+1))(\tau+1)\right]$$
$$= \frac{(L-1)\beta+L+1}{(1-\beta)^2}\max_{n=1,\cdots,N}\Gamma^n.$$

A.5 Finite horizon case

In this section we consider the finite-horizon weakly coupled dynamic program, which is the same as infinite-horizon case except that

1. The time is indexed by $t = 0, \cdots, T$.

2. The transition probability can be time-varying.

3. The linking constraint can be time-varying, and the feasible control set at time t is

$$\bar{\mathcal{A}}_t(\mathbf{x}_t) = \{\mathbf{a} = (a_t^1, \cdots, a_t^N) \in \mathcal{A}_t(\mathbf{x}_t) : \ \mathbf{B}_t(\mathbf{x}_t, \mathbf{a}_t) \triangleq \sum_{n=1}^{N} \mathbf{B}_t^n(x_t^n, a_t^n) \leq \mathbf{b}_t\},$$

   where each $\mathbf{b}_t \in \mathbb{R}^L$ for $t = 0, \cdots, T$.

4. The intermediate rewards denoted by $R_t(x_t, a_t) = \sum_{n=1}^{N} R_t^n(x_t^n, a_t^n)$ can also be time-varying.

The objective of the decision maker is to maximize the expected rewards given $\mathbf{x}_0 \in \mathcal{X}$,

$$U_0(\mathbf{x}_0) = \max_{\alpha\in\bar{\mathbb{A}}_{\mathbb{F}}(T)} U_0(\mathbf{x}_0;\alpha), \tag{97}$$

where

$$U_0(\mathbf{x}_0;\alpha) = \mathbb{E}\left[\sum_{t=0}^{T} R_t(\mathbf{x}_t, \mathbf{a}_t)\middle|\mathbf{x}_0\right],$$

and $\bar{\mathbb{A}}_{\mathbb{F}}(T)$ is the set of non-anticipative policies α that selects $\mathbf{a}_t \in \bar{\mathcal{A}}_t(\mathbf{x}_t)$ for each $t = 0, 1, \cdots, T$. Then U_0 can be solved via the dynamic programming:

$$\begin{aligned} &U_{T+1}(\mathbf{x}_{T+1}) = 0; \\ &\qquad U_t(\mathbf{x}_t) = \max_{\mathbf{a}_t \in \bar{\mathcal{A}}_t(\mathbf{x}_t)} \left\{ R_t(\mathbf{x}_t, \mathbf{a}_t) + \mathbb{E}[U_{t+1}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] \right\}. \end{aligned}$$

A.5.1 Lagrangian Relaxation

Let $\mathbb{A}_{\mathbb{F}}(T) = \{\alpha \in \mathbb{A}(T) |\ \alpha \text{ is non-anticipative}\}$. By dualizing the linking constraint with Lagrangian multipliers $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_0, \cdots, \boldsymbol{\lambda}_T) \geq 0$ with each $\boldsymbol{\lambda}_t \in \mathbb{R}^L_+$, we define for $\mathbf{x}_0 \in \mathcal{X}$,

$$J_0^{\boldsymbol{\lambda}}(\mathbf{x}_0) \triangleq \max_{\alpha \in \mathbb{A}_{\mathbb{F}}(T)} J_0^{\boldsymbol{\lambda}}(\mathbf{x}_0; \alpha), \tag{98}$$

where

$$J_0^{\boldsymbol{\lambda}}(\mathbf{x}_0; \alpha) \triangleq \mathbb{E}\left[\sum_{t=0}^{T} R_t(\mathbf{x}_t, \mathbf{a}_t) + \boldsymbol{\lambda}_t^\top \left[\mathbf{b}_t - \mathbf{B}_t(\mathbf{x}_t, \mathbf{a}_t) \right] \middle| \mathbf{x}_0 \right],$$

and $\mathbb{A}_{\mathbb{F}}(T)$ is the set of non-anticipative policies α that selects $\mathbf{a}_t \in \mathcal{A}_t(\mathbf{x}_t)$ for each $t = 0, \cdots, T$. Then $J_0^{\boldsymbol{\lambda}}$ can be solved via the dynamic programming equations:

$$\begin{aligned} &J_{T+1}^{\boldsymbol{\lambda}}(\mathbf{x}_{T+1}) = 0; \\ &\quad J_t^{\boldsymbol{\lambda}}(\mathbf{x}_t) = \max_{\mathbf{a}_t \in \mathcal{A}_t(\mathbf{x}_t)} \left\{ R_t(\mathbf{x}_t, \mathbf{a}_t) + \boldsymbol{\lambda}_t^\top [\mathbf{b}_t - \mathbf{B}_t(\mathbf{x}_t, \mathbf{a}_t)] + \mathbb{E}[J_t^{\boldsymbol{\lambda}}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] \right\}. \end{aligned} \tag{99}$$

Similar to the infinite-horizon case, the solution to (98) can be solved by decomposing (99) into N dynamic programs of lower dimensions:

$$J_0^{\boldsymbol{\lambda}}(\mathbf{x}_0; \alpha) = \sum_{t=0}^{T} \boldsymbol{\lambda}_t^\top \mathbf{b}_t + \mathbb{E}\left[\sum_{t=0}^{T} R_t(\mathbf{x}_t, \mathbf{a}_t) - \boldsymbol{\lambda}_t^\top \mathbf{B}_t(\mathbf{x}_t, \mathbf{a}_t) \middle| \mathbf{x}_0 \right] = \sum_{t=0}^{T} \boldsymbol{\lambda}_t^\top \mathbf{b}_t + \sum_{n=1}^{N} H_0^{\boldsymbol{\lambda}, n}(x_0^n),$$

where

$$\begin{aligned} &H_{T+1}^{\boldsymbol{\lambda}, n}(x_{T+1}^n) = 0, \\ &\quad H_t^{\boldsymbol{\lambda}, n}(x_t^n) = \max_{a_t^n \in \mathcal{A}_t^n(x_t^n)} \left\{ R_t^n(x_t^n, a_t^n) - \boldsymbol{\lambda}_t^\top \boldsymbol{B}_t^n(x_t^n, a_t^n) + \mathbb{E}[H_{t+1}^{\boldsymbol{\lambda}, n}(x_{t+1}^n) | x_t^n, a_t^n] \right\}. \end{aligned}$$

A.5.2 Information Relaxation

We define the space of a sequence of functions $H = (H_0, \cdots, H_{T+1})$:

$$\mathcal{D}_T \triangleq \{H = (H_0, \cdots, H_{T+1}) | H_t : \mathcal{X} \to \mathbb{R} \text{ for } t = 0, \cdots, T+1, and\ H_{T+1}(\cdot) \equiv 0\}.$$

Given $H \in \mathcal{D}_T$, we define

$$\begin{aligned}\mathcal{L}_T H(\mathbf{x}_0) \triangleq &\mathbb{E}_0 \left[\max_{\mathbf{a} \in \mathcal{A}(T)} \left\{ \sum_{t=0}^{T} \big(R_t(\mathbf{x}_t, \mathbf{a}_t) + \mathbb{E}[H_{t+1}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - H_{t+1}(\mathbf{x}_{t+1}) \big) \right\} \right] \\ = &H_0(\mathbf{x}_0) + \mathbb{E}_0 \left[\max_{\mathbf{a} \in \mathbb{A}(T)} \{ I_H(\mathbf{a}, \omega, T) \} \right],\end{aligned}$$

where we redefine $\mathbf{a} \triangleq (\mathbf{a}_0, \cdots, \mathbf{a}_T)$, and

$$I_H(\mathbf{a}, \omega, T) \triangleq \sum_{t=0}^{T} \big(R_t(\mathbf{x}_t, \mathbf{a}_t) + \mathbb{E}[H_{t+1}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - H_t(\mathbf{x}_t) \big).$$

Practical Information Relaxation Bound We further assume for each $t = 0, \cdots, T$, the function H_t is of the additively separable form

$$H_t(\mathbf{x}_t) = \theta_t + \sum_{n=1}^{N} H_t^n(x_t^n),$$

where $\theta_t \in \mathbb{R}$ and $H_t^n : \mathcal{X}^n \to \mathbb{R}$. The space of additively separable functions is denoted by

$$\begin{aligned}\mathcal{D}_T^{\circ} \triangleq \{&H = (H_0, \cdots, H_{T+1}) \in \mathcal{D}_T |\ H_t \text{ is additively separable for } t = 0, \cdots, T, \\ &\text{and } H_{T+1}(\cdot) \equiv 0\}.\end{aligned}$$

Let $\boldsymbol{\mu} \triangleq (\boldsymbol{\mu}_0, \cdots, \boldsymbol{\mu}_{\tau})$ with $\boldsymbol{\mu}_t \in \mathbb{R}_+^L$. We define the operator $\mathcal{L}_T^{\circ}$ on $\mathcal{D}_T^{\circ}$:

$$\begin{aligned}\mathcal{L}_T^{\circ} H(\mathbf{x}_0) \triangleq &\mathbb{E}_0 \Bigg[\min_{\boldsymbol{\mu} \geq 0} \max_{\mathbf{a} \in \mathcal{A}(T)} \Big\{ \sum_{t=0}^{T} \big(R_t(\mathbf{x}_t, \mathbf{a}_t) + \boldsymbol{\mu}_t^{\top} (\mathbf{b}_t - \mathbf{B}_t(\mathbf{x}_t, \mathbf{a}_t)) \\ &+ \mathbb{E}[H_{t+1}(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t] - H_{t+1}(\mathbf{x}_{t+1}) \big) \Big\} \Bigg] \qquad (100) \\ = &H_0(\mathbf{x}_0) + \mathbb{E}_0 \left[\min_{\boldsymbol{\mu} \geq 0} \max_{\mathbf{a} \in \mathcal{A}(T)} \{ I_H(\mathbf{a}, \omega, T; \boldsymbol{\mu}) \} \right],\end{aligned}$$

where

$$I_H(\mathbf{a},\omega,T;\boldsymbol{\mu}) \triangleq \sum_{t=0}^{T} \big(R_t(\mathbf{x}_t,\mathbf{a}_t)+\boldsymbol{\mu}_t^\top(\mathbf{b}_t-\mathbf{B}_t(\mathbf{x}_t,\mathbf{a}_t))+\mathbb{E}[H_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t,\mathbf{a}_t]-H_t(\mathbf{x}_{t+1})\big).$$

We list the analogous results of Theorem 1, Theorem 2, Theorem 4, and Theorem 5 for finite horizon problem in Theorem 6. Proofs are similar and hence are omitted here.

Theorem 17. *(a) (Weak Duality) For any $H \in \mathcal{D}_T$, $V_0(\boldsymbol{x}_0) \le \mathcal{L}_T H(\boldsymbol{x}_0)$ for all $\boldsymbol{x}_0 \in \mathcal{X}$.*

(b) (Tighter Bound) For any $H \in \mathcal{D}_T^*$, where

$$\mathcal{D}_T^* \triangleq \big\{H \in \mathcal{D}_T : R_t(\boldsymbol{x}_t,\boldsymbol{a}_t) + \beta\mathbb{E}[H_{t+1}(\boldsymbol{x}_{t+1})|\boldsymbol{x}_t,\boldsymbol{a}_t] \le H_t(\boldsymbol{x}_t) \\ \textit{for all } \boldsymbol{x}_t \in \mathcal{X} \textit{ and } \boldsymbol{a}_t \in \bar{\mathcal{A}}(\boldsymbol{x}_t),\ t = 0,\cdots,T\big\},$$

then $\max_{\boldsymbol{a}\in\bar{\mathcal{A}}(T)}\{I_H(\boldsymbol{a},\omega,T)\} \le 0$ for every $\omega \in \Omega$; consequently, $V_0(\boldsymbol{x}_0) \le \mathcal{L}_T H(\boldsymbol{x}_0) \le H_0(\boldsymbol{x}_0)$ for all $\boldsymbol{x}_0 \in \mathcal{X}$.

(c) (Strong Duality) $V_0(\boldsymbol{x}_0) = \mathcal{L}_T V(\boldsymbol{x}_0)$ for all $\boldsymbol{x}_0 \in \mathcal{X}$, where $V = (V_0,\cdots,V_T)$.

(d) (Comparing Lagrangian Bound) For all $\boldsymbol{x}_0 \in \mathcal{X}$, $V_0(\boldsymbol{x}_0) \le \mathcal{L}_T J^{\boldsymbol{\lambda}}((\boldsymbol{x}_0) \le J_0^{\boldsymbol{\lambda}}(\boldsymbol{x}_0)$, where $J^{\boldsymbol{\lambda}} = (J_0^{\boldsymbol{\lambda}},\cdots,J_T^{\boldsymbol{\lambda}})$.

(e) (Relaxed Inner Optimization Problem) Suppose that $H \in \mathcal{D}_T^\circ$, i.e., $H_t(\boldsymbol{x}_t) = \theta_t + \sum_{n=1}^N H_t^n(x_t^n)$, $\min_{\boldsymbol{\mu}\ge 0}\max_{\boldsymbol{a}\in\mathcal{A}(T)} I_{J^{\boldsymbol{\lambda}}}(\boldsymbol{a},\omega,T;\boldsymbol{\mu}) \le 0$ for every $\omega \in \Omega$. Consequently, $\mathcal{L}_T J^{\boldsymbol{\lambda}}(\boldsymbol{x}_0) \le \mathcal{L}_T^\circ J^{\boldsymbol{\lambda}}(\boldsymbol{x}_0) \le J_0^{\boldsymbol{\lambda}}(\boldsymbol{x}_0)$ for all $\boldsymbol{x}_0 \in \mathcal{X}$.

(f) (Duality Gap) Suppose that $H \in \mathcal{D}_T^\circ$, i.e., $H_t(\boldsymbol{x}_t) = \theta_t + \sum_{n=1}^N H_t^n(x_t^n)$, and Assumptions 1-3 in Section 3.2 hold for every $\omega \in \Omega$. Then for all $\boldsymbol{x}_0 \in \mathcal{X}$,

$$\mathcal{L}_T^\circ H(\boldsymbol{x}_0) - \mathcal{L}_T H(\boldsymbol{x}_0) \le (1 + L(T+1)) \max_{n=1,\cdots,N} \sum_{t=0}^{T} \Gamma_t^n, \tag{101}$$

where

$$\Gamma_t^n = \sup_{x_t^n \in \mathcal{X}^n, a_t^n \in \mathcal{A}^n(x_0^n)} \left\{ R^n(x_t^n, a_t^n) + \beta \mathbb{E}[H^n(x_{t+1}^n) | x_t^n, a_t^n] - H^n(x_t^n) \right\}$$

$$- \inf_{x_t^n \in \mathcal{X}^n, a_t^n \in \mathcal{A}^n(x_t^n)} \{ R^n(x_t^n, a_t^n) + \beta \mathbb{E}[H^n(x_{t+1}^n) | x_t^n, a_t^n] - H^n(x_t^n) \}.$$

APPENDIX B

CONTROLLED MARKOV DIFFUSION

In this appendix we aim to develop the value function-based penalty as a solution to the dual problem on the right side of (47), which can be viewed as the counterpart of (5) in the setting of controlled Markov diffusions. For this purpose we need to define a solution to the stochastic differential equation(SDE) (41) with an anticipative control $\mathbf{u} \in \mathcal{U}(0)$. Therefore, we introduce the Stratonovich calculus and anticipating stochastic differential equation in Appendix B.1, and present the value function-based optimal penalty in Appendix B.2. We also review the dual representation of the optimal stopping problem under the diffusion process in Appendix B.3.

B.1 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, [67, 68]). In this subsection we present the Stratonovich integral and its associated Ito formula. Then we generalize the controlled diffusion (41) to the Stratonovich sense following [28].

We first assume that $\mathbf{w} = (w_t)_{t\in[0,T]}$ is a one-dimensional Brownian Motion in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by I an arbitrary partition of the interval $[0, T]$ of the form $I = \{0 = t_0 < t_1 < \cdots < t_n = T\}$

Definition 2. *(Definition 3.1.1 in [67]) We say that a measurable process $\boldsymbol{y} = (y_t)_{t\in[0,T]}$ such that $\int_0^T |y_t|dt < \infty$ a.s. is Stratonovich integrable if the family*

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

converges in probability as $\sup_{0\le i\le n-1}(t_{i+1}-t_i)\to 0$*, and in this case the limit will be denoted by* $\int_0^T y_t \circ dw_t$*.*

Remark 7. *We can translate an Ito integral to a Stratonovich integral. If* $\boldsymbol{y} = (y_t)_{t\in[0,T]}$ *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* $\boldsymbol{y}$ *is Stratonovich integrable on any interval* $[0,t]$*, and*

$$\int_0^t y_s \circ dw_s = \int_0^t y_s \, dw_s + \langle y, w\rangle_t = \int_0^t y_s \, dw_s + \frac{1}{2}\int_0^t \zeta_s \, ds, \tag{102}$$

where $\langle y, w\rangle_t$ *denotes the joint quadrature variation of the semimartingale* $\boldsymbol{y}$ *and the Brownian motion* $\boldsymbol{w}$*. Definition 2 and the equality (102) can be naturally extended to the vector case.*

Then we present the Ito formula for Stratonovich integral in Proposition 4 (see, e.g., Section 3.2.3 of [67]).

Proposition 4 (Theorem 3.2.6 in [67])**.** *Let* $\boldsymbol{w} = (w_t^1, \cdots, w_t^m)_{t\in[0,T]}$ *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* $\boldsymbol{y} = (y_t)_{t\in[0,T]}$ *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*

$$F(y_t) = F(y_0) + \int_0^t F_y^\top(y_s)\upsilon_s \, ds + \sum_{i=1}^m \int_0^t \left[F_y(y_s)^\top \zeta_s^i\right] \circ dw_s^i, \;\; 0\le t\le T, \tag{103}$$

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

Proposition 4 basically says that the Stratonovich integral obeys the ordinary chain rule.

Based on the definition of Stratonovich integral and Remark 7, we generalize the SDE (41) to the Stratonovich sense (referred to as S-SDE) assuming that b is bounded and C^1 in (x, u); σ is bounded and C^2 in x. Then (41) is equivalent to

$$x_t = x + \int_0^t \bar{b}(t, x_t, u_t)dt + \sum_{i=1}^{m} \int_0^t \sigma^i(t, x_t) \circ dw_t^i, \quad 0 \leq t \leq T, \tag{104}$$

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 σ^{ki} is the (k, i)-th entry of σ, and $\sigma_x^i \sigma^i$ denotes an $n \times 1$ vector with $\sum_{j=1}^{n} \frac{\partial \sigma^{ki}}{\partial x_j} \sigma^{ji}$ being its k-th entry. Since the stochastic integral in (104) is in the Stratonovich sense, S-SDE (104) adopts its solution in the space of $\mathcal{B}([0, T]) \times \mathcal{F}$-measurable processes, which may not be adapted to the filtration generated by the Brownian motion. Therefore, we are allowed to consider anticipative policies $\mathbf{u} \in \mathcal{U}(0)$ in (104).

Finally, we need to ensure the existence of a solution to S-SDE (104) if the control strategy $\mathbf{u} \in \mathcal{U}(0)$ is anticipative. Following [28],[68], we have a representation of such a solution using the decomposition technique:

$$x_t = \xi_t(\eta_t), \tag{105}$$

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

$$\begin{aligned} d\xi_t &= \sum_{i=1}^{m} \sigma^i(t, \xi_t) \circ dw_t^i, \\ &= \frac{1}{2}\sum_{i=1}^{m} \sigma_x^i \sigma^i(t, \xi_t)dt + \sigma(t, \xi_t)dw_t, \quad \xi_0 = x, \end{aligned} \tag{106}$$

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

$$\frac{d\eta_t}{dt} = \left(\frac{\partial \xi_t}{\partial x}\right)^{-1} (\eta_t)\bar{b}\left(t, \xi_t(\eta_t), u_t\right), \quad \eta_0 = x, \tag{107}$$

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 [28]), the solution (105) is defined almost surely:

observe 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 $\mathbf{u} = (u_t)_{0 \leq t \leq T}$. To check that $x_t = \xi_t(\eta_t)$ satisfies (104), we need to employ a generalized Ito formula of (103) for Stratonovich integral (see Theorem 4.1 in [68]).

B.2 Value Function-Based Penalty

The tools we have introduced in the last subsection, especially the Ito formula for Stratonovich integral, enable us to show the value function-based optimal penalty for the controlled Markov diffusions that developed in Theorem 14.

Proof. [Proof of Theorem 14] Suppose $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$ and let $y_t = V_x^\top(t, x_t)\sigma^i(t, x_t)$ in Remark 7 for $i = 1, \cdots, m$. We can immediately obtain

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

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

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

and therefore, $\mathbb{E}_{0,x}[h_v^*(\mathbf{u}, \mathbf{w})] = 0$ when $\mathbf{u} \in \mathcal{U}_{\mathbb{F}}(0)$. Hence, $h_v^*(\mathbf{u}, \mathbf{w}) \in \mathcal{M}_{\mathbb{F}}(0)$. We then show the strong duality

$$V(0, x) = \mathbb{E}_{0,x}\left[\sup_{\mathbf{u} \in \mathcal{U}(0)} \left\{\Lambda(x_T) + \int_0^T g(t, x_t, u_t) dt - h_v^*(\mathbf{u}, \mathbf{w})\right\}\right]. \qquad (108)$$

According to the weak duality (i.e., Proposition 1),

$$V(0, x) \leq \mathbb{E}_{0,x}\left[\sup_{\mathbf{u} \in \mathcal{U}(0)} \left\{\Lambda(x_T) + \int_0^T g(t, x_t, u_t) dt - h_v^*(\mathbf{u}, \mathbf{w})\right\}\right]. \qquad (109)$$

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

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

where the first equality is obtained by applying Ito formula for Stratonovich integral (i.e., Proposition 4) on $V(t, x)$ with $V(T, x_T) = \Lambda(x_T)$:

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

Since we assume the value function satisfies all the assumptions in Theorem 11(b), there exists an optimal control $\mathbf{u}^* = (u_t^*)_{t\in[0,T]}$ with $u_t^* = u^*(t, x_t)$ and it satisfies

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

then we have

$$\begin{aligned}
&\sup_{\mathbf{u}\in\mathcal{U}(0)} \left\{\Lambda(x_T) + \int_0^T g(t, x_t, u_t)dt - h_v^*(\mathbf{u}, \mathbf{w})\right\} \\
&= \sup_{\mathbf{u}\in\mathcal{U}(0)} \left\{V(0, x) + \int_0^T \Bigg[g(t, x_t, u_t) + A^{u_t} V(t, x_t)\Bigg] dt\right\} \\
&\leq V(0, x) + \int_0^T \sup_{u\in\mathcal{U}} \left\{g(t, x_t, u) + A^u V(t, x_t)\right\} dt \qquad (110) \\
&= V(0, x) + \int_0^T \left[g(t, x_t^*, u_t^*) + A^{u_t^*} V(t, x_t^*)\right] dt \\
&= V(0, x). \qquad (111)
\end{aligned}$$

Taking the conditional expectation on both sides, we have

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

Together with the weak duality (109) , we reach the equality (108).

Due to the fact of the equality (108) (that is in expectation sense) and the pathwise inequality (111), we find that the only inequality (110) should be an equality in almost sure sense. So the equality (51) holds in almost sure sense. To achieve the equality in (110), the optimal control $\mathbf{u}^*$ should be applied, which implies the equality (52).

□

B.3 Optimal Stopping under Diffusion Processes and Its Dual Representation

References [7, 91] use the martingale duality approach to compute upper bounds on the prices of American options, which is a typical optimal stopping problem. By viewing the martingale-based dual approach as a case of the perfect information relaxation, [7, 91] both explored the structure of the "optimal penalty" to the dual of the optimal stopping problem under the *diffusion* process. We briefly review these results that parallel Theorem 14 for controlled diffusions.

Suppose an uncontrolled diffusion $(x_t)_{t\in[0,T]}$ follows the SDE

$$dx_t = b(t, x_t)dt + \sigma(t, x_t)dw_t, \quad 0 \leq t \leq T.$$

We still use $\mathbb{F}$ to denote the natural filtration generated by the Brownian motion $(w_t)_{t\in[0,T]}$. The primal representation of the optimal stopping problem is

$$V(t, x) = \sup_{\tau\in\mathcal{J}_t} \mathbb{E}_{t,x}\left[g(\tau, x_\tau)\right], \tag{112}$$

where $g : \bar{Q} \to \mathbb{R}$ is a reward function, and $\mathcal{J}_t$ is the set of $\mathbb{F}$-stopping times taking value in $[t, T]$. Suppose that $V(t, x)$ is uniformly bounded and is sufficiently smooth to apply Ito formula, we have the following dual representation of the optimal stopping problem.

Proposition 5 (Theorem 1 and Theorem 2 in [91])**.** *Let $\mathcal{H}_{\mathbb{F}}$ represent the space of*

$\mathbb{F}$*-martingales* $\{h_t\}_{t\in[0,T]}$ *with* $h_0 = 0$ *and* $\sup_{t\in[0,T]} \mathbb{E}[|h_t|] < \infty$. *Then*

$$V(0,x) = \min_{h\in\mathcal{H}_{\mathbb{F}}} \mathbb{E}_{0,x}\left[\max_{t\in[0,T]}\{g(t,x_t) - h_t\}\right], \tag{113}$$

In particular, the optimal martingale $\{h_t^*\}_{t\in[0,T]}$ *that achieves the minimum in (113) is of the form*

$$h_t^* = \int_0^t V_x(s,x_s)^\top \sigma(s,x_s) dw_s. \tag{114}$$

Noting that the maximization problem inside the expectation term (113) is the "inner optimization problem" in the dual representation of the optimal stopping problem, since the only control in the primal (112) is to choose "continue" or "stop" the process. The strong duality result (113) holds for general Markov processes, which relies on the the Doob-Meyer decomposition of the process $\{V(t,x_t)\}_{t\in[0,T]}$; however, the form of the optimal martingale (or penalty) h^* in (114) is true only under the diffusion process. The form of h^* exposes its connection with the value function-based penalty presented in Theorem 14.

REFERENCES

[1] Adelman, D. and Mersereau, A. J., "Relaxations of weakly coupled stochastic dynamic programs," *Operations Research*, vol. 56, no. 3, pp. 712–727, 2008.

[2] Alagoz, O., Hsu, H., Schaefer, A. J., and Roberts, M. S., "Markov decision processes: a tool for sequential decision making under uncertainty," *Medical Decision Making*, 2009.

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

[4] Bain, A. and Crisan, D., *Fundamentals of stochastic filtering*, vol. 3. Springer, 2009.

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

[6] Bellman, R., *Dynamic Programming.* Dover Publications, 1957.

[7] Belomestny, D., Bender, C., and Schoenmakers, J., "True upper bounds for Bermudan products via non-nested Monte Carlo," *Mathematical Finance*, vol. 19, pp. 53 – 71, 2009.

[8] Bertsekas, D. P., *Constrained optimization and Lagrange multiplier methods*, vol. 1. Computer Science and Applied Mathematics, Boston: Academic Press, 1982, 1982.

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

[10] Bertsekas, D. P. and Tsitsiklis, J. N., *Neuro-Dynamic Programming.* Optimization and Neural Computation Series, Athena Scientific, 1st ed., 1996.

[11] Bertsimas, D. and Mersereau, A. J., "A learning approach for interactive marketing to a customer segment," *Operations Research*, vol. 55, no. 6, pp. 1120–1135, 2007.

[12] Bertsimas, D. and Niño-Mora, J., "Restless bandits, linear programming relaxations, and a primal-dual index heuristic," *Operations Research*, vol. 48, no. 1, pp. 80–90, 2000.

[13] Birge, J. R. and Dempstert, M., "Stochastic programming approaches to stochastic scheduling," *Journal of Global Optimization*, vol. 9, no. 3-4, pp. 417–451, 1996.

[14] Brandt, M., Goyal, A., Santa-Clara, P., and Stroud, J., "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.

[15] Brooks, A. and Williams, S., "A monte carlo update for parametric POMDPs," *International Symposium of Robotics Research*, Nov. 2007.

[16] Brown, D. B. and Haugh, M. B., "Information relaxation bounds for infinite horizon markov decision processes," 2014.

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

[18] Brown, D. B. and Smith, J. E., "Optimal sequential exploration: Bandits, clairvoyants, and wildcats," *Operations research*, vol. 61, no. 3, pp. 644–665, 2013.

[19] Brown, D. B. and Smith, J. E., "Information relaxations, duality, and convex stochastic dynamic programs," *Operations Research*, vol. 62, no. 6, p-p. 1394–1415, 2014.

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

[21] Caro, F. and Gallien, J., "Dynamic assortment with demand learning for seasonal consumer goods," *Management Science*, vol. 53, no. 2, pp. 276–292, 2007.

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

[23] Chen, N. and Glasserman, P., "Additive and multiplicative duals for american option pricing," *Finance and Stochastics*, vol. 11, pp. 153 – 179, 2007.

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

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

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

[27] Davis, M. and Burstein, G., “Anticipative stochastic control,” in *Proceedings of the 30th IEEE Conference on Decision and Control*, pp. 1830–1835, 1991.

[28] Davis, M. and Burstein, G., “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.

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

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

[31] Dentcheva, D. and Römisch, W., “Duality gaps in nonconvex stochastic optimization,” *Mathematical Programming*, vol. 101, no. 3, pp. 515–535, 2004.

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

[33] Desai, V. V., Farias, V. F., and Moallemi, C. C., “Pathwise optimization for optimal stopping problems,” *Management Science*, vol. 58, no. 12, pp. 2292–2308, 2012.

[34] Devalkar, S., *Essays in optimization of commodity procurement, processing and trade operations.* PhD thesis, The University of Michigan, 2011.

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

[36] Florescu, I. and Viens, F., “Stochastic volatility: Option pricing using a multinomial recombining tree,” *Applied Mathematical Finance*, vol. 15, no. 2, pp. 151 – 181, 2008.

[37] Fox, B. and Glynn, P., “Simulating discounted costs,” *Management Science*, vol. 35, no. 11, pp. 1297–1315, 1989.

[38] Frazier, P., Powell, W., and Dayanik, S., “A knowledge-gradient policy for sequential information collection,” *SIAM Journal on Control and Optimization*, vol. 47, no. 5, pp. 2410–2439, 2008.

[39] Gittins, J. C., “Bandit processes and dynamic allocation indices,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.

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

[41] Gocgun, Y. and Ghate, A., "A Lagrangian approach to dynamic resource allocation," in *Proceedings of the Winter Simulation Conference*, pp. 3330–3340, Winter Simulation Conference, 2010.

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

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

[44] Haugh, M. and Wang, C., "Information relaxations and dynamic zero-sum games," *arXiv preprint arXiv:1405.4347*, 2014.

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

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

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

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

[49] Hawkins, J. T., *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications.* PhD thesis, Massachusetts Institute of Technology, 2003.

[50] He, H. and Pearson, N., "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.

[51] Jacko, P., "Resource capacity allocation to stochastic dynamic competitors: knapsack problem for perishable items and index-knapsack heuristic," *Annals of Operations Research*, pp. 1–25, 2013.

[52] Jones, D. M. and Gittins, J. C., *A dynamic allocation index for the sequential design of experiments.* University of Cambridge, Department of Engineering, 1972.

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

[54] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R., "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1, pp. 99–134, 1998.

[55] Kim, M. and Lim, A., "Robust multi-armed bandit problems," 2013. Working paper.

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

[57] Lai, G., Margot, F., and Secomandi, N., "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.

[58] Lamberton, D. and Lapeyre, B., *Introduction to stochastic calculus applied to finance.* Chapman & Hall/CRC, 2007.

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

[60] Longstaff, F. A. and Schwartz, E. S., "Valuing American options by simulation: A simple least-squares approach," *The Review of Financial Studies*, vol. 14, no. 1, pp. 113 – 147, 2001.

[61] Ludkovski, M., "A simulation approach to optimal stopping under partial information," *Stochastic Processes and Applications*, vol. 119, no. 12, pp. 2071 – 2087, 2009.

[62] Luenberger, D. G., "Investment science," *OUP Catalogue*, 1997.

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

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

[65] Meyn, S. P., *Control techniques for complex networks.* Cambridge University Press, 2008.

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

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

[68] Ocone, D. and Pardoux, E., "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.

[69] Pham, H., Runggaldier, W., and Sellami, A., "Approximation by quantization of the filter process and applications to optimal stopping problems under partial observation," *Monte Carlo Methods and Applicaitons*, vol. 11, no. 1, p-p. 57 – 81, 2005.

[70] Pham, H., *Continuous-time stochastic control and optimization with financial applications*, vol. 1. Springer, 2009.

[71] Porta, J. M., Vlassis, N., and amd P. Poupart, M. T. S., "Point-based value iteration for continuous POMDPs," *Journal of Machine Learning Research*, vol. 7, pp. 2329–2367, 2006.

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

[73] Puterman, M. L., *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* New York: Wiley & Sons, 1994.

[74] Rambharat, B. R. and Brockwell, A. E., "Sequential Monte Carlo pricing of American-style options under stochastic volatility models," *The Annals of Applied Statistics*, vol. 4, No. 1, 222-265, no. 1, pp. 222 – 265, 2010.

[75] Rockafellar, R. T. and Wets, R. J.-B., "Scenarios and policy aggregation in optimization under uncertainty," *Mathematics of operations research*, vol. 16, no. 1, pp. 119–147, 1991.

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

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

[78] Roy, N., *Finding Approximate* POMDP *Solutions through Belief Compression.* PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburg, PA, 2003.

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

[80] Sarimveis, H., Patrinos, P., Tarantilis, C. D., and Kiranoudis, C. T., "Dynamic modeling and control of supply chain systems: A review," *Computers & Operations Research*, vol. 35, no. 11, pp. 3530–3561, 2008.

[81] Schweitzer, P. J. and Seidmann, A., "Generalized polynomial approximations in Markovian decision processes," *Journal of mathematical analysis and applications*, vol. 110, no. 2, pp. 568–582, 1985.

[82] Secomandi, N., "Analysis and enhancement of practice-based policies for the real option management of commodity storage assets," 2014.

[83] Shapiro, A., Dentcheva, D., and Ruszczynski, A., *Lectures on stochastic programming: modeling and theory*, vol. 9. SIAM, 2009.

[84] S.Karlin and Taylor, H., *A First Course in Stochastic Process,2nd edn.* San Diego: Academic Press, 1975.

[85] Sutton, R. S. and Barto, A. G., *Introduction to reinforcement learning.* MIT Press, 1998.

[86] Talluri, K. and van Ryzin, G., "An analysis of bid-price controls for network revenue management," *Management Science*, vol. 44, no. 11-part-1, pp. 1577–1593, 1998.

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

[88] Thrun, S., "Monte Carlo POMDPs," *Advances in Neural Information Processing Systems*, vol. 12, pp. 1064–1070, 2000.

[89] Topaloglu, H., "Using Lagrangian relaxation to compute capacity-dependent bid prices in network revenue management," *Operations Research*, vol. 57, no. 3, pp. 637–649, 2009.

[90] Tsitsiklis, J. and van Roy, B., "Regression methods for pricing complex American-style options," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 694 – 703, 2001.

[91] Wang, Y. and Caflisch, R., "Fast computation of upper bounds for american-style options without nested simulation," 2010.

[92] Whittle, P., "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, pp. 287–298, 1988.

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

[94] Ye, F. and Zhou, E., "Weakly coupled dynamic program: Information and lagrangian relaxations," *arXiv preprint arXiv:1405.3363*, 2013.

[95] Ye, F. and Zhou, E., "Information relaxation and dual formulation of controlled Markov diffusions," *IEEE Transactions on Automatic Control*, 2015.

[96] Ye, F. and Zhou, E., "Pricing american options under partial observation of stochastic volatility," in *Proceedings of the Winter Simulation Conference*, pp. 3760–3771, Winter Simulation Conference, 2011.

[97] Ye, F. and Zhou, E., "Optimal stopping of partially observable markov processes: A filtering-based duality approach," *IEEE Transactions on Automatic Control*, vol. 58, no. 10, pp. 2698–2704, 2013.

[98] Ye, F. and Zhou, E., "Dual formulation of controlled markov diffusions and its application," in *World Congress*, vol. 19, pp. 7811–7818, 2014.

[99] Zhou, E., Fu, M. C., and Marcus, S. I., "Solving continuous-state POMDPs via density projection," *IEEE Transactions on Automatic Control*, vol. 55, no. 5, pp. 1101 – 1116, 2010.

[100] Zhou, E., "Optimal stopping under partial observation: Near-value iteration," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 500–506, 2013.

[101] Zhu, H., Ye, F., and Zhou, E., "Fast estimation of true bounds on bermudan option prices under jump-diffusion processes," *Quantative Finance*, 2015.