Parameterized Penalties in the Dual Representation of Markov Decision Processes

Fan Ye and Enlu Zhou

Abstract— **Duality in Markov decision processes (MDPs) has been studied recently by several researchers with the goal to derive dual bounds on the value function. In this paper we propose the idea of using parameterized penalty functions in the dual representation of MDPs, which allows us to integrate different types of penalty functions and guarantees a tighter dual bound with more penalties used. To complement and diversify the existing linear penalties developed in the literature, we also introduce a new class of nonlinear penalties that can be used for a broad class of problems and are also easy to implement in practice. Based on this new class of penalties, our framework of parameterized penalties is a promising method to produce tighter dual bounds than existing duality-based methods. We compare the performance of the dual bounds induced by different penalties on a numerical example, demonstrating the effectiveness of our method.**

I. Introduction

Markov Decision Processes (MDPs) can be used to model dynamic decision making problems under uncertainty, and hence have wide applications in diverse fields such as engineering, operations research and economics. However, the standard approach of solving for optimal policies via dynamic programming (DP) suffers from the so-called "curse of dimensionality" - the size of the state space increases exponentially with the dimension of the state, which limits the use of the exact DP to low-dimensional problems. In recent years various methods using Monte Carlo simulation have been proposed in order to combat this curse of dimensionality [1], [2], [3]. Note that it is not hard to simulate a complex stochastic system given a fixed feasible policy, which can be used to provide a lower bound (or upper bound) on the expected reward (or cost) induced by the optimal policy. However, the accuracy of the sub-optimal policies is generally unknown.

In observation of the lack of performance guarantee on sub-optimal policies, a dual representation of MDPs based on information relaxation was recently independently developed by [4] and [5] to provide dual bounds on the value functions. The main idea of this duality approach is to allow the decision maker to foresee the future uncertainty but is penalized for getting access to the information in advance. In addition, this duality approach reduces to a pathwise optimization problem and therefore is well-suited to Monte Carlo simulation, making it useful to evaluate the quality of sub-optimal policies in complex dynamic systems. However, to achieve strong duality, the optimal penalty in the dual representation involves the value function, and hence is usually intractable in practical problems. Therefore, approximation schemes for the optimal penalty were studied by recent research including [6],[7], and [8]. In particular, [6] introduced the gradient-based penalty in the context of dynamic portfolio optimization with transaction costs. [7] considered the dual formulation of controlled Markov diffusions and its application, and [8] explored the connection between the approximate linear programming and the information relaxation duality approaches. Furthermore, [9] revealed that in linear-quadratic problems the value-function-based penalty and gradient-based penalty are both optimal, but in different senses; in addition, [9] compared two penalties with the Lagrangian multiplier terms that appeared in the earlier work [10].

We note that the construction of a good penalty is usually difficult due to the tradeoff between its effectiveness and the computational cost. As a consequence, all penalties developed so far for continuous-state MDPs are linear functions for the sake of maintaining the solvability of the intermediate pathwise optimization problem. To capture the nonlinear feature of the optimal penalty in a general case, we introduce a class of simple nonlinear penalty functions that can be applied to general MDPs. This class of nonlinear penalties together with other classes of linear penalties developed in [6] and [8], all lead to dual bounds on the value function. Then some natural questions would be (i) whether we can utilize all the available penalties to derive an even tighter dual bound; and (ii) how to combine and choose different penalties. We intend to provide answers to these questions as well in this paper. In summary, our contributions are:

- We develop a framework of parameterized penalties in the dual representation of MDPs, where the optimal choice of the parameters can be determined by a convex (stochastic) optimization problem. The theoretic result guarantees a tighter dual bound if more penalties are used.
- We introduce a new class of nonlinear penalties that can be applied to general MDPs and are also very easy to implement in practice.
- We carry out some numerical experiments that provide insights into the design and choice of penalties. The numerical results show a considerable improvement on

This work was supported by National Science Foundation under Grants ECCS-0901543 and CMMI-1130273, and by the Air Force Office of Scientific Research under YIP Grant FA-9550-12-1-0250.

F. Ye is with the Department of Industrial & Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA fanye2 at illinois.edu

E. Zhou is with the Department of Industrial & Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA enluzhou at illinois.edu

the tightness of the dual bound using our parameterized penalties.

The rest of the paper is organized as follows. In Section II, we review the dual formulation of MDPs. In Section III, we develop the idea of parameterized penalties. In Section IV, we introduce different types of penalties. In Section V, we present some numerical results, and finally conclude in Section VI.

II. Dual Formulation of Markov Decision Processes

Consider a finite-horizon Markov decision process on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Time is indexed by $\mathscr{T} = \{0, 1, \cdots, T\}$. Suppose $\mathscr{X}_t$ is the state space and $\mathscr{A}_t$ is the control space at time t. The state $\{x_t\}$ follows the equation

$$x_{t+1} = f(x_t, a_t, w_{t+1}),\ t = 0, 1, \cdots, T-1, \tag{1}$$

where $a_t \in \mathscr{A}_t$ is the control whose value is decided at time t, and $\{w_t\}$ is a sequence of independent random variables taking values in the set $\mathscr{W}_t$ with known distribution. The evolution of the information is described by the filtration $\mathbb{F} = \{\mathscr{F}_0, \cdots, \mathscr{F}_T\}$ with $\mathscr{F} = \mathscr{F}_T$.

Denote by $\mathbb{A}$ the set of feasible strategies $\mathbf{a} \triangleq (a_1, \cdots, a_{T-1})$, i.e., $a_t \in \mathscr{A}_t$ for each $t \in \mathscr{T}$. Let $\mathbb{A}_{\mathbb{F}}$ be the set of feasible strategies that are adapted to the filtration $\mathbb{F}$, i.e., a_t is $\mathscr{F}_t$-adapted for every t. We also call any $\mathbf{a} \in \mathbb{A}_{\mathbb{F}}$ a *nonanticipative* policy. Given an $x_0 \in \mathscr{X}_0$, the objective is to maximize the expected reward by selecting a nonanticipative policy $\mathbf{a} \in \mathbb{A}_{\mathbb{F}}$:

$$\begin{aligned} V_0^*(x_0) &= \sup_{\mathbf{a}\in\mathbb{A}_{\mathbb{F}}} V_0(x_0, \mathbf{a}), \\ \text{where}\ \ V_0(x_0, \mathbf{a}) &\triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} g_t(x_t, a_t) + g_T(x_T)|x_0\right]. \end{aligned} \tag{2}$$

The expectation in (2) is taken with respect to the random sequence $\mathbf{w} \triangleq (w_1, \cdots, w_T)$. To avoid technical issues we assume that $V_0(x_0, \mathbf{a})$ has a uniform bound for all $\mathbf{a} \in \mathbb{A}_{\mathbb{F}}$.

The value function V_0^* is a solution to the following dynamic programming recursion:

$$\begin{aligned} &V_T^*(x_T) \triangleq g_T(x_T); \\ &V_t^*(x_t) \triangleq \sup_{a_t\in\mathscr{A}_t} \{g_t(x_t, a_t) + \mathbb{E}[V_{t+1}^*(x_{t+1})|x_t, a_t]\} \\ &\text{for}\ \ t = T-1, \cdots, 0. \end{aligned} \tag{3}$$

Hence, the optimal control $\mathbf{a}^* = (a_0^*, \cdots, a_{T-1}^*)$ satisfies

$$\begin{aligned} a_t^*(x_t) = \arg \sup_{a_t\in\mathscr{A}_t} \{g_t(x_t, a_t) + \mathbb{E}[V_{t+1}^*(f(x_t, a_t, w_{t+1}))|x_t, a_t]\}, \\ t = 0, \cdots, T-1. \end{aligned}$$

The exact computation of DP recursion in (3) is intractable or time-consuming for complex dynamic systems in practice, especially when the state space $\mathscr{X}_t$ is continuous or high-dimensional, or the optimization with respect to a_t is difficult to handle. In these situations we are seeking some heuristic policy, and by simulating the dynamics under this policy we obtain a lower bound on the value function V_0^*.

We now introduce a dual formulation of the value functions, which is developed by [4] and [5] to obtain an upper bound on V_0^*. Throughout this paper we only consider the *perfect information relaxation*, i.e., we are allowed to have full knowledge of all future randomness, since this relaxation is usually more applicable in practice.

Define $\mathbf{x} \triangleq (x_1, \cdots, x_T)$. Let $\mathscr{M}$ denote the space of functions $M(\mathbf{x}, \mathbf{a}, \mathbf{w})$ satisfying

$$\mathbb{E}[M(\mathbf{x}, \mathbf{a}, \mathbf{w})|x_0] = 0 \ \text{ for all }\ \mathbf{a} \in \mathbb{A}_{\mathbb{F}}.$$

Here $M(\mathbf{x}, \mathbf{a}, \mathbf{w})$ can be written more explictly as $M(\mathbf{x}(x_0, \mathbf{a}, \mathbf{w}), \mathbf{a}, \mathbf{w})$ according to the state equation (1).

Denote by $\mathscr{D}_0$ the space of real-valued functions on $\mathscr{X}_0$. Then we define an operator $\mathscr{L}: \mathscr{M} \to \mathscr{D}_0$ by

$$(\mathscr{L}M)(x_0) = \mathbb{E}\left[\sup_{\mathbf{a}\in\mathbb{A}}\{\sum_{t=0}^{T-1} g_t(x_t, a_t) + g_T(x_T) - M(\mathbf{x}, \mathbf{a}, \mathbf{w})\}|x_0\right]. \tag{4}$$

Note that the supremum in (4) is over the set of feasible strategies $\mathbb{A}$ not the set of nonanticipative policies $\mathbb{A}_{\mathbb{F}}$. The optimization problem inside the expectation in (4) is usually referred to as the *inner optimization problem*. In particular, the right hand side of (4) is well-suited to Monte Carlo simulation: we can simulate a realization of $\mathbf{w} = \{w_1, \cdots, w_T\}$ and solve the inner optimization problem:

$$\begin{aligned} I(x_0, M, \mathbf{w}) \triangleq \max_{\mathbf{a}} \quad & \sum_{t=0}^{T-1} g_t(x_t, a_t) + g_T(x_T) - M(\mathbf{x}, \mathbf{a}, \mathbf{w}) && (5) \\ \text{s.t.}\ & x_t = f(x_{t-1}, a_{t-1}, w_t),\ t = 1, \cdots, T, && (6) \\ & a_t \in \mathscr{A}_t,\ t = 0, \cdots, T-1, && (7) \end{aligned}$$

which is in fact a *deterministic* dynamic program. The optimal value $I(x_0, M, \mathbf{w})$ is an unbiased estimator of $(\mathscr{L}M)(x_0)$.

The function $M \in \mathscr{M}$ can be constructed using a martingale difference operator that is defined as follows. Suppose $h = (h_1, \cdots, h_T)$ is a vector of functions, where each h_t is a real-valued function defined on $\mathscr{X}_t$. Define a martingale difference operator Δ that maps each h_{t+1} to a real-valued function on $\mathscr{X}_t \times \mathscr{A}_t \times \mathscr{W}_{t+1}$ for $t = 0, \cdots, T-1$:

$$\begin{aligned} &(\Delta h_{t+1})(x_t, a_t, w_{t+1}) \\ \triangleq &h_{t+1}(x_{t+1}) - \mathbb{E}[h_{t+1}(x_{t+1})|x_t, a_t] \\ = &h_{t+1}(f(x_t, a_t, w_{t+1})) - \mathbb{E}[h_{t+1}(f(x_t, a_t, w_{t+1}))|x_t, a_t]. \end{aligned} \tag{8}$$

In particular, $\mathbb{E}[\Delta h_{t+1}(x_t, a_t, w_{t+1})|x_0] = 0$ for $t = 0, \cdots, T-1$. We also define

$$(\Delta h)(\mathbf{x}, \mathbf{a}, \mathbf{w}) \triangleq \sum_{t=0}^{T-1} \Delta h_{t+1}(x_t, a_t, w_{t+1}).$$

Then $\Delta h \in \mathscr{M}$.

Theorem 1(a) below suggests that $\mathscr{L}M$ can be used to derive an upper bound on the value function V_0^* given any $M \in \mathscr{M}$. Hence, $I(x_0, M, \mathbf{w})$ is a high-biased estimator of $V_0^*(x_0)$ for all $x_0 \in \mathscr{X}_0$. Theorem 1(b) is a strong duality result which states that the duality gap vanishes as the dual problem is solved by taking $M(\mathbf{x}, \mathbf{a}, \mathbf{w}) = \sum_{t=0}^{T-1} \Delta V_{t+1}^*(x_t, a_t, w_{t+1})$.

Theorem 1 (Theorem 2.1 in [5])

(a)For all $M \in \mathscr{M}$ and all $x_0 \in \mathscr{X}_0$, $V_0^*(x_0) \le (\mathscr{L}M)(x_0)$.

(b)For all $x_0 \in \mathscr{X}_0$, $V_0^*(x_0) = (\mathscr{L}M^*)(x_0)$, where

$$M^*(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{w}) = \sum_{t=0}^{T-1} \Delta V_{t+1}^*(x_t, a_t, w_{t+1}) \in \mathscr{M}.$$

Proof: (a) Note that for any $M \in \mathscr{M}$, $x_0 \in \mathscr{X}_0$ and $\mathbf{a} \in \mathbb{A}_{\mathbb{F}}$,

$$\begin{aligned} V_0(x_0, \mathbf{a}) &= \mathbb{E}[\sum_{t=0}^{T-1} g_t(x_t, a_t) + g_T(x_T) - M(\mathbf{x}, \mathbf{a}, \mathbf{w}))|x_0] \\ &\le \mathbb{E}[\sup_{\mathbf{a} \in \mathbb{A}} \{\sum_{t=0}^{T-1} g_t(x_t, a_t) + g_T(x_T) - M(\mathbf{x}, \mathbf{a}, \mathbf{w})\}|x_0] \\ &= (\mathscr{L}M)(x_0). \end{aligned}$$

Hence, for all $M \in \mathscr{M}$ and $x_0 \in \mathscr{X}_0$,

$$V_0^*(x_0) = \sup_{\mathbf{a} \in \mathbb{A}_{\mathbb{F}}} V_0(x_0, \mathbf{a}) \le (\mathscr{L}M)(x_0).$$

(b) The result can be established by showing $V_0^*(x_0) \ge (\mathscr{L}M^*)(x_0)$. By the definition of M^* and $(\mathscr{L}M^*)(x_0)$,

$$\begin{aligned} &(\mathscr{L}M^*)(x_0) \\ =&\mathbb{E}[\sup_{\mathbf{a} \in \mathbb{A}} \{\sum_{t=0}^{T-1} (g_t(x_t, a_t) - \Delta V_{t+1}^*(x_t, a_t, w_{t+1})) + g_T(x_T)\}|x_0] \\ =&\mathbb{E}[\sup_{\mathbf{a} \in \mathbb{A}} \{V_0^*(x_0) + \sum_{t=0}^{T-1} (g_t(x_t, a_t) + \mathbb{E}[V_{t+1}^*(x_{t+1})|x_t, a_t] - V_t^*(x_t)) \\ &- V_T^*(x_T) + g_T(x_T)\}|x_0] \\ \le& V_0^*(x_0), \end{aligned}$$

where the last equality follows from (3), i.e.,

$$V_t^*(x_t) \ge g_t(x_t, a_t) + E[V_{t+1}^*(x_{t+1})|x_t, a_t] \quad \text{for all} \quad a_t \in \mathscr{A}_t,$$

and $V_T^*(x_T) = g_T(x_T)$. ■

Note that the relaxation of the nonanticipative control in (4) is compensated by using M^* according to (b). This is the reason why we call $M \in \mathscr{M}$ the *penalty function*. On the other hand, the zero penalty $M = 0$ is trivially in the space $\mathscr{M}$, which is equivalent to finding the optimal control with perfect knowledge of the future information. It can be expected that the dual bound induced by the zero penalty is typically weak. After all, we are looking for penalties that can offset the benefit brought by the foreknowledge of the future uncertainty.

III. Parameterized Penalties

In this section we propose the idea of parameterized penalties in the context of dual representation of MDPs. We note that Theorem 1 is usually of no practical use since the exact V_t^* in M^* is not known. A natural idea is to derive penalty functions by approximating the optimal value function or the optimal policy. Methods based on these ideas have been successfully implemented in the American option pricing problem in [11], [12], and [13], and in the examples of [5]. However, these approaches can not be extended immediately to general MDPs in parallel with the American options pricing problem. The first difficulty, as pointed out in [9], is that the approximate value functions $\{\hat{V}_t\}$ are not always available even if a good suboptimal policy is available. What's worse, $\mathbb{E}[\hat{V}_{t+1}(x_{t+1})|x_t, a_t]$ usually cannot be written as an analytic function of x_t and a_t, which makes the inner optimization problem intractable.

Hence, it may be more reasonable to develop a pure penalty approach that does not depend on the form of the approximate value functions. We may choose the penalty functions of any structure that can simplify the inner optimization problem. Later in Section IV-B we will develop a new and simple class of penalty functions that can be viewed as an alternative to the existing linear penalties. Furthermore, we hope that some linear combinations over available penalties would lead to a tighter dual bound. A natural and consequent question would be how to choose such good linear combinations. To answer these questions, we will formalize our idea of parameterized penalties in the rest of this section. It generalizes the idea in [8], where the parametrization is imposed on the approximate value functions together with (8) to derive penalties. Since it is not the only way to obtain effective penalties as can be seen in [6],[7], and this paper, our framework seems more general to incorporate different types of penalties.

Suppose that we are given a set of basis penalties $\Phi \triangleq \{\phi_1, \cdots, \phi_N\} \subset \mathscr{M}$. Let Θ_i be a convex set in $\mathbb{R}$ containing 0. Define an N-dimensional vector $\mathbf{r} = (r_1, \cdots, r_N) \in \Theta \subset \mathbb{R}^N$, where $\Theta \triangleq \prod_{i=1}^N \Theta_i$. Then we define a product operation:

$$\mathbf{\Phi r} \triangleq \sum_{i=1}^N \phi_i r_i \in \mathscr{M}.$$

Let $\mathscr{P}$ denote the subset of $\mathscr{M}$ spanned by Φ:

$$\mathscr{P} \triangleq \{\mathbf{\Phi r};\ \mathbf{r} \in \Theta\} \subset \mathscr{M}. \tag{9}$$

With the intention to find the tightest upper bound induced by this family of penalties $\mathscr{P}$, we consider the following optimization of parameterized penalties (OPP) problem:

$$\inf_{M \in \mathscr{P}} (\mathscr{L}M)(x) = \inf_{\mathbf{r} \in \Theta} (\mathscr{L}\mathbf{\Phi r})(x), \tag{10}$$

According to Theorem 1, it is obvious to see that the optimal objective function of (10) is $V_0^*(x)$ if $M^* = \sum_{t=0}^{T-1} \Delta V_{t+1}^*$ lies in $\mathscr{P}$. Moreover, the following theorem shows that the OPP problem is a convex optimization problem:

Theorem 2 *Let $\Phi \triangleq \{\phi_1, \cdots, \phi_N\} \subset \mathscr{M}$. Suppose $\Theta \subset \mathbb{R}^N$ is a convex set. Then for any $x \in \mathscr{X}_0$,*

$$\min_{\boldsymbol{r} \in \Theta} \quad (\mathscr{L}\mathbf{\Phi}\boldsymbol{r})(x) \tag{11}$$

is a convex optimization problem.

Proof: First note that $(\mathscr{L}M)(x) = \mathbb{E}[I(x, M, \mathbf{w})]$ for any $x \in \mathscr{X}_0$. The definition of $I(x, M, \mathbf{w})$ in (5) implies that for any $M_1, M_2 \in \mathscr{M}$ and $0 \le \alpha \le 1$,

$$I(x, (1-\alpha)M_1 + \alpha M_2, \mathbf{w}) \le (1-\alpha)I(x, M_1, \mathbf{w}) + \alpha I(x, M_2, \mathbf{w}). \tag{12}$$

Taking expectation on both sides yields

$$(\mathscr{L}((1-\alpha)M_1+\alpha M_2))(x) \leq (1-\alpha)(\mathscr{L}M_1)(x)+\alpha(\mathscr{L}M_2)(x).$$

Since Θ is convex, we can show Theorem 2 simply by replacing M_1 and M_2 with $\Phi\mathbf{r}_1$ and $\Phi\mathbf{r}_2$. ■

Since OPP is a convex optimization problem, we may develop some local-minima-free algorithm to find the optimal solution $\mathbf{r}^*$. The OPP problem in our context is also referred to as the *outer optimization problem*. The following theorem shows that a larger set of basis functions always leads to a tighter dual bound on V_0^*.

Theorem 3 *Suppose Φ_1 and Φ_2 are two finite subsets of $\mathscr{M}$, where $\Phi_1 = \{\phi_1,\cdots,\phi_{N_1}\}$ and $\Phi_2 = \Phi_1 \cup \{\phi_{N_1+1},\cdots,\phi_{N_2}\}$ with $0 < N_1 < N_2$. Let $\Theta^k = \prod_{i=1}^{N_k}\Theta_i$ for $k = 1,2$, where each Θ_i is a convex set in $\mathbb{R}$ containing 0. Define $\mathscr{P}_k = \{\Phi_k r_k; r_k \in \Theta^k\}$ for $k = 1,2$. Then for all $x \in \mathscr{X}_0$,*

$$\inf_{M\in\mathscr{P}_2}(\mathscr{L}M)(x) \leq \inf_{M\in\mathscr{P}_1}(\mathscr{L}M)(x).$$

Proof: Fixed an $x \in \mathscr{X}_0$ and define

$$J_k \triangleq \inf_{M\in\mathscr{P}_k}(\mathscr{L}M)(x) = \inf_{\mathbf{r}_k\in\Theta^k}(\mathscr{L}\Phi_k\mathbf{r}_k)(x), \quad k = 1,2.$$

For any $\varepsilon > 0$, there exists $\mathbf{r}_1 \in \Theta^1$ such that $(\mathscr{L}\Phi_1\mathbf{r}_1)(x) < J_1+\varepsilon$. With $\mathbf{r}_2 = (\mathbf{r}_1, \mathbf{0}^{N_2-N_1})$, where $\mathbf{0}^d$ is a d-dimensional zero vector, it is straightforward to obtain

$$(\mathscr{L}\Phi_2\mathbf{r}_2)(x) = (\mathscr{L}\Phi_1\mathbf{r}_1)(x) < J_1+\varepsilon.$$

Note that $\Theta^1 \times \{\mathbf{0}^{N_2-N_1}\} \subset \Theta^2$ implies $\mathbf{r}_2 \in \Theta^2$. So $J_2 \leq (\mathscr{L}\Phi_2\mathbf{r}_2)(x) < J_1+\varepsilon$ for any $\varepsilon > 0$. Hence, $J_2 \leq J_1$. ■

In order to seek a numerical solution to the outer optimization problem (i.e., OPP) that is indeed a stochastic optimization problem with convex structure, we can use either the stochastic approximation (SA) method (see [14] for reference) or the sample average approximation (SAA) method (see [15] for reference). In this paper we apply the SAA method to approximate the original problem (11), to be precise, we consider the optimal solution $\tilde{\mathbf{r}}^*$ of the following optimization problem

$$\min_{\mathbf{r}\in\Theta} \quad Q(\mathbf{r}) \triangleq \frac{1}{L}\sum_{l=1}^{L} I(x_0,\Phi\mathbf{r},\mathbf{w}^l) \tag{13}$$

as the approximation to $\mathbf{r}^*$, where $\{\mathbf{w}^l, l = 1,\cdots,L\}$ is a set of i.i.d. samples of $\mathbf{w}$. Once these samples are fixed, (13) becomes a deterministic optimization problem, and thus can be solved using an appropriate optimization algorithm. We note that $I(x_0,\Phi\mathbf{r},\mathbf{w})$ is convex in $\mathbf{r}$ due to (12), so $Q(\mathbf{r})$ is also convex in $\mathbf{r}$. However, the differentiability of $I(x_0,\Phi\mathbf{r},\mathbf{w})$ and $Q(\mathbf{r})$ in $\mathbf{r}$ is generally unknown, which means some approximate subgradient method should be considered to solve (13). In this paper we employ the simultaneous perturbation (SP) method (see [16] for reference) for subgradient estimation and to determine the optimal solution $\tilde{\mathbf{r}}^*$ to (13), since the underlying subgradient approximation only requires two measurements of $Q(\mathbf{r})$ regardless of the dimension of $\mathbf{r}$. Starting with an initial guess $\mathbf{r} = \mathbf{r}_0$, SP method updates the value of $\mathbf{r}$ iteratively through the following formula with fixed basis penalties Φ and $x_0 \in \mathscr{X}_0$:

$$\mathbf{r}_{n+1} := \Pi_\Theta(\mathbf{r}_n - a_n\widehat{\nabla}_n),$$

where Π_Θ denotes a projection back into the feasible region Θ if the updated $\mathbf{r}$ lies outside of Θ, $\{a_n\}$ is an appropriate sequence satisfying

$$a_n > 0, \quad \sum_n a_n = \infty, \text{ and } \sum_n a_n^2 < \infty,$$

and $\widehat{\nabla}_n = (\widehat{\nabla}_{n,1},\cdots,\widehat{\nabla}_{n,N})$ denotes an estimate of the (sub)gradient of $Q(\mathbf{r})$ given by

$$\widehat{\nabla}_{n,i} = \frac{Q(\mathbf{r}_n+c_n\delta_n)-Q(\mathbf{r}_n-c_n\delta_n)}{2c_n\delta_{n,i}} \tag{14}$$

where $\delta_n = (\delta_{n,1},\cdots,\delta_{n,N})$ is an N-dimensional random perturbation vector, which is independently generated from a zero-mean probability distribution.

Under the assumption that the inner optimization problem $I(x_0,\Phi\mathbf{r},\mathbf{w})$ is solvable for each $\mathbf{w} = (w_1,\cdots,w_T)$ and each $\mathbf{r} \in \Theta$ with the basis penalties Φ we have chosen, the following algorithm provides an approximate solution to (11).

```
Algorithm 1 Numerical solution of $r^*$ by SAA+SP
Input: MaxItr $\in \mathbb{N}, \varepsilon, a, \beta_1, A, c, \beta_2, C \in \mathbb{R}^+$, $\mathbf{r}_0 \in \Theta$.
Generate a set of random sequences $\{\mathbf{w}^l, l = 1,\cdots,L\}$ that are used to determine the value of $Q$. Set $n = 0$.
While $n \leq$ MaxItr
1.Set $a_n = a/(n+A)^{\beta_1}$, $c_n = c/(n+C)^{\beta_2}$.
2.Generate an $N$-dimensional random vector $\delta_n$ from Bernoulli $\pm 1$ distribution with probability 1/2 for each $\pm 1$.
3.Set $\mathbf{r}_n^+ = \mathbf{r}_n + c_n\delta_n$ and $\mathbf{r}_n^- = \mathbf{r}_n - c_n\delta_n$.
4.Set $\widehat{\nabla}_n = (\widehat{\nabla}_{n,1},\cdots,\widehat{\nabla}_{n,N})$, where
    $\widehat{\nabla}_{n,i} = (Q(\mathbf{r}_n^+) - Q(\mathbf{r}_n^-))/(2c_n\delta_{n,i}),\ i = 1,\cdots,N.$
5.Set $\mathbf{r}_{n+1} = \mathbf{r}_n - a_n\widehat{\nabla}_n$.
6.If some stopping criterion is satisfied, break.
7.Set $n = n+1$.
End
```

Remark 1 *If $\mathbf{r}_n^+$,$\mathbf{r}_n^-$ or $\mathbf{r}_{n+1}$ in Algorithm 1 takes value outside Θ, it should be redefined as its closest point in Θ.*

We terminate Algorithm 1 if there is little change in several successive iterates or the maximum number of the iterations is reached. It is commonly known that the choices of a_n and c_n are critical to the performance of Algorithm 1. The details and the convergence result of SAA and SP can be found in [15], [16], and [14] respectively.

IV. Penalty Functions

The general form of penalty functions in the inner optimization problem (5) can make it extremely difficult to solve; hence, the form of the penalty function is usually restricted to a narrow class. In particular, [8] approximate the value function using quadratic functions in a linear system with convex constraints on the controls, which induces penalties that are linear functions of **a** and **x**. However, this approach may not provide a good approximation in general considering that the ideal penalty M_t^* can be highly nonlinear in **a** and **x**. The aforementioned drawbacks motivate us to develop tractable penalty functions besides linear ones. All these classes of penalty functions can be incorporated in the framework of parameterized penalties developed in Section III.

A. Canonical Example

As an illustration, we begin with an example (see [17]) to show the difficulty of designing penalty functions. Consider the following state equation:

$$x_{t+1} = 2x_t - a_t + w_{t+1},$$

where the state x_t and the control a_t take value in $\mathbb{R}$, and the random variables $\{w_t\}$ are identically and independently distributed. The natural filtration is $\mathscr{F}_t$, where $\mathscr{F}_t = \sigma\{w_1, \cdots, w_t\}$ for $t = 1, \cdots, T$. The objective is to maximize

$$V_0(x_0, \mathbf{a}) = \mathbb{E}\left[\sum_{t=0}^{T-1} -\exp(-\gamma a_t) - \alpha \exp(-\gamma x_T)|x_0\right] \quad (15)$$

over $\mathbf{a} \in \mathbb{A}_{\mathbb{F}}$ for a given x_0, where α and γ are given positive numbers. Under the assumption that $\mu = \mathbb{E}[\exp(-\gamma w_t)] < \infty$, we can show that

$$V_t^*(x) = -\alpha_t \exp(-\gamma x), \quad (16)$$

where α_t satisfies the backward recursion:

$$\alpha_T = \alpha;$$
$$\alpha_t = 2\sqrt{\alpha_{t+1}\mu}, \quad \text{for} \quad t = T-1, \cdots, 0.$$

The optimal control is $a_t^*(x_t) = x_t - \frac{\ln(\alpha_{t+1}\mu)}{2\gamma}$ for $t = 0, \cdots, T-1$.

Note that the optimal penalty is $M^* = \sum_{t=0}^{T-1} \Delta V_{t+1}^*$, where

$$\begin{aligned}&\Delta V_t^*(x_t, a_t, w_{t+1})\\ =&V_{t+1}^*(x_{t+1}) - \mathbb{E}[V_{t+1}^*(x_{t+1})|x_t, a_t]\\ =&-\alpha_{t+1}\exp(-\gamma(2x_t - a_t)) \cdot (\exp(-\gamma w_{t+1}) - \mathbb{E}[\exp(-\gamma w_{t+1})]).\end{aligned}$$

By plugging the optimal penalty M^* and a fixed sequence $\mathbf{w} = (w_1, \cdots, w_T)$ into (4), we have the following inner optimization problem:

$$\begin{aligned}\max_{\mathbf{a}} \quad & -\alpha \exp(-\gamma x_T) - \sum_{t=0}^{T-1} \exp(-\gamma a_t)\\ & - \sum_{t=0}^{T-1} \Delta V_t^*(x_t, a_t, w_{t+1}) \qquad (17)\\ \text{s.t. } & x_t = 2x_{t-1} - a_{t-1} + w_t, \quad t = 1, \cdots, T.\end{aligned}$$

Though the function inside the expectation in (15) is jointly concave in **x** and **a**, the objective function in the inner optimization problem (17) is not always concave. In particular, in the case that

$$\exp(-\gamma w_{t+1}) - \mathbb{E}[\exp(-\gamma w_{t+1})] > 0,$$

the objective function is no longer concave. Therefore, the global optimal solution may be extremely hard to find.

B. Coefficient-based Penalty

The difficulty in solving the above inner optimization problem can be circumvented by ensuring the concavity of its objective function. To this end, we develop a class of coefficient-based penalties that are general for a broad class of problems and are also very easy to implement. We come back to the general case (2) and in the rest of paper adopt the following assumptions, which were also implicitly or explicitly used in [6] and [8].

Assumption 1
(i) $g_t(x_t, a_t)$ is jointly concave in x_t and a_t, and $g_T(x_T)$ is concave in x_T.
(ii) The set of $(\boldsymbol{x}, \boldsymbol{a})$ constrained by (6)-(7) is convex.

Assumption 1 (ii) is satisfied, for example, by assuming that the dynamics (1) satisfies the linear form

$$x_t = \sum_{j=1}^{t-1} f_{t,j}^1(x_0, \mathbf{w}_t)x_j + \sum_{j=1}^{t-1} f_{t,j}^2(x_0, \mathbf{w}_t)a_j + f_t^3(x_0, \mathbf{w}_t)$$

for some functions $f_{t,j}^1$, $f_{t,j}^2$, and f_t^3, $j = 1, \cdots, t-1$, where $\mathbf{w}_t \triangleq (w_0, \cdots, w_t)$, and each $\mathscr{A}_t$ is convex.

Generally we can choose a set of basis penalties in the form $\Phi = \{\phi_1, \cdots, \phi_T\}$, where

$$\begin{aligned}&\phi_{t+1}(\mathbf{x}, \mathbf{a}, \mathbf{w})\\ =&l_t(x_t, a_t) \cdot (k_{t+1}(x_t, a_t, w_{t+1}) - \mathbb{E}[k_{t+1}(x_t, a_t, w_{t+1})|x_t, a_t])\end{aligned} \quad (18)$$

with arbitrary functions l_t and k_{t+1}, for $t = 0, \cdots, T-1$. In the ideal case that $l_t = 1$ and $k_{t+1} = V_{t+1}^* \circ f$ we reach the optimal penalty function.

To make the inner optimization problem in $(\mathscr{L}\Phi\mathbf{r})(x_0)$ computationally tractable, we introduce a simple class of nonlinear penalties in the form (18). Since this type of penalty derived from a "trick" on the coefficient of the reward function g_t, we refer to it as the *coefficient-based penalty*. Suppose $c_{t+1}(\cdot)$ is any nonzero function. Then we define

$$\begin{aligned}&\phi_{t+1}(\mathbf{x}, \mathbf{a}, \mathbf{w})\\ =&\begin{cases}0, & \text{if } c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})] \notin (q_{1,t+1}, q_{2,t+1}),\\ g_t(x_t, a_t) \cdot (c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})]), & \text{otherwise,}\end{cases}\end{aligned} \quad (19)$$

where $q_{1,t+1}$ and $q_{2,t+1}$ are two numbers satisfying

$$q_{1,t+1} < 0 < q_{2,t+1}$$

such that

$$\begin{aligned}&\mathbb{E}\big[(c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})])\\ &\quad \cdot 1_{\{c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})] \in [q_{1,t+1}, q_{2,t+1}]\}}\big] = 0,\end{aligned} \quad (20)$$

where $1_{\{.\}}$ is the indicator function. With assumption (20) ϕ_{t+1} is automatically a penalty function. With this set of penalty functions the inner optimization problem of $(\mathscr{L}\Phi\mathbf{r})(x_0)$ becomes

$$\max_{\mathbf{a}} \quad \sum_{t=0}^{T-1} g_t(x_t, a_t) \cdot \left[1 - r_{t+1} \cdot (c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})]) \right.$$
$$\left. \cdot 1_{\{c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})] \in [q_{1,t+1}, q_{2,t+1}]\}}\right] + g_T(x_T)$$

If the parameter r_{t+1} is further constrained in the set $\Theta_{t+1} \triangleq [\frac{1}{q_{1,t+1}}, \frac{1}{q_{2,t+1}}]$, the objective function remains concave. Under the condition that the random variable $c_{t+1}(w_{t+1})$ is bounded, we can simply choose

$$q_{1,t+1} = \inf\{c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})]\},$$
$$\text{and} \quad q_{2,t+1} = \sup\{c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})]\}.$$

As a result, ϕ_{t+1} is degenerated to

$$\phi_{t+1}(x_t, a_t, w_{t+1}) = g_t(x_t, a_t) \cdot (c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})]).$$

The coefficient-based penalty is very easy to implement. In particular, there is no need to explore the heuristic policy. Further, it can be used to derive a dual bound in a broad class of MDPs: as long as the inner optimization problem with zero penalty is solvable, so is that with coefficient-based penalty. Though the structures of the coefficient-based penalty and the optimal penalty may not be quite similar, by solving the associated OPP (10), we can detect the best performance of the dual bounds induced within this family of penalties. As a simple illustration, the candidate coefficient-based penalties for (15) may be chosen as

$$\phi^1_{t+1} = -\exp(-\gamma a_t) \cdot z_{t+1}(w_{t+1}) \cdot 1_{\{z_{t+1}(w_{t+1}) \in [q_{1,t+1}, q_{2,t+1}]\}},$$

where $z_{t+1}(w_{t+1}) \triangleq \exp(-\gamma w_{t+1}) - \mathbb{E}[\exp(-\gamma w_{t+1})]$, and $[q_{1,t+1}, q_{2,t+1}]$ is properly chosen such that

$$\mathbb{E}\left[z_{t+1}(w_{t+1}) \cdot 1_{\{z_{t+1}(w_{t+1}) \in [q_{1,t+1}, q_{2,t+1}]\}}\right] = 0.$$

The second class of penalties are linear in $\mathbf{x}$ and $\mathbf{a}$, some variants of which have been studied in [6] and [8] . The linear penalty seems a promising and effective penalty in many circumstances. One big advantage of linear penalty is that it always preserves the concavity of the objective function in the inner optimization problem of $(\mathscr{L}\Phi\mathbf{r})(x_0)$ without any constraint on the parameter $\mathbf{r}$. There are several ways to construct such linear penalties. The simplest one, which is an immediate generalization of the penalties considered in [8], is

$$\phi_{t+1} = a_t \cdot (c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})]), \ t = 0, \cdots, T-1,$$
$$\text{or} \quad \phi_{t+1} = x_t \cdot (c_{t+1}(w_{t+1}) - \mathbb{E}[c_{t+1}(w_{t+1})]), \ t = 1, \cdots, T-1.$$

Another method is based on the first-order approximation of $\Delta V^*_{t+1} = V^*_{t+1}(x_{t+1}) - \mathbb{E}[V^*_{t+1}(x_{t+1}) | x_t, a_t]$ around some fixed heuristic policy $\hat{\mathbf{a}} = (\hat{a}_1, \cdots, \hat{a}_{T-1}) \in \mathbb{A}_{\mathbb{F}}$, which is introduced in [18]. To be specific, we can approximate ΔV^*_{t+1} by

$$\Delta V^*_{t+1} \approx (\hat{V}_{t+1}(x_{t+1}) - \mathbb{E}[\hat{V}_{t+1}(x_{t+1}) | x_t, \hat{a}_t])$$
$$+ \nabla_{a_t} \left(\hat{V}_{t+1}(x_{t+1}) - \mathbb{E}[\hat{V}_{t+1}(x_{t+1}) | x_t, a_t]\right)|_{a_t = \hat{a}_t} \cdot (a_t - \hat{a}_t),$$

where $\hat{V}_t$ is the approximate value function at time t and $\{x_t\}$ follows the state equation $x_{t+1} = f(x_t, \hat{a}_t, w_{t+1})$. A careful look reveals that these linear penalties are special cases or have similar structure of (18).

So far we have derived two different classes of penalties: the coefficient-based penalties and linear penalties, which make the idea of parameterized penalties very promising in obtaining tight dual bounds for practical problems.

V. Numerical Example

In this section we implement our algorithm of parameterized penalties on the canonical example in Section IV-A, since the exact value function is available as a benchmark. The numerical results are presented in Table I, together with the choice of the parameters in (15). To solve for the dual bounds, we employ the following basis of coefficient-based penalties

$$\Phi_1 = \{\phi^1_{t+1} = -\exp(-\gamma a_t) \cdot z_{t+1}(w_{t+1}), \ t = 0, \cdots, T-1.\}$$

The dual bounds induced by (11) with $\Phi = \Phi_1$ are denoted by *Coeff.* in Table I. Note that for each $t = 0, \cdots, T-1$, w_{t+1} is uniformly distributed over $[-3, 0]$. So $z_{t+1}(w_{t+1}) = \exp(-\gamma w_{t+1}) - \mathbb{E}[\exp(-\gamma w_{t+1})]$ is bounded. Therefore we can choose $\Theta^1_{t+1} = [\frac{1}{q_{1,t+1}}, \frac{1}{q_{2,t+1}}]$ with $q_{1,t+1} = \inf\{z_{t+1}(w_{t+1})\} = -5.3618$ and $q_{2,t+1} = \sup\{z_{t+1}(w_{t+1})\} = 13.7237$. We also employ the following basis of linear penalties

$$\Phi_2 = \{\phi^2_{t+1} = a_t \cdot z_{t+1}(w_{t+1}), \ t = 0, \cdots, T-1;$$
$$\phi^3_{t+1} = x_t \cdot z_{t+1}(w_{t+1}), \ t = 1, \cdots, T-1.\}$$

to derive dual bounds induced by (11) with $\Phi = \Phi_2$, which are denoted by *Linear* in Table I. In this case the parameter domain $\Theta^2_{t+1} = \Theta^3_{t+1} = (-\infty, +\infty)$. The dual bounds induced by $\Phi = \{0\}$ and $\Phi = \Phi_1 \cup \Phi_2$ are denoted by *Zero* and *Combined* respectively. We will explain the implementation details in the following two paragraphs.

TABLE I

Dual Bounds with Different Penalties

$(T = 3, \alpha = 2, \gamma = 1, \, w_t \sim Unif[-3, 0])$

Initial	Exact	Dual Bounds with Different Penalties			
x_0	V^*_0	Zero	Coeff.	Linear	Combined
0	−18.517	−15.457 (0.060)	−16.974 (0.169)	−17.440 (0.171)	−17.863 (0.235)
−1	−50.334	−41.695 (0.164)	−45.782 (0.737)	−46.338 (0.617)	−48.049 (0.543)
−2	−136.822	−113.155 (0.717)	−123.878 (1.598)	−126.988 (1.727)	−128.978 (1.550)

To approximate the outer optimization problem, we first generate $\{\mathbf{w}^l, l = 1, \cdots, 100\}$ independently from the uniform distribution over $[-3, 0]$ and its antithetic pair (see [19] for reference on antithetic variates), which are used to define the function

$$Q(\mathbf{r}) = \frac{1}{2L} \sum_{l=1}^{L} [I(x_0, \Phi\mathbf{r}, \mathbf{w}^l) + I(x_0, \Phi\mathbf{r}, -3 - \mathbf{w}^l)].$$

Then we employ Algorithm 1 to find the minimizer of Q with maximum iteration number $MaxItr = 300$ and other parameters $a = 1, \beta_1 = 1, A = 100, \beta_2 = 0.4, C = 100$. Particularly, we use $c = 5, c = 100$, and $c = 150$ for the basis penalties Φ_1, Φ_2, and $\Phi_1 \cup \Phi_2$ respectively, considering the gradient estimation in different parameter domains. We initialize the parameters $\mathbf{r}_0 = \mathbf{0}^3$, $\mathbf{0}^5$, and $(\mathbf{0}^3, \tilde{\mathbf{r}}_2^*)$ for the basis penalties Φ_1, Φ_2, and $\Phi_1 \cup \Phi_2$ respectively, where $\mathbf{0}^d$ is a d-dimensional zero vector and $\tilde{\mathbf{r}}_2^*$ is the numerical solution to the outer optimization problem with basis penalties Φ_2 (*i.e.*, linear penalties). With such an initial choice the subsequent dual bound will be at least as good as the dual bound associated with the linear penalties. The inner optimization problems are solved using the MOSEK optimization toolbox 6.0 for Matlab. We record the approximate solutions corresponding to different sets of penalty functions.

With the approximate solution to the outer optimization problem, we estimate each dual bound by generating 100 independent sample paths of $\mathbf{w}$ and its antithetic pair, and solving the inner optimization problem with parameterized penalties. This procedure is repeated for 10 independent runs. We present our numerical results in Table I, where each entry shows the sample average and standard error (in parentheses) of the 10 independent runs.

The criterion of examining the quality of dual bounds is obvious: the smaller the gap between the exact value V_0^* and the dual bound, the better the bound. We observe that the gap between V_0^* and the zero-penalty bound increases quickly as x_0 decreases, which calls for more effective penalties to help reduce the dual bound. It can be seen that the dual bounds with coefficient-based penalties and linear penalties make considerable improvement compared with zero-penalty bounds. In our numerical experiments the linear penalties outperform the coefficient-based penalties, probably due to the following reasons: the number of the basis penalties in the linear class is greater than that in the coefficient-based class; in addition, the range of Θ_{t+1}^1 is much more constrained compared with Θ_{t+1}^2 and Θ_{t+1}^3.

Finally, we note that the dual bounds induced by combined penalties significantly reduce the duality gap and perform consistently the best among all the dual bounds, which confirms the theoretical result in Section III.

In practice, we may start with only one class of penalties to see whether the resulting dual bound is close enough to the value function induced by some suboptimal policy. If the gap is under the tolerance level, we are done; otherwise, we should consider improving the suboptimal policy as well as the dual bounds, for the sake of which an additional class of penalties can be introduced.

VI. Conclusion

In this paper we develop a numerical approach to find the dual bound on the value function of Markov Decision Processes. We present a framework of parameterized penalties under the dual representation of MDPs in order to tighten the dual bound. This approach has two layers: at the first level, we formulate the outer optimization problem as a convex optimization problem, solving which provides the solution to the optimal choice of the parameters; at the second level, the Monte Carlo simulation provides a high-biased estimator on the value function. To diversify the penalties, we propose a new class of nonlinear penalties that are easy to implement and can be used in a broad class of MDPs. We test our algorithm on a numerical example and compare the performance of the dual bounds induced by different sets of penalties. The numerical results show: (i) our proposed coefficient-based penalties can be used to tighten the dual bound; (ii) moreover, an appropriate combination of different types of penalties can considerably improve the quality of the dual bound as we expect. Some future direction includes the extension to the dual representation of continuous-time controlled Markov processes based on information relaxation [7].

References

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

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

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

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

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

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

[7] F. Ye and E. Zhou, "Dual formulation of controlled Markov diffusions and its application," 2012, working paper.

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

[9] M. Haugh and A. Lim, "Linear-quadratic control and information relaxations," working paper.

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

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

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

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

[14] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[15] S. Kim, R. Pasupathy, and S. Henderson, "A guide to sample-average approximation," 2011.

[16] J. C. Spall, "Overview of the simultaneous perturbation method for efficient optimization," *Johns Hopkins APL Technical Digest*, vol. 19, no. 4, pp. 482–492, 1998.

[17] A. Seierstad, *Stochastic Control in Discrete and Continuous Time*. Springer, 2009.

[18] D. B. Brown, "Gradient-based bounds for stochastic dynamic program," 2011, INFORMS Annual Meeting, Charlotte, NC.

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