MANAGEMENT SCIENCE
Vol. 58, No. 12, December 2012, pp. 2292–2308
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1551

Pathwise Optimization for Optimal Stopping Problems

Vijay V. Desai
Industrial Engineering and Operations Research, Columbia University, New York, New York 10027, vvd2101@columbia.edu

Vivek F. Farias
MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, vivekf@mit.edu

Ciamac C. Moallemi
Graduate School of Business, Columbia University, New York, New York 10027, ciamac@gsb.columbia.edu

We introduce the pathwise optimization (PO) method, a new convex optimization procedure to produce upper and lower bounds on the optimal value (the "price") of a high-dimensional optimal stopping problem. The PO method builds on a dual characterization of optimal stopping problems as optimization problems over the space of martingales, which we dub the martingale duality approach. We demonstrate via numerical experiments that the PO method produces upper bounds of a quality comparable with state-of-the-art approaches, but in a fraction of the time required for those approaches. As a by-product, it yields lower bounds (and suboptimal exercise policies) that are substantially superior to those produced by state-of-the-art methods. The PO method thus constitutes a practical and desirable approach to high-dimensional pricing problems. Furthermore, we develop an approximation theory relevant to martingale duality approaches in general and the PO method in particular. Our analysis provides a guarantee on the quality of upper bounds resulting from these approaches and identifies three key determinants of their performance: the quality of an input value function approximation, the square root of the effective time horizon of the problem, and a certain spectral measure of "predictability" of the underlying Markov chain. As a corollary to this analysis we develop approximation guarantees specific to the PO method. Finally, we view the PO method and several approximate dynamic programming methods for high-dimensional pricing problems through a common lens and in doing so show that the PO method dominates those alternatives.

Key words: dynamic programming; optimal control; optimal stopping; American options; Bermudian options
History: Received October 5, 2010; accepted December 18, 2011, by Wei Xiong, stochastic models and simulation. Published online in *Articles in Advance* June 15, 2012.

1. Introduction

Consider the following optimal control problem: A Markov process evolves in discrete time over the state space $\mathscr{X}$. Denote this process by $\{x_t, t \geq 0\}$. The process is associated with a state-dependent reward function $g: \mathscr{X} \to \mathbb{R}$. Our goal is to solve the optimization problem

$$\sup_{\tau} \mathsf{E}[\alpha^\tau g(x_\tau) \mid x_0 = x],$$

where the optimization is over stopping times τ adapted to the $\{x_t\}$ process, and $\alpha \in [0, 1)$ is a discount factor. In other words, we wish to pick a stopping time that maximizes the expected discounted reward. Such *optimal stopping* problems arise in a myriad of applications, most notably in the pricing of financial derivatives.

In principle, the above stopping problem can be solved via the machinery of dynamic programming. However, the applicability of the dynamic programming approach is typically curtailed by the size of the state space $\mathscr{X}$. In particular, in many applications of interest, $\mathscr{X}$ is a high-dimensional space and thus intractably large.

Because high-dimensional stopping problems are important from a practical perspective, a number of alternative approaches that contend with the so-called "curse of dimensionality" have emerged. There are two broad classes of methods by which one can develop bounds on the optimal value of a stopping problem, motivated essentially by distinct characterizations of the optimal solution to the stopping problem:

• *Lower Bounds/Approximate Dynamic Programming (ADP)*. The optimal control is characterized by an optimal value function, which, in turn, is the unique solution to the so-called Bellman equation. A natural goal is to attempt to approximate this value function by finding "approximate" solutions to the Bellman equation. This is the central goal of ADP algorithms such as *regression pricing* methods of the type pioneered by Carriere (1996), Longstaff and Schwartz (2001), and Tsitsiklis and Van Roy (2001). Such an approximate solution can then be used to both define a control

policy and, via simulation of that (suboptimal) policy, a lower bound on the optimal value function.

• *Upper Bounds/Martingale Duality.* At a high level, this approach may be thought of as relaxing the requirement of causality, while simultaneously introducing a penalty for this relaxation. The appropriate penalty function is itself a stochastic process (a martingale), and by selecting the "optimal" martingale, one may in fact solve the original stopping problem. In the context of stopping problems, part of this characterization appears to date back at least to the work by Davis and Karatzas (1994), and this idea was subsequently fully developed by Rogers (2002) and Haugh and Kogan (2004).

Not surprisingly, finding such an optimal martingale is no easier than solving the original stopping problem. As such, the martingale duality approach consists of heuristically selecting "good" martingale penalty functions, using these to compute upper bounds on the price (i.e., the optimal value of the stopping problem). Here, two techniques are commonly employed. The first, which we will call a *dual value function* approach, derives a martingale penalty function from an approximation to the optimal value function. Such an approximation will typically be generated, for example, along the course of regression pricing procedures such as those described above. Alternatively, in what we will call a *dual policy* approach, a martingale penalty function can be derived from a heuristic control policy. This latter approach was proposed by Andersen and Broadie (2004). A good control policy will typically also be generated using a regression pricing procedure.

A combination of these methods has come to represent the state of the art in financial applications (see, e.g., Glasserman 2004). There, practitioners typically use regression pricing to derive optimal policies for the exercise of American and Bermudan options, and to derive lower bounds on prices. The martingale duality approach is then applied in a complementary fashion to generate upper bounds, using either the dual value function approach or the dual policy approach. Taken together, these methods provide a "confidence bound" on the true price. In this area, the development of such methodologies is thought to be worth considerable financial value, and thus may represent the greatest practical success of approximate dynamic programming.

The present paper, in a nutshell, introduces a new approach to solving high-dimensional stopping problems that draws on techniques from both of the methodologies above, and ultimately unifies our understanding of the two approaches. This new method is ultimately seen to be desirable from the practical perspective of rapidly pricing high-dimensional financial derivatives. In addition, we develop a theory that allows us to characterize the quality of the solutions produced by the approaches above.

In greater detail, we make the following contributions

• *A New Algorithm.* ADP algorithms systematically explore approximations to the optimal value function within the span of some predefined set of basis functions. The duality approach, on the other hand, relies on an ad hoc specification of an appropriate martingale penalty process. We introduce a new approach, which we call the *pathwise optimization* (*PO*) method. The PO method systematizes the search for a good martingale penalty process. In particular, given a set of basis functions whose linear span is expected to contain a good approximation to the optimal value function, we posit a family of martingales. As it turns out, finding a martingale within this family that produces the best possible upper bound to the value function is a convex optimization problem. The PO method seeks to solve this problem. We show that this method has several merits relative to extant schemes:

1. The PO method is a specific instance of the dual value function approach. By construction, however, the PO method produces an upper bound that is provably tighter than *any* other dual value function approach that employs a value function approximation contained in the span of the same basis function set. These latter approximations are analogous to what is typically found using regression methods of the type proposed by Longstaff and Schwartz (2001) and Tsitsiklis and Van Roy (2001). We demonstrate this fact in numerical experiments, where we will show that, given a fixed set of basis functions, the benefit of the PO method over the dual value function approach in concert with regression pricing can be substantial. We also see that the incremental computational overhead of the PO method over the latter method is manageable.

2. We compare the PO method to upper bounds generated using the dual policy approach in concert with policies derived from regression pricing. Given a fixed set of basis functions, we will see in numerical experiments that the PO method yields upper bounds that are comparable to but not as tight as those from the latter approach. However, the PO method does so in a substantially shorter amount of time, typically requiring a computational budget that is smaller by an order of magnitude.

3. The aforementioned regression techniques are the mainstay for producing control policies and lower bounds in financial applications. We illustrate that the PO method yields a continuation value approximation that can subsequently be used to derive control policies and lower bounds. In computational experiments, these control policies and lower bounds are

substantially superior to those produced by regression methods.

In summary, the PO method is quite attractive from a practical perspective.

• *Approximation Theory.* We offer new guarantees on the quality of upper bounds of martingale penalty approaches in general, as well as specific guarantees for the PO method. We compare these guarantees favorably to guarantees developed for other ADP methods. Our guarantees characterize the structural properties of an optimal stopping problem that are general determinants of performance for these techniques. Specifically:

1. In an infinite horizon setting, we show that the quality of the upper bound produced by the generic martingale duality approach depends on three parameters: the error in approximating the value function (measured in a root-mean-squared error sense), the square root of the effective time horizon (as also observed by Chen and Glasserman 2007), and a certain measure of the "predictability" of the underlying Markov process. We believe that this latter parameter provides valuable insight on aspects of the underlying Markov process that make a particular pricing problem easy or hard.

2. In an infinite horizon setting, we produce *relative* upper bound guarantees for the PO method. In particular, we produce guarantees on the upper bound that scale linearly with the approximation error corresponding to the *best possible* approximation to the value function within the span of the basis functions employed in the approach. Note that the latter approximation is typically not computable. This result makes precise the intuition that the PO method produces good price approximations if there exists *some* linear combination of the basis functions that is able to describe the value function well.

3. Upper bounds produced by the PO methods can be directly compared to upper bounds produced by linear programming-based ADP algorithms of the type introduced by Schweitzer and Seidmann (1985), de Farias and Van Roy (2003), and Desai et al. (2012). In particular, we demonstrate that the PO method produces provably tighter upper bounds than the latter methods. Although these methods have achieved considerable success in a broad range of large-scale dynamic optimization problems, they are dominated by the PO method for optimal stopping problems.

The literature on ADP algorithms is vast, and we make no attempt to survey it here. Bertsekas (2007, Chap. 6) provides a good, brief overview. ADP algorithms are usually based on an approximate approach for solving Bellman's equation. In the context of optimal stopping, methods have been proposed that are variations of approximate value iteration (Tsitsiklis and Van Roy 1999, Yu and Bertsekas 2007), approximate policy iteration (Longstaff and Schwartz 2001, Clément et al. 2002), and approximate linear programming (ALP) (Borkar et al. 2009).

Martingale duality-based upper bounds for the pricing of American and Bermudan options, which rely on Doob's decomposition to generate the penalty process, were introduced by Rogers (2002) and Haugh and Kogan (2004). Rogers (2002) suggested the possibility of determining a good penalty process by optimizing linear combinations of martingales; our method is a special case of this that uses a specific parametrization of candidate martingales in terms of basis functions. Andersen and Broadie (2004) showed how to compute martingale penalties from rules and obtain upper bounds; practical improvements to these technique were studied by Broadie and Cao (2008). An alternative "multiplicative" approach to duality was introduced by Jamshidian (2003). Its connections with martingale duality approach were explored by Chen and Glasserman (2007), who also developed approximation guarantees for martingale duality upper bounds. Belomestny et al. (2009) described a variation of the martingale duality procedure that does not require inner simulation. Rogers (2010) described a pure dual algorithm for pricing. Generalizations of the martingale duality approach to control problems other than optimal stopping have been studied by Rogers (2008), Lai et al. (2010), Brown et al. (2010), and Brown and Smith (2011). Furthermore, Brown et al. (2010) generalized martingale duality to a broader class of information relaxations.

2. Formulation

Our framework will be that of an optimal stopping problem over a finite time horizon. Specifically, consider a discrete-time Markov chain with state $x_t \in \mathcal{X} \subset \mathbb{R}^n$ at each time $t \in \mathcal{T} \triangleq \{0, 1, \ldots, d\}$. Denote by P the transition kernel of the chain. Without loss of generality, we will assume that P is time invariant. Let $\mathcal{F} \triangleq \{\mathcal{F}_t\}$ be the natural filtration generated by the process $\{x_t\}$, i.e., for each time t, $\mathcal{F}_t \triangleq \sigma(x_0, x_1, \ldots, x_t)$.

Given a measurable function $g\colon \mathcal{X} \to \mathbb{R}$, we define the payoff of stopping when the state is x_t as $g(x_t)$. For each $t \in \mathcal{T}$, let $\mathcal{S}_t$ be the space of real-valued measurable functions $J_t\colon \mathcal{X} \to \mathbb{R}$ defined on state space $\mathcal{X}$, with $\mathsf{E}[J_t(x_t)^2 \mid x_0] < \infty$, for all $x_0 \in \mathcal{X}$. Assume that $g \in \mathcal{S}_t$ for all t. Define $\mathcal{P}$ to be the set of functions $J\colon \mathcal{X} \times \mathcal{T} \to \mathbb{R}$ such that, for each $t \in \mathcal{T}$, $J_t \triangleq J(\cdot, t)$ is contained in the set $\mathcal{S}_t$. In other words, $\mathcal{P}$ is the set of Markovian processes (i.e., time-dependent functionals of the state) that possess second moments.

A stationary exercise policy $\mu \triangleq \{\mu_t,\ t \in \mathcal{T}\}$ is a collection of measurable functions where each $\mu_t\colon \mathcal{X} \to \{\mathsf{STOP}, \mathsf{CONTINUE}\}$ determines the choice of action

at time t as a function of the state x_t. Without loss of generality, we will require that $\mu_d(x) = \mathsf{STOP}$ for all $x \in \mathscr{X}$, i.e., the process is always stopped at the final time d.

We are interested in finding a policy that maximizes the expected discounted payoff of stopping. The value of a policy μ assuming one starts at state x in period t is given by

$$J_t^\mu(x) \triangleq \mathsf{E}[\alpha^{\tau_\mu(t)-t} g(x_{\tau_\mu(t)}) \mid x_t = x],$$

where $\tau_\mu(t)$ is the stopping time $\tau_\mu(t) \triangleq \min\{s \geq t\colon \mu_s(x_s) = \mathsf{STOP}\}$. Our goal is to find a policy μ that simultaneously maximizes the value function $J_t^\mu(x)$ for all t and x. We will denote such an *optimal policy* by μ^* and the corresponding *optimal value function* by J^*.

In principle, J^* may be computed via the following dynamic programming backward recursion: for all $x \in \mathscr{X}$ and $t \in \mathscr{T}$,

$$J_t^*(x) \triangleq \begin{cases} \max\{g(x), \alpha\mathsf{E}[J_{t+1}^*(x_{t+1}) \mid x_t = x]\} \\ \qquad \text{if } t < d, \\ g(x) \quad \text{if } t = d. \end{cases} \tag{1}$$

The corresponding optimal stopping policy μ^* is "greedy" with respect to J^* and given by

$$\mu_t^*(x) \triangleq \begin{cases} \mathsf{CONTINUE} \\ \quad \text{if } t < d \text{ and } g(x) < \alpha\mathsf{E}[J_{t+1}^*(x_{t+1}) \mid x_t = x], \\ \mathsf{STOP} \quad \text{otherwise.} \end{cases} \tag{2}$$

2.1. The Martingale Duality Approach

We begin by defining the *martingale difference operator* Δ. The operator Δ maps a function $V \in \mathscr{S}_1$ to the function $\Delta V\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ according to $(\Delta V)(x_1, x_0) \triangleq V(x_1) - \mathsf{E}[V(x_1) \mid x_0]$. Given an arbitrary function $J \in \mathscr{P}$, define the process

$$M_t \triangleq \sum_{s=1}^{t} \alpha^s (\Delta J_s)(x_s, x_{s-1}), \quad \forall\, t \in \mathscr{T}.$$

Then, M is a martingale adapted to the filtration $\mathscr{F}$. Hence, we view Δ as a projection onto the space of martingale differences.

Next, we define, for each $t \in \mathscr{T}$, the *martingale duality upper bound operator* $F_t\colon \mathscr{P} \to \mathscr{S}_t$ according to

$$(F_t J)(x) \triangleq \mathsf{E}\left[\max_{t \leq s \leq d}\left(\alpha^{s-t} g(x_s) - \sum_{p=t+1}^{s} \alpha^{p-t} \Delta J_p(x_p, x_{p-1})\right) \,\middle|\, x_t = x\right].$$

Finally, we define $J^* \in \mathscr{P}$ according to $J^*(x, t) \triangleq J_t^*(x)$. We are now ready to state the following key lemma, due to Rogers (2002) and Haugh and Kogan (2004). A proof is provided in §A.1 of the online supplement (available at http://moallemi.com/ciamac/papers/po-2010-supplement.pdf) for completeness.

Lemma 1 (Martingale Duality). (i) (Weak Duality) *For any $J \in \mathscr{P}$ and all $x \in \mathscr{X}$ and $t \in \mathscr{T}$, $J_t^*(x) \leq F_t J(x)$.*

(ii) (Strong Duality) *For all $x \in \mathscr{X}$ and $t \in \mathscr{T}$, $J_t^*(x) = F_t J^*(x)$.*

The above result may be succinctly stated as follows: For any $t \in \mathscr{T}$, $x \in \mathscr{X}$,

$$J_t^*(x) = \inf_{J \in \mathscr{P}} F_t J(x). \tag{3}$$

This is an alternative (and somewhat convoluted) characterization of the optimal value function J^*. Its value, however, lies in the fact that *any* $J \in \mathscr{P}$ yields an upper bound, and evaluating this upper bound for a given J is for all practical purposes *not* impacted by the size of $\mathscr{X}$. Indeed, extant approaches to using the above characterization to produce upper bounds on J^* use, as surrogates for J, an approximation of the optimal value function J^* (see, e.g., Glasserman 2004). This approximation can be derived over the course of a regression pricing method of the type introduced by Longstaff and Schwartz (2001) or Tsitsiklis and Van Roy (2001). We call this the *dual value function* approach. Alternatively, an approximating value function corresponding to a suboptimal policy (Andersen and Broadie 2004) can be used, where the policy is typically produced by a regression pricing method. We call this the *dual policy* approach.

3. The Pathwise Optimization Method

Motivated by the (in general, intractable) optimization problem (3), we are led to consider the following: what if one chose to optimize over functions $J \in \hat{\mathscr{P}} \subset \mathscr{P}$, where $\hat{\mathscr{P}}$ is compactly parametrized and easy to optimize over? Motivated by ADP algorithms that seek approximations to the optimal value function that are linear combinations of some set of basis functions, we are led to the following parametrization: Assume we are given a collection of K *basis functions* $\Phi \triangleq \{\phi_1, \phi_2, \ldots, \phi_K\} \subset \mathscr{P}$. Ideally these basis functions capture features of the state space or optimal value function that are relevant for effective decision making, but frequently generic selections work well (e.g., all monomials up to a fixed degree). We may then consider restricting attention to functions that are linear combinations of elements of Φ, i.e., functions of the form

$$(\Phi r)_t(x) \triangleq \sum_{l=1}^{K} r_l \phi_l(x, t), \quad \forall\, x \in \mathscr{X},\ t \in \mathscr{T}.$$

Here, $r \in \mathbb{R}^K$ is known as a *weight vector*. Denote this subspace of $\mathscr{P}$ by $\hat{\mathscr{P}}$ and note that $\hat{\mathscr{P}}$ is compactly parameterized by K parameters (as opposed to $\mathscr{P}$, which is infinite dimensional in general). Setting

the starting epoch to $t = 0$ for convenience, we may rewrite the optimization problem (3) restricted to $\hat{\mathcal{P}}$ as

$$\inf_r F_0 \Phi r(x). \tag{4}$$

We call this problem the *pathwise optimization* problem. The lemma below demonstrates that (4) is, in fact, a *convex* optimization problem.

LEMMA 2. *For every $t \in \mathcal{T}$ and $x \in \mathcal{X}$, the function $r \mapsto F_t \Phi r(x)$ is convex in r.*

PROOF. Observe that, given a fixed (x, t) and as a function of r, $F_t \Phi r(x)$ is a nonnegative linear combination of a set of pointwise suprema of affine functions of r, and hence must be convex as each of these operations preserves convexity. □

Before devising a practical approach to solving (4), let us reflect on what solving this program accomplishes. We have devised a means to systematically and, anticipating the developments in the sequel, practically, find a martingale penalty process within a certain parametrized family of martingales. To appreciate the value of this approach, we note that it is guaranteed, by construction, to produce tighter upper bounds on price than *any* dual value function methods derived from value function approximations that are within the span of the same basis function set. These latter approximations are analogous to what is typically found using regression methods of the type proposed by Longstaff and Schwartz (2001) and Tsitsiklis and Van Roy (2001).[1]

Now, from a practical perspective, the optimization problem (4) is an unconstrained minimization of a convex function over a relatively low-dimensional space. Algorithmically, the main challenge is evaluating the objective, which is the expectation of a functional over paths in a high-dimensional space. We will demonstrate that this can be efficiently approximated via sampling.

3.1. Solution via Sampling

Consider sampling S independent *outer* sample paths of the underlying Markov process starting at some given state x_0; denote path i by $x^{(i)} \triangleq \{x_s^{(i)}, s \in \mathcal{T}\}$ for $i = 1, 2, \ldots, S$. Given a fixed weight vector r and initial state x_0, define a sampled approximation to the upper bound $F_0 \Phi r(x_0)$ by

$$\hat{F}_0^S \Phi r(x_0) \triangleq \frac{1}{S} \sum_{i=1}^{S} \max_{0 \le s \le d} \left(\alpha^s g(x_s^{(i)}) - \sum_{p=1}^{s} \alpha^p \Delta(\Phi r)_p(x_p^{(i)}, x_{p-1}^{(i)}) \right). \tag{5}$$

[1] Strictly speaking, the regression pricing approaches of Longstaff and Schwartz (2001) and Tsitsiklis and Van Roy (2001) seek linearly parameterized approximations to the optimal continuation value function, as is described in §4. However, the same ideas could easily be applied to find linearly parameterized approximations to the optimal value function.

By the strong law of large numbers, almost surely, $\hat{F}_0^S \Phi r(x_0) \to F_0 \Phi r(x_0)$, as $S \to \infty$. This suggests $\hat{F}_0^S \Phi r(x_0)$ as a useful proxy for the objective in the pathwise optimization problem (4).

However, consider the quantities that appear in the left-hand side of (5),

$$\Delta(\Phi r)_p(x_p^{(i)}, x_{p-1}^{(i)}) = (\Phi r)_p(x_p^{(i)}) - \mathsf{E}[(\Phi r)_p(x_p) \mid x_{p-1} = x_{p-1}^{(i)}].$$

The expectation in the above expression may, in certain cases, be computed in closed form (see, e.g., Glasserman and Yu 2002, Belomestny et al. 2009). More generally, however, to achieve a tractable objective, we can instead replace the conditional expectation by its empirical counterpart. In particular, we generate I independent *inner* samples $\{x_p^{(i,j)}, j = 1, \ldots, I\}$, conditional on $x_{p-1} = x_{p-1}^{(i)}$. In other words, these inner samples are generated according to the one-step transition distribution $P(x_{p-1}^{(i)}, \cdot)$. Then, consider the approximation

$$\hat{\Delta}(\Phi r)_p(x_p^{(i)}, x_{p-1}^{(i)}) \triangleq (\Phi r)_p(x_p^{(i)}) - \frac{1}{I} \sum_{j=1}^{I} (\Phi r)_p(x_p^{(i,j)}). \tag{6}$$

Note that, almost surely, $\hat{\Delta}(\Phi r)_p(x_p^{(i)}, x_{p-1}^{(i)}) \to \Delta(\Phi r)_p \cdot (x_p^{(i)}, x_{p-1}^{(i)})$ as $I \to \infty$. This suggests the *nested* Monte Carlo approximation

$$\hat{F}_0^{S,I} \Phi r(x_0) \triangleq \frac{1}{S} \sum_{i=1}^{S} \max_{0 \le s \le d} \left(\alpha^s g(x_s^{(i)}) - \sum_{p=1}^{s} \alpha^p \hat{\Delta}(\Phi r)_p(x_p^{(i)}, x_{p-1}^{(i)}) \right) \tag{7}$$

to the objective in the pathwise optimization problem (4). Having thus replaced expectations by their empirical counterparts, we are ready to state a general, implementable, sampled variant of the optimization problem (4):

$$\inf_r \hat{F}_0^{S,I} \Phi r(x). \tag{8}$$

The following theorem establishes that, subject to technical conditions and given a sufficiently large number of outer sample paths S and one-stage inner samples I, the upper bound achieved by minimizing the nested Monte Carlo approximation $\hat{F}_0^{S,I} \Phi r(x_0)$ can be made arbitrarily close to that of the pathwise optimization problem (4).

THEOREM 1. *Let $\mathcal{N} \subset \mathbb{R}^K$ be a compact set. Fix an initial state x_0 and $\epsilon > 0$. Then, almost surely, if S is sufficiently large, for all I sufficiently large,*

$$\left| \min_{r \in \mathcal{N}} F_0 \Phi r(x_0) - \min_{r \in \mathcal{N}} \hat{F}_0^{S,I} \Phi r(x_0) \right| \le \epsilon.$$

The proof of Theorem 1 is provided in §A.2 of the online supplement. It relies on establishing the *uniform* convergence of $\hat{F}_0^{S,I}\Phi r(x_0) \to F_0\Phi r(x_0)$ over all r in the compact set $\mathcal{N}$.[2]

Now, observe that the sampled optimization problem (7) can be written as

$$\begin{aligned} \underset{r,\,u}{\text{minimize}} \quad & \frac{1}{S}\sum_{i=1}^{S} u_i \\ \text{subject to} \quad & u_i + \sum_{p=1}^{s} \alpha^p \hat{\Delta}(\Phi r)_p(x_p^{(i)}) \geq \alpha^s g(x_s^{(i)}), \\ & \qquad \forall\, 1 \leq i \leq S,\ 0 \leq s \leq d. \end{aligned} \tag{9}$$

The optimization problem (9) is a linear program (LP) that can be solved with standard commercial LP solvers. It has $K+S$ variables and $S(d+1)$ constraints. Because no two variables $\{u_i, u_j\}$ with $i \neq j$ appear in the same constraint, it is easy to see that the Hessian corresponding to a logarithmic barrier function for the problem has block arrow structure. Inverting this matrix will require $O(K^2S)$ floating point operations (see, e.g., Boyd and Vandenberghe 2004, Appendix C, p. 675). Consequently, one may argue that the complexity of solving this LP via an interior point method essentially scales linearly with the number of outer sample paths S.

3.2. Unbiased Upper Bound Estimation

Denote by $\hat{r}_{\mathsf{PO}}$ a solution to the sampled pathwise problem (9). An obvious upper bound on $J_0^*(x_0)$ is given by the optimal value of quantity $F_0\Phi\hat{r}_{\mathsf{PO}}(x_0)$. Because we can't compute this quantity directly, it makes sense to approximate it via sampling to obtain an estimated upper bound. Note that the optimal objective value of the problem (9) itself is a biased upper bound estimate. This bias comes from the fact that the fact that the expected value of the minimum of the sample average in (9) is less than the minimum of the expected value and is essentially a consequence of Jensen's inequality (see, e.g., Glasserman 2004, §8.2). To obtain an unbiased upper bound estimate, given $\hat{r}_{\mathsf{PO}}$, we use a second, independent Monte Carlo procedure to estimate an upper bound as follows:

1. Generate a second set of S outer sample paths, each with I inner samples, obtained independently of the samples used in solving (9).

2. Compute the sampled martingale differences associated with value function approximation $\Phi\hat{r}_{\mathsf{PO}}$ using (6), with the new set of samples. As discussed by Glasserman (2004, §8.7, p. 473), because (6) involves an unbiased estimate of the conditional expectation, this expression indeed yields a martingale difference.

3. Using the new sample paths and the new sampled martingale differences, compute the quantity

$$\frac{1}{S}\sum_{i=1}^{S} \max_{0\leq s\leq d}\left(\alpha^s g(x_s^{(i)}) - \sum_{p=1}^{s} \alpha^p \hat{\Delta}(\Phi\hat{r}_{\mathsf{PO}})_p(x_p^{(i)}, x_{p-1}^{(i)})\right). \tag{10}$$

By Lemma 1, the expected value of (10) is an upper bound on the optimal value. By the strong law of large numbers, (10) will thus converge to an upper bound as $S \to \infty$. Finally, the central limit theorem can be applied to compute confidence intervals for the upper bound estimator of (10).

3.3. Lower Bounds and Policies

The PO method generates upper bounds on the performance of an optimal policy. We are also interested in generating good stopping policies, which, in turn, will yield lower bounds on optimal performance. Here, we describe a method that does so by computing a continuation value approximation.

In particular, for $0 \leq t < d$ and $x_t \in \mathcal{X}$, denote by $C_t^*(x_t)$ the optimal continuation value, or the best value the can be achieved by any policy at time t and state x_t that does not immediately stop. Mathematically,

$$C_t^*(x_t) \triangleq \alpha\mathsf{E}[J_{t+1}^*(x_{t+1}) \mid x_t].$$

Note that the optimal policy μ^* can be expressed succinctly in terms of C^* via

$$\mu_t^*(x) \triangleq \begin{cases} \mathsf{CONTINUE} & \text{if } t < d \text{ and } g(x) < C_t^*(x), \\ \mathsf{STOP} & \text{otherwise,} \end{cases} \tag{11}$$

for all $t \in \mathcal{T}$ and $x \in \mathcal{X}$. In other words, μ^* decides to stop or not by acting greedily using C^* to assess the value of not stopping. Inspired by this, given a good approximation $\tilde{C}$ to the optimal continuation value, we can attempt to construct a good policy by replacing C^* with $\tilde{C}$ in (11).

Now, given a solution to (9), $\hat{r}_{\mathsf{PO}}$, we can generate upper bounds on continuation value and regress these against basis functions to generate a continuation value approximation. In particular, it follows from Lemma 1 that

$$C_t^*(x_t) \leq \mathsf{E}\left[\max_{t+1\leq s\leq d} \alpha^{s-t} g(x_s) - \sum_{p=t+2}^{s} \alpha^{p-t}\Delta(\Phi\hat{r}_{\mathsf{PO}})_p(x_p, x_{p-1}) \,\middle|\, x_t\right], \tag{12}$$

[2] Note that the restriction of the weight vectors to a compact set is a standard technical assumption in the theoretical analysis of sample average approximations to optimization problems (see, e.g., Shapiro et al. 2009). In practice, this bounding set can be chosen to be sufficiently large so as to be likely to include the optimal solution of the unconstrained pathwise optimization problem (4), or it can simply be omitted.

for all $0 \le t < d$ and $x_t \in \mathscr{X}$. Thus, at time t along the ith sample path, a point estimate of an upper bound on $C_t^*(x_t^{(i)})$ is given by

$$\bar{c}_t^{(i)} \triangleq \max_{t+1 \le s \le d} \alpha^{s-t} g_s(x_s^{(i)}) - \sum_{p=t+2}^{s} \alpha^{p-t} \left\{ (\Phi \hat{r}_{\mathsf{PO}})_p(x_p^{(i)}) - \hat{\mathsf{E}}\left[(\Phi \hat{r}_{\mathsf{PO}})_p(x_p) \mid x_{p-1}^{(i)} \right] \right\}.$$

For each $0 \le t < d-1$, the values $\{\bar{c}_t^{(i)}, 1 \le i \le S\}$ can now be regressed against basis functions to obtain a continuation value approximation. In particular, defining a set of K basis functions of the state x_t, $\Psi_t \triangleq \{\psi_{1,t}, \psi_{2,t}, \ldots, \psi_{K,t}\} \subset \mathscr{S}_t$, we can consider linear combinations of the form

$$(\Psi_t \kappa_t)(x) \triangleq \sum_{l=1}^{K} \kappa_{l,t} \psi_{l,t}(x), \quad \forall x \in \mathscr{X},$$

where $\kappa_t \in \mathbb{R}^K$ is a weight vector.[3] The weight vectors $\{\kappa_t, 0 \le t < d\}$ can be computed efficiently in a recursive fashion as follows:

1. Iterate backward over times $t = d-1, d-2, \ldots, 0$.

2. For each sample path $1 \le i \le S$, we need to compute the continuation value estimate $\bar{c}_t^{(i)}$. If $t = d-1$, this is simply $\bar{c}_{d-1}^{(i)} = \alpha g(x_d^{(i)})$. If $t < d-1$, this can be computed recursively as

$$\bar{c}_t^{(i)} = \alpha \max\left\{ g(x_{t+1}^{(i)}), \bar{c}_{t+1}^{(i)} - \alpha((\Phi \hat{r}_{\mathsf{PO}})_{t+2}(x_{t+2}^{(i)}) - \hat{\mathsf{E}}[(\Phi \hat{r}_{\mathsf{PO}})_{t+2}(x_{t+2}) \mid x_{t+1}^{(i)}]) \right\}.$$

3. Compute the weight vector κ_t via the regression

$$\kappa_t \in \arg\min_{\kappa} \frac{1}{S} \sum_{i=1}^{S} (\Psi_t \kappa(x_t^{(i)}) - \bar{c}_t^{(i)})^2.$$

We may then use the suboptimal policy that is greedy with respect to the continuation value approximation given by $\Psi_t \kappa_t$, for each $0 \le t \le d-1$.

Observe that, at a high-level, our algorithm is reminiscent of the regression pricing approach of Longstaff and Schwartz (2001). Both methods proceed backward in time over a collection of sample paths, regressing basis functions against point estimates of continuation values. Longstaff and Schwartz (2001) use point estimates of lower bounds derived from suboptimal future policies. We, on the other hand, use point estimates of upper bounds derived from the PO linear program (9). As we shall see in §4, despite the similarities, the PO-derived policy can offer significant improvements in practice.

[3] In our experimental work we used $\psi_{l,t}(\cdot) = \phi_l(\cdot, t)$. In other words, we used the same basis function architecture to approximate continuation values as was used for value functions.

4. Computational Results

In this section, we will illustrate the performance of the PO method versus a collection of competitive benchmark algorithms in numerical experiments. We begin by defining the benchmark algorithms in §4.1. In §4.2, we define the problem setting, which is that of pricing a high-dimensional Bermudan option. Implementation details such as the choice of basis functions and the state sampling parameters are given in §4.3. Finally, the results are presented in §4.4.

4.1. Benchmark Methods

The landscape of techniques available for pricing high-dimensional options is rich; a good overview of these is available from Glasserman (2004, Chap. 8). We consider the following benchmarks, representative of mainstream methods, for purposes of comparison with the PO method:

- *Lower Bound Benchmark.* The line of work developed by Carriere (1996), Tsitsiklis and Van Roy (2001), and Longstaff and Schwartz (2001) seeks to produce approximations to the optimal continuation value function. These approximations are typically weighted combinations of prespecified basis functions that are fit via a regression-based methodology. The greedy policies with respect to these approximations yield lower bounds on price.

We generate a continuation value approximation $\hat{C}$ using the Longstaff and Schwartz (2001) (LS) method. Details are available from Glasserman (2004, Chap. 8, p. 461). We simulate the greedy policy with respect to this approximation to generate lower bounds. We refer to this approach as LS-LB.

- *Upper Bound Benchmarks.* The martingale duality approach, originally proposed for this task by Rogers (2002) and Haugh and Kogan (2004), is widely used for upper bounds. Recall from §2.1 that a martingale for use in the duality approach is computed using the optimal value function, and extant heuristics use surrogates that approximate the optimal value function. We consider the following surrogates:

1. DVF-UB. This is a dual value function approach that derives a value function approximation from the continuation value approximation of the LS-LB regression pricing procedure. In particular, given the LS-LB continuation value approximation, $\hat{C}$, we generate a value function approximation $\hat{V}$ according to

$$\hat{V}_t(x) \triangleq \max\{g(x), \hat{C}_t(x)\}, \quad \forall x \in \mathscr{X},\ t \in \mathscr{T}.$$

This approach is described by Glasserman (2004, §8.7, p. 473).

2. DP-UB. This is a dual policy approach that derives a value function approximation from the policy suggested by the LS-LB regression pricing procedure. In particular, let $\hat{\mu}$ denote the greedy policy

derived from the LS-LB continuation value approximation $\hat{C}$, i.e., for all states x and times t,

$$\hat{\mu}_t(x) \triangleq \begin{cases} \text{CONTINUE} & \text{if } t < d \text{ and } g(x) < \hat{C}_t(x), \\ \text{STOP} & \text{otherwise.} \end{cases}$$

Define $V_t^{\hat{\mu}}(x)$ as the value of using the policy $\hat{\mu}$ starting at state x in time t. The quantity $V_t^{\hat{\mu}}(x)$ can be computed via an inner Monte Carlo simulation over paths that start at time t in state x. This can then be used as a value function surrogate to derive a martingale for the duality approach. This approach was introduced by Andersen and Broadie (2004), and a detailed description is available from Glasserman (2004, §8.7, pp. 474–475).

The LS-LB, DVF-UB, and DP-UB methods described above will be compared with upper bounds computed with the PO method (PO-UB) and their corresponding lower bounds (PO-LB), as described in §3. Further implementation details for each of these techniques will be provided in §4.3.

4.2. Problem Setting

We consider a Bermudan option over a calendar time horizon T defined on multiple assets. The option has a total of d exercise opportunities at calendar times $\{\delta, 2\delta, \ldots, \delta d\}$, where $\delta \triangleq T/d$. The payoff of the option corresponds to that of a call option on the maximum of n non-dividend-paying assets with an *up-and-out* barrier. We assume a Black–Scholes framework, where risk-neutral asset price dynamics for each asset j are given by a geometric Brownian motion, i.e., the price process $\{P_s^j, s \in \mathbb{R}_+\}$ follows the stochastic differential equation

$$dP_s^j = rP_s^j\,ds + \sigma_j P_s^j\,dW_s^j. \tag{13}$$

Here, r is the continuously compounded risk-free interest rate, σ_j is the volatility of asset j, W_s^j is a standard Brownian motion, and the instantaneous correlation of each pair W_s^j and $W_s^{j'}$ is $\rho_{jj'}$. Let $\{p_t, 0 \le t \le d\}$ be the discrete-time process obtained by sampling P_s at intervals of length δ, i.e., $p_t^j \triangleq P_{\delta t}^j$ for each $0 \le t \le d$. On the discrete time scale indexed by t, the possible exercise times are given by $\mathcal{T} \triangleq \{1, 2, \ldots, d\}$, and the discount factor is given by $\alpha \triangleq e^{-r\delta}$.

The option is "knocked out" (and worthless) at time t if, at any of the times preceding and including t, the maximum of the n asset prices exceeded the barrier B. We let $y_t \in \{0, 1\}$ serve as an indicator that the option is knocked out at time t. In particular, $y_t = 1$ if the option has been knocked out at time t or at some time prior, and $y_t = 0$ otherwise. The $\{y_t\}$ process evolves according to

$$y_t = \begin{cases} \mathbb{I}_{\{\max_{1\le j\le n} p_0^j \ge B\}} & \text{if } t = 0, \\ y_{t-1} \vee \mathbb{I}_{\{\max_{1\le j\le n} p_t^j \ge B\}} & \text{otherwise.} \end{cases}$$

A state in the associated stopping problem is then given by the tuple $x \triangleq (p, y) \in \mathbb{R}^n \times \{0, 1\}$, and the payoff function is defined according to

$$g(x) \triangleq \Big(\max_j p_i(x) - K\Big)^+ (1 - y(x)).$$

where $y(x)$ and $p_j(x)$, respectively, are the knock-out indicator and the jth price coordinates of the composite state x.

4.3. Implementation Details

4.3.1. Basis Functions. We use the following set of $n+2$ basis functions:

$$\phi_1(x, t) = (1 - y(x)), \quad \phi_2(x, t) = g(x),$$
$$\phi_{j+2}(x, t) = (1 - y(x))p_j(x), \quad \forall\, 1 \le j \le n.$$

Described succinctly, our basis function architecture consists of a constant function, the payoff function, and linear functions of each asset price, where we have further ensured that each basis function takes the value zero in states where the option is knocked out. This is because zero is known to be the exact value of the option in such states. Note that many other basis functions are possible. For instance, the prices of barrier options on each of the individual stocks seems like a particularly appropriate choice. We have chosen a relatively generic basis architecture, however, to disentangle the study of the pricing methodology from the goodness of a particular tailor-made architecture.

4.3.2. State Sampling. Both the PO method as well as the benchmark methods require sampling states from the underlying Markov chain; however, their requirements tend to be different. In particular, the LS-LB procedure requires only outer sample paths, DVF-UB and PO-UB require outer sample paths with shallow inner sampling (next state samples), and DP-UB requires outer sample paths with deep inner sampling (sample paths simulated till the option expires or gets exercised). In general, it may be possible to judiciously choose the sampling parameters to, for example, optimize the accuracy of a method given a fixed computational budget, and that such a good choice of parameters will likely vary from method to method. We have not attempted such an optimization. For LS-LB and DP-UB, we have chosen parameters that generally follow those chosen by Andersen and Broadie (2004), and for DVF-UB and PO-UB, parameters were chosen so that the resulting standard error is comparable to that for DP-UB. In this sense, our choice of parameters represents an "apples-to-apples" comparison. Our parameter settings are listed below:

- LS-LB. This approach requires sample paths of the underlying Markov process to run the regression

procedure. We used 200,000 sample paths for the regression. The greedy policy with respect to the regressed continuation values was evaluated over 2,000,000 sample paths.

• PO-UB. In the notation of §3.1, we solved the LP (9) using $S = 30{,}000$ outer sample paths and $I = 500$ next state inner samples for one-step expectation computations. Given a solution, $\hat{r}_{\mathsf{PO}}$, we evaluated $F_0\Phi\hat{r}_{\mathsf{PO}}(x_0)$ using a distinct set of $S = 30{,}000$ outer sample paths, with $I = 500$ inner samples for one-step expectations.

• PO-LB. The policy here is constructed using computations entailed in the PO-UB method. We evaluate this policy to compute the lower bound using the same set of 2,000,000 sample paths used for the evaluation of LS-LB above.

• DVF-UB. As discussed earlier, a value function estimate $\hat{V}$ is obtained from the continuation value estimates of the regression procedure used for LS-LB above. We then estimate the DVF-UB upper bound, $F_0\hat{V}(x_0)$, using the same set of 30,000 sample paths and one-step samples in the evaluation of PO-UB above.

• DP-UB. As discussed earlier, this approach uses the value function approximation $V^{\hat{\mu}}$. We obtain continuation value estimates $\hat{C}$ via the regression computation for LS-LB. We estimate the upper bound $F_0V^{\hat{\mu}}(x_0)$ using 3,000 sample paths;[4] we evaluate $V^{\hat{\mu}}$ at each point along these sample paths using 10,000 inner sample paths.

4.4. Results

In the numerical results that follow, the following common problem settings were used:[5]

• strike price $K = 100$; knock-out barrier price $B = 170$; time horizon, $T = 3$ years;

• risk-free rate $r = 5\%$ (annualized); volatility $\sigma_j = 20\%$ (annualized).

In Table 1, we see the upper and lower bounds produced by the PO approach and the benchmark schemes described above. Here, we vary the number of assets n and the initial price $p_0^j = \bar{p}_0$ common to all assets, and the assets are uncorrelated ($\rho_{jj'} = 0$). Standard errors are in parentheses. We report average upper and lower bounds on the option price over 10 trials. In §C of the online supplement, we provide additional results where the number of exercise opportunities d and the asset price correlation matrix ρ are varied. Taken together, we make the following broad conclusions from these experimental results:

• *Lower Bound Quality.* The PO-LB method provides substantially better exercise policies than does the LS-LB procedure and consequently tighter lower bounds. The exercise policies provide an improvement of over 100 basis points in most of the experiments; in some cases the gain was as much as 200 basis points.

• *Upper Bound Quality.* The DVF-UB upper bounds are the weakest, whereas the DP-UB upper bounds are typically the strongest. The gap between these two bounds was typically on the order of 100 basis points. The upper bound produced via the PO-UB method was of intermediate quality, but typically recovered approximately 60% of the gap between the DVF-UB and DP-UB upper bounds.

Table 2 summarizes relative computational requirements of each method. Note that for the dual upper bound methods we report the time to compute both upper and lower bounds. This is for consistency, because for the DVF-UB and DP-UB methods, the LS-LB continuation value estimate is required and must be computed first. The running times are typically dominated by sampling requirements and can be broken down as follows:

• LS-LB. The LS-LB method requires only the generation of outer sample paths and is thus the fastest.

• LS-LB + DVF-UB. Along each outer sample path, the DVF-UB method requires generation of inner samples for the next state.

• PO-LB + PO-UB. For the PO-UB method, the structure of the LP (9) permits extremely efficient solution via an interior point method as discussed in §3.1; the computation time is dominated by sampling rather than optimization. Qualitatively, the sampling requirements for the PO-UB method are the same as that of DVF-UB: next state inner samples are needed. However, to generate an unbiased estimate, the PO-UB method requires one set of sample paths for optimization and a second set of sample paths for evaluation of the upper bound estimate. Hence, PO-UB takes about twice the computational time of DVF-UB.

• LS-LB + DP-UB. The inner simulation requirements for DP-UB result in that method requiring an order of magnitude more time than either of the other upper bound approaches. This is because, along each outer sample path, inner samples are needed not just for one time step, but for an entire trajectory until the option is knocked out or exercised.

To summarize, these experiments demonstrate the two primary merits to using the PO method to produce upper and lower bounds:

1. The PO-UB method produces upper bounds that are superior to the DVF-UB method and, in many

[4] Andersen and Broadie (2004) used 1,500 sample paths. We chose the larger number to obtain standard errors comparable to the other approaches in the study.

[5] Note that all the parameter choices here are symmetric across assets, and hence the assets are identical in the problems we consider. However, this symmetry was not exploited in our implementations.

Table 1 A Comparison of the Lower and Upper Bound Estimates of the PO and Benchmarking Methods as a Function of the Common Initial Asset Price $p_0^j = \bar{p}_0$ and the Number of Assets n

(a) Upper and lower bounds, with standard errors

$\bar{p}_0$	LS-LB	S.E.	PO-LB	S.E.	DP-UB	S.E.	PO-UB	S.E.	DVF-UB	S.E.
					$n = 4$ assets					
90	32.754	(0.005)	33.011	(0.011)	34.989	(0.014)	35.117	(0.026)	35.251	(0.013)
100	40.797	(0.003)	41.541	(0.009)	43.587	(0.016)	43.853	(0.027)	44.017	(0.011)
110	46.929	(0.003)	48.169	(0.004)	49.909	(0.016)	50.184	(0.017)	50.479	(0.008)
					$n = 8$ assets					
90	43.223	(0.005)	44.113	(0.009)	45.847	(0.016)	46.157	(0.037)	46.311	(0.015)
100	49.090	(0.004)	50.252	(0.006)	51.814	(0.023)	52.053	(0.027)	52.406	(0.014)
110	52.519	(0.005)	53.488	(0.007)	54.890	(0.020)	55.064	(0.019)	55.513	(0.005)
					$n = 16$ assets					
90	49.887	(0.003)	50.885	(0.006)	52.316	(0.020)	52.541	(0.010)	52.850	(0.011)
100	52.879	(0.001)	53.638	(0.004)	54.883	(0.020)	55.094	(0.016)	55.450	(0.013)
110	54.620	(0.002)	55.146	(0.003)	56.201	(0.009)	56.421	(0.016)	56.752	(0.007)

(b) Relative values of bounds

$\bar{p}_0$	(PO-LB) − (LS-LB)	(%)	(PO-UB) − (DP-UB)	(%)	(DVF-UB) − (PO-UB)	(%)
			$n = 4$ assets			
90	0.257	0.78	0.127	0.39	0.134	0.41
100	0.744	1.82	0.266	0.65	0.164	0.40
110	1.240	2.64	0.275	0.59	0.295	0.63
			$n = 8$ assets			
90	0.890	2.06	0.310	0.72	0.154	0.36
100	1.162	2.37	0.239	0.49	0.353	0.72
110	0.970	1.85	0.174	0.33	0.450	0.86
			$n = 16$ assets			
90	0.998	2.00	0.225	0.45	0.308	0.62
100	0.759	1.43	0.210	0.40	0.356	0.67
110	0.526	0.96	0.220	0.40	0.331	0.61

Notes. For each algorithm, the mean and standard error (S.E.; over 10 independent trials) is reported. The number of exercise opportunities was $d = 54$, and the common correlation was $\rho_{jj'} 0 = \bar{\rho} = 0$. Percentage relative values are expressed relative to the LS-LB lower bound.

cases, of comparable quality to the state-of-the-art DP-UB method. However, the PO-UB method requires an order of magnitude less computational effort than the DP-UB approach and is highly practical.

2. The PO-LB method produces substantially superior exercise policies relative to the LS-LB method. These policies are effectively a by-product of the upper bound computation.

Table 2 Relative Time Values for Different Algorithms for the Stopping Problem Setting of Table 1 with $n = 16$ Assets

Method	Time (normalized)
LS-LB (lower bound only)	1.0
LS-LB + DVF-UB (upper and lower bounds)	3.6
PO-LB + PO-UB (upper and lower bounds)	6.8
LS-LB + DP-UB (upper and lower bounds)	51.7

Note. Here, all times are normalized relative to that required for the computation of the LS-LB lower bound. All computations were single threaded and performed on an Intel Xeon E5620 2.40 GHz CPU with 64 GB of RAM. The PO-UB linear program was solved with IBM ILOG CPLEX 12.1.0 optimization software.

5. Theory

In this section, we will seek to provide theoretical guarantees for the martingale penalty approach in general as well as specific guarantees for the PO method.

Note that our setting here will be that of an optimal stopping problem that is discounted, stationary, and has an infinite horizon. This will yield us considerably simpler notation and easier statement of results and is also consistent with other theoretical literature on ADP for optimal stopping problems (e.g., Tsitsiklis and Van Roy 1999, Van Roy 2010). Many of our results in the aforementioned setting have finite horizon, nonstationary analogues, and the intuition derived from these results carries over to the nonstationary setting. In particular, we establish two theorems for the stationary setting and outline their nonstationary analogues in §B of the online supplement. Our stationary setting is introduced in §5.1.

Our first class of theoretical results are *approximation guarantees*. These guarantee the quality of an upper

bound derived from the martingale duality approach, relative to error in approximating the value function. A crucial parameter for our bounds measures the "predictability" of a Markov chain; this is introduced in §5.2. In §5.4, we develop an approximation guarantee that applies generically to martingale duality upper bounds and discuss the structural properties of optimal stopping problems that impact this bound. In §5.5, we develop a *relative* guarantee that is specific to the PO method; this guarantees the quality of the PO upper bound relative to the best approximation of the true value function within the span of the basis functions. In §5.6, we compare our guarantees to similar guarantees that have been developed for ADP lower bounds.

Our second class of theoretical results are *comparison bounds*, developed in §5.7. Here, we compare the upper bounds arising to the PO approach to other upper bounds that have been developed using ADP techniques based in linear programming. In this case, the upper bounds can be compared on a problem instance by problem instance basis, and we show that the PO method dominates the alternatives.

5.1. Preliminaries

Consider a discrete-time Markov chain with state $x_t \in \mathcal{X} \subset \mathbb{R}^n$ at each time $t \in \{0, 1, \ldots\}$. Denote by P the transition kernel of the chain. Assume that the chain is ergodic, with stationary distribution π. Let $\mathcal{F} \triangleq \{\mathcal{F}_t\}$ be the natural filtration generated by the process $\{x_t\}$, i.e., for each time t, $\mathcal{F}_t \triangleq \sigma(x_0, x_1, \ldots, x_t)$.

Given a function $g\colon \mathcal{X} \to \mathbb{R}$, we define the payoff of stopping when the state is x_t as $g(x_t)$. We define $\mathcal{P}$ to be the set[6] of real-valued functions $V\colon \mathcal{X} \to \mathbb{R}$ of the state space with $\mathsf{E}_\pi[V(x_0)^2] < \infty$. Here, E_π denotes expectation with respect to the stationary distribution. We assume that $g \in \mathcal{P}$. We are interested in maximizing the expected discounted payoff of stopping. In particular, given an initial state $x \in \mathcal{X}$, define the optimal value function

$$J^*(x) \triangleq \sup_\tau \mathsf{E}[\alpha^\tau g(x_\tau) \mid x_0 = x].$$

Here, the supremum is taken over all $\mathcal{F}$-adapted stopping times τ, and $\alpha \in [0, 1)$ is the discount factor.

We will abuse notation to also consider the transition kernel as a one-step expectation operator $P\colon \mathcal{P} \to \mathcal{P}$, defined by

$$(PJ)(x) \triangleq \mathsf{E}[J(x_{t+1}) \mid x_t = x], \quad \forall x \in \mathcal{X}.$$

Given a function $J \in \mathcal{P}$, define the *Bellman operator* $T\colon \mathcal{P} \to \mathcal{P}$ by

$$(TJ)(x) \triangleq \max\{g(x), \alpha PJ(x)\}, \quad \forall x \in \mathcal{X}.$$

[6] Note that earlier we defined $\mathcal{P}$ to be the set of real-valued functions of state and time. In the stationary infinite horizon setting, it suffices to consider only functions of state.

Observe that the optimal value function is the unique fixed point $TJ^* = J^*$.

To define the pathwise optimization approach in this setting, we first define the *martingale difference operator* Δ. The operator Δ maps a function $J \in \mathcal{P}$ to a function $\Delta J\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where

$$\Delta J(x_t, x_{t-1}) \triangleq J(x_t) - PJ(x_{t-1}), \quad \forall x_{t-1}, x_t \in \mathcal{X}.$$

Observe that, for any J, the process $\{\Delta J(x_t, x_{t-1}), t \geq 1\}$ is a martingale difference sequence. Now, for each J, the *martingale duality upper bound operator* $F\colon \mathcal{P} \to \mathcal{P}$ is given by

$$(FJ)(x) \triangleq \mathsf{E}\left[\sup_{s \geq 0} \alpha^s g(x_s) - \sum_{t=1}^{s} \alpha^t \Delta J(x_t, x_{t-1}) \,\middle|\, x_0 = x\right], \quad \forall x \in \mathcal{X}.$$

The following lemma establishes that the F operator yields dual upper bounds to the original problem; the proof follows along the lines of the proof of Lemma 1, found in §A.1 of the online supplement, and is omitted:

Lemma 3 (Infinite Horizon Martingale Duality). (i) (Weak Duality). *For any function $J \in \mathcal{P}$ and all $x \in \mathcal{X}$, $J^*(x) \leq FJ(x)$.*

(ii) (Strong Duality). *For all $x \in \mathcal{X}$, $J^*(x) = FJ^*(x)$.*

To find a good upper bound, we begin with collection of K basis functions

$$\Phi \triangleq \{\phi_1, \phi_2, \ldots, \phi_K\} \subset \mathcal{P}.$$

Given a weight vector $r \in \mathbb{R}^K$, define the function $\Phi r \in \mathcal{P}$ as the linear combination

$$(\Phi r)(x) \triangleq \sum_{l=1}^{k} r_l \phi_l(x), \quad \forall x \in \mathcal{X}.$$

We will seek to find functions within the span of the basis Φ that yields the tightest *average* upper bound. In other words, we will seek to solve the optimization problem

$$\underset{r}{\text{minimize}}\ \mathsf{E}_\pi[F\Phi r(x_0)]. \tag{14}$$

As before, this optimization problem is an unconstrained minimization of a convex function.

5.2. Predictability

Our approximation guarantees incorporate a notion of predictability of the underlying Markov chain, which we will define in this section. First, we begin with some notation. For functions $J, J' \in \mathcal{P}$, define the inner product

$$\langle J, J' \rangle_\pi \triangleq \mathsf{E}_\pi[J(x_0)J'(x_0)].$$

Similarly, define the norms

$$\|J\|_{p,\pi} \triangleq (\mathsf{E}_\pi[|J(x_0)|^p])^{1/p}, \quad \forall p \in \{1,2\},$$
$$\|J\|_\infty \triangleq \sup_{x\in\mathscr{X}} |J(x)|;$$

define $\mathrm{Var}_\pi(J)$ to be the variance of $J(x)$ under the distribution π, i.e.,

$$\mathrm{Var}_\pi(J) \triangleq \mathsf{E}_\pi[(J(x_0) - \mathsf{E}_\pi[J(x_0)])^2];$$

and define the one-step conditional variance

$$\mathrm{Var}(J(x_1) \mid x_0) \triangleq \mathsf{E}\big[(J(x_1) - \mathsf{E}[J(x_1) \mid x_0])^2 \mid x_0\big].$$

The following quantity, a property of the transition kernel P, will be important for our analysis:

$$\lambda(P) = \sup_{J\in\mathscr{P},\, J\neq 0} \left(\frac{\mathsf{E}_\pi[\mathrm{Var}(J(x_1) \mid x_0)]}{\mathrm{Var}_\pi(J)}\right)^{1/2}. \tag{15}$$

To interpret $\lambda(P)$, note that, by the law of total variance and the fact that π is the stationary distribution, for $J \in \mathscr{P}$,

$$\begin{aligned}\mathsf{E}_\pi[\mathrm{Var}(J(x_1) \mid x_0)] &= \mathrm{Var}_\pi(J(x_1)) - \mathrm{Var}_\pi(\mathsf{E}[J(x_1) \mid x_0]) \\ &= \mathrm{Var}_\pi(J(x_0)) - \mathrm{Var}_\pi(\mathsf{E}[J(x_1) \mid x_0]) \\ &\leq \mathrm{Var}_\pi(J); \end{aligned} \tag{16}$$

thus, $\lambda(P) \in [0,1]$. By the definition (15), for all $J \in \mathscr{P}$,

$$\mathsf{E}_\pi[\mathrm{Var}(J(x_1) \mid x_0)] \leq \lambda(P)^2 \mathrm{Var}_\pi(J).$$

Suppose that $\lambda(P) \approx 0$. Then, for all J,

$$\mathsf{E}_\pi[\mathrm{Var}(J(x_1) \mid x_0)] \ll \mathrm{Var}_\pi(J).$$

In this case, for *all* J, the average uncertainty of $J(x_1)$ conditioned on the previous state x_0 is much less than the unconditional uncertainty of $J(x_1)$, i.e., when x_1 is distributed according to its prior distribution, which is the stationary distribution. For such chains, the state of the Markov chain x_0 gives significant information about all functionals of the subsequent process state x_1, and thus, for all intents and purposes, significant information about the subsequent state x_1 itself. Alternatively, suppose that $\lambda(P) \approx 1$. Then, there exists *some* J such that

$$\mathsf{E}_\pi[\mathrm{Var}(J(x_1) \mid x_0)] \approx \mathrm{Var}_\pi(J).$$

In this case, knowledge of the state x_0 does not meaningfully reduce the uncertainty of $J(x_1)$. Motivated by these cases, we interpret $\lambda(P)$ as a measure of *predictability*, and we will call Markov chains where $\lambda(P) \approx 0$ *predictable*.

Predictability is important because it provides a bound on the operator norm of the martingale difference operator Δ. When a Markov chain is predictable, it may be possible to approximate a particular martingale difference, say ΔJ^*, by some other martingale difference, say ΔJ, even if J^* is not particularly well approximated by J. This is captured in the following lemma:

Lemma 4. *Given functions $J, J' \in \mathscr{P}$, define a distance between the martingale differences ΔJ, $\Delta J'$ by*

$$\|\Delta J - \Delta J'\|_{2,\pi} \triangleq \sqrt{\mathsf{E}_\pi[|\Delta J(x_1,x_0) - \Delta J'(x_1,x_0)|^2]}.$$

Then,

$$\|\Delta J - \Delta J'\|_{2,\pi} \leq \lambda(P)\sqrt{\mathrm{Var}_\pi(J - J')}.$$

Proof. Set $W \triangleq J - J'$, and observe that

$$\begin{aligned}\|\Delta W\|_{2,\pi}^2 &= \mathsf{E}_\pi[(W(x_1) - \mathsf{E}[W(x_1) \mid x_0])^2] \\ &= \mathsf{E}_\pi[\mathrm{Var}(W(x_1) \mid x_0)] \leq \lambda(P)^2 \mathrm{Var}_\pi(W).\end{aligned}$$

The result follows. □

5.3. Example of a Predictable Chain

In this section, we will provide an alternative, spectral characterization of predictability. We will use this characterization to illustrate a naturally arising example of a predictable Markov chain, namely, a chain where the calendar time between transitions is short.

To begin, recall that P is the transition kernel of the Markov chain, which we also interpret as a one-step expectation operator. Define P^* to be the adjoint of P with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$. In the case of a finite or countable state space, P^* can be written explicitly according to

$$P^*(y,x) \triangleq \frac{\pi(x)P(x,y)}{\pi(y)}, \quad \forall x, y \in \mathscr{X}.$$

Note that P^* is the *time reversal* of P; it corresponds to the transition kernel of the Markov chain running backward in time.

The following lemma, the proof of which is provided in §A.3 of the online supplement, provides a spectral characterization of the predictability of P:

Lemma 5. *Suppose that the state space $\mathscr{X}$ is finite. Then,*

$$\lambda(P) = \sqrt{\rho(I - P^*P)},$$

where $\rho(\cdot)$ is the spectral radius. Furthermore, if P is time-reversible (i.e., if $P = P^*$), then

$$\lambda(P) = \sqrt{\rho(I - P^2)} \leq \sqrt{2\rho(I - P)}.$$

Observe that the matrix P^*P, known as a *multiplicative reversiblization* (Fill 1991), corresponds to a transition one step backward in time in the original Markov chain, followed by an independent step forward in time. Suppose for the moment that the Markov chain is reversible, i.e., that $P = P^*$. Then, by Lemma 5, $\lambda(P)$ will be small when $I \approx P$, or, the state x_{t+1} at time $t+1$ in the Markov chain is approximated well by the current state x_t. In other words, the Markov chain is closer to a deterministic process.

The spectral analysis of $I - P^*P$ is also important in the study of mixing times, or the rate of convergence of a Markov chain to stationarity. In that context, one is typically concerned with the *smallest* nonzero eigenvalue (see, e.g., Montenegro and Tetali 2006); informally, if this is large, the chain is said to be *fast mixing*. In the present context, we are interested in the *largest* eigenvalue, which is small in the case of a predictable chain. Thus, our predictable chains necessarily mix slowly.

One class of predictable Markov chains occurs when the calendar time scale between successive stopping opportunities is small:

Example 1 (Sampled State Dynamics). Suppose that the Markov chain $\{x_t\}$ takes the form $x_t = z_{t\delta}$ for all integers $t \geq 0$, where $\delta > 0$ and $\{z_s \in \mathscr{X}, s \in \mathbb{R}_+\}$ is a continuous time Markov chain with generator Q over a finite state space $\mathscr{X}$. In other words, $\{x_t\}$ are discrete-time samples of an underlying continuous time chain over time scales of length δ. In this case, the transition probabilities take the form $P = e^{Q\delta}$ and $P^* = e^{Q^*\delta}$. As $\delta \to 0$,

$$\lambda(P) = \sqrt{\rho(I - e^{Q^*\delta}e^{Q\delta})} = \sqrt{\delta\rho(Q^* + Q)} + o(\sqrt{\delta}) \to 0.$$

5.4. Upper Bound Guarantees

Lemma 3 establishes that, given a function $J \in \mathscr{P}$, FJ is an upper bound on J^*, and that if $J = J^*$, this upper bound is tight. Hence, it seems reasonable to pick J to be a good approximation of the optimal value function J^*. In this section, we seek to make this intuition precise. In particular, we will provide a guarantee on the quality of the upper bound, that is, a bound on the distance between FJ and J^*, as a function of the quality of the value function approximation J and other structural features of the optimal stopping problem.

The following lemma provides the key result for our guarantee. It characterizes the difference between two upper bounds FJ and FJ' that arise from two different value function approximations $J, J' \in \mathscr{P}$. The proof is provided in §A.3 of the online supplement.

Lemma 6. *For any pair of functions $J, J' \in \mathscr{P}$,*

$$\|FJ - FJ'\|_{2,\pi} \leq \frac{R(\alpha)\alpha}{\sqrt{1-\alpha}}\lambda(P)\sqrt{\mathrm{Var}_\pi(J - J')},$$

where $R\colon [0,1) \to [1, \sqrt{5/2}]$ is a bounded function given by

$$R(\alpha) \triangleq \min\left\{\frac{1}{\sqrt{1-\alpha}}, \frac{2}{\sqrt{1+\alpha}}\right\}.$$

Taking $J' = J^*$ in Lemma 6, we immediately have the following:

Theorem 2. *For any function $J \in \mathscr{P}$,*

$$\|FJ - J^*\|_{2,\pi} \leq \frac{R(\alpha)\alpha}{\sqrt{1-\alpha}}\lambda(P)\sqrt{\mathrm{Var}_\pi(J - J^*)}. \tag{17}$$

Theorem 2 provides a guarantee on the upper bound FJ arising from an arbitrary function J. It is reminiscent of the upper bound guarantee of Chen and Glasserman (2007). In the present (discounted and infinite horizon) context, their upper bound guarantee can be stated as

$$\|FJ - J^*\|_\infty \leq \frac{4\alpha}{\sqrt{1-\alpha^2}}\|J - J^*\|_\infty. \tag{18}$$

It what follows, we will compare these two bounds, as well as identify the structural features of the optimal stopping problem and the function J that lead to a tight upper bound FJ. In particular, notice that the right-hand side of the guarantee in Theorem 2 can be decomposed into three distinct components:

- *Value Function Approximation Quality.* Theorem 2 guarantees that the closer the value function approximation J is to J^*, the tighter the upper bound FJ will be. Importantly, the distance between J and J^* is measured in terms of the standard deviation of their difference. Under this metric, the relative importance of accurately approximating J^* in two different states is commensurate to their relative probabilities. On the other hand, the guarantee (18) requires a *uniformly* good approximation of J^*. In a large state space, this can be challenging.
- *Time Horizon.* Theorem 2 has dependence on the discount factor α. In typical examples, $\alpha \approx 1$, and hence we are most interested in this regime.

One way to interpret α is as defining an effective time horizon. To be precise, consider an undiscounted stopping problem with the same state dynamics and reward function, but with a random finite horizon that is geometrically distributed with parameter α. We assume that the random time horizon is unknown to the decision maker, and that if the process is not stopped before the end of this time horizon, the reward is zero. This undiscounted, random but finite horizon formulation is mathematically equivalent to our discounted, infinite horizon problem. Hence, we define the *effective time horizon* $\mathsf{T}_{\mathsf{eff}}$ to be the expected length of the random finite time horizon, or

$$\mathsf{T}_{\mathsf{eff}} \triangleq \frac{1}{1-\alpha}. \tag{19}$$

The guarantee of Theorem 2 is $O(\sqrt{\mathsf{T}_{\mathsf{eff}}})$, i.e., it grows as the square root of the effective time horizon. This matches (18), as well as the original finite horizon bound of Chen and Glasserman (2007).

- *Predictability.* Theorem 2 isolates the dynamics of the Markov chain through the $\lambda(P)$ term; if $\lambda(P)$ is small, then the upper bound FJ will be tight. In other words, all else being equal, chains that are more predictable yield better upper bounds. In some sense, optimal stopping problems on predictable Markov chains are closer to deterministic problems to begin

with, and hence less care is needed in relaxing nonanticipativity constraints.

The dependence of Theorem 2 on predictability can be interpreted in the sampled state dynamics of Example 1. In this case, we assume that the transition probabilities of the Markov chain take the form $P = e^{Q\delta}$, where Q is the generator for a continuous time Markov chain, and $\delta > 0$ is the calendar time between successive stopping opportunities. In this setting, it is natural that the discount factor also scale as a function of the time interval δ, taking the form $\alpha = e^{-r\delta}$, where $r > 0$ is a continuously compounded interest rate. Then, as $\delta \to 0$,

$$\frac{R(\alpha)\alpha}{\sqrt{1-\alpha}}\lambda(P) = \sqrt{\frac{2\rho(Q^* + Q)}{r}} + o(1).$$

In this way, the premultiplying constants on the right-hand side of Theorem 2 remain bounded as the number of stopping opportunities is increased. This is *not* the case for (18).

5.5. Pathwise Optimization Approximation Guarantee

The result of §5.4 provides a guarantee on the upper bounds produced by the martingale duality approach given an arbitrary value function approximation J as input. When the value function approximation J arises from the PO method, we have the following result:

THEOREM 3. *Suppose that r_{PO} is an optimal solution for* (14). *Then*,

$$\|F\Phi r_{\mathsf{PO}} - J^*\|_{1,\pi} \le \frac{R(\alpha)\alpha}{\sqrt{1-\alpha}}\lambda(P)\min_r \sqrt{\mathrm{Var}_\pi(\Phi r - J^*)}.$$

PROOF. Observe that, for any $r \in \mathbb{R}^K$, by the optimality of r_{PO} and Lemma 3,

$$\begin{aligned}\|F\Phi r_{\mathsf{PO}} - J^*\|_{1,\pi} &= \mathsf{E}_\pi[F\Phi r_{\mathsf{PO}}(x_0) - J^*(x_0)] \\ &\le \mathsf{E}_\pi[F\Phi r(x_0) - J^*(x_0)] \\ &= \|F\Phi r - J^*\|_{1,\pi}.\end{aligned}$$

Because π is a probability distribution, $\|\cdot\|_{1,\pi} \le \|\cdot\|_{2,\pi}$, thus, applying Theorem 2,

$$\begin{aligned}\|F\Phi r_{\mathsf{PO}} - J^*\|_{1,\pi} &\le \|F\Phi r - J^*\|_{2,\pi} \\ &\le \frac{R(\alpha)\alpha}{\sqrt{1-\alpha}}\lambda(P)\sqrt{\mathrm{Var}_\pi(\Phi r - J^*)}.\end{aligned}$$

The result follows after minimizing the right-hand side over r. □

To compare Theorems 2 and 3, observe that Theorem 2 provides a guarantee that is a function of the distance between the value function approximation J and the optimal value function J^*. Theorem 3, on the other hand, provides a guarantee relative to the distance between the *best possible* approximation given the basis functions Φ and the optimal value function J^*. Note that it is not possible, in general, to directly compute this best approximation, which is the projection of J^* on to the subspace spanned by Φ, because J^* is unknown to begin with.

5.6. Comparison to Lower Bound Guarantees

It is instructive to compare the guarantees provided on upper bounds by Theorems 2 and 3 to guarantees that can be obtained on lower bounds derived from ADP methods. In general, the ADP approach to lower bounds involves identifying approximations to the optimal continuation value function C^*, which is related to the optimal value function J^* via

$$\begin{aligned}C^*(x) &= \alpha\mathsf{E}[J^*(x_{t+1}) \mid x_t = x], \\ J^*(x) &= \max\{g(x), C^*(x)\}, \quad \forall x \in \mathscr{X}.\end{aligned}$$

Given the optimal continuation function C^*, an optimal policy is defined via

$$\mu^*(x) \triangleq \begin{cases}\mathsf{CONTINUE} & \text{if } g(x) < C^*(x), \\ \mathsf{STOP} & \text{otherwise.}\end{cases}$$

In other words, μ^* stops when $g(x) \ge C^*(x)$.

Similarly, given an approximate continuation value function C, we can define the policy

$$\mu(x) \triangleq \begin{cases}\mathsf{CONTINUE} & \text{if } g(x) < C(x), \\ \mathsf{STOP} & \text{otherwise.}\end{cases}$$

The value function J^μ for this policy can be estimated via Monte Carlo simulation. Because J^* is the optimal value function, we have that $J^\mu(x) \le J^*(x)$ for every state x. In other words, J^μ is a lower bound to J^*.

Analogous to Theorem 2, Tsitsiklis and Van Roy (1999) establish that

$$\|J^* - J_\mu\|_{2,\pi} \le \frac{1}{1-\alpha}\|C - C^*\|_{2,\pi}. \tag{20}$$

Given a set of basis functions Φ, there are a number of ways to select a weight vector r so that the linear function Φr can be used as an approximate continuation value function. Methods based on approximate value iteration are distinguished by the availability of theoretical guarantees. Indeed, Van Roy (2010) establishes a result analogous to Theorem 3 for approximate value iteration, that

$$\begin{aligned}\|J^* - J_\mu\|_{1,\pi} &\le \|J^* - J_\mu\|_{2,\pi} \\ &\le \frac{L^*}{1-\alpha}\min_r \|\Phi r - C^*\|_{2,\pi},\end{aligned} \tag{21}$$

where $L^* \approx 2.17$.

Comparing (20) and (21) to Theorems 2 and 3, we see broad similarities: both sets of results provide guarantees on the quality of the lower (respectively,

upper) bounds produced, as a function of the quality of approximation of C^* (respectively, J^*). There are key differences, however. Defining the effective time horizon $\mathsf{T}_{\mathsf{eff}} \triangleq (1-\alpha)^{-1}$ as in §5.4, the premultiplying constants in the lower bound guarantees are $O(\mathsf{T}_{\mathsf{eff}})$, whereas the corresponding terms in our upper bound guarantees are $O(\sqrt{\mathsf{T}_{\mathsf{eff}}})$. Furthermore, Van Roy (2010) established that, for *any* ADP algorithm, a guarantee of the form (21) that applies over all problem instances must be linear in the effective time horizon. In this way, the upper bound guarantees of Theorems 2 and 3 have better dependence on the effective time horizon than is possible for lower bounds, independent of the choice of ADP algorithm. Furthermore, the upper bound guarantees highlight the importance of a structural property of the Markov chain, namely, predictability. There is no analogous term in the lower bound guarantees.

5.7. Comparison to Linear Programming Methods

We can compare upper bounds derived from the pathwise method directly to upper bounds derived from two other approximate dynamic programming techniques.

First, we consider the *approximate linear programming* approach. The ALP approach to ADP was introduced by Schweitzer and Seidmann (1985) and analyzed and further developed by de Farias and Van Roy (2003, 2004). The ALP is based on the LP formulation for the exact solution of a dynamic program due to Manne (1960). A testament to the success of the ALP approach is the number of applications it has seen in recent years in large-scale dynamic optimization problems. In our discounted, infinite horizon optimal stopping setting, the ALP approach involves finding a value function approximation within the span of the basis by solving the optimization program

$$\begin{aligned} \underset{r}{\text{minimize}} \quad & \mathsf{E}_c[\Phi r(x_0)] \\ \text{subject to} \quad & \Phi r(x) \geq g(x), \quad \forall x \in \mathscr{X}, \\ & \Phi r(x) \geq \alpha \mathsf{E}[\Phi r(x_{t+1}) \mid x_t = x], \quad \forall x \in \mathscr{X}. \end{aligned} \tag{22}$$

Here, c is a positive probability distribution over the state space know as the *state-relevance* distribution; it is natural (but not necessary) to take $c = \pi$. Note that (22) is a linear program and that, for each state x, the pair of linear constraints in (22) are equivalent to the Bellman inequality $\Phi r(x) \geq T\Phi r(x)$. Denote the set of feasible r by $\mathscr{C}_{\mathsf{ALP}} \subset \mathbb{R}^K$.

As we shall see momentarily, if $r \in \mathscr{C}_{\mathsf{ALP}}$ is feasible for the ALP (22), then Φr is a pointwise upper bound to the optimal value function J^*. The following theorem establishes that the martingale duality upper bound $F\Phi r$ is at least as good:

Theorem 4. *Suppose $r \in \mathscr{C}_{\mathsf{ALP}}$ is feasible for the ALP approach* (22). *Then, for all $x \in \mathscr{X}$,*

$$J^*(x) \leq F\Phi r(x) \leq \Phi r(x).$$

Proof. Using Lemma 3 and the definition of the constraint set $\mathscr{C}_{\mathsf{ALP}}$,

$$\begin{aligned} J^*(x) &\leq F\Phi r(x) \\ &= \mathsf{E}\Bigg[\sup_{s\geq 0} \alpha^s g(x_s) \\ &\qquad - \sum_{t=1}^{s} \alpha^t \big(\Phi r(x_t) - \mathsf{E}[\Phi r(x_t) \mid x_{t-1}]\big) \,\Big|\, x_0 = x\Bigg] \\ &= \mathsf{E}\Bigg[\sup_{s\geq 0} \alpha^s (g(x_s) - \Phi r(x_s)) + \Phi r(x_0) \\ &\qquad + \sum_{t=0}^{s-1} \alpha^t \big(\alpha \mathsf{E}[\Phi r(x_{t+1}) \mid x_t] - \Phi r(x_t)\big) \,\Big|\, x_0 = x\Bigg] \\ &\leq \mathsf{E}\Big[\sup_{s\geq 0} \Phi r(x_0) \,\Big|\, x_0 = x\Big] = \Phi r(x). \quad \square \end{aligned}$$

We can interpret the ALP approach (22) as finding an upper bound in the set $\{\Phi r, r \in \mathscr{C}_{\mathsf{ALP}}\}$ that is smallest on average, as measured according to the state-relevance distribution c. Alternatively, consider solving the pathwise optimization problem

$$\underset{r}{\text{minimize}} \, \mathsf{E}_c[F\Phi r(x_0)]. \tag{23}$$

Theorem 4 implies that the resulting martingale duality upper bound will be, on average, at least as good. In this way, the PO method dominates ALP.

Similarly, *smoothed approximate linear programming* (SALP) was recently introduced by Desai et al. (2012). In our present context, this seeks to solve the linear program

$$\begin{aligned} \underset{r,\, s}{\text{minimize}} \quad & \mathsf{E}_\pi\Big[\Phi r(x_0) + \frac{1}{1-\alpha} s(x_0)\Big] \\ \text{subject to} \quad & \Phi r(x) + s(x) \geq g(x), \quad \forall x \in \mathscr{X}, \\ & \Phi r(x) + s(x) \\ & \quad \geq \alpha \mathsf{E}[\Phi r(x_{t+1}) \mid x_t = x], \quad \forall x \in \mathscr{X}, \\ & s(x) \geq 0, \quad \forall x \in \mathscr{X}. \end{aligned} \tag{24}$$

Observe that (24) is a relaxation of (22) when $c = \pi$ that is formed by introducing a vector of slack variables $s \in \mathbb{R}^{\mathscr{X}}$. Desai et al. (2012) argue that this relaxation yields a number of theoretical benefits relative to the ALP approach, and demonstrate superior practical performance in a computational study.

The following lemma allows us to interpret SALP as an unconstrained convex minimization problem:

Lemma 7. *Given $J \in \mathscr{P}$, define the operator F_{SALP}: $\mathscr{P} \to \mathscr{P}$ by*

$$(F_{\mathsf{SALP}} J)(x) \triangleq \mathsf{E}\Big[J(x_0) + \sum_{t=0}^{\infty} \alpha^t (TJ(x_t) - J(x_t))^+ \,\Big|\, x_0 = x\Big], \quad \forall x \in \mathscr{X}.$$

Then, SALP (24) *is equivalent to the convex optimization problem*

$$\underset{r}{\text{minimize}}\ \mathsf{E}_\pi[F_{\mathsf{SALP}}\Phi r(x_0)]. \tag{25}$$

Proof. Suppose (r, s) is feasible for SALP (24). Then,

$$\begin{aligned}
&\mathsf{E}_\pi\left[\Phi r(x_0)+\frac{1}{1-\alpha}s(x_0)\right]\\
&\quad\geq \mathsf{E}_\pi\left[\Phi r(x_0)+\frac{1}{1-\alpha}(T\Phi r(x_0)-\Phi r(x_0))^+\right]\\
&\quad= \mathsf{E}_\pi\left[\Phi r(x_0)+\sum_{t=0}^{\infty}\alpha^t(T\Phi r(x_t)-\Phi r(x_t))^+\right]\\
&\quad= \mathsf{E}_\pi[F_{\mathsf{SALP}}\Phi r(x_0)],
\end{aligned}\tag{26}$$

where we use the constraints of (24) and the fact that π is the stationary distribution. Hence, r achieves at least the same objective value in (25). Conversely, for any r, define $s \triangleq (T\Phi r-\Phi r)^+$ componentwise. Then, (r, s) is feasible for (24), and (26) holds with equality. Thus, (r, s) achieves the same objective value in (24) as r in (25). □

The following theorem shows that the F_{SALP} operator also yields dual upper bounds to the optimal value function, analogous to the F operator in the pathwise method. Critically, however, the upper bounds of the pathwise method pointwise dominate those of the SALP method, which in turn pointwise dominate that of those ALP method.

Theorem 5. *For an arbitrary weight vector* $r \in \mathbb{R}^K$,

$$J^*(x) \leq F\Phi r(x) \leq F_{\mathsf{SALP}}\Phi r(x), \quad \forall x \in \mathscr{X}.$$

In addition, if $r \in \mathscr{C}_{\mathsf{ALP}}$, *i.e., r is feasible for the ALP approach* (22), *then*

$$J^*(x) \leq F\Phi r(x) \leq F_{\mathsf{SALP}}\Phi r(x) = \Phi r(x), \quad \forall x \in \mathscr{X}.$$

Proof. Given a weight vector $r \in \mathbb{R}^K$, by Lemma 3,

$$\begin{aligned}
J^*(x) &\leq F\Phi r(x) = \mathsf{E}\left[\sup_{s\geq 0}\alpha^s g(x_s)\right.\\
&\qquad\left.-\sum_{t=1}^{s}\alpha^t\left(\Phi r(x_t)-\mathsf{E}[\Phi r(x_t)\,|\,x_{t-1}]\right)\Big|\,x_0=x\right]\\
&= \mathsf{E}\left[\sup_{s\geq 0}\alpha^s(g(x_s)-\Phi r(x_s))+\Phi r(x_0)\right.\\
&\qquad\left.+\sum_{t=0}^{s-1}\alpha^t\left(\alpha\mathsf{E}[\Phi r(x_{t+1})\,|\,x_t]-\Phi r(x_t)\right)\Big|\,x_0=x\right]\\
&\leq \mathsf{E}\left[\sup_{s\geq 0}\alpha^s(g(x_s)-\Phi r(x_s))^+ +\Phi r(x_0)\right.\\
&\qquad\left.+\sum_{t=0}^{s-1}\alpha^t\left(\alpha\mathsf{E}[\Phi r(x_{t+1})\,|\,x_t]-\Phi r(x_t)\right)^+\Big|\,x_0=x\right]\\
&\leq \mathsf{E}\left[\sup_{s\geq 0}\Phi r(x_0)\right.\\
&\qquad\left.+\sum_{t=0}^{s}\alpha^t\left(T\Phi r(x_t)-\Phi r(x_t)\right)^+\Big|\,x_0=x\right]\\
&= F_{\mathsf{SALP}}\Phi r(x),
\end{aligned}$$

which completes the first part of the result. If $r \in \mathscr{C}_{\mathsf{ALP}}$, it immediately follows that $F_{\mathsf{SALP}}\Phi r(x) = \Phi r(x)$. □

In the context of the ALP and SALP optimization problems (22) and (24), Theorem 5 yields that

$$\begin{aligned}
\underset{r}{\text{minimize}}\ \mathsf{E}_\pi[F\Phi r(x_0)] &\leq \underset{r}{\text{minimize}}\ \mathsf{E}_\pi[F_{\mathsf{SALP}}\Phi r(x_0)]\\
&\leq \underset{r\in\mathscr{C}_{\mathsf{ALP}}}{\text{minimize}}\ \mathsf{E}_\pi[\Phi r(x_0)].
\end{aligned}$$

In other words, given a fixed set of basis functions, the PO method yields an upper bound that is on average at least as tight as that of the SALP method, which in turn yields an upper bound that is on average at least as tight at that of the ALP method.

6. Conclusion

We have presented what we believe is a practical scheme for high-dimensional pricing problems based on the martingale duality approach. In particular, we have attempted to show that the PO method can be used to compute upper bound on price of a quality comparable with state-of-the-art methods in a fraction of the time required for those methods. In addition, the approach yields, as a by-product, exercise policies that yield substantial improvements over policies derived via generic regression-based methods. There are several directions that merit further investigation; we point out two:

- *Implementation.* As opposed to solving an LP, one may imagine solving the minimization problem over weight vectors r in the PO method via a stochastic (sub)gradient method. In particular, define

$$\delta_l(r) \triangleq \mathsf{E}\left[-\sum_{p=1}^{s^*(r)}\alpha^p\Delta\phi_l(x_p, x_{p-1})\,\Big|\,x_0=x\right], \quad \forall 1\leq l\leq K,$$

where $s^*(r)$ is a random variable that, along each sample path, is a time that maximizes the inner optimization problem in the definition of $F_0\Phi r(x)$. It is not difficult to see that the vector $\delta(r)$ is a subgradient of $F_0\Phi r(x)$ with respect to r. Thus, very roughly, one might imagine a method that would update the r vector incrementally with each sampled path, $x^{(i)}$, according to an update rule of the form $r \leftarrow r+\gamma_i\delta^{(i)}(r)$. Here, $\gamma_i > 0$ is a step size, and $\delta^{(i)}(r)$ is a point estimate of the subgradient $\delta(x)$ evaluated over the single sample path $x^{(i)}$. Such a method has the advantage of not requiring an LP solver in addition to being online—the approach optimizes the upper bound simultaneously with sampling.

• *Policy Generation.* The policy used to generate our lower bounds required that we regress continuation value upper bounds implied by our approach against a set of basis functions. It is natural to ask whether a more direct method is possible—for instance, the greedy policy with respect to the Φr_{PO}. This appears to be a nontrivial question. In particular, it is not hard to see that if the constant function were a basis function, then the PO method could not identify a unique optimal coefficient for that basis function. On the other hand, if one chose to use a policy that was greedy with respect to Φr_{PO}, it is clear that the coefficient corresponding to this basis function could dramatically alter the nature of the policy.

Acknowledgments

The third author thanks Mark Broadie and Paul Glasserman for helpful discussions. The research of the second author was partially supported by a National Science Foundation CAREER award.

References

Andersen L, Broadie M (2004) Primal-dual simulation algorithm for pricing multidimensional American options. *Management Sci.* 50(9):1222–1234.

Belomestny D, Bender C, Schoenmakers J (2009) True upper bounds for Bermudan products via non-nested Monte Carlo. *Math. Finance* 19(1):53–71.

Bertsekas DP (2007) *Dynamic Programming and Optimal Control*, Vol. 2, 3rd ed. (Athena Scientific, Belmont, MA).

Borkar VS, Pinto J, Prabhu T (2009) A new learning algorithm for optimal stopping. *Discrete Event Dynamic Systems* 19(1):91–113.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Broadie M, Cao M (2008) Improved lower and upper bound algorithms for pricing American options by simulation. *Quant. Finance* 8(8):845–861.

Brown DB, Smith JE (2011) Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds. *Management Sci.* 57(10):1752–1770.

Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 58(4):785–801.

Carriere JF (1996) Valuation of the early-exercise price for derivative securities using simulations and splines. *Insurance: Math. Econom.* 19(1):19–30.

Chen N, Glasserman P (2007) Additive and multiplicative duals for American option pricing. *Finance Stochastics* 11(2):153–179.

Clément E, Lamberton D, Protter P (2002) An analysis of a least squares regression method for American option pricing. *Finance Stochastics* 6(4):449–471.

Davis M, Karatzas I (1994) A deterministic approach to optimal stopping. Kelly FP, ed. *Probability, Statistics, and Optimization: A Tribute to Peter Whittle* (John Wiley & Sons, New York), 455–466.

de Farias DP, Van Roy B (2003) The linear programming approach to approximate dynamic programming. *Oper. Res.* 51(6): 850–865.

de Farias DP, Van Roy B (2004) On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.* 293(3):462–478.

Desai VV, Farias VF, Moallemi CC (2012) Approximate dynamic programming via a smoothed linear program. *Oper. Res.* 60(3): 655–674.

Fill JA (1991) Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Probab.* 1(1):62–87.

Glasserman P (2004) *Monte Carlo Methods in Financial Engineering* (Springer-Verlag, New York).

Glasserman P, Yu B (2002) Simulation for American options: Regression now or regression later? Niederreiter H, ed. *Monte Carlo and Quasi-Monte Carlo Methods* (Springer-Verlag, New York), 213–226.

Haugh MB, Kogan L (2004) Pricing American options: A duality approach. *Oper. Res.* 52(2):258–270.

Jamshidian F (2003) Minimax optimality of Bermudan and American claims and their Monte-Carlo upper bound approximation. Technical report, NIB Capitial, The Hague, The Netherlands.

Lai G, Margot F, Secomandi N (2010) An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Oper. Res.* 58(3):564–582.

Longstaff FA, Schwartz ES (2001) Valuing American options by simulation: A simple least-sqaures approach. *Rev. Financial Stud.* 14(1):113–147.

Manne AS (1960) Linear programming and sequential decisions. *Management Sci.* 60(3):259–267.

Montenegro R, Tetali P (2006) *Mathematical Aspects of Mixing Times in Markov Chains* (NOW Publishers, Boston).

Rogers LCG (2002) Monte Carlo valuation of American options. *Math. Finance* 12(3):271–286.

Rogers LCG (2008) Pathwise stochastic optimal control. *SIAM J. Control Optim.* 46(3):1116–1132.

Rogers LCG (2010) Dual valuation and hedging of Bermudan options. *SIAM J. Financial Math.* 1(1):604–608.

Schweitzer P, Seidmann A (1985) Generalized polynomial approximations in Markovian decision processes. *J. Math. Anal. Appl.* 110(2):568–582.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM, Philadelphia).

Tsitsiklis JN, Van Roy B (1999) Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Trans. Automatic Control* 44(10):1840–1851.

Tsitsiklis JN, Van Roy B (2001) Regression methods for pricing complex American-style options. *IEEE Trans. Neural Networks* 12(4):694–703.

Van Roy B (2010) On regression-based stopping times. *Discrete Event Dynamic Systems* 20(3):307–324.

Yu H, Bertsekas DP (2007) A least squares Q-learning algorithm for optimal stopping problems. Technical Report 2731, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge.