Optimal Stopping of Markov Processes: Hilbert Space Theory, Approximation Algorithms, and an Application to Pricing High-Dimensional Financial Derivatives

John N. Tsitsiklis, *Fellow, IEEE*, and Benjamin Van Roy

***Abstract*—The authors develop a theory characterizing optimal stopping times for discrete-time ergodic Markov processes with discounted rewards. The theory differs from prior work by its view of per-stage and terminal reward functions as elements of a certain Hilbert space. In addition to a streamlined analysis establishing existence and uniqueness of a solution to Bellman's equation, this approach provides an elegant framework for the study of approximate solutions. In particular, the authors propose a stochastic approximation algorithm that tunes weights of a linear combination of basis functions in order to approximate a value function. They prove that this algorithm converges (almost surely) and that the limit of convergence has some desirable properties. The utility of the approximation method is illustrated via a computational case study involving the pricing of a path-dependent financial derivative security that gives rise to an optimal stopping problem with a 100-dimensional state space.**

***Index Terms*— Complex systems, curse of dimensionality, dynamic programming, function approximation, optimal stopping, stochastic approximation.**

I. Introduction

THE PROBLEM of optimal stopping is that of determining an appropriate time at which to terminate a process in order to maximize expected rewards. Examples arise in sequential analysis, the timing of a purchase or sale of an asset, and the analysis of financial derivatives. In this paper, we introduce a class of optimal stopping problems, provide a characterization of optimal stopping times, and develop a computational method for approximating solutions to problems for which classical methods become intractable. To illustrate the method, we present a computational case study involving the pricing of a (fictitious) high-dimensional financial derivative instrument.

Shiryaev [16] provides a fairly comprehensive treatment of optimal stopping problems. Under each of a sequence of increasingly general assumptions, he characterizes optimal stopping times and optimal rewards. We consider a rather restrictive class of problems relative to those captured by Shiryaev's analysis, but we employ a new line of analysis that leads to a simple characterization of optimal stopping times and, most importantly, the development of approximation algorithms. Furthermore, this line of analysis can be applied to other classes of optimal stopping problems, though the full extent of its breadth is not yet known.

In addition to providing a means for addressing large-scale optimal stopping problems, the approximation algorithm we develop plays a significant role in the broader context of stochastic control. In particular, the algorithm exemplifies simulation-based optimization techniques from the field of neuro-dynamic programming, pioneered by Barto, Sutton [17], and Watkins [22] that have been successfully applied to a variety of large-scale stochastic control problems; see Bertsekas and Tsitsiklis [6]. The practical success of these algorithms is not fully explained by existing theory, and our analysis represents progress toward an improved understanding. In particular, we prove the first convergence result involving the use of a variant of temporal-difference learning [17] to tune weights of general basis functions in order to approximately solve a control problem.

This paper is organized as follows. The next section defines the class of problems we consider (involving ergodic Markov processes with discounted rewards) and develops some basic theory concerning optimal stopping times and optimal rewards for such problems. Section III introduces and analyzes the approximation algorithm. A computational case-study involving the pricing of a financial derivative instrument is described in Section IV. Finally, extensions and connections between the ideas in this paper and the neuro-dynamic programming and reinforcement learning literature are discussed in a closing section. A preliminary version of some of the results of this paper, for the case of a finite state space, have been presented in [20] and are also included in [6].

II. An Optimal Stopping Problem and Its Solution

In this section, we define a class of optimal stopping problems involving stochastic processes that are Markov and ergodic, and we present an analysis that characterizes corresponding value functions and optimal stopping times. Though the results of this section are standard in flavor, the

Manuscript received May 30, 1997; revised July 2, 1998 and October 15, 1998. Recommended by Associate Editor, G. G. Yin. This work was supported by the NSF under Grant DMI-9625489 and the ARO under Grant DAAL-03-92-G-0115.

The authors are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jnt@mit.edu).

Publisher Item Identifier S 0018-9286(99)07846-0.

assumptions are not, as they are designed to accommodate the study of approximations, which will be the subject of Section III.

A. Assumptions and Main Result

We consider a stochastic process $\{x_t \mid t = 0, 1, 2, \cdots\}$ that evolves in a state space $\Re^d$, defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Each random variable x_t is measurable with respect to the Borel σ-algebra associated with $\Re^d$, which is denoted by $\mathcal{B}(\Re^d)$. We denote the σ-algebra of events generated by the random variables $\{x_0, x_1, \cdots, x_t\}$ by $\mathcal{F}_t \subset \mathcal{F}$.

We define a stopping time to be a random variable τ that takes on values in $\{0, 1, 2, \cdots, \infty\}$ and satisfies $\{\omega \in \Omega \mid \tau(\omega) \leq t\} \in \mathcal{F}_t$ for all finite t. The set of all such random variables is denoted by U. Since we have defined $\mathcal{F}_t$ to be the σ-algebra generated by $\{x_0, x_1, \cdots, x_t\}$, the stopping time is determined solely by the already available samples of the stochastic process. In particular, we do not consider stopping times that may be influenced by random events other than the stochastic process itself. This preclusion is not necessary for our analysis, but it is introduced to simplify the exposition.

An optimal stopping problem is defined by the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, stochastic process $\{x_t \mid t = 0, 1, 2, \cdots\}$, reward functions $g : \Re^d \mapsto \Re$ and $G : \Re^d \mapsto \Re$ associated with continuation and termination, and a discount factor α. The expected reward associated with a stopping time τ is defined by

$$E\left[\sum_{t=0}^{\tau-1} \alpha^t g(x_t) + \alpha^\tau G(x_\tau)\right]$$

where $G(x_\tau)$ is taken to be 0 if $\tau = \infty$. An optimal stopping time τ^* is one that satisfies

$$E\left[\sum_{t=0}^{\tau^*-1} \alpha^t g(x_t) + \alpha^{\tau^*} G(x_{\tau^*})\right] = \sup_{\tau \in U} E\left[\sum_{t=0}^{\tau-1} \alpha^t g(x_t) + \alpha^\tau G(x_\tau)\right].$$

Certain conditions ensure that an optimal stopping time exists. When such conditions are met, the optimal stopping problem is that of finding an optimal stopping time.

We now state a few assumptions that define the class of optimal stopping problems that will be addressed in this paper. Our first assumption places restrictions on the underlying stochastic process.

Assumption 1: The process $\{x_t \mid t = 0, 1, 2, \cdots\}$ is ergodic and Markov.

By ergodicity, we mean that the process is stationary and every invariant random variable of the process is almost surely equal to a constant. Hence, expectations $E[\cdot]$ are always with respect to a stationary distribution (e.g., $E[J(x_0)] = E[J(x_t)]$ for any function J and any t). The Markov condition corresponds to the existence of a transition probability kernel $P : \Re^d \times \mathcal{B}(\Re^d) \mapsto [0, 1]$ satisfying

$$\text{Prob}[x_{t+1} \in A \mid \mathcal{F}_t] = P(x_t, A)$$

for any $A \in \mathcal{B}(\Re^d)$ and any time t. Therefore, for any Borel function $J : \Re^d \mapsto \Re$ that is either nonnegative or absolutely integrable with respect to $P(x_t, \cdot)$, we have

$$E[J(x_{t+1}) \mid \mathcal{F}_t] = \int J(y) P(x_t, dy).$$

We define an operator P, mapping a function J to a new function PJ, by

$$(PJ)(x) = \int J(y) P(x, dy).$$

Since the process is stationary, there exists a probability measure $\pi : \mathcal{B}(\Re^d) \mapsto [0, 1]$ such that $\text{Prob}[x_t \in A] = \pi(A)$ for any $A \in \mathcal{B}(\Re^d)$ and any time t. Ergodicity implies that this is a unique invariant distribution. We define a Hilbert space $L_2(\pi)$ of real-valued functions on $\Re^d$ with inner product $\langle J, \bar{J} \rangle_\pi = E[J(x_0)\bar{J}(x_0)]$ and $\text{norm}\|J\|_\pi = \sqrt{E[J^2(x_0)]}$. This Hilbert space plays a central role in our analysis, and its use is the main feature that distinguishes our analysis from previous work on optimal stopping. To avoid confusion of equality in the sense of $L_2(\pi)$ with pointwise equality, we will employ the notation $J \overset{ae(\pi)}{=} \bar{J}$ to convey the former notion, whereas $J = \bar{J}$ will represent the latter.

Our second assumption ensures that the per-stage and terminal reward functions are in the Hilbert space of interest.

Assumption 2: The reward functions g and G are in $L_2(\pi)$.

Our final assumption is that future rewards are discounted.

Assumption 3: The discount factor α is in $(0, 1)$.

We will provide a theorem that characterizes value functions and optimal stopping times for the class of problems under consideration. However, before doing so, let us introduce some useful notation. We define an operator T by

$$TJ = \max\{G, g + \alpha PJ\}$$

where the max denotes pointwise maximization. This is the so-called "dynamic programming operator," specialized to the case of an optimal stopping problem. To each stopping time τ, we associate a value function J^τ defined by

$$J^\tau(x) = E\left[\sum_{t=0}^{\tau-1} \alpha^t g(x_t) + \alpha^\tau G(x_\tau) \mid x_0 = x\right].$$

Because g and G are in $L_2(\pi)$, J^τ is also an element of $L_2(\pi)$ for any τ. Hence, a stopping time τ^* is optimal if and only if

$$E[J^{\tau^*}(x_0)] = \sup_{\tau \in U} E[J^\tau(x_0)].$$

It is not hard to show that optimality in this sense corresponds to pointwise optimality for all elements x of some set A with $\pi(A) = 1$. However, this fact will not be used in our analysis.

The main results of this section are captured by the following theorem.

Theorem 1: Under Assumptions 1–3, the following statements hold.

1) There exists a function $J^* \in L_2(\pi)$ uniquely satisfying

$$J^* \overset{ae(\pi)}{=} TJ^*.$$

2) The stopping time τ^*, defined by

$$\tau^* = \min\{t \mid G(x_t) \geq J^*(x_t)\}$$

is an optimal stopping time. (The minimum of an empty set is taken to be ∞.)

3) The function J^{τ^*} is equal to J^* [in the sense of $L_2(\pi)$].

B. Preliminaries

Our first lemma establishes that the operator P is a nonexpansion in $L_2(\pi)$.

Lemma 1: Under Assumption 1, we have

$$||PJ||_\pi \leq ||J||_\pi, \qquad \forall J \in L_2(\pi).$$

Proof: The proof of the lemma involves Jensen's inequality and the Tonelli–Fubini theorem. In particular, for any $J \in L_2(\pi)$, we have

$$\begin{aligned} ||PJ||_\pi^2 &= E[(PJ)^2(x_0)] \\ &= E[(E[J(x_1) \mid x_0])^2] \\ &\leq E[E[J^2(x_1) \mid x_0]] \\ &= E[J^2(x_1)] \\ &= ||J||_\pi^2. \end{aligned}$$

□

The following lemma establishes that T is a contraction on $L_2(\pi)$.

Lemma 2: Under Assumptions 1–3, the operator T satisfies

$$||TJ - T\bar{J}||_\pi \leq \alpha||J - \bar{J}||_\pi, \qquad \forall J, \bar{J} \in L_2(\pi).$$

Proof: For any scalars c_1, c_2, and c_3

$$|\max\{c_1, c_3\} - \max\{c_2, c_3\}| \leq |c_1 - c_2|.$$

It follows that for any $x \in \Re^d$ and $J, \bar{J} \in L_2(\pi)$

$$|(TJ)(x) - (T\bar{J})(x)| \leq \alpha|(PJ)(x) - (P\bar{J})(x)|.$$

Given this fact, the result easily follows from Lemma 1. □

The fact that T is a contraction implies that it has a unique fixed point in $J^* \in L_2(\pi)$ (by unique here, we mean unique up to the equivalence classes of $L_2(\pi)$). This establishes part 1) of the theorem.

Let J^* denote the fixed point of T. Let us define a second operator T^* by

$$(T^*J)(x) = \begin{cases} G(x), & \text{if } G(x) \geq J^*(x), \\ g(x) + (\alpha PJ)(x), & \text{otherwise.} \end{cases}$$

(Note that T^* is the dynamic programming operator corresponding to the case of a fixed policy, namely, the policy corresponding to the stopping time τ^* defined in the statement of the above theorem.) The following lemma establishes that T^* is also a contraction, and furthermore, the fixed point of this contraction is equal to J^* (in the sense of $L_2(\pi)$).

Lemma 3: Under Assumptions 1–3, the operator T^* satisfies

$$||T^*J - T^*\bar{J}||_\pi \leq \alpha||J - \bar{J}||_\pi, \qquad \forall J, \bar{J} \in L_2(\pi).$$

Furthermore, $J^* \in L_2(\pi)$ is the unique fixed point of T^*.

Proof: We have

$$\begin{aligned} ||T^*J - T^*\bar{J}||_\pi &\leq ||\alpha PJ - \alpha P\bar{J}||_\pi \\ &\leq \alpha||J - \bar{J}||_\pi \end{aligned}$$

where the final inequality follows from Lemma 1.

Recall that J^* uniquely satisfies $J^* \overset{ae(\pi)}{=} TJ^*$, or written differently

$$J^* \overset{ae(\pi)}{=} \max\{G, g + \alpha PJ^*\}.$$

This equation can also be rewritten as

$$J^*(x) = \begin{cases} G(x), & \text{if } G(x) \geq g(x) + (\alpha PJ^*)(x) \\ g(x) + (\alpha PJ^*)(x), & \text{otherwise} \end{cases}$$

almost surely with respect to π. Note that for almost all x (a set $A \in \mathcal{B}(\Re^d)$ with $\pi(A) = 1$), $G(x) \geq g(x) + (\alpha PJ^*)(x)$ if and only if $G(x) = J^*(x)$. Hence, J^* satisfies

$$J^*(x) = \begin{cases} G(x), & \text{if } G(x) \geq J^*(x) \\ g(x) + (\alpha PJ^*)(x), & \text{otherwise} \end{cases}$$

almost surely with respect to π, or more concisely, $J^* \overset{ae(\pi)}{=} T^*J^*$. Since T^* is a contraction, it has a unique fixed point in $L_2(\pi)$, and this fixed point is J^*. □

C. Proof of Theorem 1

Part 1) of the result follows from Lemma 2. As for Part 3), we have

$$\begin{aligned} J^{\tau^*}(x) &= \begin{cases} G(x), & \text{if } G(x) \geq J^*(x) \\ g(x) + (\alpha PJ^{\tau^*})(x), & \text{otherwise} \end{cases} \\ &= (T^*J^{\tau^*})(x) \end{aligned}$$

and since T^* is a contraction with fixed point J^* (Lemma 3), it follows that

$$J^{\tau^*} \overset{ae(\pi)}{=} J^*.$$

We are left with the task of proving Part 2). For any nonnegative integer n, we have

$$\begin{aligned} &\sup_{\tau \in U} E[J^\tau(x_0)] \\ &\quad \leq \sup_{\tau \in U} E[J^{\tau \wedge n}(x_0)] + E\left[\sum_{t=n}^{\infty} \alpha^t(|g(x_t)| + |G(x_t)|)\right] \\ &\quad = \sup_{\tau \in U} E[J^{\tau \wedge n}(x_0)] + \frac{\alpha^n}{1-\alpha} E[|g(x_0)| + |G(x_0)|] \\ &\quad \leq \sup_{\tau \in U} E[J^{\tau \wedge n}(x_0)] + \alpha^n C \end{aligned}$$

for some scalar C that is independent of n, where the equality follows from the Tonelli–Fubini theorem and stationarity. By arguments standard to the theory of finite-horizon dynamic programming

$$\sup_{\tau \in U} J^{\tau \wedge n}(x) = (T^n G)(x), \qquad \forall x \in \Re^d.$$

(This equality is simply saying that the optimal reward for an n-horizon problem is obtained by applying n iterations of the

dynamic programming recursion.) It is easy to see that T^nG, and therefore also $\sup_{\tau\in U} J^{\tau\wedge n}(\cdot)$, is measurable. It follows that

$$\sup_{\tau\in U} E[J^{\tau\wedge n}(x_0)] \le E\left[\sup_{\tau\in U} J^{\tau\wedge n}(x_0)\right] = E[(T^nG)(x_0)].$$

Combining this with the bound on $\sup_{\tau\in U} E[J^{\tau}(x_0)]$, we have

$$\sup_{\tau\in U} E[J^{\tau}(x_0)] \le E[(T^nG)(x_0)] + \alpha^n C.$$

Since T is a contraction on $L_2(\pi)$ (Lemma 2), T^nG converges to J^* in the sense of $L_2(\pi)$. It follows that

$$\lim_{n\to\infty} E[(T^nG)(x_0)] = E[J^*(x_0)]$$

and we therefore have

$$\sup_{\tau\in U} E[J^{\tau}(x_0)] \le \lim_{n\to\infty} E[(T^nG)(x_0)] = E[J^*(x_0)] = E[J^{\tau^*}(x_0)].$$

Hence, the stopping time τ^* is optimal. □

III. An Approximation Scheme

In addition to establishing the existence of an optimal stopping time, Theorem 1 offers an approach to obtaining one. In particular, the function J^* can be found by solving the equation

$$J^* \overset{ae(\pi)}{=} TJ^*$$

and then used to generate an optimal stopping time. However, for most problems, it is not possible to derive a "closed-form" solution to this equation. In this event, one may resort to the discretization of a relevant portion of $\Re^d$ and then use numerical algorithms to approximate J^* over this discretized space. Unfortunately, this approach becomes infeasible as d grows, since the number of points in the discretized space grows exponentially with the dimension. This phenomenon, known as the "curse of dimensionality," plagues the field of stochastic control and gives rise to the need for parsimonious approximation schemes.

One approach to approximation involves selecting a set of basis functions $\{\phi_k : \Re^d \mapsto \Re \mid k = 1, 2, \cdots, K\}$ and computing weights $r(1), \cdots, r(k) \in \Re$ such that the weighted combination $\sum_{k=1}^{K} r(k)\phi_k$ is "close" to J^*. Much like the context of statistical regression, the basis functions should be selected based on engineering intuition and/or analysis concerning the form of the function J^*, while numerical algorithms may be used to generate appropriate weights. Also, similarly with linear regression, a good choice of basis functions is critical for accurate approximations. In this section, we introduce an algorithm for computing basis function weights and provide an analysis of its behavior.

We begin by presenting our algorithm and a theorem that establishes certain desirable properties. Sections III-B and III-C provide the analysis required to prove this theorem. Our algorithm is stochastic and relies in a fundamental way on the use of a simulated trajectory, as is discussed in Section III-D.

A. The Approximation Algorithm

In our analysis of optimal stopping problems, the function J^* played a central role in characterizing an optimal stopping time and the rewards it would generate. The algorithm we will develop approximates a different, but closely related, function Q^*, defined by

$$Q^* = g + \alpha P J^*. \tag{1}$$

Functions of this type were first employed by Watkins in conjunction with his Q-learning algorithm [22]. Intuitively, for each state x, $Q^*(x)$ represents the optimal attainable reward, starting at state $x_0 = x$, if stopping times are constrained to be greater than zero. An optimal stopping time can be generated according to

$$\tau^* = \min\{t \mid G(x_t) \ge Q^*(x_t)\}.$$

Note that this is equivalent to the generation of an optimal stopping time based on a value function J^*, since $J^* = \max\{G, Q^*\}$ and therefore

$$\begin{aligned}\min\{t \mid G(x_t) \ge J^*(x_t)\} &= \min\{t \mid G(x_t) \ge \max\{G(x_t), Q^*(x_t)\}\} \\ &= \min\{t \mid G(x_t) \ge Q^*(x_t)\}.\end{aligned}$$

Our approximation algorithm employs a set of basis functions $\phi_1, \cdots, \phi_K \in L_2(\pi)$ that are hand-crafted prior to execution. To condense notation, let us define an operator $\Phi : \Re^K \mapsto L_2(\pi)$ by $\Phi r = \sum_{k=1}^{K} r(k)\phi_k$, for any vector of weights $r = (r(1), \cdots, r(K))'$. Also, let $\phi(x) \in \Re^K$ be the vector of basis function values, evaluated at x, so that $(\Phi r)(x) = \phi'(x)r$.

The algorithm is initialized with a weight vector $r_0 = (r_0(1), \cdots, r_0(K))' \in \Re^K$. During the simulation of a trajectory $\{x_t \mid t = 0, 1, 2, \cdots\}$ of the Markov chain, the algorithm generates a sequence of weight vectors $\{r_t \mid t = 1, 2, \cdots\}$ according to

$$\begin{aligned} r_{t+1} = r_t + \gamma_t \phi(x_t)(&g(x_t) \\ &+ \alpha \max\{(\Phi r_t)(x_{t+1}), G(x_{t+1})\} - (\Phi r_t)(x_t)) \end{aligned} \tag{2}$$

where each γ_t is a positive scalar step size. One (heuristic) interpretation of this update equation is as one that tries to make the approximate Q-value $(\Phi r_t)(x_t)$ closer to an "improved approximation" $g(x_t) + \alpha \max\{(\Phi r_t)(x_{t+1}), G(x_{t+1})\}$. In this context, $\phi(x_t)$, which is equal to the gradient of the aproximate Q-value with respect to the weights, provides a direction in which to alter the parameter vector, and this direction is scaled by the difference between the current and improved approximation. We will prove that, under certain conditions, the sequence r_t converges to a vector r^*, and Φr^* approximates Q^*. Furthermore, the stopping time $\tilde{\tau}$, given by

$$\tilde{\tau} = \min\{t \mid G(x_t) \ge (\Phi r^*)(x_t)\}$$

approximates the performance of τ^*.

Let us now introduce our assumptions so that we can formally state results concerning the approximation algorithm. Our first assumption pertains to the basis functions.

Assumption 4:

1) The basis functions $\phi_1, \cdots, \phi_K$ are linearly independent.
2) For each k, the basis function ϕ_k is in $L_2(\pi)$.

The requirement of linear independence is not truly necessary, but simplifies the exposition. The assumption that the basis functions are in $L_2(\pi)$ limits their rate of growth and is essential to the convergence of the algorithm.

Our next assumption requires that the Markov chain exhibits a certain "degree of stability" and that certain functions do not grow to quickly. (We use $\|\cdot\|$ to denote the Euclidean norm on finite-dimensional spaces.)

Assumption 5:

1) For any positive scalar q, there exists a scalar μ_q such that for all x and t

$$E[1 + \|x_t\|^q \mid x_0 = x] \leq \mu_q(1 + \|x\|^q).$$

2) There exist scalars C_1, q_1 such that, for any function J satisfying $|J(x)| \leq C_2(1 + \|x\|^{q_2})$, for some scalars C_2 and q_2

$$\sum_{t=0}^{\infty} |E[J(x_t) \mid x_0 = x] - E[J(x_0)]| \leq C_1 C_2(1 + \|x\|^{q_1 q_2}), \qquad \forall x \in \Re^d.$$

3) There exist scalars C and q such that for all $x \in \Re^d$, $|g(x)| \leq C(1 + \|x\|^q)$, $|G(x)| \leq C(1 + \|x\|^q)$, and $\|\phi(x)\| \leq C(1 + \|x\|^q)$.

Our final assumption places constraints on the sequence of step sizes. Such constraints are fairly standard to stochastic approximation algorithms.

Assumption 6: The step sizes γ_t are nonincreasing and predetermined (chosen prior to execution of the algorithm). Furthermore, they satisfy $\sum_{t=0}^{\infty} \gamma_t = \infty$, and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$.

Before stating our results concerning the behavior of the algorithm, let us introduce some notation that will make the statement concise. We define a "projection operator" Π that projects onto the subspace $\{\Phi r \mid r \in \Re^K\}$ of $L_2(\pi)$. In particular, for any function $Q \in L_2(\pi)$, let

$$\Pi Q = \arg\min_{\bar{Q} \in \{\Phi r \mid r \in \Re^K\}} \|\bar{Q} - Q\|_\pi.$$

We define an additional operator F by

$$FQ = g + \alpha P \max\{G, Q\} \tag{3}$$

for any $Q \in L_2(\pi)$.

The main result of this section follows.

Theorem 2: Under Assumptions 1–6, the following hold.

1) The approximation algorithm converges almost surely.
2) The limit of convergence r^* is the unique solution of the equation

$$\Pi F(\Phi r^*) \overset{ae(\pi)}{=} \Phi r^*.$$

3) Furthermore, r^* satisfies

$$\|\Phi r^* - Q^*\|_\pi \leq \frac{1}{\sqrt{1 - \alpha^2}} \|\Pi Q^* - Q^*\|_\pi.$$

4) Let $\tilde{\tau}$ be defined by

$$\tilde{\tau} = \min\{t \mid G(x_t) \geq (\Phi r^*)(x_t)\}.$$

Then

$$E[J^*(x_0)] - E[J^{\tilde{\tau}}(x_0)] \leq \frac{2}{(1 - \alpha)\sqrt{1 - \alpha^2}} \|\Pi Q^* - Q^*\|_\pi.$$

Note that the bounds provided by parts 3) and 4) involve a term $\|\Pi Q^* - Q^*\|_\pi$. This term represents the smallest approximation error (in terms of $\|\cdot\|_\pi$) that can be achieved given the choice of basis functions. Hence, as the subspace spanned by the basis functions comes closer to Q^*, the error generated by the algorithm diminishes to zero and the performance of the resulting stopping time approaches optimality.

B. Preliminaries

Our next lemma establishes that F is a contraction in $L_2(\pi)$ and that Q^* is its fixed point.

Lemma 4: Under Assumptions 1–3, the operator F satisfies

$$\|FQ - F\bar{Q}\|_\pi \leq \alpha \|Q - \bar{Q}\|_\pi, \qquad \forall Q, \bar{Q} \in L_2(\pi).$$

Furthermore, Q^* is the unique fixed point of F in $L_2(\pi)$.

Proof: For any $Q, \bar{Q} \in L_2(\pi)$, we have

$$\begin{aligned} \|FQ - F\bar{Q}\|_\pi &= \alpha \|P \max\{G, Q\} - P \max\{G, \bar{Q}\}\|_\pi \\ &\leq \alpha \|\max\{G, Q\} - \max\{G, \bar{Q}\}\|_\pi \\ &\leq \alpha \|Q - \bar{Q}\|_\pi \end{aligned}$$

where the first inequality follows from Lemma 1 and the second makes use of the fact

$$|\max\{c_1, c_3\} - \max\{c_2, c_3\}| \leq |c_1 - c_2|$$

for any scalars c_1, c_2, and c_3. Hence, F is a contraction on $L_2(\pi)$. It follows that F has a unique fixed point. By Theorem 1, we have

$$\begin{aligned} J^* &\overset{ae(\pi)}{=} TJ^* \\ g + \alpha P J^* &\overset{ae(\pi)}{=} g + \alpha P \max\{G, g + \alpha P J^*\} \\ Q^* &\overset{ae(\pi)}{=} g + \alpha P \max\{G, Q^*\} \\ Q^* &\overset{ae(\pi)}{=} FQ^* \end{aligned}$$

and therefore, Q^* is the fixed point. □

The next lemma establishes that the composition ΠF is a contraction on $L_2(\pi)$ and that its fixed point is equal to Φr^* for a unique $r^* \in \Re^K$. The lemma also places a bound on the magnitude of the approximation error $\Phi r^* - Q^*$. We will later establish that this vector is the limit of convergence of our approximation algorithm.

Lemma 5: Under Assumptions 1–4, the composition ΠF satisfies

$$\|\Pi F Q - \Pi F \bar{Q}\|_\pi \leq \alpha \|Q - \bar{Q}\|_\pi, \qquad \forall Q, \bar{Q} \in L_2(\pi).$$

Furthermore ΠF has a unique fixed point of the form Φr^* for a unique vector $r^* \in \Re^K$, and this vector satisfies

$$\|\Phi r^* - Q^*\|_\pi \leq \frac{1}{\sqrt{1 - \alpha^2}} \|\Pi Q^* - Q^*\|_\pi.$$

Proof: Since Π is a nonexpansion in $L_2(\pi)$ (by virtue of being a projection operator), we have

$$\|\Pi FQ - \Pi F\bar{Q}\|_\pi \leq \|FQ - F\bar{Q}\|_\pi \leq \alpha\|Q - \bar{Q}\|_\pi$$

by Lemma 4. Since the range of Π is the same as that of Φ, the fixed point of ΠF is of the form Φr^* for some $r^* \in \Re^K$. Furthermore, because the basis functions are linearly independent, this fixed point is associated with a unique r^*.

Note that by the orthogonality properties of projections, we have $\langle \Phi r^* - \Pi Q^*, \Pi Q^* - Q^* \rangle_\pi = 0$. Using also the Pythagorean theorem and Lemma 4, we have

$$\begin{aligned}\|\Phi r^* - Q^*\|_\pi^2 &= \|\Phi r^* - \Pi Q^*\|_\pi^2 + \|\Pi Q^* - Q^*\|_\pi^2 \\ &= \|\Pi F\Phi r^* - \Pi Q^*\|_\pi^2 + \|\Pi Q^* - Q^*\|_\pi^2 \\ &\leq \|F\Phi r^* - Q^*\|_\pi^2 + \|\Pi Q^* - Q^*\|_\pi^2 \\ &\leq \alpha^2\|\Phi r^* - Q^*\|_\pi^2 + \|\Pi Q^* - Q^*\|_\pi^2\end{aligned}$$

and it follows that

$$\|\Phi r^* - Q^*\|_\pi \leq \frac{1}{\sqrt{1-\alpha^2}}\|\Pi Q^* - Q^*\|_\pi. \qquad \square$$

Given r^* (which would be obtained by running the algorithm on a simulated trajectory), we define a stopping time $\tilde{\tau} = \min\{t \mid G(x_t) \geq (\Phi r^*)(x_t)\}$. Let us define operators H and $\tilde{F}$ by

$$(HQ)(x) = \begin{cases} G(x), & \text{if } G(x) \geq (\Phi r^*)(x) \\ Q(x), & \text{otherwise} \end{cases}$$

and

$$\tilde{F}Q = g + \alpha PHQ. \tag{4}$$

The next lemma establishes that $\tilde{F}$ is a contraction on $L_2(\pi)$ with a fixed point $\tilde{Q} = g + \alpha PJ^{\tilde{\tau}}$.

Lemma 6: Under Assumptions 1–4, for any $Q, \bar{Q} \in L_2(\pi)$

$$\|\tilde{F}Q - \tilde{F}\bar{Q}\|_\pi \leq \alpha\|Q - \bar{Q}\|_\pi.$$

Furthermore, $\tilde{Q} = g + \alpha PJ^{\tilde{\tau}}$ is the unique fixed point of $\tilde{F}$.

Proof: For any $Q, \bar{Q} \in L_2(\pi)$, we have

$$\begin{aligned}\|\tilde{F}Q - \tilde{F}\bar{Q}\|_\pi &= \|(g + \alpha PHQ) - (g + \alpha PH\bar{Q})\|_\pi \\ &\leq \alpha\|HQ - H\bar{Q}\|_\pi \\ &\leq \alpha\|\max\{G, Q - \bar{Q}\}\|_\pi \\ &\leq \alpha\|Q - \bar{Q}\|_\pi\end{aligned}$$

where the first inequality follows from Lemma 1.

To prove that $\tilde{Q} = g + \alpha PJ^{\tilde{\tau}}$ is the fixed point, observe that

$$\begin{aligned}(H\tilde{Q})(x) &= (H(g + \alpha PJ^{\tilde{\tau}}))(x) \\ &= \begin{cases} G(x), & \text{if } G(x) \geq (\Phi r^*)(x), \\ \tilde{Q}(x), & \text{otherwise}, \end{cases} \\ &= \begin{cases} G(x), & \text{if } G(x) \geq (\Phi r^*)(x) \\ g(x) + (\alpha PJ^{\tilde{\tau}})(x), & \text{otherwise} \end{cases} \\ &= J^{\tilde{\tau}}(x).\end{aligned}$$

Therefore

$$\tilde{F}\tilde{Q} = g + \alpha PH\tilde{Q} = g + \alpha PJ^{\tilde{\tau}} = \tilde{Q}$$

as desired. $\square$

The next lemma places a bound on the loss in performance incurred when using the stopping time $\tilde{\tau}$ instead of an optimal stopping time.

Lemma 7: Under Assumptions 1–4, the stopping time $\tilde{\tau}$ satisfies

$$E[J^*(x_0)] - E[J^{\tilde{\tau}}(x_0)] \leq \frac{2}{(1-\alpha)\sqrt{1-\alpha^2}}\|\Pi Q^* - Q^*\|_\pi.$$

Proof: By stationarity and Jensen's inequality, we have

$$\begin{aligned}E[J^*(x_0)] - E[J^{\tilde{\tau}}(x_0)] &= E[(PJ^*)(x_0)] - E[(PJ^{\tilde{\tau}})(x_0)] \\ &\leq |E[(PJ^*)(x_0)] - E[(PJ^{\tilde{\tau}})(x_0)]| \\ &\leq \|PJ^* - PJ^{\tilde{\tau}}\|_\pi.\end{aligned}$$

Recall that $Q^* = g + \alpha PJ^*$ and $\tilde{Q} = g + \alpha PJ^{\tilde{\tau}}$. We therefore have

$$\begin{aligned}E[J^*(x_0)] - E[J^{\tilde{\tau}}(x_0)] &\leq \frac{1}{\alpha}\|(g + \alpha PJ^*) - (g + \alpha PJ^{\tilde{\tau}})\|_\pi \\ &= \frac{1}{\alpha}\|Q^* - \tilde{Q}\|_\pi.\end{aligned}$$

Hence, it is sufficient to place a bound on $\|Q^* - \tilde{Q}\|_\pi$.

It is easy to show that $F(\Phi r^*) = \tilde{F}(\Phi r^*)$ [compare definitions (3) and (4)]. Using this fact, the triangle inequality, the equality $FQ^* \overset{ae(\pi)}{=} Q^*$ (Lemma 4), and the equality $\tilde{F}\tilde{Q} \overset{ae(\pi)}{=} \tilde{Q}$ (Lemma 6), we have

$$\begin{aligned}\|Q^* - \tilde{Q}\|_\pi &\leq \|Q^* - F(\Phi r^*)\|_\pi + \|\tilde{Q} - \tilde{F}(\Phi r^*)\|_\pi \\ &\leq \alpha\|Q^* - \Phi r^*\|_\pi + \alpha\|\tilde{Q} - \Phi r^*\|_\pi \\ &\leq 2\alpha\|Q^* - \Phi r^*\|_\pi + \alpha\|Q^* - \tilde{Q}\|_\pi\end{aligned}$$

and it follows that

$$\begin{aligned}\|Q^* - \tilde{Q}\|_\pi &\leq \frac{2\alpha}{1-\alpha}\|Q^* - \Phi r^*\|_\pi \\ &\leq \frac{2\alpha}{(1-\alpha)\sqrt{1-\alpha^2}}\|Q^* - \Pi Q^*\|_\pi\end{aligned}$$

where the final inequality follows from Lemma 5. Finally, we obtain

$$E[J^*(x_0)] - E[J^{\tilde{\tau}}(x_0)] \leq \frac{2}{(1-\alpha)\sqrt{1-\alpha^2}}\|\Pi Q^* - Q^*\|_\pi.$$

$\square$

We now continue with the analysis of the stochastic algorithm. Let us define a stochastic process $\{z_t \mid t = 0, 1, 2, \cdots\}$ taking on values in $\Re^{2d}$ where $z_t = (x_t, x_{t+1})$. It is easy to see that z_t is ergodic and Markov (recall that, by our definition, ergodic processes are stationary). Furthermore, the iteration given by (2) can be rewritten as

$$r_{t+1} = r_t + \gamma_t s(z_t, r_t)$$

for a function $s : \Re^{2d} \times \Re^K \mapsto \Re^K$ given by

$$s(z, r) = \phi(x)(g(x) + \alpha\max\{(\Phi r)(y), G(y)\} - (\Phi r)(x))$$

for any r and $z = (x, y)$. We define a function $\bar{s} : \Re^K \mapsto \Re^K$ by

$$\bar{s}(r) = E[s(z_0, r)], \qquad \forall r.$$

(Note that this is an expectation over z_0 for a fixed r. It is easy to show that the random variable $s(z_0, r)$ is absolutely integrable and $\bar{s}(r)$ is well-defined as a consequence of Assumption 5.) Note that each component $\bar{s}_k(r)$ can be represented in terms of an inner product according to

$$\begin{aligned}\bar{s}_k(r) &= E[\phi_k(x_0)(g(x_0) + \alpha \max\{(\Phi r)(x_1), G(x_1)\} \\ &\quad - (\Phi r)(x_0))] \\ &= E[\phi_k(x_0)(g(x_0) + \alpha E[\max\{(\Phi r)(x_1), G(x_1)\} \mid x_0] \\ &\quad - (\Phi r)(x_0))] \\ &= E[\phi_k(x_0)(g(x_0) + \alpha P \max\{\Phi r, G\}(x_0) \\ &\quad - (\Phi r)(x_0))] \\ &= \langle \phi_k, F\Phi r - \Phi r\rangle_\pi\end{aligned}$$

where the definition of the operator P is used.

Lemma 8: Under Assumptions 1–4, we have

$$(r - r^*)'\bar{s}(r) < 0, \qquad \forall r \neq r^*$$

and

$$\bar{s}(r^*) = 0.$$

Proof: For any r, we have

$$\begin{aligned}(r - r^*)'\bar{s}(r) &= \sum_{k=1}^{K} (r(k) - r^*(k))\langle \phi_k, F\Phi r - \Phi r\rangle_\pi \\ &= \langle \Phi r - \Phi r^*, F\Phi r - \Phi r\rangle_\pi \\ &= \langle \Phi r - \Phi r^*, (I - \Pi)F\Phi r + \Pi F\Phi r - \Phi r\rangle_\pi \\ &= \langle \Phi r - \Phi r^*, \Pi F\Phi r - \Phi r\rangle_\pi\end{aligned}$$

where the final equality follows because Π projects onto the range of Φ, and the range of $(I - \Pi)$ is therefore orthogonal to that of Φ. Since Φr^* is the fixed point of ΠF, Lemma 5 implies that

$$\|\Pi F\Phi r - \Phi r^*\|_\pi \leq \alpha \|\Phi r - \Phi r^*\|_\pi.$$

Using the Cauchy–Schwartz inequality together with this fact, we obtain

$$\begin{aligned}&\langle \Phi r - \Phi r^*, \Pi F\Phi r - \Phi r\rangle_\pi \\ &\quad = \langle \Phi r - \Phi r^*, (\Pi F\Phi r - \Phi r^*) + (\Phi r^* - \Phi r)\rangle_\pi \\ &\quad \leq \|\Phi r - \Phi r^*\|_\pi \cdot \|\Pi F\Phi r - \Phi r^*\|_\pi - \|\Phi r^* - \Phi r\|_\pi^2 \\ &\quad \leq (\alpha - 1)\|\Phi r - \Phi r^*\|_\pi^2.\end{aligned}$$

By Assumption 4-1), for any $r \neq r^*$, we have $\|\Phi r - \Phi r^*\|_\pi \neq 0$. Since $\alpha < 1$, the first part of the result follows.

As for the second part, we have to complete the proof, thus we have

$$\bar{s}_k(r^*) = \langle \phi_k, F\Phi r^* - \Phi r^*\rangle = \langle \phi_k, \Pi F\Phi r^* - \Phi r^*\rangle = 0.$$

□

We now state without proof a result concerning stochastic approximation, which will be used in the proof of Theorem 2. This is a special case of a general result on stochastic approximation algorithms [3, Th. 17, p. 239]. It is straightforward to check that all of the assumptions in the result of [3] follow from the assumptions imposed in the result below. We do not show here the assumptions of [3] because the list is long and would require a lot in terms of new notation. However, we note that in our setting here, the potential function $U(\cdot)$ that would be required to satisfy the assumptions of the theorem from [3] is given by $U(r) = \|r - r^*\|^2$.

Theorem 3: Consider a process r_t taking values in $\Re^K$, initialized with an arbitrary vector r_0, that evolves according to

$$r_{t+1} = r_t + \gamma_t s(z_t, r_t)$$

for some $s : \Re^{2d} \times \Re^K \mapsto \Re^K$, where we have the following.

1) $\{z_t \mid t = 0, 1, 2, \cdots\}$ is a (stationary) ergodic Markov process taking values in $\Re^{2d}$.
2) For any positive scalar q, there exists a scalar μ_q such that $E[1 + \|z_t\|^q \mid z_0 = z] \leq \mu_q(1 + \|z\|^q)$, for any time t and $z \in \Re^{2d}$.
3) The (predetermined) step size sequence γ_t is nonincreasing and satisfies $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$.
4) There exist scalars C and q such that
$$\|s(z, r)\| \leq C(1 + \|r\|)(1 + \|z\|^q), \qquad \forall z, r.$$
5) There exist scalars C and q such that
$$\sum_{t=0}^{\infty} \|E[s(z_t, r) \mid z_0 = z] - E[s(z_0, r)]\| \leq C(1 + \|r\|)(1 + \|z\|^q), \qquad \forall z, r.$$
6) There exists a scalar C such that
$$\|E[s(z_0, r)] - E[s(z_0, \bar{r})]\| \leq C\|r - \bar{r}\|, \qquad \forall r, \bar{r}.$$
7) There exist scalars C and q such that
$$\sum_{t=0}^{\infty} \|E[s(z_t, r) - s(z_t, \bar{r}) \mid z_0 = z] - E[s(z_0, r) - s(z_0, \bar{r})]\| \leq C\|r - \bar{r}\|(1 + \|z\|^q), \qquad \forall z, r, \bar{r}.$$
8) There exists some $r^* \in \Re^K$ such that $\bar{s}(r)'(r - r^*) < 0$, for all $r \neq r^*$, and $\bar{s}(r^*) = 0$. Then, r_t almost surely converges to r^*.

C. Proof of Theorem 2

We will prove Part 1) of Theorem 2 by establishing that the conditions of Theorem 3 are valid. Conditions 1) and 2) pertain to the dynamics of the process $z_t = (x_t, x_{t+1})$. The former condition follows easily from Assumption 1, while the latter is a consequence of Assumption 5-1). Condition 3), concerning the step size sequence, is the same as Assumption 6.

To establish validity of Condition 4), for any r and $z = (x, y)$, we have

$$\begin{aligned}\|s(z, r)\| &= \|\phi(x)(g(x) + \alpha \max\{(\Phi r)(y), G(y)\} \\ &\quad - (\Phi r)(x))\| \\ &\leq \|\phi(x)\|(|g(x)| + \alpha(\|\phi(y)\|\|r\| \\ &\quad + |G(y)|) + \|\phi(x)\|\|r\|) \\ &\leq \|\phi(x)\|(|g(x)| + \alpha|G(y)|) \\ &\quad + \|\phi(x)\|(\alpha\|\phi(y)\| + \|\phi(x)\|)\|r\|.\end{aligned}$$

Condition 4) then easily follows from the polynomial bounds of Assumption 5-3). Given that Condition 4) is valid, Condition 5) follows from Assumptions 5-1) and 5-2) in a straightforward manner. (Using these assumptions, it is easy to show that a condition analogous to Assumption 5-2) holds for functions of $z_t = (x_t, x_{t+1})$ that are bounded by polynomials in x_t and x_{t+1}.)

Let us now address Conditions 6) and 7). We first note that for any r, $\bar{r}$, and z, we have

$$\begin{aligned}
&\|s(z,r) - s(z,\bar{r})\| \\
&\quad = \|\phi(x)(\alpha \max\{(\Phi r)(y), G(y)\} \\
&\qquad - \alpha \max\{(\Phi \bar{r})(y), G(y)\} - (\Phi r)(x) + (\Phi \bar{r})(x))\| \\
&\quad \le \alpha \|\phi(x)\| |\max\{\phi'(y) r, G(y)\} \\
&\qquad - \max\{\phi'(y)\bar{r}, G(y)\}| + \|\phi(x)\| |\phi'(x) r - \phi'(x)\bar{r}| \\
&\quad \le \alpha \|\phi(x)\| |\phi'(y) r - \phi'(y)\bar{r}| + \|\phi(x)\|^2 \|r - \bar{r}\| \\
&\quad \le \alpha \|\phi(x)\| \|\phi(y)\| \|r - \bar{r}\| + \|\phi(x)\|^2 \|r - \bar{r}\|.
\end{aligned}$$

It then follows from the polynomial bounds of Assumption 5-3) that there exist scalars C_2 and q_2 such that for any r, $\bar{r}$, and z

$$\|s(z,r) - s(z,\bar{r})\| \le C_2 \|r - \bar{r}\| (1 + \|z\|^{q_2}).$$

Validity of Condition 6) follows. Finally, it follows from Assumptions 5-1) and 5-2) that there exist scalars C_1, q_1, C_2, and q_2, such that for any r, $\bar{r}$, and z

$$\begin{aligned}
\sum_{t=0}^{\infty} \| &E[s(z_t, r) - s(z_t, \bar{r}) \mid z_0 = z] - E[s(z_0, r) \\
&- s(z_0, \bar{r})]\| \le C_1 C_2 \|r - \bar{r}\| (1 + \|z\|^{q_1 q_2}).
\end{aligned}$$

This establishes Condition 7).

Validity of Condition 8) is assured by Lemma 8. This completes the proof for Part 1) of the theorem. To wrap up the proof, Parts 2) and 3) of the theorem follow from Lemma 5, while Part 4) is established by Lemma 7.

D. On the Importance of Simulated Trajectories

The approximation algorithm we analyzed can be thought of as a variant of the temporal-difference learning, also known as $\mathrm{TD}(\lambda)$, with the parameter λ set to zero. The $\mathrm{TD}(0)$ algorithm approximates the value function for an autonomous system using an iteration of the form

$$r_{t+1} = r_t + \gamma_t \phi(x_t)(g(x_t) + \alpha(\Phi r_t)(x_{t+1}) - (\Phi r_t)(x_t))$$

which replaces the term $\max\{(\Phi r_t)(x_{t+1}), G(x_{t+1})\}$ from the algorithm we have proposed with $(\Phi r_t)(x_{t+1})$. Intuitively, this is like applying our algorithm to a stopping problem for which the reward $G(x)$ for stopping is always a large negative number, making stopping undesirable.

An interesting characteristic of temporal-difference learning, first conjectured by Sutton [18] and later elucidated by the analysis of Tsitsiklis and Van Roy [19], is that the use of simulated trajectories is critical for convergence. The same is true for the algorithm proposed in the current paper. Consider, for example, an algorithm that, on each tth step, samples a state $y_t \in \Re^d$ according to a probability measure $\bar{\pi} : \mathcal{B}(\Re^d) \mapsto [0,1]$ and a state $\bar{y}_t \in \Re^d$ according to $\mathrm{Prob}[\bar{y}_t \in A] = P(y_t, A)$, and updates the weight vector according to

$$\begin{aligned}
r_{t+1} = r_t + \gamma_t \phi(y_t)(g(y_t) \\
+ \alpha \max\{G(\bar{y}_t), (\Phi r_t)(\bar{y}_t)\} - (\Phi r_t)(y_t)).
\end{aligned} \tag{5}$$

Such an algorithm does not generally converge. We refer the reader to Tsitsiklis and Van Roy [19] for a more detailed discussion of this phenomenon.

IV. Pricing Financial Derivatives

In this section, we illustrate the steps required in applying our algorithm by describing a simple case study. The problem is representative of high-dimensional derivatives pricing problems arising in the rapidly growing structured products (a.k.a. "exotics") industry [14]. Our approach involving the approximation of a value function is similar in spirit to the earlier experimental work of Barraquand and Martineau [2]. However, the algorithm employed in that study is different from ours, and the approximations were comprised of piecewise constant functions.

Another notable approach to approximating solutions of optimal stopping problems that arise in derivatives pricing is the "stochastic mesh" methods of Broadie and Glasserman [8], [9]. These methods can be thought of as variants of Rust's algorithm [15], which like traditional grid techniques, approximates values at points in a mesh over the state space. The innovation of Rust's approach, however, is that the mesh includes a tractable collection of randomly sampled states, rather than the intractable grid that would arise in standard state space discretization. Unfortunately, when the state space is high-dimensional, except for cases that satisfy restrictive assumptions as those presented in [15], the randomly sampled states may not generally be sufficiently representative for effective value function approximation.

We will begin by providing some background and references to standard material on derivatives pricing. Section IV-B then introduces the particular security we consider and a related optimal stopping problem. Section IV-C presents the performance of some simple stopping strategies. Finally, the selection of basis functions and computational results generated by our approximation algorithm are discussed in Section IV-D.

A. Background

Financial derivative securities (or derivatives, for short) are contracts that promise payoffs contingent on the future prices of basic assets such as stocks, bonds, and commodities. Certain types of derivatives, such as put and call options, are in popular demand and traded alongside stocks in large exchanges. Other more exotic derivatives are tailored by banks and other financial intermediaries in order to suit specialized needs of various institutions and are sold in "over-the-counter" markets.

When there is a fixed date at which payments are made and certain common simplified models of stock price movements and trading are employed, it is possible to devise a hedging strategy that perfectly replicates the payoffs of a derivative security. Hence, the initial investment required to operate this

hedging strategy must be equal to the value of the security. This approach to replication and valuation, introduced by Black and Scholes [7] and Merton [13] and presented in its definitive form by Harrison and Kreps [10] and Harrison and Pliska [11], has met wide application and is the subject of much subsequent research.

When there is a possibility of early exercise (i.e., the contract holder can decide at any time to terminate the contract and receive payments based on prevailing market conditions), the value of the derivative security depends on how the client chooses a time to exercise. Given that the bank cannot control the client's behavior, it must prepare for the worst by assuming that the client will employ an exercising strategy that maximizes the value of the security. Pricing the derivative security in this context generally requires solving an optimal stopping problem.

In the next few sections, we present one fictitious derivative security that leads to a high-dimensional optimal stopping problem, and we employ the algorithm we have developed in order to approximate its price. Our focus here is to demonstrate the use of the algorithm, rather than to solve a real-world problem. Hence, we employ very simple models and ignore details that may be required in order to make the problem realistic.

B. Problem Formulation

The financial derivative instrument we will consider generates payoffs that are contingent on prices of a single stock. At the end of any given day, the holder may opt to exercise. At the time of exercise, the contract is terminated, and a payoff is received in an amount equal to the current price of the stock divided by the price prevailing 100 days beforehand.

We will employ a standard continuous-time economic model involving a stochastic stock price process and deterministic returns generated by short-term bonds. Given this model, under certain technical conditions, it is possible to replicate derivative securities that are contingent on the stock price process by rebalancing a portfolio of stocks and bonds. This portfolio needs only an initial investment and is self-financing thereafter. Hence, to preclude arbitrage, the price of the derivative security must be equal to the initial investment required by such a portfolio. Karatzas [12] provides a comprehensive treatment of this pricing methodology in the case where early exercising is allowed. In particular, the value of the security is equal to the optimal reward for a particular optimal stopping problem. The framework of [12] does not explicitly capture our problem at hand (the framework allows early exercise at any positive time, while our security can only be exercised at the end of each day), but the extension is immediate. Since our motivation is to demonstrate the use of our algorithm, rather than dwelling on the steps required to formally reduce pricing to an optimal stopping problem, we will simply present the underlying economic model and the optimal stopping problem it leads to, omitting the technicalities needed to formally connect the two.

We model time as a continuous variable $t \in [-100, \infty)$ and assume that the derivative security is issued at time $t = 0$. Each unit of time is taken to be a day, and the security can be exercised at times $t \in \{0, 1, 2, \cdots\}$. We model the stock price process $\{p_t \mid t \geq -100\}$ as a geometric Brownian motion

$$p_t = p_{-100} + \int_{s=-100}^{t} \mu p_s \, ds + \int_{s=-100}^{t} \sigma p_s \, dw_s$$

for some positive scalars p_{-100}, μ, and σ and a standard Brownian motion w_t. The payoff received by the security holder is equal to $p_\tau / p_{\tau-100}$ where $\tau \geq 0$ is the time of exercise. Note that we consider negative times because the stock prices up to 100 days prior to the date of issue may influence the payoff of the security. We assume that there is a constant continuously compounded short-term interest rate ρ. In other words, D_0 dollars invested in the money market at time 0 grows to a value

$$D_t = D_0 e^{\rho t}$$

at time t.

We will now characterize the price of the derivative security in a way that gives rise to a related optimal stopping problem. Let $\{\tilde{p}_t \mid t \geq -100\}$ be a stochastic process that evolves according to

$$d\tilde{p}_t = \rho \tilde{p}_t \, dt + \sigma \tilde{p}_t \, dw_t.$$

Define a discrete-time process $\{x_t \mid t = 0, 1, 2, \cdots\}$ taking values in $\Re^{100}$, with

$$x_t = \left(\frac{\tilde{p}_{t-99}}{\tilde{p}_{t-100}}, \frac{\tilde{p}_{t-98}}{\tilde{p}_{t-100}}, \cdots, \frac{\tilde{p}_t}{\tilde{p}_{t-100}} \right)'.$$

Intuitively, the ith component $x_t(i)$ of x_t represents the amount a one-dollar investment made in the stock at time $t - 100$ would grow to at time $t - 100 + i$ if the stock price followed $\{\tilde{p}_t\}$. It is easy to see that this process $\{x_t \mid t = 0, 1, 2, \cdots\}$ is Markov. Furthermore, it is ergodic since, for any $t \in \{0, 1, 2, \cdots\}$, the random variables x_t and x_{t+100} are independent and identically distributed. Letting $\alpha = e^{-\rho}$, $G(x) = x(100)$, and

$$x = \left(\frac{p_{-99}}{p_{-100}}, \frac{p_{-98}}{p_{-100}}, \cdots, \frac{p_t}{p_{-100}} \right)'$$

the value of the derivative security is given by

$$\sup_{\tau \in U} E[\alpha^\tau G(x_\tau) \mid x_0 = x].$$

If τ^* is an optimal stopping time, we have

$$E[\alpha^{\tau^*} G(x_{\tau^*}) \mid x_0 = x] = \sup_{\tau \in U} E[\alpha^\tau G(x_\tau) \mid x_0 = x]$$

for almost every x_0. Hence, given an optimal stopping time, we can price the security by evaluating an expectation, possibly through use of Monte Carlo simulation. However, because the state space is so large, it is unlikely that we will be able to compute an optimal stopping time. Instead, we must resort

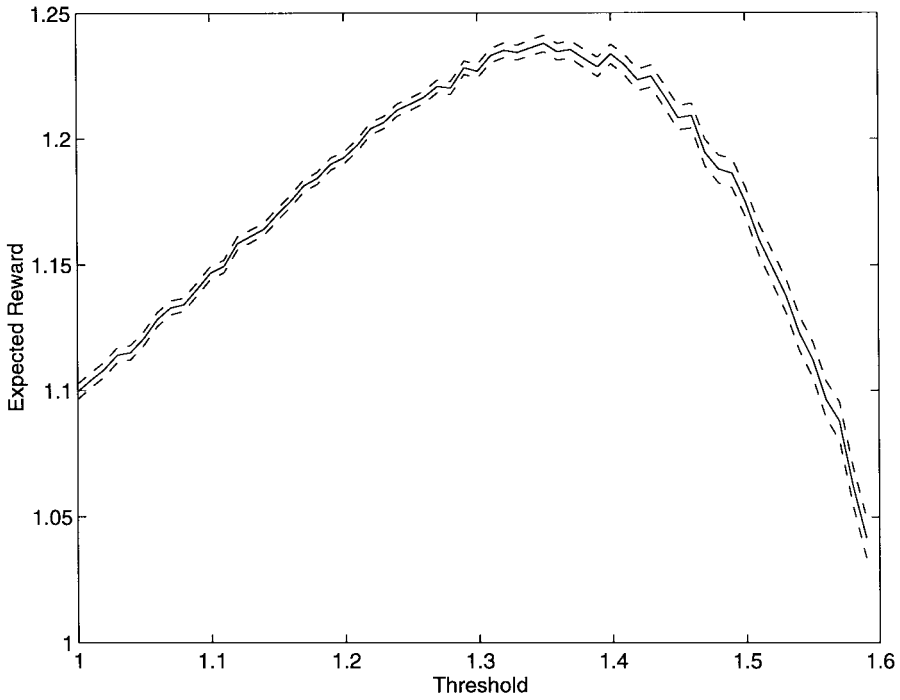


Fig. 1. Expected reward as a function of threshold. The values plotted are estimates generated by averaging rewards obtained over 10 000 simulated trajectories, each initialized according to the steady-state distribution and terminated according to the stopping time dictated by the thresholding strategy. The dashed lines represent confidence bounds generated by estimating the standard deviation of each sample mean, and adding/subtracting twice this estimate to/from the sample mean.

to generating a suboptimal stopping time $\tilde{\tau}$ and computing

$$E[\alpha^{\tilde{\tau}} G(x_{\tilde{\tau}}) \mid x_0 = x]$$

as an approximation to the security price. Note that this approximation is a lower bound for the true price. The approximation generally improves with the performance of the optimal stopping strategy. In the next two sections, we present computational results involving the selection of stopping times for this problem and the assessment of their performance. In the particular example we will consider, we use the settings $\sigma = 0.02$ and $\rho = 0.0004$ (the value of the drift μ is inconsequential). Intuitively, these choices correspond to a stock with a daily volatility of 2% and an annual interest rate of about 10% (assuming that interest only compounds while the market is open).

C. A Thresholding Strategy

In order to provide a baseline against which we can compare the performance of our approximation algorithm, let us first discuss the performance of a simple heuristic stopping strategy. In particular, consider the stopping time $\tau_B = \min\{t \mid G(x_t) \geq B\}$ for a scalar threshold $B \in \Re$. We define the performance of such a stopping time in terms of the expected reward $E[J^{\tau_B}(x_0)]$. In the context of our pricing problem, this quantity represents the average price of the derivative security (averaged over possible initial states). Expected rewards generated by various threshold values are presented in Fig. 1. The optimal expected reward over the thresholds tried was 1.238.

It is clear that a thresholding strategy is not optimal. For instance, if we know that there was a large slump and recovery in the process $\{\tilde{p}_t\}$ within the past 100 days, we should probably wait until we are about 100 days past the low point in order to reap potential benefits. However, the thresholding strategy, which relies exclusively on the ratio between $\tilde{p}_t$ and $\tilde{p}_{t-100}$, cannot exploit such information.

What is not clear is the *degree* to which the thresholding strategy can be improved. In particular, it may seem that events in which such a strategy makes significantly inadequate decisions are rare, and it therefore might be sufficient, for practical purposes, to limit attention to thresholding strategies. In the next section, we rebut this hypothesis by generating a substantially superior stopping time using our approximation methodology.

D. Using the Approximation Algorithm

Perhaps the most important step prior to applying our approximation algorithm is selecting an appropriate set of basis functions. Though analysis can sometimes help, this task is largely an art form, and the process of basis function selection typically entails repetitive trial and error.

We were fortunate in that our first choice of basis functions for the problem at hand delivered promising results relative to thresholding strategies. To generate some perspective, along with describing the basis functions, we will provide brief discussions concerning our (heuristic) rationale for selecting them. The first two basis functions were simply a constant function $\phi_1(x) = 1$ and the reward function $\phi_2(x) = G(x)$. Next, thinking that it might be important to know the maximal and minimal returns over the past 100 days, and how long ago they occurred, we constructed the following four basis

functions:

$$\phi_3(x) = \min_{i=1,\cdots,100} x(i) - 1$$
$$\phi_4(x) = \max_{i=1,\cdots,100} x(i) - 1$$
$$\phi_5(x) = \frac{1}{50} \arg\min_{i=1,\cdots,100} x(i) - 1$$
$$\phi_6(x) = \frac{1}{50} \arg\max_{i=1,\cdots,100} x(i) - 1.$$

Note that the basis functions involve constant scaling factors and/or offsets. The purpose of these transformations is to maintain the ranges of basis function values within the same regime. Though this is not required for convergence of our algorithm, it can speed up the process significantly.

As mentioned previously, if we invested one dollar in the stock at time $t-100$ and the stock price followed the process $\{\tilde{p}_t\}$, then the sequence $x_t(1), \cdots, x_t(100)$ represents the daily values of the investment over the following 100-day period. Conjecturing that the general shape of this 100-day sample path is of importance, we generated four basis functions aimed at summarizing its characteristics. These basis functions represent inner products of the sample path with Legendre polynomials of degrees one through four. In particular, letting $j = i/50 - 1$, we defined

$$\phi_7(x) = \frac{1}{100} \sum_{i=1}^{100} \frac{x(i) - 1}{\sqrt{2}}$$
$$\phi_8(x) = \frac{1}{100} \sum_{i=1}^{100} x(i) \sqrt{\frac{3}{2}} j$$
$$\phi_9(x) = \frac{1}{100} \sum_{i=1}^{100} x(i) \sqrt{\frac{5}{2}} \left(\frac{3j^2}{2} - \frac{1}{2} \right)$$
$$\phi_{10}(x) = \frac{1}{100} \sum_{i=1}^{100} x(i) \sqrt{\frac{7}{2}} \left(\frac{5j^3}{2} - \frac{3j}{2} \right).$$

So far, we have constructed basis functions in accordance with "features" of the state that might be pertinent to effective decision-making. Since our approximation of the value function will be composed of a weighted sum of the basis functions, the nature of the relationship between these features and approximated values is restricted to linear. To capture more complex tradeoffs between features, it is useful to consider nonlinear combinations of certain basis functions. For our problem, we constructed six additional basis functions using products of the original features. These basis functions are given by

$$\phi_{11}(x) = \phi_2(x)\phi_3(x)$$
$$\phi_{12}(x) = \phi_2(x)\phi_4(x)$$
$$\phi_{13}(x) = \phi_2(x)\phi_7(x)$$
$$\phi_{14}(x) = \phi_2(x)\phi_8(x)$$
$$\phi_{15}(x) = \phi_2(x)\phi_9(x)$$
$$\phi_{16}(x) = \phi_2(x)\phi_{10}(x).$$

Using our 16 basis functions, we generated a sequence of parameters $r_0, r_1, \cdots, r_{10^6}$ by initializing each component of r_0 to zero and iterating the update equation 1 000 000 times with a step size of $\gamma_t = 0.001$. The evolution of the iterates is illustrated in Fig. 2.

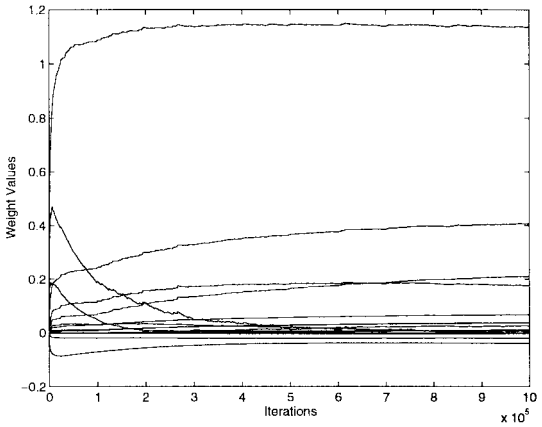


Fig. 2. The evolution of weights during execution of the algorithm. The value of the security under the resulting strategy was 1.282.

The weight vector r_{10^6} resulting from our numerical procedure was used to generate a stopping time $\tilde{\tau} = \min\{t \mid G(x_t) \geq (\Phi r_{10^6})(x_t)\}$. The corresponding expected reward $E[J^{\tilde{\tau}}(x_0)]$, estimated by averaging the results of 10 000 trajectories each initialized according to the steady-state distribution and terminated according to the stopping time $\tilde{\tau}$, was 1.282 (the estimated standard deviation for this sample mean was 0.0022). This value is significantly greater than the expected reward generated by the optimized threshold strategy of the previous section. In particular, we have

$$E[J^{\tilde{\tau}}(x_0) - J^{\tau_B}(x_0)] \approx 0.044.$$

As a parting note, we mention that each stopping time τ corresponds to an exercising strategy that the holder of the security may follow, and $J^{\tau}(x_0)$ represents the value of the security under this exercising strategy. Hence, the difference between $E[J^{\tilde{\tau}}(x_0)]$ and $E[J^{\tau_B}(x_0)]$ implies that, on average (with respect to the steady-state distribution of x_t), the fair price of the security is about 4% higher when exercised according to $\tilde{\tau}$ instead of τ_B. In the event that a bank assumes that τ_B is optimal and charges a price of $J^{\tau_B}(x_0)$, an arbitrage opportunity may become available.

V. Conclusion

We have introduced a theory and algorithm pertaining to approximate solutions of optimal stopping problems. The algorithm involves Hilbert space approximation of the value function via a linear combination of user-selected basis functions and can be thought of as a "regression method" for optimal stopping rather than statistical modeling. We believe that the methodology provides a systematic approach to dealing with complex optimal stopping problems such as those arising in the exotic derivatives trade, and our computational study provides some preliminary support for this view.

Though the algorithm and theory developed in this paper are useful in their own right, they represent contributions to a broader context. In particular, our algorithms exemplify methods from the emerging fields of neuro-dynamic programming and reinforcement learning that have been successful in

solving a variety of large-scale stochastic control problems [6]. We hope that our treatment of optimal stopping problems will serve as a starting point for further analysis of methods with broader scope.

Indeed, many ideas in this paper were motivated by research in neuro-dynamic programming and reinforcement learning. The benefits of switching the order of expectation and maximization by employing "Q-functions" instead of value functions were first recognized by Watkins [22] and Watkins and Dayan, [23]. The type of stochastic approximation update rule that we use to tune weights of a linear combination of basis functions resembles temporal-difference methods originally proposed by Sutton [17], who also conjectured that the use of simulated trajectories in conjunction with such algorithms could be important for convergence [18]. This observation was later formalized by Tsitsiklis and Van Roy [19], who analyzed temporal-difference methods and provided a counterexample as discussed in Section III-A (a related counter-example has also been proposed by Baird [1]). Bertsekas and Tsitsiklis [6] summarize much work directed at understanding such algorithms.

Finally, we should mention that the line of analysis developed in this paper extends easily to additional classes of optimal stopping problems. Several such extensions, including those involving finite horizons, independent increment processes, and stopping games, are treated in [21].

ACKNOWLEDGMENT

The authors would like to thank D. Bertsekas and V. Borkar for useful discussions. They also thank the anonymous reviewers for many detailed comments that have improved the paper.

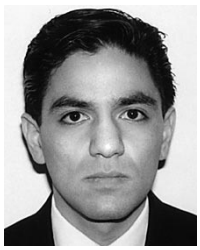
REFERENCES

[1] L. C. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Machine Learning: Proc. Twelfth Int. Conf.*, Prieditis and Russell, Eds. San Francisco, CA: Morgan Kaufman, 1995.
[2] J. Barraquand and D. Martineau, "Numerical valuation of high dimensional multivariate american securities," to be published.
[3] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. Berlin, Germany: Springer-Verlag, 1990.
[4] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.
[5] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*. Belmont, MA: Athena Scientific, 1996.
[6] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.
[7] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *J. Political Economy*, vol. 81, pp. 637–654, 1973.
[8] M. Broadie and P. Glasserman, "Pricing american-style securities using simulation," *J. Economic Dynam. Contr.*, vol. 21, pp. 1323–1352, 1997.
[9] ——, "A stochastic mesh method for pricing high-dimensional american options," to be published.
[10] J. M. Harrison and D. Kreps, "Martingales and arbitrage in multiperiod securities markets," *J. Economic Theory*, vol. 20, pp. 381–408, 1979.
[11] J. M. Harrison and S. Pliska, "Martingales and stochastic integrals in the theory of continuous trading," *Stochastic Processes and Their Appl.*, vol. 11, pp. 215–260, 1981.
[12] I. Karatzas, "On the pricing of american options," *Applied Math. Operations Res.*, pp. 37–60, 1988.
[13] R. C. Merton, "Theory of rational option pricing," *Bell J. Economics and Management Sci.*, vol. 4, pp. 141–183, 1973.
[14] M. Parsley, "Exotics enter the mainstream," *Euromoney*, pp. 127–130, Mar. 1997.
[15] J. Rust, "Using randomization to break the curse of dimensionality," *Econometrica*, 1996.
[16] A. N. Shiryaev, *Optimal Stopping Rules*. New York: Springer-Verlag, 1978.
[17] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, 1988.
[18] ——, "On the virtues of linear learning and trajectory distributions," in *Proc. Workshop on Value Function Approximation, Machine Learning Conf. 1995*, Boyan, Moore, and Sutton, Eds. Technical Rep. CMU-CS-95-206, Carnegie Mellon Univ., Pittsburgh, PA 15213, 1995, p. 85.
[19] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. Automat. Contr.*, May 1997.
[20] ——, "Approximate solutions to optimal stopping problems," in *Advances in Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997.
[21] B. Van Roy, "Learning and value function approximation in complex decision processes," Ph.D. dissertation, MIT, June 1998.
[22] C. J. C. H. Watkins, "Learning from delayed rewards," Doctoral dissertation, Univ. Cambridge, Cambridge, U.K., 1989.
[23] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

John N. Tsitsiklis (S'80–M'81–SM'97–F'99) was born in Thessaloniki, Greece, in 1958. He received the B.S. degree in mathematics in 1980 and the B.S. degree in 1980, M.S. degree in 1981, and Ph.D. degree in 1984 in electrical engineering, all from the Massachusetts Institute of Technology, Cambridge, Massachusetts.

During the academic year 1983 to 1984, he was an Acting Assistant Professor of Electrical Engineering at Stanford University, Stanford, CA. Since 1984, he has been with the Massachusetts Institute of Technology, where he is currently a Professor of Electrical Engineering. He has also been a Visitor with the Department of Electrical Engineering and Computer Science at the University of California, Berkeley, and the Institute for Computer Science in Iraklion, Greece. His research interests include the fields of systems, optimization, control, and operations research. He is a coauthor of *Parallel and Distributed Computation: Numerical Methods* (with D. Bertsekas, 1989), *Neuro-Dynamic Programming* (with Dimitri Bertsekas, 1996), and *Introduction to Linear Optimization* (with Dimitris Bertsimas, 1997).

Dr. Tsitsiklis has been a recipient of an IBM Faculty Development Award (1983), an NSF Presidential Young Investigator Award (1986), an Outstanding Paper Award by the IEEE Control Systems Society, the M.I.T. Edgerton Faculty Achievement Award (1989), the Bodossakis Foundation Prize (1995), and the INFORMS/CSTS prize (1997). He was a plenary speaker at the 1992 IEEE Conference on Decision and Control. He is an Associate Editor of *Applied Mathematics Letters* and has been an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and *Automatica*. He is a member of SIAM and INFORMS.



Benjamin Van Roy received the S.B. degree in computer science and engineering and the S.M. and Ph.D. degrees in electrical engineering and computer science, all from the Massachusetts Institute of Technology, Cambridge, in 1993, 1995, and 1998, respectively.

From 1993 to 1997, he worked with Unica Technologies. He also co-authored a book, *Solving Pattern Recognition Problems*, with several members of Unica's technical staff. During the summer of 1997, he worked with the Equity Derivatives Group at Morgan Stanley. He is currently an Assistant Professor and Terman Fellow in the Department of Engineering-Economic Systems and Operations Research at Stanford University, with courtesy appointments in the Departments of Electrical Engineering and Computer Science. His research interests include the control of complex systems, computational learning, and financial economics.

During his time at MIT, Dr. Vam Roy received a Digital Equipment Corporation Scholarship, the George C. Newton (undergraduate laboratory project) Award, the Morris J. Levin Memorial (Master's thesis) Award, and the George M. Sprowls (Ph.D. dissertation) Award.