May, 1997 LIDS-P 2389

Research Supported By:

NSF grant DMI-9625489
ARO grant DAAL-03-92-G-0115

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

Tsitsiklis, J.N.
Roy, B.V.

May 15, 1997

LIDS-P-2389

Optimal Stopping of Markov Processes: Hilbert Space Theory, Approximation Algorithms, and an Application to Pricing High–Dimensional Financial Derivatives[1]

John N. Tsitsiklis and Benjamin Van Roy

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
e-mail: jnt@mit.edu, bvr@mit.edu

[1] This research was supported by the NSF under grant DMI-9625489 and the ARO under grant DAAL-03-92-G-0115.

ABSTRACT

We 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, we propose a stochastic approximation algorithm that tunes weights of a linear combination of basis functions in order to approximate a value function. We prove that this algorithm converges (almost surely) and that the limit of convergence has some desirable properties. We discuss how variations on this line of analysis can be used to develop similar results for other classes of optimal stopping problems, including those involving independent increment processes, finite horizons, and two–player zero–sum games. We illustrate the approximation method with a computational case study involving the pricing of a path–dependent financial derivative security that gives rise to an optimal stopping problem with a one–hundred–dimensional state space.

1 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 illustrate the method, we present a computational case study involving the pricing of a (fictitious) financial derivative instrument.

Shiryaev (1978) 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 important, 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 (Bertsekas, 1995; Bertsekas and Shreve, 1996). In particular, the algorithm exemplifies simulation–based optimization techniques from the field of neuro-dynamic programming, pioneered by Barto, Sutton (1988), and Watkins (1989), that have been successfully applied to a variety of large–scale stochastic control problems (Bertsekas and Tsitsiklis, 1996). The practical success of these algorithms is not fully explained by existing theory, and our analysis represents progress towards an improved understanding.

This paper is organized as follows. The next section introduces our problem formulation. Section 3 defines a restricted class of problems we consider (ergodic Markov processes with discounted rewards) and develops some basic theory concerning optimal stopping times and optimal rewards for such problems. Section 4 introduces and analyzes the approximation algorithm. A computational case–study involving the pricing of a financial derivative instrument is described in Section 5. Section 6 presents several additional classes of optimal stopping problems to which our analysis can be extended, including independent increment processes, finite–horizon problems, and Markov games. Finally, 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 (Tsitsiklis and Van Roy, 1997) and are also included in (Bertsekas and Tsitsiklis, 1996).

2 The Optimal Stopping Problem

In this section, we present a rather general problem formulation. Our analysis later in the paper actually requires significant assumptions that will further constrain the characteristics of the problem. However, beginning with a more general formulation allows for clearer exposition.

We consider a stochastic process $\{x_t | t = 0, 1, 2, \ldots\}$ 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 $\{x_0, x_1, \ldots, x_t\}$ by $\mathcal{F}_t \subset \mathcal{F}$.

We define a stopping time to be a random variable τ that takes on values in $\{0, 1, 2, \ldots, \infty\}$ and satisfies $\{\omega \in \Omega | \tau(\omega) \leq t\} \in \mathcal{F}_t$ 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 $\{x_0, x_1, \ldots, x_t\}$, 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 $\{x_t | t = 0, 1, 2, \ldots\}$, reward functions $g : \Re^d \mapsto \Re$ and $G : \Re^d \mapsto \Re$ associated with continuation and termination, and a discount factor $\alpha \in (0, 1]$. 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.

3 Basic Theory

In this section, we develop some basic theory about a limited class of optimal stopping problems. Perhaps the most severe restriction we will impose is that the underlying stochastic process is Markov and ergodic. In addition, we require that the discount factor is strictly less than one and that the reward functions satisfy certain technical conditions. Together, these restrictions enable an elegant analysis, establishing existence and constructive characterizations of optimal stopping times. Most importantly, however, this analysis leads to the development of approximation algorithms, as will be presented in Section 4.

3.1 Assumptions and Main Result

We begin by stating our assumptions and a theorem characterizing optimal stopping times for problems satisfying the assumptions. Our first assumption places restrictions on the underlying stochastic process.

Assumption 1 *The process $\{x_t | t = 0, 1, 2, \ldots\}$ 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. The Markov condition corresponds to the existence of a transition probability kernel $P : \Re^d \times \mathcal{B}(\Re^d) \mapsto [0, 1]$ satisfying

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

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

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

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 probability measure, which can be interpreted as a "steady–state distribution," is unique. 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 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* $(0, 1)$.

Before stating the main result of this section, we 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

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

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 through 3, the following statements hold:*
(a) There exists a function $J^* \in L_2(\pi)$ uniquely satisfying

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

(b) The stopping time τ^*, defined by

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

is an optimal stopping time. (The minimum of an empty set is taken to be ∞.)
(c) The function J^{τ^*} is equal to J^* (in the sense of $L_2(\pi)$).

3.2 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

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

q.e.d.

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

Lemma 2 *Under Assumptions 1 and 2, 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. **q.e.d.**

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

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

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

(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, 2, and 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^*. **q.e.d.**

3.3 Proof of Theorem 1

Part (a) of the result follows from Lemma 2. As for Part (c), 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 (b). For any nonnegative integer n, we have

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

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 $T^n G$, and therefore also $\sup_{\tau \in U} J^{\tau \wedge n}(\cdot)$, is measurable. It follows that

$$\sup_{\tau \in U} E[J^{\tau \wedge n}(x_0)] \leq E[\sup_{\tau \in U} J^{\tau \wedge n}(x_0)] = E[(T^n G)(x_0)].$$

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

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

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

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

and we therefore have

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

Hence, stopping time τ^* is optimal. **q.e.d.**

4 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 $\{\phi_k : \Re^d \mapsto \Re | k = 1, 2, \ldots, K\}$ and computing weights $r(1), \ldots, r(k) \in \Re$ 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. In this section, we introduce one such algorithm and provide an analysis of its behavior.

We begin by presenting our algorithm and a theorem that establishes certain desirable properties. Sections 4.2 and 4.3 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 of the Markov process. To illustrate this fact, in Section 4.4, we propose a generalization of the algorithm that does not employ a simulated trajectory, and we demonstrate through a counterexample that this new algorithm is not sound.

4.1 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 (Watkins, 1989). 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 0. An optimal stopping time can be generated according to

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

Our approximation algorithm employs a set of basis functions $\phi_1, \ldots, \phi_K \in L_2(\pi)$ 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 $r = (r(1), \ldots, r(K))'$. 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 $r_0 = (r_0(1), \ldots, r_0(K))' \in \Re^K$. During the simulation of a trajectory $\{x_t | t = 0, 1, 2, \ldots\}$ of the Markov chain, the algorithm generates a sequence of weight vectors $\{r_t | t = 1, 2, \ldots\}$ according to

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

where each γ_t is a positive scalar step size. 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

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

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 *(a) The basis functions $\phi_1, \ldots, \phi_K$ are linearly independent.*
(b) 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 *(a) For any positive scalar q, there exists a scalar μ_q such that for all x and t,*

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

(b) 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,

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

(c) 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 $\{\Phi r | r \in \Re^K\}$ of $L_2(\pi)$. In particular, for any function $Q \in L_2(\pi)$, let

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

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 through 6, the following hold:*
(a) The approximation algorithm converges almost surely.
(b) The limit of convergence r^* is the unique solution of the equation

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

(c) Furthermore, r^* satisfies

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

(d) Let $\tilde{\tau}$ be defined by

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

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 (c) and (d) 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.

4.2 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, 2, and 3, the operator F satisfies*

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

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

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

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

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. **q.e.d.**

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, 2, 3, and 4, the composition* ΠF *satisfies*

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

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

$$\|\Pi F Q - \Pi F \overline{Q}\|_\pi \leq \|F Q - F \overline{Q}\|_\pi \leq \alpha \|Q - \overline{Q}\|_\pi,$$

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

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

q.e.d.

Given, r^*, we define a stopping time $\tilde{\tau} = \min\{t | G(x_t) \geq (\Phi r^*)(x_t)\}$. 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

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

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

Lemma 6 *Under Assumptions 1, 2, 3, and 4, for any* $Q, \overline{Q} \in L_2(\pi)$,

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

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

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

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

where the first inequality follows from Lemma 1.

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

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

Therefore,

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

as desired. **q.e.d.**

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, 2, 3, and 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

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

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, and the equality $FQ^* \overset{ae(\pi)}{=} Q^*$ (Lemma 4), 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

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

q.e.d.

We now continue with the analysis of the stochastic algorithm. Let us define a stochastic process $\{z_t | t = 0, 1, 2, \ldots\}$ 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 Equation (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

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

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

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

(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 $\overline{s}_k(r)$ can be represented in terms of an inner product according to

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

Lemma 8 *Under Assumptions 1, 2, 3, and 4, we have*

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

Proof: For any r, we have

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

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

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

By Assumption 4(a), for any $r \neq r^*$, we have $\|\Phi r - \Phi r^*\|_\pi \neq 0$. Since $\alpha < 1$, the result follows. **q.e.d.**

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 (Theorem 17, on page 239 of (Benveniste et al., 1990)). It is straightforward to check that all of the assumptions in the result of (Benveniste et al., 1990) follow from the assumptions imposed in the result below. We do not show here the assumptions of (Benveniste et al., 1990) 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 (Benveniste et al., 1990) 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 $s : \Re^N \times \Re^K \mapsto \Re^K$, where:
(a) $\{z_t | t = 0, 1, 2, \ldots\}$ is a (stationary) ergodic Markov process taking values in $\Re^N$.
(b) For any positive scalar q, there exists a scalar μ_q such that $E[1 + \|z_t\|^q | z_0 = z] \leq \mu_q(1 + \|z\|^q)$, for any time t and $z \in \Re^N$.
(c) The (predetermined) step size sequence γ_t is nonincreasing and satisfies $\sum_{t=0}^\infty \gamma_t = \infty$ and $\sum_{t=0}^\infty \gamma_t^2 < \infty$.
(d) There exist scalars C and q such that

$$\|s(z,r)\| \leq C(1 + \|r\|)(1 + \|z\|^q), \qquad \forall z, r.$$

(e) There exist scalars C and q such that

$$\sum_{t=0}^{\infty} \Big\| E[s(z_t, r) | z_0 = z] - E[s(z_0, r)] \Big\| \leq C(1 + \|r\|)(1 + \|z\|^q), \qquad \forall z, r.$$

(f) There exist scalars C and q such that

$$\sum_{t=0}^{\infty} \Big\| E[s(z_t, r) - s(z_t, \overline{r}) | z_0 = z] - E[s(z_0, r) - s(z_0, \overline{r})] \Big\| \leq C\|r - \overline{r}\|(1 + \|z\|^q), \qquad \forall z, r, \overline{r}.$$

(g) There exists some $r^* \in \Re^K$ such that $\overline{s}(r)'(r - r^*) < 0$, for all $r \neq r^*$, and $\overline{s}(r^*) = 0$. Then, r_t almost surely converges to r^*.

4.3 Proof of Theorem 2

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

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

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

Condition (d) then easily follows from the polynomial bounds of Assumption 5(c). Given that Condition (d) is valid, Condition (e) follows from Assumptions 5(a) and 5(b) in a straightforward manner. (Using these assumptions, it is easy to show that a condition analogous to Assumption 5(b) 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 Condition (f). We first note that for any r, $\bar{r}$, and z, we have

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

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

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

Finally, it follows from Assumptions 5(a) and 5(b) that there exist scalars C_1 and q_1 such that for any r, $\bar{r}$, and c,

$$\sum_{t=0}^{\infty} \Big\|E[s(z_t,r) - s(z_t,\bar{r})|z_0 = z] - E[s(z_0,r) - s(z_0,\bar{r})]\Big\| \leq C_1C_2\|r - \bar{r}\|)(1 + \|z\|^{q_1q_2}).$$

This establishes the validity of Condition (f).

Validity of Condition (g) is assured by Lemma 8. This completes the proof for Part (a) of the theorem. To wrap up the proof, Parts (b) and (c) of the theorem follow from Lemma 5, while Part (d) is established by Lemma 7. **q.e.d.**

4.4 On the Importance of Simulated Trajectories

The approximation algorithm we analyzed iteratively updates a weight vector based on states visited during a simulated trajectory. There are fundamental reasons for using an entire simulated trajectory. To illustrate this fact, we will introduce a variant of the algorithm that only simulates individual transitions originating from states sampled in some prespecified manner. It turns out that this new algorithm can diverge. We demonstrate this shortcoming by presenting a simple counterexample.

Consider an algorithm that, on each tth step, samples a state $y_t \in \Re^d$ according to a probability measure $\overline{\pi} : \mathcal{B}(\Re^d) \mapsto [0,1]$ and a state $\overline{y}_t \in \Re^d$ according to $\text{Prob}[\overline{y}_t \in A] = P(y_t, A)$, and updates the weight vector according to

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

At first sight, this algorithm may be expected to deliver results similar to those offered by the one discussed in the previous section. However, this algorithm can actually lead to very different behavior and may not even converge.

To illustrate the potential for divergence, we provide a simple example involving only two states. Indexing the states as 1 and 2, let $\alpha \in (5/6, 1)$, and let the probability measure π satisfy $1 > \pi(2) > 5/6\alpha$. We define the transition probability kernel P by $P(y, \{1\}) = \pi(1)$ and $P(y, \{2\}) = \pi(2)$. To conclude the definition of our example, let all continuation and termination rewards be zero (i.e., $g = 0$ and $G = 0$). Note that $J^* = 0$, since no rewards are ever obtained.

Prior to executing the algorithm, we must define a sampling distribution and basis functions. Suppose we define the sampling distribution $\overline{\pi}$ such that $\overline{\pi}(1) \geq \overline{\pi}(2)$. Let us employ one basis function $\phi : \{1, 2\} \mapsto \Re$, defined by $\phi(1) = 1$ and $\phi(2) = 2$. Given that we only have one basis function, r_t is a sequence of scalars.

For the example we have described, using algebra identical to that of Section 9 of (Tsitsiklis and Van Roy, 1997), we can show that

$$E[r_{t+1}|r_0 = r] \geq (1 + \gamma_t \overline{\pi}(1)\epsilon) E[r_t|r_0 = r],$$

and since $\sum_{t=0}^{\infty} \gamma_t = \infty$, we have

$$\lim_{t\to\infty} E[r_{t+1}|r_0 = r] = \infty, \qquad \forall r > 0.$$

Let us close this section by reflecting on the implications of this counterexample. It demonstrates that, if the sampling distribution $\overline{\pi}$ is chosen independently of the dynamics of the Markov process, there is no convergence guarantee. Clearly, this does not imply that divergence will always occur when such a random sampling scheme is employed in practice. In fact, for any problem, there is a set of sampling distributions that lead to convergence. Our current understanding indicates that π is an element of this set, so it seems sensible to take advantage of this knowledge by setting $\overline{\pi} = \pi$. However, since it is often difficult to model the ergodic measure, one generally must resort to simulation in order to generate the desired samples. This leads back to the algorithm considered in Section 4.

5 Pricing Financial Derivatives

In this section, we illustrate the steps required in applying our algorithm by describing a simple case study involving the pricing of a fictitious high–dimensional financial derivative security. In this context, our approach involving the approximation of a value function is similar in spirit to the earlier experimental work of Barraquand and Martineau (1995). However, our algorithms are different from the ones they used.

We will begin by providing some background on financial derivative securities. Section 5.2 then introduces the particular security we consider and a related optimal stopping problem. Section 5.3 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 5.4.

5.1 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.

Exotic derivatives tend to be illiquid relative to securities that are traded in mainstream markets. Consequently, it may be difficult for an institution to "cash in" on the worth of the contract when the need arises unless such a situation is explicitly accommodated by the terms of the contract. Because institutions desire flexibility, derivatives typically allow the possibility of "early exercise." In particular, an institution may "exercise" the security at various points during the lifetime of the contract, thereby settling with the issuer according to certain prespecified terms.

Several important considerations come into play when a bank designs a derivative security. First, the product should well suit the needs of clients, incurring low costs for large gains in customer satisfaction. Second, it is necessary to devise a hedging strategy, which is a plan whereby the bank can be sure to fulfill the terms of the contract without assuming significant risks. Finally, the costs of implementing the hedging strategy must be computed in order to determine an appropriate price to charge clients.

When there is no possibility of early exercise and certain technical conditions are satisfied, 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, 1973) and (Merton, 1973) and presented in its definitive form by (Harrison and Kreps, 1979) and (Harrison and Pliska, 1981), has met wide application and is the subject of much subsequent research.

When there is a possibility of early exercise, the value of the derivative security depends on how the client chooses a time to exercise. Given that the bank can not 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.

5.2 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 one hundred days beforehand.

One interesting interpretation of this derivative security is as an oracle that offers a degree of foresight. The payoff is equal to the amount that would accrue from a one Dollar investment in the stock made one hundred days prior to exercise. However, the holder of the security can base her choice of the time at which this Dollar is invested on knowledge of the returns over the one hundred days. The price of this security should in some sense represent the value of this foresight.

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 (1988) 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 (Karatzas, 1988) 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 $t \in \{0, 1, 2, \ldots\}$. We model the stock price process $\{p_t | t \geq -100\}$ 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_0, 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 a hundred 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 $\{\tilde{p}_t | t \geq -100\}$ be a stochastic process that evolves according to

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

Define a discrete–time process $\{x_t | t = 0, 1, 2, \ldots\}$ taking values in $\Re^{100}$, with

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

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 $\{x_t | t = 0, 1, 2, \ldots\}$ is Markov. Furthermore, it is ergodic since, for any $t \in \{0, 1, 2, \ldots\}$, the random variables x_t and x_{t+100} are independent and identically distributed. Letting $\alpha = e^{-\rho}$, $G(x) = x(100)$, and

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

the value of the derivative security is given by

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

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

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

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

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

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%.

5.3 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 $\tau_B = \min\{t | G(x_t) \geq B\}$ 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 Figure 1. The optimal expected reward over the thresholds tried was 1.238.

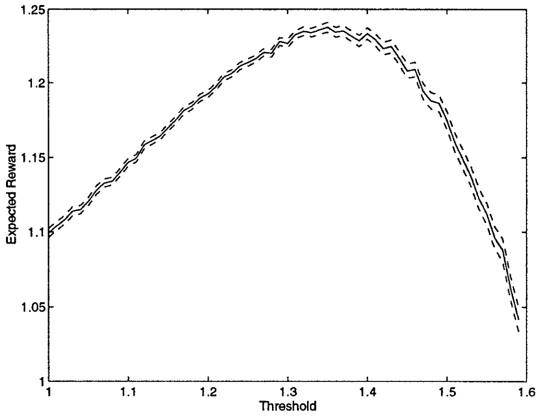

Figure 1: Expected reward as a function of threshold. The values plotted are estimates generated by averaging rewards obtained over ten thousand simulated trajectories, each initialized according to the steady–state distribution and terminated according to the stopping time dictated by the thresholding 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.

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 hundred days, we should probably wait until we are about a hundred 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.

5.4 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 hundred days, and how long ago they occurred, we constructed the following four basis functions

$$\begin{aligned}
\phi_3(x) &= \min_{i=1,\ldots,100} x(i) - 1, \\
\phi_4(x) &= \max_{i=1,\ldots,100} x(i) - 1, \\
\phi_5(x) &= \frac{1}{50} \arg \min_{i=1,\ldots,100} x(i) - 1, \\
\phi_6(x) &= \frac{1}{50} \arg \max_{i=1,\ldots,100} x(i) - 1.
\end{aligned}$$

Note that that the basis functions involve constant scaling factors and/or offsets. The purpose of these transformation 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 $x_t(1), \ldots, x_t(100)$ represents the daily values of the investment over the following hundred day period. Conjecturing that the general shape of this hundred–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

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

So far, we have constructed basis functions in accordance to "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 trade-offs 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

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

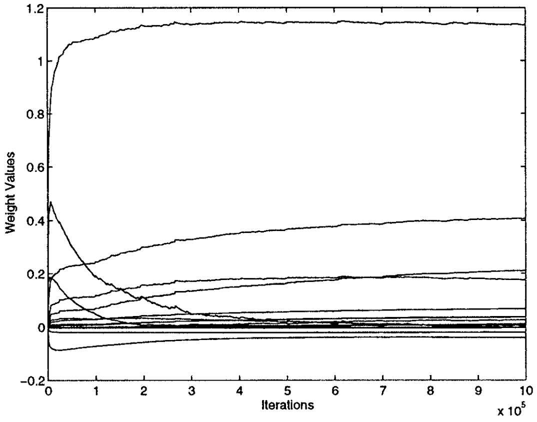

Figure 2: The evolution of weights during execution of the algorithm. The value of the security under the resulting strategy was 1.282.

Using our sixteen basis functions, we generated a sequence of parameters $r_0, r_1, \ldots, r_{10^6}$ by initializing each component of r_0 to 0 and iterating the update equation one million times with a step size of $\gamma_t = 0.001$. The evolution of the iterates is illustrated in Figure 2.

The weight vector r_{10^6} resulting from our numerical procedure was used to generate a stopping time $\tilde{\tau} = \min\{t | G(x_t) \geq (\Phi r_{10^6})(x_t)\}$. The corresponding expected reward $E[J^{\tilde{\tau}}(x_0)]$, estimated by averaging the results of ten thousand 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 four percent 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.

6 Extensions

Our line of analysis can be extended to encompass additional classes of optimal stopping problems. In this section, we describe several such classes and applicable approximation algorithms. Our discussion is less rigorous than that of previous sections. In particular, we do not prove formal results. Instead, we provide coarse overviews of approaches to extending the theory and summarize the types of results that can be obtained.

6.1 Processes with Independent Increments

In our original problem formulation, we assumed that the Markov process of interest is ergodic. This assumption ensures a certain sense of "stability" in the underlying system. In particular, the probability distribution over states is stationary. In this subsection, we examine a class of "unstable" Markov processes in which the distribution over states becomes increasingly diffuse over time. Specifically, we will investigate stopping problems involving Markov processes with independent increments.

6.1.1 Problem Formulation

We assume that the underlying process is Markov and that the transition probability kernel satisfies

$$P(x, A) = P(0, \{y - x | y \in A\}) \qquad \forall x \in \Re^d, A \in \mathcal{B}(\Re^d).$$

Hence, each increment $x_{t+1} - x_t$ is independent of x_t and t.

Instead of assuming that the reward functions g and G are in $L_2(\pi)$, as we did in the case of ergodic processes, we will assume that they are elements of L_2 – the Hilbert space with inner product $\langle J, \overline{J} \rangle = \int J(x)\overline{J}(x)dx$ and norm $\|J\| = \sqrt{\int J^2(x)dx}$. This Hilbert space will substitute for the role played by $L_2(\pi)$ in the analysis of Sections 3 and 4.

6.1.2 Basic Theory

The keystone of our basic theory in Section 3 was Lemma 1, which stated that P is a nonexpansion in $L_2(\pi)$. Since processes with independent increments do not possess invariant distributions, a new notion is required. It turns out that the appropriate object is a new lemma, stating that for such processes, P is a nonexpansion in L_2. This fact can be established by an argument analogous to that used in proving Lemma 1. In particular, by Jensen's inequality, for any $J \in L_2$, we have

$$\begin{aligned} \|PJ\|^2 &= \int (PJ)^2(x)dx \\ &= \int \left(E[J(x_1)|x_0 = x]\right)^2 dx \\ &\leq \int E[J^2(x_1)|x_0 = x]dx, \end{aligned}$$

and noting that the increment $\Delta = x_1 - x_0$ is independent of x_0, it follows that

$$\begin{aligned} \|PJ\|^2 &\leq E\left[\int J^2(x + \Delta)dx\right] \\ &= E\left[\|J\|^2\right] \\ &= \|J\|^2. \end{aligned}$$

Using this fact and the arguments from Section 3, it is possible to prove an analog of Theorem 1 under a new set of assumptions: (a) the Markov process has independent increments; (b) the discount factor α is in $(0, 1)$; and (c) the reward functions g and G are in L_2. The results would be the same as those of Theorem 1, except for the fact that each statement that is true "almost everywhere with respect to π" is now true "almost everywhere with respect to Lebesgue measure." Furthermore, the stopping time τ^* satisfies

$$\int_A J^{\tau^*}(x)dx = \sup_{\tau \in U} \int_A J^{\tau}(x)dx,$$

for every set A with $0 < \int_A dx < \infty$.

6.1.3 An Approximation Algorithm

We will now discuss an approximation algorithm that is suitable for the new class of optimal stopping problems. Similar to the algorithm of Section 4, we start by selecting a set of linearly independent basis functions $\phi_1, \ldots, \phi_K$. We now require, however, that the basis functions are in L_2 and have compact support. In particular, there exists a set $A \in \mathcal{B}(\Re^d)$ such that $\int_A dx < \infty$ and $\phi_i(x) = 0$ for all $i \in \{1, \ldots, K\}$ and $x \notin A$.

Since the Markov process is "unstable," it is no longer viable to generate weight updates based solely on a single simulated trajectory. Instead, given the basis functions and an initial weight vector r_0, we generate a sequence according to

$$r_{m+1} = r_m + \gamma_t \phi(x_m)\Big(g(x_m) + \alpha \max\Big\{(\Phi r_m)(x_m + \Delta_m), G(x_m + \Delta_m)\Big\} - (\Phi r_m)(x_m)\Big),$$

where the x_m's are independent identically distributed random variables drawn from a uniform distribution over a compact set A that supports the basis functions, and each Δ_m is drawn independently from all other random variables according to $\text{Prob}[\Delta_m \in B] = P(0, B)$ for all $B \in \mathcal{B}(\Re^d)$.

Once more, we define the operator F by $FQ = g + \alpha P \max\{G, Q\}$. Defining Π to be the operator that projects in L_2 onto the span of the basis functions, we can establish that the composition ΠF is a contraction on L_2 using arguments from the proofs of Lemmas 4 and 5. Hence, ΠF has a unique fixed point Φr^*. Furthermore, bounds on approximation error and the quality of the resulting stopping time can be generated using arguments from the proofs of Lemmas 5 and 7.

Following the line of reasoning from Section 4, we rewrite the above update equation as

$$r_{m+1} = r_m + \gamma_m s(z_m, r_m),$$

this time with s defined by

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

for any r and $z = (x, \Delta)$. Furthermore, defining $\bar{s}(r) = \int s(z_0, r)dx$, we now have

$$\bar{s}_k(r) = \Big\langle \phi_k, F\Phi r - \Phi r \Big\rangle.$$

This relation enables us to prove an analog of Lemma 8, showing that $(r - r^*)'\bar{s}(r) < 0$ for all $r \neq r^*$. Using this fact and the theorem on stochastic approximation (Theorem 3), we can then establish that r_t almost surely converges to r^*.

In summary, under our new assumptions concerning the Markov chain and basis functions together with a technical step size condition (Assumption 6), we can establish results analogous to those of Theorem 2 for our new algorithm. The only differences are that the norm $\|\cdot\|_\pi$ is replaced by $\|\cdot\|$ and the fixed point equation is now true almost everywhere with respect to the Lebesgue measure. Note that the fact that each x_m is drawn independently alleviates the need for a counterpart to Assumption 5.

6.2 Finite Horizon Problems

In certain practical situations, one may be interested in optimizing rewards over only a finite time horizon. Such problems are generally simpler to analyze than their infinite horizon counterparts, but at the same time, involve an additional complication because expected rewards will generally depend on the remaining time. In the problem we consider, we fix the time horizon $h < \infty$ and we look for an optimal stopping time τ^* that satisfies

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

In this section, we develop an approximation algorithm that is suitable for such problems.

6.2.1 Basic Theory

We assume that Assumptions 1, 2, and 3, from Section 3, hold. The standard results in the finite–horizon dynamic programming literature apply. In particular, for any nonnegative integer h,

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

where

$$TJ = \max\{G, \alpha PJ\}, \qquad \forall J,$$

and $\tau_h \in U$ is defined by

$$\tau_h = \min\Big\{t \leq h | G(x_t) \geq (T^{h-t}G)(x_t)\Big\}.$$

(We let $\tau_h = \infty$ if the set is empty.) Hence, $\tau^* = \tau_h$ is an optimal stopping time.

6.2.2 An Approximation Algorithm

As in Section 4, let the operator F be defined by $FQ = g + \alpha P \max\{G, Q\}$. It is easy to verify that the optimal stopping time τ^* can alternatively be generated according to

$$\tau^* = \min\Big\{t \leq h | G(x_t) \geq (F^{h-t}G)(x_t)\Big\}.$$

Note that this construction relies on knowledge of $FG, F^2G, \dots, F^hG$. A suitable approximation algorithm should be designed to accommodate such needs.

Our approximation algorithm here employs a set of basis functions $\phi_1, \dots, \phi_K$, each mapping $\Re^d \times \{0, \dots, h\}$ to $\Re$ and satisfying $\phi_k(x, h) = G(x)$ for all k and x. Beginning with an arbitrary initial weight vector r_0, a sequence is generated according to

$$r_{t+1} = r_t + \gamma_t \sum_{i=0}^{h-1} \phi(x_t, i)\Big(g(x_t) + \alpha \max\{(\Phi r_t)(x_{t+1}, i+1), G(x_{t+1})\} - (\Phi r_t)(x_t, i)\Big),$$

where x_t is a trajectory of the Markov process and the step size sequence satisfies Assumption 6.

We now discuss how the ideas of Section 4 can be extended to this new context. Let the measure μ over the product space $(\mathcal{B}(\Re^d))^h$ be defined by

$$\mu(A) = \pi(A_0) + \pi(A_1) + \cdots + \pi(A_{h-1}), \qquad \forall A = A_0 \times A_1 \times \cdots \times A_{h-1},$$

and let $L_2(\mu)$ be the Hilbert space defined with respect to this measure. Note that $(Q_0, Q_1, \ldots, Q_{h-1}) \in L_2(\mu)$ if and only if $Q_0, Q_1, \ldots, Q_{h-1} \in L_2(\pi)$. We define an operator $H : L_2(\mu) \mapsto L_2(\mu)$ according to

$$H(Q_0, \ldots, Q_{h-1}) = (FQ_1, FQ_2, \ldots, FQ_{h-1}, FG), \qquad \forall Q_0, Q_1, \ldots, Q_{h-1} \in L_2(\pi).$$

Using Lemma 1, it is easy to show that H is a contraction in $L_2(\mu)$ with a contraction factor α. Furthermore, the unique fixed point is given by

$$(Q_0^*, Q_1^*, \ldots, Q_{h-1}^*) \overset{ae(\mu)}{=} (F^h G, F^{h-1} G, \ldots, F^2 G, FG),$$

and can therefore be used to generate an optimal stopping time.

We assume that the basis functions are linearly independent and in $L_2(\mu)$. Let Π be the operator that projects in $L_2(\mu)$ onto the subspace spanned by the basis functions. Since Π is nonexpansive, the composition ΠH is a contraction on $L_2(\mu)$. Hence, it has a unique fixed point of the form Φr^* for some $r^* \in \Re^K$. Furthermore, it is possible to establish a bound of the form

$$\|\Phi r^* - (Q_0^*, \ldots, Q_{h-1}^*)\|_\mu \leq \frac{1}{\sqrt{1-\alpha^2}} \|\Pi(Q_0^*, \ldots, Q_{h-1}^*) - (Q_0^*, \ldots, Q_{h-1}^*)\|_\mu,$$

using the same arguments as in the proof of Lemma 5. A bound on the performance of a stopping time $\tilde{\tau} = \min\left\{t | G(x_t) \geq (\Phi r^*)(x_t, t)\right\}$ can also be established:

$$E[J_{0,h}^*(x_0)] - E[J_{0,h}^{\tilde{\tau}}(x_0)] \leq \frac{2}{(1-\alpha)\sqrt{1-\alpha^2}} \|\Pi(Q_0^*, \ldots, Q_{h-1}^*) - (Q_0^*, \ldots, Q_{h-1}^*)\|_\mu.$$

(Both bounds can be strengthened, if we allow coefficients on the right–hand–sides to depend on h, but we will not pursue this issue further here.)

Once again, Theorem 3 can be used to prove convergence. Following the approach used in Section 4, we rewrite the update equation in the form

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

and we define a function $\bar{s}(r) = E[s(z_0, r)]$. Some algebra gives us

$$(r - r^*)' \bar{s}(r) = \left\langle \Phi r - \Phi r^*, \Pi H \Phi r - \Phi r \right\rangle_\mu.$$

Since the composition ΠH is a contraction with fixed point Φr^*, it follows that $(r - r^*)' \bar{s}(r) < 0$. Together with technical assumptions (an analog of Assumption 5), this fact enables the application of Theorem 3, which establishes that r_t almost surely converges to r^*.

6.3 A Two–Player Zero–Sum Game

Many interesting phenomena arise when multiple participants make decisions within a single system. In this section, we consider a simple two–player zero–sum game in which a reward–maximizing player ("player 1") is allowed to stop a process at any even time step and a reward–minimizing player ("player 2") can opt to terminate during odd time steps.

6.3.1 Problem Formulation

We consider an ergodic Markov process with a steady–state distribution π together with reward functions $g, G_1, G_2 \in L_2(\pi)$ and a discount factor $\alpha \in (0,1)$. Prior to termination, a reward of $g(x_t)$ is obtained during each time step, and upon termination, a reward of either $G_1(x_t)$ or $G_2(x_t)$ is generated depending on which player opted to terminate. We define sets $U_1 = \{\tau \in U | \tau \text{ even}\}$ and $U_2 = \{\tau \in U | \tau \text{ odd}\}$ corresponding to admissible strategies for players 1 and 2, respectively. For each pair of stopping times $\tau_1 \in U_1$ and $\tau_2 \in U_2$, we define a value function

$$J^{\tau_1,\tau_2}(x) = E\left[\sum_{t=0}^{\tau_1 \wedge \tau_2 - 1} g(x_t) + \psi_1 G_1(x_{\tau_1}) + \psi_2 G_2(x_{\tau_2}) \Big| x_0 = x\right],$$

where ψ_1 and ψ_2 are indicators of the events $\{\tau_1 < \tau_2, \tau_1 < \infty\}$ and $\{\tau_2 < \tau_1, \tau_2 < \infty\}$, respectively. Hence, if players 1 and 2 take $\tau_1 \in U_1$ and $\tau_2 \in U_2$ as their strategies, the expected reward for the game is $E[J^{\tau_1,\tau_2}(x_0)]$. We consider sup-inf and inf-sup expected rewards

$$\sup_{\tau_1 \in U_1} \inf_{\tau_2 \in U_2} E\left[J^{\tau_1,\tau_2}(x_0)\right] \quad \text{and} \quad \inf_{\tau_2 \in U_2} \sup_{\tau_1 \in U_1} E\left[J^{\tau_1,\tau_2}(x_0)\right].$$

which correspond to different orders in which the players select their strategies. When both of these expressions take on the same value, this is considered to be the *equilibrium value* of the game. A pair of stopping times $\tau_1^* \in U_1$ and $\tau_2^* \in U_2$ are optimal if

$$E\left[J^{\tau_1^*,\tau_2^*}(x_0)\right] = \sup_{\tau_1 \in U_1} \inf_{\tau_2 \in U_2} E\left[J^{\tau_1,\tau_2}(x_0)\right] = \inf_{\tau_2 \in U_2} \sup_{\tau_1 \in U_1} E\left[J^{\tau_1,\tau_2}(x_0)\right].$$

The problem of interest is that of finding such stopping times.

6.3.2 Basic Theory

We define operators $T_1 J = \max\{G_1, g + \alpha PJ\}$ and $T_2 J = \min\{G_2, g + \alpha PJ\}$. By the same argument as that used to prove Lemma 2, both these operators are contractions on $L_2(\pi)$. It follows that the compositions $T_1 T_2$ and $T_2 T_1$ are also contractions on $L_2(\pi)$. We will denote the fixed points of $T_1 T_2$ and $T_2 T_1$ by J_1^* and J_2^*, respectively.

We define stopping times $\tau_1^* = \min\{\text{even } t | G(x_t) \geq J_1^*(x_t)\}$ and $\tau_2^* = \min\{\text{odd } t | G(x_t) \leq J_2^*(x_t)\}$. Using the fact that $T_1 T_2$ and $T_2 T_1$ are contractions, the arguments of Section 3 can be generalized to prove that

$$J_1^* \overset{ae(\pi)}{=} J^{\tau_1^*,\tau_2^*},$$

and

$$\sup_{\tau_1 \in U_1} \inf_{\tau_2 \in U_2} E[J^{\tau_1,\tau_2}(x_0)] = \inf_{\tau_2 \in U_2} \sup_{\tau_1 \in U_1} E[J^{\tau_1,\tau_2}(x_0)] = E[J^{\tau_1^*,\tau_2^*}(x_0)].$$

In other words, the pair of stopping times τ_1^* and τ_2^* is optimal.

6.3.3 An Approximation Algorithm

We now present an algorithm for approximating a pair of optimal stopping times and the equilibrium value of the game. Given a set of linearly independent basis functions

$\phi_1, \ldots, \phi_K \in L_2(\pi)$, we begin with initial weight vectors $r_{1,0}, r_{2,0} \in \Re^K$ and generate two sequences according to

$$r_{1,t+1} = r_{1,t} + \gamma_t \phi(x_t)\Big(g(x_t) + \alpha \min\{(\Phi r_{2,t})(x_{t+1}), G_2(x_{t+1})\} - (\Phi r_{1,t})(x_t)\Big),$$

and

$$r_{2,t+1} = r_{2,t} + \gamma_t \phi(x_t)\Big(g(x_t) + \alpha \max\{(\Phi r_{1,t})(x_{t+1}), G_1(x_{t+1})\} - (\Phi r_{2,t})(x_t)\Big),$$

where the step sizes satisfy Assumption 6.

To generalize the analysis of Section 4, we define operators $F_1 Q = g + \alpha P \min\{G_2, Q\}$ and $F_2 Q = g + \alpha P \max\{G_1, Q\}$. It is easy to show that these operators are contractions on $L_2(\pi)$, and so are their compositions $F_1 F_2$ and $F_2 F_1$. It is also easy to verify that the fixed points of $F_1 F_2$ and $F_2 F_1$ are given by $Q_1^* = g + \alpha P J_2^*$ and $Q_2^* = g + \alpha P J_1^*$, respectively. Furthermore, $Q_1^* \overset{ae(\pi)}{=} F_1 Q_2^*$, $Q_2^* \overset{ae(\pi)}{=} F_2 Q_1^*$, and the stopping times τ_1^* and τ_2^* can alternatively be generated according to $\tau_1^* = \min\{\text{even } t | G(x_t) \geq Q_1^*(x_t)\}$ and $\tau_2^* = \min\{\text{odd } t | G(x_t) \leq Q_2^*(x_t)\}$.

Let us define a measure μ over the product space $(\mathcal{B}(\Re^d))^2$ by $\mu(A_1, A_2) = \pi(A_1) + \pi(A_2)$ and an operator $H : L_2(\mu) \mapsto L_2(\mu)$, given by

$$H(Q_1, Q_2) = (F_1 Q_2, F_2 Q_1).$$

It is easy to show that H is a contraction on $L_2(\mu)$ with fixed point (Q_1^*, Q_2^*).

Let Π be the operator that projects in $L_2(\mu)$ onto the subspace $\{(\Phi r, \Phi \bar{r}) | r, \bar{r} \in \Re^K\}$. The composition ΠH is a contraction in $L_2(\mu)$ with a fixed point of the form $(\Phi r_1^*, \Phi r_2^*)$. Using arguments along the lines of Lemma 5, we can establish a bound of the form

$$\|(\Phi r_1^*, \Phi r_2^*) - (Q_1^*, Q_2^*)\|_\mu \leq \frac{1}{\sqrt{1-\alpha^2}} \|\Pi(Q_1^*, Q_2^*) - (Q_1^*, Q_2^*)\|_\mu.$$

Furthermore, the value of the game under stopping times $\tilde{\tau}_1 = \min\{\text{even } t | G(x_t) \geq (\Phi r_1^*)(x_t)\}$ and $\tilde{\tau}_2 = \min\{\text{odd } t | G(x_t) \leq (\Phi r_2^*)(x_t)\}$ deviates by a bounded amount from the equilibrium value. In particular,

$$|E[J_1^*(x_0)] - E[J^{\tilde{\tau}_1, \tilde{\tau}_2}(x_0)]| \leq \frac{2}{(1-\alpha)\sqrt{1-\alpha^2}} \|\Pi(Q_1^*, Q_2^*) - (Q_1^*, Q_2^*)\|_\mu.$$

To establish convergence, we rewrite the update equation in the form

$$(r_{1,t+1}, r_{2,t+1}) = (r_{1,t}, r_{2,t}) + \gamma_t s(z_t, (r_{1,t}, r_{2,t})),$$

where $z_t = (x_t, x_{t+1})$, and note that $\bar{s}(r_1, r_2) = E[s(z_0, (r_1, r_2))]$ is given by

$$\bar{s}_k(r_1, r_2) = \begin{cases} \Big\langle \phi_k, F_1 \Phi r_2 - \Phi r_1 \Big\rangle_\pi, & \text{if } k \leq K, \\ \Big\langle \phi_{k-K}, F_2 \Phi r_1 - \Phi r_2 \Big\rangle_\pi, & \text{otherwise,} \end{cases}$$

for any $r_1, r_2 \in \Re^K$. Letting $r^* = (r_1^*, r_2^*)$ and $r = (r_1, r_2)$, it follows that

$$(r - r^*)' \bar{s}(r) = \Big\langle (\Phi r_1, \Phi r_2) - (\Phi r_1^*, \Phi r_2^*), \Pi H(\Phi r_2, \Phi r_1) - (\Phi r_1, \Phi r_2) \Big\rangle_\mu.$$

Since ΠH is a contraction on $L_2(\mu)$ with fixed point $(\Phi r_1^*, \Phi r_2^*)$, we have

$$(r - r^*)'\overline{s}(r) < 0.$$

Combining this fact with an analog of Assumption 5, the technical requirements of Theorem 3 can be verified. It can then be deduced that $r_{1,t}$ and $r_{2,t}$ almost surely converge to r_1^* and r_2^*, respectively.

7 Conclusion

We have introduced a theory and algorithms pertaining to approximate solutions of optimal stopping problems. Though these developments are useful in their own right, as demonstrated by our application to pricing derivative securities, 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 (Bertsekas and Tsitsiklis, 1996). 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 (Watkins, 1989; Watkins and Dayan, 1992). 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 (1988), who also conjectured that the use of simulated trajectories in conjunction with such algorithms could be important for convergence (Sutton, 1995). This observation was later formalized by Tsitsiklis and Van Roy (1997), who analyzed temporal–difference methods and provided a counterexample in the same spirit as that of Section 4.4 (a related counter–example has also been proposed by Baird (1995)). Bertsekas and Tsitsiklis (1996) summarize much work directed at understanding such algorithms.

Acknowledgments

We thank Dimitri Bertsekas and Vivek Borkar for useful discussions.

References

Baird, L. C. (1995) "Residual Algorithms: Reinforcement Learning with Function Approximation," in Prieditis & Russell, eds. Machine Learning: Proceedings of the Twelfth International Conference, 9-12 July, Morgan Kaufman Publishers, San Francisco, CA.

Barraquand, J. & Martineau, D. (1995) "Numerical Valuation of High Dimensional Multivariate American Securities," unpublished manuscript.

Benveniste, A., Metivier, M., & Priouret, P. (1990) *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin.

Bertsekas, D. P. (1995) *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA.

Bertsekas, D. P. & Shreve, S. E. (1996) *Stochastic Optimal Control: The Discrete Time Case*, Athena Scientific, Belmont, MA.

Bertsekas, D. P. & Tsitsiklis, J. N. (1996) *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA.

Black, F. & Scholes, M. (1973) "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81:637–654.

Harrison, J. M. & Kreps, D. (1979) "Martingales and Arbitrage in Multiperiod Securities Markets," *Journal of Economic Theory*, 20:381–408.

Harrison, J. M. & Pliska, S. (1981) "Martingales and Stochastic Integrals in the Theory of Continuous Trading," *Stochastic Processes and Their Applications*, 11:215–260.

Karatzas, I. (1988) "On the Pricing of American Options," *Applied Mathematics and Operations Research*, pp. 37–60.

Merton, R. C. (1973) "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science*, 4: 141–183.

Shiryaev, A. N. (1978) *Optimal Stopping Rules*, Springer-Verlag, New York, NY.

Sutton, R. S. (1988) "Learning to Predict by the Methods of Temporal Differences," Machine Learning, vol. 3, pp. 9-44.

Sutton, R.S. (1995) "On the Virtues of Linear Learning and Trajectory Distributions," Proceedings of the Workshop on Value Function Approximation, Machine Learning Conference 1995, Boyan, Moore, and Sutton, Eds., p. 85. Technical Report CMU-CS-95-206, Carnegie Mellon University, Pittsburgh, PA 15213.

Tsitsiklis, J. N. & Van Roy, B., (1997) "An Analysis of Temporal-Difference Learning with Function Approximation," to appear in the *IEEE Transactions on Automatic Control*, May 1997.

Tsitsiklis, J. N. & Van Roy, B., (1997) "Approximate Solutions to Optimal Stopping Problems," in Advances in Neural Information Processing Systems 9, M.C. Mozer, M.I. Jordan, T. Petsche, eds., MIT Press.

Watkins, C. J. C. H. (1989) Learning from Delayed Rewards. Doctoral dissertation, University of Cambridge, Cambridge, United Kingdom.

Watkins, C. J. C. H. & Dayan, P. (1992) "Q–learning," Machine Learning, vol. 8, pp. 279-292.