Bounds for Markov Decision Processes

Vijay V. Desai
Industrial Engineering and Operations Research
Columbia University
email: vvd2101@columbia.edu

Vivek F. Farias
Sloan School of Management
Massachusetts Institute of Technology
email: vivekf@mit.edu

Ciamac C. Moallemi
Graduate School of Business
Columbia University
email: ciamac@gsb.columbia.edu

November 5, 2011

Abstract

We consider the problem of producing lower bounds on the optimal cost-to-go function of a Markov decision problem. We present two approaches to this problem: one based on the methodology of approximate linear programming (ALP) and another based on the so-called martingale duality approach. We show that these two approaches are intimately connected. Exploring this connection leads us to the problem of finding 'optimal' martingale penalties within the martingale duality approach which we dub the pathwise optimization (PO) problem. We show interesting cases where the PO problem admits a tractable solution and establish that these solutions produce tighter approximations than the ALP approach.

1. Introduction

Markov decision processes (MDPs) provide a general framework for modeling sequential decision-making under uncertainty. A large number of practical problems from diverse areas can be viewed as MDPs and can, in principle, be solved via dynamic programming. However, for many problems of interest, the state space of the corresponding dynamic program is intractably large. This phenomenon, referred to as the *curse of dimensionality*, renders exact approaches to solving Markov decision problems impractical.

Solving an MDP may be viewed as equivalent to the problem of computing an *optimal cost-to-go function*. As such, approximation algorithms for solving MDPs whose state space is intractably large frequently treat the task of computing an approximation to this optimal cost-to-go function as the key algorithmic task; given such an approximation, the greedy policy with respect to the approximation is a canonical candidate for an approximate policy. The collective research area devoted to the development of such algorithms is frequently referred to as *approximate dynamic programming*; see, Van Roy (2002) or Bertsekas (2007, Chap. 6) for brief surveys of this area of research. Now consider a Markov decision problem wherein we wish to minimize expected costs,

Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, First Edition. Edited by Frank L. Lewis and Derong Liu.

discounted over an infinite time horizon and consider the problem of producing upper and lower bounds on the costs incurred under an optimal policy starting at a specific state (the 'cost-to-go' of that state). By simulating an arbitrary feasible policy starting at that state, we obtain an upper bound on the cost-to-go of the state. Given a complementary *lower* bound on the cost-to-go of this state, one may hope to construct a 'confidence interval' of sorts on the cost-to-go of the state in question.[1] The task of finding a lower bound on the cost-to-go of a state is not quite as straightforward. Moreover, we are interested in *good* bounds. The literature offers us two seemingly disparate alternatives to serve this end:

- **Lower bounds via approximate linear programming (ALP).** This approach was introduced by Schweitzer and Seidmann (1985) and later developed and analyzed by de Farias and Van Roy (2003, 2004). Given a set of 'basis functions', the ALP produces an approximation to the optimal cost-to-go function spanned by these basis functions that is provably a pointwise lower bound to the optimal cost-to-go function. The quality of the cost-to-go function approximation produced by the ALP can be shown to compete, in an appropriate sense, with the best possible approximation afforded by the basis function architecture. The ALP approach is attractive for two reasons: First, from a practical standpoint, the availability of reliable linear programming solvers allows the solution of large ADP problems. Second, the structure of the linear program allows strong theoretical guarantees to be established.

- **Lower bounds via martingale duality.** A second approach to computing lower bounds, which constitutes an active area of research, relies on 'information relaxations'. As a trivial example, consider giving the optimizer *a priori* knowledge of all randomness that will be realized over time; clearly this might be used to compute a 'clairvoyant' lower bound on the optimal cost-to-go. These approaches introduce, in the spirit of Lagrangian duality, a penalty for relaxing the restrictions on information available to the controller. The penalty function is itself a stochastic process and, frequently, is a martingale adapted to the natural filtration of the MDP; hence the nomenclature martingale duality. An important application of these approaches can be found in the context of pricing high dimensional American options following the work of Rogers (2002) and Haugh and Kogan (2004). Generalizations of this approach to control problems other than optimal stopping, have also been studied (see, e.g., Brown et al., 2010; Rogers, 2008).

The two approaches above are, at least superficially, fairly distinct from each other. Computing a good cost-to-go function approximation via the ALP relies on finding a good set of basis functions. The martingale duality approach on the other hand requires that we identify a suitable martingale

[1] Equivalently, in problems where reward is maximized, the quantity of interest is the value of rewards achieved under an optimal policy, starting from a specific state. Lower bounds are available from the simulation of suboptimal policies, and one might seek complimentary upper bounds. We will choose between the objectives of cost minimization and reward maximization in this chapter, according to what is most natural to the immediate setting.

to serve as the penalty process. The purpose of this chapter is to present a simple unified view of the two approaches through the lens of, what we call, the *pathwise optimization* (PO) method. This method was introduced in the context of high-dimensional optimal stopping problems by Desai et al. (2010) and later extended to a larger class of problems (optimizing convex cost functionals subject to linear system dynamics) in Desai et al. (2011).

We will shortly present a brief literature review. Following that, the remainder of the chapter is organized as follows: In Section 2, we formulate our problem and state the Bellman equation. Sections 3 and 4 introduce the ALP and martingale duality approaches, respectively, for the problem. The PO approach is described in Section 5 and its applications to optimal stopping and linear convex systems are described in Section 6.

1.1. Related Literature

The landscape of ADP algorithms is rich and varied; we only highlight some of the literature related to ALP. Bertsekas and Tsitsiklis (1996) and Powell (2007) are more detailed references on the topic. The ALP approach was introduced by Schweitzer and Seidmann (1985) and further developed by de Farias and Van Roy (2003, 2004) who established approximation guarantees for this approach. This method has seen a number of applications, which includes scheduling in queueing networks (Moallemi et al., 2008; Morrison and Kumar, 1999; Veatch, 2005), revenue management (Adelman, 2007; Farias and Van Roy, 2007; Zhang and Adelman, 2008), portfolio management (Han, 2005), inventory problems (Adelman, 2004; Adelman and Klabjan, 2009), and algorithms for solving stochastic games (Farias et al., 2011), among others.

Martingale duality methods for the pricing of American and Bermudan options, which rely on Doob's decomposition to generate the penalty process, were introduced by Rogers (2002) and Haugh and Kogan (2004). Andersen and Broadie (2004) show how to compute martingale penalties using stopping rules and are able to obtain tight bounds. An alternative 'multiplicative' approach to duality was introduced by Jamshidian (2003) and its connections with the above 'additive' duality approaches was explored in Chen and Glasserman (2007). Beyond stopping problems, these methods are applicable for general control problems as discussed in Rogers (2008) and Brown et al. (2010). Further, Brown et al. (2010) consider a broader class of information relaxations than the typical case of a perfect information relaxation. Applications of these methods were considered in portfolio optimization (Brown and Smith, 2010) and valuation of natural gas storage (Lai et al., 2010a,b), among others.

2. Problem Formulation

Consider a discounted, infinite horizon problem with state space $\mathcal{X}$ and action set $\mathcal{A}$. At time t, given state x_t and action a_t, the per stage cost is given by $g(x_t, a_t)$. The state evolves according to

$$x_{t+1} = h(x_t, a_t, w_t),$$

where $\{w_t\}$ are independent and identically distributed random variables taking values in the set $\mathcal{W}$. Let $\mathcal{F} \triangleq \{\mathcal{F}_t\}$ be the natural filtration generated by the process $\{w_t\}$, i.e., for each time t, $\mathcal{F}_t \triangleq \sigma(w_0, w_1, \ldots, w_t)$. So as to avoid discussion of technicalities which are not central to our main ideas, for ease of exposition, we assume finite state and control spaces.

A stationary policy $\mu \colon \mathcal{X} \to \mathcal{A}$ maps the state space $\mathcal{X}$ to the set of actions $\mathcal{A}$. In other words, given a state x_t, the action taken at that state under policy μ is $a_t = \mu(x_t)$. The cost-to-go function J_μ associated with a stationary policy μ is given by

$$J_\mu(x) = \mathsf{E}\left[\sum_{t=0}^{\infty} \alpha^t g(x_t, \mu(x_t)) \middle| x_0 = x\right],$$

where α is the discount factor.

We define the Bellman operator associated with policy μ according to

$$(T_\mu J)(x) \triangleq g(x, \mu(x)) + \alpha \mathsf{E}[J(h(x, \mu(x), w))].$$

Given this defintion, J_μ is given as the unique solution to the Bellman's equation $T_\mu J = J$. We further define the optimal cost-to-go function J^* according to $J^*(x) = \min_\mu J_\mu(x),\ \forall\ x \in \mathcal{X}$. J^* may be computed as the unique solution to *Bellman's equation*. In particular, define the Bellman operator $T : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{X}|}$ according to $TJ = \min_\mu T_\mu J$. Bellman's equation is simply the fixed point equation $TJ = J$.

Given the optimal cost-to-go function, the optimal policy is obtained by acting greedily with respect to the optimal cost-to-go function, i.e.,

$$\mu^*(x) \in \operatorname*{argmin}_a\ g(x, a) + \alpha \mathsf{E}[J^*(h(x, a, w))]. \tag{1}$$

The Problem: Computing J^* is in general intractable for state spaces $\mathcal{X}$ that are intractably large. As such our goal in this paper will be to compute *lower bounds* to the optimal cost-to-go function of a specific state x, $J^*(x)$. We will particularly be interested in issues of tractability and the tightness of the resulting bounds.

3. The Linear Programming Approach

This section describes an approximate dynamic programming approach (dubbed approximate linear programming) to solving the above problem. The approach relies on solving a linear program motivated largely by a certain 'exact' linear program for the exact solution of Bellman's equation. We begin by describing the exact linear program.

3.1. The Exact Linear Program

Given any vector $\nu \in \mathbb{R}^{|\mathcal{X}|}$ with positive components, the exact linear program, credited to Manne (1960), is given by:

$$\begin{aligned} \underset{J}{\text{maximize}} \quad & \nu^\top J \\ \text{subject to} \quad & J \leq TJ. \end{aligned} \tag{2}$$

Although the Bellman operator T is nonlinear, this program can be easily transformed into a linear program. Consider a state $x \in \mathcal{X}$, the constraint $J(x) \leq (TJ)(x)$ is equivalent to $|\mathcal{A}|$ linear constraints given by

$$J(x) \leq g(x,a) + \alpha \mathsf{E}[J(h(x,a,w))], \quad \forall\ a \in \mathcal{A}.$$

Using this transformation, the exact linear program has as many variables as the state space size $|\mathcal{X}|$ and as many constraints as $|\mathcal{X} \times \mathcal{A}|$.

We recall the following basic properties of the Bellman operator T. The interested reader is referred to Bertsekas (2006) for details of the proof.

Proposition 1. *Let $J, J' \in \mathbb{R}^{|\mathcal{X}|}$.*

1. *(Monotonicity) If $J \geq J'$, then $TJ \geq TJ'$.*
2. *(Max-norm contraction) $\|TJ - TJ'\|_\infty \leq \alpha \|J - J'\|_\infty$.*

The following theorem establishes that the program (2) yields, as its unique optimal solution, the optimal cost to go J^*. We provide a proof of this fact for completeness.

Theorem 1.

1. *For all $J \in \mathbb{R}^{|\mathcal{X}|}$ such that $J \leq TJ$, we have $J \leq J^*$.*
2. *J^* is the unique optimal solution to the exact linear program (2).*

Proof. Now by the monotonicity of T, for any J satisfying $J \leq TJ$, we must also have $J \leq TJ \leq \ldots \leq T^k J$, for any integer $k \geq 1$. Since T is a contraction mapping, however, we have that, as $k \to \infty$, $T^k J \to J^*$, the unique fixed point of the operator T. It follows that any feasible solution to (2), J, satisfies $J \leq J^*$. This is the first part of the theorem. Further, since J^* is itself a

feasible solution, and since the components of ν are strictly positive, we have the second part of the theorem. ■

Of course, the exact linear program has $|\mathcal{X}|$ variables and $|\mathcal{X} \times \mathcal{A}|$ constraints and, as such, we must still contend with the curse of dimensionality. This motivates an effort to reduce the dimensionality of the problem by permitting approximations to the cost-to-go function.

3.2. Cost-To-Go Function Approximation

Cost-to-go function approximations address the curse of dimensionality through the use of parameterized function approximations. In particular, it is common to focus on linear parameterizations. Consider a collection of *basis functions* $\{\phi_1, \ldots, \phi_K\}$ where each $\phi_i \colon \mathcal{X} \to \mathbb{R}$ is a real-valued function on the state space. ADP algorithms seek to find linear combinations of the basis functions that provide good approximations to the optimal cost-to-go function. In particular, we seek a vector of weights $r \in \mathbb{R}^K$ so that

$$\Phi r(x) \triangleq \sum_{\ell=1}^{K} \phi_\ell(x) r_\ell \approx J^*(x).$$

Here, we define $\Phi \triangleq [\phi_1 \ \phi_2 \ \ldots \ \phi_K]$ to be a matrix with columns consisting of the basis functions. Given such an approximation to the cost-to-go function, a natural policy to consider is simply the policy that acts greedily with respect to the cost-to-go function approximation. Such a policy is given by:

$$\mu^r(x) \in \operatorname*{argmin}_{a \in \mathcal{A}} \ g(x,a) + \mathsf{E}[\Phi r(h(x,a,w)]. \tag{3}$$

Notice that such a policy is eminently implementable. In contrast with the optimal policy which would generally require a lookup table for the optimal cost-to-go function (and consequently, storage space on the order of the size of the state space), the policy μ^r simply requires that we store K numbers corresponding to the weights r and have access to an oracle that for a given state x computes the basis functions at that state. The approximations to the cost-to-go function can then be computed online, as and when needed.

3.3. The Approximate Linear Program

In light of the approximation described above, a natural idea would be to restrict attention to solutions of the exact linear program that lie in the lower dimensional space spanned by the basis functions (i.e, span(Φ)). The Approximate Linear Program (ALP) does exactly this:

$$\begin{array}{ll} \underset{r}{\text{maximize}} & \nu^\top \Phi r \\ \text{subject to} & \Phi r \leq T \Phi r. \end{array} \tag{4}$$

Notice that the above program continues to have a large number of constraints but a substantially smaller number of variables, K.

For any feasible solution r to this program, we must have, by Theorem 1, that the approximation implied by r provides a lower bound to the optimal cost-to-go. That is, $\Phi r \leq J^*$. This observation also allows us to rewrite ALP as

$$\begin{array}{ll} \underset{r}{\text{minimize}} & \|J^* - \Phi r\|_{1,\nu} \\ \text{subject to} & \Phi r \leq T\Phi r, \end{array} \tag{5}$$

where the weighted 1-norm in the objective is defined by

$$\|J^* - \Phi r\|_{1,\nu} \triangleq \sum_{x \in \mathcal{X}} \nu(x)|J^*(x) - \Phi r(x)|.$$

This representation of the ALP makes it clear that ν can be used to emphasize regions of the state space where we would like a good approximation and consequently, the components of ν are referred to as the state-relevance weights.

Now, for a fixed state $x \in \mathcal{X}$, the best lower bound to $J^*(x)$ we might compute using this approach simply calls for us to choose the state-relevance weights such that $\nu(x)$ is large. Moreover, if J^* is in the linear span of Φ, then it is clear from (5) that the approximation error would be zero. Apart from obtaining lower bounds, the cost-to-go function approximation obtained by solving the ALP can be used to generate policies, simply by acting greedily with respect to the approximation as shown in (3).

4. The Martingale Duality Approach

Every feasible solution to the ALP constitues a lower bound to the optimal cost-to-go function; the quality of this bound is determined largely by our choice of basis functions. A different approach to obtaining lower bounds is via an information relaxation. The idea is to allow policies to have knowledge of all future randomness and 'penalize' this relaxation in the spirit of Lagrangian duality. The penalties are themselves stochastic processes, and typically martingales. We describe this approach next.

Let $\mathcal{P}$ be the space of real-valued functions defined on $\mathcal{X}$. Intuitively, one can think of this as the space of cost-to-go functions. Let us begin with defining the martingale difference operator Δ that maps a function $J \in \mathcal{P}$ to a real-valued function ΔJ on $\mathcal{X} \times \mathcal{X} \times \mathcal{A}$ according to

$$(\Delta J)(x_{t+1}, x_t, a_t) \triangleq J(x_{t+1}) - \mathsf{E}[J(x_{t+1})|x_t, a_t].$$

We are interested in computing lower bounds by considering a perfect information relaxation. Let $\mathcal{A}^\infty$ be the set of infinite sequences of elements of $\mathcal{A}$. For an arbitrary sequence of actions

$\mathbf{a} \in \mathcal{A}^\infty$, define the process $M_t^{\mathbf{a}}(J)$ by

$$M_0^{\mathbf{a}}(J) \triangleq 0, \qquad M_t^{\mathbf{a}}(J) \triangleq \sum_{s=1}^{t} \alpha^s \Delta J(x_s, x_{s-1}, a_{s-1}), \quad \forall\, t \geq 1.$$

Clearly $M_t^{\mathbf{a}}(J)$ is adapted to the filtration $\mathcal{F}$. Further, if actions are chosen according to $\mathbf{a}$, then $M_t^{\mathbf{a}}(J)$ is a martingale. Using the fact that the state space $\mathcal{X}$ and action space $\mathcal{A}$ are finite, there exists a constant C_J such that

$$|\Delta J(x_s, x_{s-1}, a_{s-1})| < C_J, \quad \forall\, (x_s, x_{s-1}, a_{s-1}) \in \mathcal{X} \times \mathcal{X} \times \mathcal{A}.$$

It then follows from the orthogonality of martingale increments that

$$\mathsf{E}\left[M_t^{\mathbf{a}}(J)^2\right] = \sum_{s=1}^{t} \alpha^{2s} \mathsf{E}\left[|\Delta J(x_s, x_{s-1}, a_{s-1})|^2\right] < \frac{C_J^2 \alpha^2}{1-\alpha^2}.$$

Thus, $M_t^{\mathbf{a}}(J)$ is a $\mathcal{L}^2$-martingale. By the martingale convergence theorem, the limit

$$M_\infty^{\mathbf{a}}(J) \triangleq \sum_{s=1}^{\infty} \alpha^s \Delta J(x_s, x_{s-1}, a_{s-1}) \tag{6}$$

is well-defined.

We now define the *martingale duality* operator $F\colon \mathcal{P} \to \mathcal{P}$ according to:

$$(FJ)(x) \triangleq \mathsf{E}\left[\inf_{\mathbf{a} \in \mathcal{A}^\infty} \sum_{t=0}^{\infty} \alpha^t g(x_t, a_t) - M_\infty^{\mathbf{a}}(J) \,\middle|\, x_0 = x\right], \tag{7}$$

where the expectation is with respect the infinite sequence of disturbances $(w_0, w_1, \ldots)$. The deterministic minimization problem embedded inside the expectation will be referred to as the *inner problem*.

Given any $J \in \mathcal{P}$, $FJ(x)$ can be used to obtain lower bounds on the optimal cost-to-go function $J^*(x)$. Moreover, there exists $J \in \mathcal{P}$ for which the lower bounds are tight, and one such choice of J is the optimal cost-to-go function J^*. The following theorem justifies these claims.

Theorem 2.

(i) (Weak duality) For any $J \in \mathcal{P}$ and all $x \in \mathcal{X}$, $FJ(x) \leq J^*(x)$.

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

Proof. (i) For each state $x \in \mathcal{X}$,

$$\begin{aligned} J^*(x) &= \min_{\mu} \ \mathsf{E}\left[\sum_{t=0}^{\infty} \alpha^t g(x_t, \mu(x_t)) \ \middle| \ x_0 = x\right] \\ &\overset{(a)}{=} \min_{\mu} \ \mathsf{E}\left[\sum_{t=0}^{\infty} \alpha^t g(x_t, \mu(x_t)) - M_{\infty}^{\mu}(J) \ \middle| \ x_0 = x\right] \\ &\overset{(b)}{\geq} \mathsf{E}\left[\inf_{\mathbf{a} \in \mathcal{A}^{\infty}} \ \sum_{t=0}^{\infty} \alpha^t g(x_t, a_t) - M_{\infty}^{\mathbf{a}}(J) \ \middle| \ x_0 = x\right] \\ &= FJ(x). \end{aligned}$$

Here, (a) follows from the fact that $M_{\infty}^{\mu}(J)$ is zero mean, and (b) follows from that fact that the objective value can only be decreased given knowledge of the entire sample path of disturbances.

(ii) From (i), we have that $FJ^*(x) \leq J^*(x)$. We will establish the result by showing $FJ^*(x) \geq J^*(x)$. Using the definition of $FJ^*(x)$, we have

$$\begin{aligned} FJ^*(x) &= \mathsf{E}\left[\inf_{\mathbf{a} \in \mathcal{A}^{\infty}} \ \sum_{t=0}^{\infty} \alpha^t \big(g(x_t, a_t) - \alpha \Delta J^*(x_{t+1}, x_t, a_t)\big) \ \middle| \ x_0 = x\right] \\ &= \mathsf{E}\left[\inf_{\mathbf{a} \in \mathcal{A}^{\infty}} \ \sum_{t=0}^{\infty} \alpha^t \left(g(x_t, a_t) + \alpha \mathsf{E}\big[J^*(x_{t+1}) | x_t, a_t\big] - \alpha J^*(x_{t+1})\right) \ \middle| \ x_0 = x\right] \\ &= \mathsf{E}\left[\inf_{\mathbf{a} \in \mathcal{A}^{\infty}} \ \sum_{t=0}^{\infty} \alpha^t \left(g(x_t, a_t) + \alpha \mathsf{E}\big[J^*(x_{t+1}) | x_t, a_t\big] - J^*(x_t)\right) + J^*(x_0) \ \middle| \ x_0 = x\right] \\ &\geq J^*(x). \end{aligned}$$

The last inequality follows from the fact that J^* satisfies the Bellman equation, thus $J^*(x) \leq g(x, a) + \alpha \mathsf{E}[J^*(x_{t+1}) | x_t = x, a]$ for all $a \in \mathcal{A}$ and $x \in \mathcal{X}$. ■

We can succinctly state the above result as:

$$J^*(x) = \sup_{J \in \mathcal{P}} \ FJ(x), \tag{8}$$

which we refer to as the *dual problem*. Although this dual problem is typically thought of as an optimization over an appropriate space of martingales, our exposition suggests that as an alternative, we may think of the dual problem as optimizing over the space of cost-to-go functions. This view will be crucial in unifying the ALP and martingale duality approaches. The optimization over the space $\mathcal{P}$ will be referred to as the *outer problem* to distinguish it from the inner problem, which is a deterministic minimization problem embedded inside the F operator.

The dual problem is challenging for various reasons. In particular, optimizing over $\mathcal{P}$ is non-trivial when the state space is high-dimensional. It has nevertheless inspired heuristic methods for computing lower bounds. Given a cost-to-go function approximation J, one can use Monte Carlo

simulation to estimate $FJ(x)$ and this serves as a lower bound on $J^*(x)$. The approximation J itself could be the product of an ADP method. Alternatively, it could be obtained by simplifying the original problem with the goal of being able to compute a surrogate to the cost-to-go function. These approaches have been successfully applied in a wide variety of settings. In the context of American option pricing, for example, Andersen and Broadie (2004) use regression based approaches to obtain a cost-to-go function approximation, which can then be used to construct martingale penalties which yield remarkably tight bounds. Beyond American option pricing problem, such approaches have been used in portfolio optimization (Brown and Smith, 2010) and the valuation of natural gas storage (Lai et al., 2010a), among other applications.

5. The Pathwise Optimization Method

Observe that the dual problem entails optimization over a very high-dimensional space (namely, $\mathcal{P} \triangleq \mathbb{R}^{|\mathcal{X}|}$). This is reminiscent of the challenge with the exact linear program. Analogous to our derivation of the ALP then, we are led to restrict the optimization problem to a lower dimensional subspace. In particular, given a set of basis functions, Φ, define $\hat{\mathcal{P}} \triangleq \{\Phi r \ : \ r \in \mathbb{R}^K\} \subset \mathcal{P}$. We consider finding a good approximation to the cost-to-go function of the form FJ, with $J \in \hat{\mathcal{P}}$ restricted to the subspace spanned by the basis. To accomplish this, given a state-relevance vector $\nu \in \mathbb{R}^{|\mathcal{X}|}$ with positive components, we define the *pathwise optimization* (PO) problem by

$$\sup_r \ \nu^\top F\Phi r \triangleq \sup_r \ \sum_{x\in\mathcal{X}} \nu(x) F\Phi r(x). \tag{9}$$

Several remarks are in order. Observe that from Theorem 2, for any r, $F\Phi r(x) \leq J^*(x)$ for all states x. Therefore, the PO program (9) is equivalent to

$$\inf_r \ \|J^* - F\Phi r\|_{1,\nu}.$$

Thus, the PO program will seek to find $\Phi r \in \hat{\mathcal{P}}$, so that the resulting lower bound $F\Phi r(x)$ will be close to the true optimal cost-to-go $J^*(x)$, measured on average across states x according to the state-relevance weight ν.

Similar to the ALP, if J^* is in the span of Φ, it is clear that the optimal solution to the above problem will yield the optimal cost-to-go function J^*. In addition, as the following theorem establishes, the PO problem is a *convex* optimization problem[2] over a low-dimensional space:

Theorem 3. *The function $r \mapsto \nu^\top F\Phi r$ is concave in $r \in \mathbb{R}^K$.*

Proof. Observe that, as a function of r, $\nu^\top F\Phi r$ is a nonnegative linear combination of a set of pointwise infima of affine functions of r, and hence must be concave in r as each of these operations

[2]Here, we refer to an optimization problem as convex if it involves the minimization of a convex function over a convex feasible set, or, equivalently, the maximization of a concave function over a convex feasible set.

preserves concavity. ■

The PO problem puts the martingale duality and ALP approaches on a common footing: both approaches can now be seen to require a set of basis function Φ whose span ideally contains a good approximation to the optimal cost-to-go function. Given such a set of basis functions, both approaches require the solution of a convex optimization problem over a low-dimensional space of weight vectors $r \in \mathbb{R}^K$: (4) for the ALP, and (9) for the pathwise approach. Given an optimal solution r, both methods can produce a lower bound on the optimal cost-to-go $J^*(x)$ at an arbitrary state x: $\Phi r(x)$ for the ALP, and $F\Phi r(x)$ for the pathwise approach.The natural question one might then ask is: how do these approaches relate to each other in terms of the lower bounds they produce? We answer this question next:

Theorem 4. *Let r be any feasible solution to the ALP, i.e., r satisfies $\Phi r \leq T\Phi r$. Then, for all $x \in \mathcal{X}$,*

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

Proof. Using the weak duality result in Theorem 2,

$$\begin{aligned} J^*(x) \geq F\Phi r(x) &= \mathsf{E}\left[\inf_{\mathbf{a}\in\mathcal{A}^\infty} \sum_{t=0}^{\infty} \alpha^t (g(x_t,a_t) - \alpha\Delta\Phi r(x_{t+1},x_t,a_t)) \;\middle|\; x_0 = x\right] \\ &= \mathsf{E}\left[\inf_{\mathbf{a}\in\mathcal{A}^\infty} \sum_{t=0}^{\infty} \alpha^t \left(g(x_t,a_t) + \alpha\mathsf{E}[\Phi r(x_{t+1})|x_t,a_t] - \Phi r(x_t)\right) + \Phi r(x_0) \;\middle|\; x_0 = x\right] \\ &\geq \Phi r(x), \end{aligned}$$

where the final inequality follows since r is feasible for the ALP. ■

Theorem 5 establishes a strong relationship between the lower bounds arising from the ALP and PO methods. For *any* feasible candidate weight vector r, the corresponding ALP lower bound $\Phi r(x)$ is dominated by the PO lower bound $F\Phi r(x)$, at *every* state x. Since the PO program (9) further considers a large set of feasible r, it immediately follows that the optimal solution of the PO program will provide an lower bound that is, in an appropriately weighted sense, tighter than that of the ALP method. The fact that the PO method provably dominates the ALP method is the content of the following theorem:

Theorem 5. *Suppose that r_{PO} is an optimal solution to the PO program* (9)*, while r_{ALP} is an optimal solution to the ALP* (4)*. Then,*

$$\|J^* - F\Phi r_{PO}\|_{1,\nu} \leq \|J^* - \Phi r_{ALP}\|_{1,\nu}.$$

Proof. Note that, using Theorems 2 and 5,

$$\begin{aligned}\|J^* - F\Phi r_{\mathsf{PO}}\|_{1,\nu} &= \nu^\top J^* - \nu^\top F\Phi r_{\mathsf{PO}} \leq \nu^\top J^* - \nu^\top F\Phi r_{\mathsf{ALP}} \\ &\leq \nu^\top J^* - \nu^\top \Phi r_{\mathsf{ALP}} = \|J^* - \Phi r_{\mathsf{ALP}}\|_{1,\nu}.\end{aligned}$$

■

6. Applications

The results of the prior section establish that the PO method is a convex optimization problem over a low-dimensional space that delivers provably stronger bounds than the ALP approach. However, challenges remain in implementing the PO method. The PO objective in (9) is the expectation of a complicated random variable, namely, the objective value of the inner optimization problem. We can use a sample average approximation to estimate the outer expectation. However, for each sample path, the inner optimization problem will correspond to a potentially high dimensional *deterministic* dynamic program. This program may be no easier to solve than the original *stochastic* dynamic program. In particular, for example, solution of the deterministic problem via exact dynamic programming would be subject to the same curse-of-dimensionality as the stochastic problem. Hence, we expect that solving the PO problem in a tractable fashion is likely to call for additional problem structure. In this section, we present two broad classes of problems whose structure admit a tractable PO problem.

Our discussion thus far has focused on the infinite horizon, discounted case. We chose to do so for two reasons: simplicity on the one hand, and the fact that in such a setting, results such as Theorem 5 demonstrate that the approximations produced by the PO method inherit approximation guarantees established for the ALP in the discounted, infinite horizon setting. In what follows we will consider two concrete classes of problems that are more naturally studied in a *finite* horizon setting. As it turns out, the PO problem has a natural analog in such a setting and the following examples will serve to illustrate this analog in addition to specifying broad classes of problems where the PO approach is tractable.

6.1. Optimal Stopping

Optimal stopping problems are a fundamental class of stochastic control problems. The problem of valuing American options is among the more significant examples of such a control problem. It is most natural to consider dealing with the finite horizon case here. As such, time becomes a relevant state variable and the PO method as stated earlier needs to be adapted. Further, our problem formulation and development of the ALP and PO method, so far, has been couched in a discounted infinite horizon setting where one seeks to minimize cost. However, these methods are equally applicable to the finite horizon case where one seeks to maximize reward. Motivated by the

application of option pricing, we will consider this latter setting in the context of optimal stopping.

In particular, consider a discounted problem over the finite horizon $\mathcal{T} \triangleq \{0, 1, \ldots, T\}$. The state evolves as a Markov process, so that

$$x_{t+1} = h(x_t, w_t),$$

where w_t is an i.i.d. disturbance. The action at each time step is either to stop or to continue and thus $\mathcal{A} \triangleq \{\mathsf{STOP},\ \mathsf{CONTINUE}\}$. On choosing to stop at time t in state x_t, the discounted reward is $\alpha^t g(x_t)$, where α is the discount factor. An exercise policy $\mu \triangleq \{\mu_t,\ t \in \mathcal{T}\}$, is a sequence of functions where each $\mu_t \colon \mathcal{X} \to \{\mathsf{STOP},\ \mathsf{CONTINUE}\}$ specifies the stopping decision at time t, as a function of state x_t. We require that stopping occur at some time in $\mathcal{T}$, and our goal is to obtain an exercise policy that maximizes the expected discounted reward.

In principle, J^* may be computed via the following dynamic programming backward recursion

$$J_t^*(x) \triangleq \begin{cases} \max\ \{g(x),\ \alpha \mathsf{E}\left[J_{t+1}^*(x_{t+1}) \mid x_t = x\right]\} & \text{if } t < T. \\ g(x) & \text{if } t = T, \end{cases} \tag{10}$$

for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$, The corresponding optimal stopping policy μ^* that acts 'greedily' with respect to J^* is given by

$$\mu_t^*(x) \triangleq \begin{cases} \mathsf{CONTINUE} & \text{if } t < T \text{ and } g(x) < \alpha \mathsf{E}[J_{t+1}^*(x_{t+1}) \mid x_t = x], \\ \mathsf{STOP} & \text{otherwise.} \end{cases} \tag{11}$$

6.1.1. The Martingale Duality Approach

Let $\mathcal{S}$ be the space of real-valued functions defined on the state space $\mathcal{X}$, i.e., functions of the form $V \colon \mathcal{X} \to \mathbb{R}$. Define $\mathcal{P}$ to be the set of functions $J \colon \mathcal{X} \times \mathcal{T} \to \mathbb{R}$ of state and time, and, for notational convenience, denote $J_t \triangleq J(\cdot, t)$. One can think of $\mathcal{P}$ as the space of value functions. We begin by defining the *martingale difference operator* Δ. The operator Δ maps a function $V \in \mathcal{S}$ to the function $\Delta V \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ according to

$$(\Delta V)(x_{t+1}, x_t) \triangleq V(x_{t+1}) - \mathsf{E}[V(x_{t+1}) | x_t].$$

Given an arbitrary function $J \in \mathcal{P}$, and a time $\tau \in \mathcal{T}$, define the process

$$M_t^{(\tau)}(J) \triangleq \sum_{s=1}^{t \wedge \tau} \alpha^s (\Delta J_s)(x_s, x_{s-1}), \quad \forall\ t \in \mathcal{T}. \tag{12}$$

Then, $M^{(\tau)}$ is a martingale adapted to the filtration $\mathcal{F}$. Next, we define the *martingale duality operator* $F\colon \mathcal{P} \to \mathcal{S}$ according to:

$$(FJ)(x) \triangleq \mathsf{E}\left[\max_{t\in\mathcal{T}} \ \alpha^t g(x_t) - M_T^{(t)}(J) \ \middle|\ x_0 = x\right]. \tag{13}$$

Observe that the martingale penalty (12) is a natural analog of the penalty (6) introduced earlier. In the stopping problem, the sequence of actions simply corresponds to a choice of time $t \in \mathcal{T}$ at which to stop. Beyond that time, the optimal value function will take the value zero. Hence, when constructing a martingale penalty according to an optimal value function surrogate, it is not necessary to consider times after the stopping time. With these observations, it is clear that the penalty (6) simplifies to the penalty (12) for a stopping problem, and hence the operator (13) is a natural generalization of the operator (7).

For any given $J \in \mathcal{P}$ and a state $x_0 \in \mathcal{X}$, an analog to Theorem 2 establishes that $FJ(x_0)$ provides an upper bound on the optimal value $J_0^*(x_0)$. With the intention of optimizing the bound $FJ(x_0)$ over a parameterized subspace $\hat{\mathcal{P}} \subset \mathcal{P}$, we introduce the collection of K basis functions

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

Each vector $r \in \mathbb{R}^K$ determines a value function approximation of the form

$$(\Phi r)_t(x) \triangleq \sum_{\ell=1}^{K} \phi_\ell(x,t) r_\ell, \quad \forall\ x \in \mathcal{X},\ t \in \mathcal{T}.$$

Thus, the PO problem of finding the tightest upper bound of the form $F\Phi r(x_0)$ can be defined as

$$\inf_r \ F\Phi r(x_0). \tag{14}$$

The problem (14) is an unconstrained convex optimization problem over a low-dimensional space. However, the challenge is that the objective involves expectation over an inner optimization problem. Further, the inner optimization problem, in its use of the Δ operator, implicitly relies on the ability to take one-step conditional expectation of the basis functions. We approximate these expectations by sample averages.

In particular, consider sampling a set of S *outer* sample paths denoted by $x^{(i)} \triangleq \{x_s^{(i)},\ s \in \mathcal{T}\}$ for $i = 1, 2, \ldots, S$, each sampled independently, conditional on the initial state x_0. Along each of these sample paths, we approximate the Δ operator by generating one-step *inner* samples. In particular, for each time $p \in \{1, \ldots, T\}$, we generate I independent *inner* samples $\{x_p^{(i,j)},\ j = 1, \ldots, I\}$, conditional on $x_{p-1} = x_{p-1}^{(i)}$, resulting in the approximation

$$\hat{\Delta}(\Phi r)_p(x_p^{(i)}, x_{p-1}^{(i)}) \triangleq (\Phi r)_p(x_p^{(i)}) - \frac{1}{I}\sum_{j=1}^{I} (\Phi r)_p(x_p^{(i,j)}). \tag{15}$$

Having thus replaced the expectations by their empirical counterparts, we obtain the following *nested* Monte Carlo approximation to the objective:

$$\hat{F}^{S,I}\Phi r(x_0) \triangleq \frac{1}{S}\sum_{i=1}^{S} \max_{0\le s\le d} \left(\alpha^s g(x_s^{(i)}) - \sum_{p=1}^{s} \alpha^p \hat{\Delta}(\Phi r)_p(x_p^{(i)}, x_{p-1}^{(i)}) \right). \tag{16}$$

Consequently, the sampled variant of PO is given by

$$\inf_r \ \hat{F}^{S,I}\Phi r(x_0),$$

which is equivalent to the following linear program

$$\begin{array}{ll} \underset{r,u}{\text{minimize}} & \dfrac{1}{S}\displaystyle\sum_{i=1}^{S} u_i \\ \text{subject to} & u_i + \displaystyle\sum_{p=1}^{s} \alpha^p \hat{\Delta}(\Phi r)_p(x_p^{(i)}) \ge \alpha^s g(x_s^{(i)}), \quad \forall\ 1 \le i \le S, \quad 0 \le s \le d, \\ & r \in \mathbb{R}^K,\ u \in \mathbb{R}^S. \end{array} \tag{17}$$

Desai et al. (2010) establish the convergence of this sampled LP, as the number of samples (S, I) tend to infinity.

The linear program (17) has $K + S$ variables and $S(d + 1)$ constraints. Since the u_i variables appear only 'locally', the Hessian corresponding to the logarithmic barrier function can be inverted in $O(K^2 S)$ floating point operations (see, for example, Boyd and Vandenberghe, 2004). Therefore, one may argue that the complexity of solving this LP via an interior point method essentially scales linearly with the number of outer sample paths S.

The PO method is a specific instance of a method that uses value function approximations to compute the martingale penalty. Further, the method can be shown to enjoy strong approximation guarantees. The quality of the upper bound produced by the PO method depends on three parameters: the error due to the *best possible* approximation afforded by the chosen basis function architecture, the square root of the effective time horizon, and a certain measure of the 'predictability' of the underlying Markov process. The latter parameter provides valuable insight on aspects of the underlying Markov process that make a particular pricing problem easy or hard. This result, described in Desai et al. (2010), also makes precise the intuition that the PO method produces good price approximations if the chosen basis function architecture contains a good approximation to the value function.

6.2. Linear Convex Control

In this section, we consider yet another class of MDPs, which we refer to as *linear convex control problems*. These problems essentially call for the minimization of some convex cost function of

the state trajectory subject to linear dynamics and, potentially, convex constraints on the control sequence. A number of interesting problems ranging from inventory control to portfolio optimization to network revenue management can be addressed using this framework.

Consider an MDP over the finite time horizon $\mathcal{T} \triangleq \{0, 1, \ldots, T\}$. For the purpose of this section, we assume that the state space $\mathcal{X} \triangleq \mathbb{R}^m$, the action space $\mathcal{A} \triangleq \mathbb{R}^n$ and the disturbance space $\mathcal{W} \triangleq \mathbb{R}^m$. The cost of taking some action a in state x at time t is given by a function $g_t \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ that is assumed jointly convex in its arguments. Further, the dynamics governing the evolution of x_t are assumed to be linear:

$$x_{t+1} = h(x_t, a_t, w_t) = A_t x_t + B_t a_t + w_t,$$

where $A_t \in \mathbb{R}^{m \times m}$ and $B_t \in \mathbb{R}^{m \times n}$ are deterministic matrices that govern the system dynamics, and $w_t \in \mathbb{R}^m$ is an i.i.d. disturbance. We allow for constraints on controls of the form $a_t \in \mathcal{K}_t$, where $\mathcal{K}_t \subset \mathbb{R}^n$ is a convex set. While we do not discuss this here, both the cost function and the nature of the constraints can be substantially relaxed: we may consider cost functions that are general convex functionals of the state and control trajectories and, under some technical conditions, can permit general convex constraints on the sequence of control actions employed; see Desai et al. (2011) for further details.

Let the sequence of policies, actions, states, and disturbances be denoted by $\boldsymbol{\mu}_T \triangleq (\mu_0, \mu_1 \ldots, \mu_T)$, $\mathbf{a}_T \triangleq (a_0, a_1, \ldots, a_T)$, $\mathbf{x}_T \triangleq (x_0, x_1 \ldots, x_T)$, and $\mathbf{w}_T \triangleq (w_0, \ldots, w_{T-1})$, respectively. Define the set of feasible nonanticipative policies by

$$\mathcal{A}^{\mathbb{F}} \triangleq \{\boldsymbol{\mu}_T \ : \ \mu_t \in \mathcal{K}_t, \ \forall\, t \in \mathcal{T}, \text{ and } \boldsymbol{\mu}_T \text{ is adapted to filtration } \mathcal{F}\}.$$

We are interested in the following undiscounted, finite horizon optimization problem

$$\inf_{\boldsymbol{\mu}_T \in \mathcal{A}^{\mathbb{F}}} \ \mathsf{E}\left[\sum_{t=0}^{T} g_t(x_t, \mu_t)\right]. \tag{18}$$

Under mild technical conditions (for details, see Desai et al., 2011) the optimal cost-to-go function J^* satisfies the Bellman equation

$$J_t^*(x) = \begin{cases} \inf\limits_{a_t \in \mathcal{K}_t} \ g_t(x, a_t) + \mathsf{E}\left[\, J_{t+1}^*(x_{t+1}) \middle|\, x_t = x, \ a_t\right] & \text{if } t < T, \\ \inf\limits_{a_T \in \mathcal{K}_T} \ g_T(x, a_T) & \text{if } t = T. \end{cases} \tag{19}$$

6.2.1. The Martingale Duality Approach

Let $\mathcal{S}$ be the space of real-valued functions defined on state space $\mathbb{R}^m$ and $\mathcal{P}$ be the space of real-valued functions on $\mathbb{R}^m \times \mathcal{T}$, such that $J_t \triangleq J(\cdot, t)$ belongs to $\mathcal{S}$. Define the *martingale difference*

operator Δ that maps a function $V \in \mathcal{S}$ to the function $\Delta V \colon \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ according to

$$(\Delta V)(x_{t+1}, x_t, a_t) \triangleq V(x_{t+1}) - \mathsf{E}[V(x_{t+1})|x_t, a_t].$$

We are interested in computing lower bounds by considering a perfect information relaxation. Define $\mathbf{K} \triangleq \mathcal{K}_0 \times \mathcal{K}_1 \times \ldots \times \mathcal{K}_T$ to be the set of all feasible control sequences $\mathbf{a}_T$. Given a feasible sequence of actions $\mathbf{a}_T \in \mathbf{K}$ and a function $J \in \mathcal{P}$, define the martingale $M_t^{\mathbf{a}_T}(J)$ by

$$M_0^{\mathbf{a}_T}(J) \triangleq 0, \qquad M_t^{\mathbf{a}_T}(J) \triangleq \sum_{s=1}^{t} \Delta J_s(x_s, x_{s-1}, a_{s-1}), \quad \forall\ 1 \leq t \leq T.$$

Then, we can define the *martingale duality operator* $F \colon \mathcal{P} \to \mathcal{S}$ according to:

$$(FJ)(x) \triangleq \mathsf{E}\left[\inf_{\mathbf{a}_T \in \mathbf{K}} \sum_{t=0}^{T} g_t(x_t, a_t) - M_T^{\mathbf{a}_T}(J) \;\middle|\; x_0 = x \right]. \tag{20}$$

In order for the deterministic inner optimization problem in (20) to be tractable, we need to impose special structure on the function J. To this end, given a sequence of matrices $\mathbf{\Gamma} \triangleq (\Gamma_1, \ldots, \Gamma_T)$, define the function $J^{\mathbf{\Gamma}} \in \mathcal{P}$ by

$$J_0^{\mathbf{\Gamma}}(x) \triangleq 0, \qquad J_t^{\mathbf{\Gamma}}(x) \triangleq x^\top \Gamma_t x, \quad \forall\ 1 \leq t \leq T.$$

Denote by $\mathcal{C} \subset \mathcal{P}$ the set of all functions of the form $J^{\mathbf{\Gamma}}$. The following theorem establishes that, for this class of quadratic functions, the inner optimization problem in (20) is a convex optimization problem, and therefore is tractable:

Theorem 6. *For all $J \in \mathcal{C}$, the inner optimization problem of* (20) *is a convex optimization problem.*

Proof. Suppose that $J = J^{\mathbf{\Gamma}} \in \mathcal{C}$. For each time t, apply the martingale difference operator Δ to $J_t^{\mathbf{\Gamma}}$ to obtain

$$\Delta J_t(x_t, x_{t-1}, a_{t-1}) = 2w_{t-1}^\top \Gamma_t (A_{t-1}x_{t-1} + B_{t-1}a_{t-1}) + w_{t-1}^\top \Gamma_t w_{t-1} - \mathsf{E}\left[w_{t-1}^\top \Gamma_t w_{t-1}\right] \tag{21}$$

Observe that the quantity $w_{t-1}^\top \Gamma_t w_{t-1} - \mathsf{E}\left[w_{t-1}^\top \Gamma_t w_{t-1}\right]$ is zero mean and independent of the control $\mathbf{a}_T$. We may consequently eliminate those terms from the inner optimization problem. In particular, given a fixed sequence of disturbances $\mathbf{w}_T$, the inner optimization problem becomes:

$$\begin{array}{ll} \underset{\mathbf{a}_T, \mathbf{x}_T}{\text{minimize}} & g_0(x_0, u_0) + \displaystyle\sum_{t=1}^{T} \Big\{ g_t(x_t, a_t) - 2w_{t-1}^\top \Gamma_t (A_{t-1}x_{t-1} + B_{t-1}a_{t-1}) \Big\} \\ \text{subject to} & x_{t+1} = A_t x_t + B_t a_t + w_t, \quad \forall\ 0 \leq t \leq T-1, \\ & a_t \in \mathcal{K}_t, \quad \forall\ 0 \leq t \leq T. \end{array} \tag{22}$$

This is clearly a convex optimization problem. ■

Theorem 6 suggested that for a quadratic[3] cost-to-go function surrogate $J^{\mathbf{\Gamma}} \in \mathcal{C}$, the lower bound $FJ^{\mathbf{\Gamma}}(x)$ on the optimal cost-to-go $J_0^*(x)$ can be efficiently computed. Finding the *tightest* such lower bound suggests the optimization problem

$$\sup_{\mathbf{\Gamma}} \; FJ^{\mathbf{\Gamma}}(x). \tag{23}$$

We now establish that this is also a convex optimization problem:

Theorem 7. *$FJ^{\mathbf{\Gamma}}(x)$ is concave in $\mathbf{\Gamma}$.*

Proof. Using the definition of the F operator given by (20) and the expression for $\Delta J(x_t, x_{t-1}, a_{t-1})$ given by (21), we obtain

$$FJ^{\mathbf{\Gamma}}(x) = \mathsf{E}\left[\inf_{a_T \in \mathbf{K}} \; g_0(x_0, a_0) + \sum_{t=1}^{T} \left\{ g_t(x_t, a_t) - 2w_{t-1}^\top \Gamma_t (A_{t-1}x_{t-1} + B_{t-1}a_{t-1}) \right\} \;\middle|\; x_0 = x \right].$$

Observe that $FJ^{\mathbf{\Gamma}}(x)$ is given by nonnegative linear combinations of infima of affine functions of $\mathbf{\Gamma}$. Since each of these operations preserves concavity, we obtain the desired result. ■

The PO problem given by (23) can be viewed as a stochastic optimization problem. This suggests two methods of solution:

- Iterative methods based on stochastic gradient descent can be used to solve (23). Starting from an initial guess for $\mathbf{\Gamma}$, the gradient of $FJ^{\mathbf{\Gamma}}(x)$ can be estimated along a single sample path $\mathbf{w}_T$ of random disturbances. The stochastic gradient estimate is then used to update the choice of $\mathbf{\Gamma}$, and the procedure is repeated until convergence. These steps together give rise to a simple online method that can be used to handle large problems with a low memory requirement.

- Alternatively, a sample average approximation can be used. Here, the objective function $FJ^{\mathbf{\Gamma}}(x)$ is approximated with a sample average over sequences of random disturbances $\mathbf{w}_T$. For a given realization of this sequence, the objective cost-to-go of the inner optimization problem, (22) can be expressed (using the appropriate conjugate functions) as a convex function of Γ. In several special cases this representation allows us to rewrite the overall optimization problem in a form suitable for direct optimization.

The details of both these approaches, along with application to a high-dimensional financial application, namely, an optimal execution problem, can be found in Desai et al. (2011).

Observe that the classic convex *linear quadratic control* (LQC) problem is an example of a linear convex problem. It is well-known that the optimal cost-to-go function for the convex LQC problem

[3]In fact, a broader class of cost-to-go functions including constant and linear terms could also be considered. However, such constant and linear terms are eliminated in the evaluation of the martingale difference operator in (21). Hence, they do not enter into the lower bound and can be ignored.

takes a positive semi-definite quadratic form and can be computed recursively (and efficiently) by solving the so-called Ricatti equation (see, e.g., Bertsekas, 1995). This tractability breaks down under seemingly innocuous constraints such as requiring non-negative control actions. Loosely speaking, the PO method bootstraps our ability to solve convex LQC problems to the task of producing good approximations to linear convex problems. It does so by seeking martingale penalty functions derived from quadratic approximations to the cost-to-go function. In particular, if convex quadratic forms are likely to provide a reasonable approximation to the cost-to-go function of the linear convex problem at hand, then one can expect the PO method to produce good lower bounds.

7. Conclusion

This chapter set out with the task of producing lower bounds on the optimal cost-to-go for high-dimensional Markov decision problems. We considered two seemingly disparate approaches to this task: the approximate linear programming (ALP) methodology and an approach based on finding martingale 'penalties' in a certain dual problem. In concluding, we observe that these two methodologies are intimately connected:

1. We have observed that given an approximation architecture for the ALP approach, one is naturally led to consider a corresponding family of martingale penalties derived from the same architecture. This consideration suggests an optimization problem that produces a martingale penalty yielding the tightest lower bound possible within the corresponding family of martingale penalties. We referred to this problem as the pathwise optimization (PO) problem.

2. We established that solving the PO problem yields approximations to the cost-to-go that are no worse than those produced by the ALP approach. This provided an elegant unification of the two approaches.

3. Finally, we demonstrated the algorithmic value of the PO method in the context of two broad classes of MDPs.

Moving forward, we believe that much remains to be done in developing the pathwise optimization approach described in this chapter. In particular, developing the approach successfully for a given class of problems requires that one first identify a suitable approximation architecture for that class of problems. This architecture should admit tractable PO problems and simultaneously be rich enough that it captures essential features of the true cost-to-go function. A number of problems from areas such as financial engineering, revenue management and inventory management are ripe for precisely this sort of study.

On an orthogonal note, while we have not studied this issue here, much remains to be done in using the solution of the PO problem to generate good heuristic *policies*. Desai et al. (2010) discuss

this in the context of optimal stopping, and demonstrate in numerical examples that PO-derived policies can be superior to policies derived from more conventional ADP methods. In general, some careful thought is needed here since optimal solutions to the PO problem are not unique. For example, in linear convex setting, the optimal solutions are only identified up to affine translations.

References

D. Adelman. A price-directed approach to stochastic inventory/routing. *Operations Research*, 52(4):499–514, 2004.

D. Adelman. Dynamic bid prices in revenue management. *Operations Research*, 55(4):647–661, 2007.

D. Adelman and D. Klabjan. Computing near optimal policies in generalized joint replenishment. Working paper, January 2009.

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

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

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 3rd edition, 2006.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 3rd edition, 2007.

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

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

D. B. Brown and J. E. Smith. Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds. *Management Science*, Forthcoming, 2010.

D. B. Brown, J. E. Smith, and P. Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4):785–801, July-August 2010.

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

D. P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 293(3):462–478, 2004.

V. V. Desai, V. F. Farias, and C. C. Moallemi. Pathwise optimization for optimal stopping problems. *Submitted*, 2010.

V. V. Desai, V. F. Farias, and C. C. Moallemi. Pathwise optimization for linear convex systems. *Working paper*, 2011.

V. F. Farias and B. Van Roy. An approximate dynamic programming approach to network revenue management. Working paper, 2007.

V. F. Farias, D. Saure, and G. Y. Weintraub. An approximate dynamic programming approach to solving dynamic oligopoly models. Working paper, 2011.

J. Han. *Dynamic Portfolio Management - An Approximate Linear Programming Approach*. PhD thesis, Stanford University, 2005.

M. B. Haugh and L. Kogan. Pricing American options: A duality approach. *Operations Research*, 52(2): 258–270, 2004.

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

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

Guoming Lai, Mulan X. Wang, Sunder Kekre, Alan Scheller-Wolf, and Nicola Secomandi. Valuation of storage at a liquefied natural gas terminal. *Operations Research*, Forthcoming, 2010b.

A. S. Manne. Linear programming and sequential decisions. *Management Science*, 60(3):259–267, 1960.

C. C. Moallemi, S. Kumar, and B. Van Roy. Approximate and data-driven dynamic programming for queueing networks. Working paper, 2008.

J. R. Morrison and P. R. Kumar. New linear program performance bounds for queueing networks. *Journal of Optimization Theory and Applications*, 100(3):575–597, 1999.

W. B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley and Sons, 2007.

L. C. G. Rogers. Monte Carlo valuation of American options. *Mathematical Finance*, 12(3):271–286, 2002.

L. C. G. Rogers. Pathwise stochastic optimal control. *SIAM Journal on Control and Optimization*, 46(3): 1116–1132, 2008.

P. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110:568–582, 1985.

B. Van Roy. Neuro-dynamic programming: Overview and recent trends. In A. Shwartz E. Feinberg, editor, *Handbook of Markov Decision Processes*. Kluwer, Boston, 2002.

M. H. Veatch. Approximate dynamic programming for networks: Fluid models and constraint reduction. Working paper, 2005.

D. Zhang and D. Adelman. An approximate dynamic programming approach to network revenue management with customer choice. Working paper, 2008.