

# Information Relaxations and Duality in Stochastic Dynamic Programs

David B. Brown, James E. Smith, Peng Sun

Fuqua School of Business, Duke University, Durham, North Carolina 27708  
{dbbrown@duke.edu, jes9@duke.edu, psun@duke.edu}

We describe a general technique for determining upper bounds on maximal values (or lower bounds on minimal costs) in stochastic dynamic programs. In this approach, we relax the nonanticipativity constraints that require decisions to depend only on the information available at the time a decision is made and impose a “penalty” that punishes violations of nonanticipativity. In applications, the hope is that this relaxed version of the problem will be simpler to solve than the original dynamic program. The upper bounds provided by this dual approach complement lower bounds on values that may be found by simulating with heuristic policies. We describe the theory underlying this dual approach and establish weak duality, strong duality, and complementary slackness results that are analogous to the duality results of linear programming. We also study properties of good penalties. Finally, we demonstrate the use of this dual approach in an adaptive inventory control problem with an unknown and changing demand distribution and in valuing options with stochastic volatilities and interest rates. These are complex problems of significant practical interest that are quite difficult to solve to optimality. In these examples, our dual approach requires relatively little additional computation and leads to tight bounds on the optimal values.

*Subject classifications:* dynamic programming; duality; inventory control; option pricing.

*Area of review:* Stochastic Models.

*History:* Received November 2007; revisions received November 2008, July 2009, September 2009; accepted October 2009. Published online in *Articles in Advance* April 9, 2010.

## 1. Introduction

In principle, dynamic programming provides a powerful framework for determining optimal policies in complex decision problems where uncertainty is resolved and decisions are made over time. However, the widespread use of dynamic programming is hampered by the so-called “curse of dimensionality”—the size of the state space typically grows exponentially in the number of state variables considered. In contrast, Monte Carlo simulation methods typically scale well with the number of state variables considered and, given a control policy, it is not difficult to simulate a complex dynamic system with many uncertainties. Simulating with a feasible policy provides a lower bound on the expected value (or upper bound on the expected costs) of an optimal policy, but Monte Carlo simulation typically does not provide a good way to identify an optimal policy or provide an upper bound on the value of an optimal policy.

In this paper, we describe a dual approach for studying stochastic dynamic programs (DPs) that focuses on providing an upper bound on the optimal expected value. This dual approach consists of two elements: (1) we relax the nonanticipativity constraints that require decisions to depend only on the information available at the time a decision is made and (2) we impose a penalty that punishes violations of the nonanticipativity constraints. By relaxing the

nonanticipativity constraints, we can often greatly simplify the DP. For example, we study an adaptive inventory control problem with an unknown and changing demand distribution and stochastic ordering costs. Here a “perfect information relaxation” assumes the decision maker (DM) knows all demands and costs before placing any orders. With this information, the problem of choosing an optimal ordering schedule is a deterministic DP that can be solved quite easily. In another example, we study an option-pricing model with stochastic volatilities and stochastic interest rates and consider an “imperfect information relaxation” where volatilities and interest rates are known in advance but the stock price is not: with the volatilities and interest rates known, we can value the option using standard lattice methods.

Because these relaxations assume the DM has more information than is truly available, they lead to an upper bound on value. Without any penalty for using this additional information, the bound obtained is often quite weak. Informally, we say a penalty is dual feasible if it does not penalize any policy that is nonanticipative; the penalties may, however, punish policies that do not satisfy the nonanticipativity constraints. We will show that in principle we can always find a dual feasible penalty that provides a tight bound, i.e., strong duality holds.

We view this dual approach as a complement to the use of simulation methods and modern approximate

dynamic programming methods for studying DPs (see, e.g., Bertsekas and Tsitsiklis 1996, de Farias and Van Roy 2003, Powell 2007, or Adelman and Merseuer 2008). As mentioned earlier, given a candidate policy (perhaps identified using a heuristic approach or using approximate DP techniques), we can use standard simulation techniques to estimate the expected value with this policy and thereby generate a lower bound on the expected value with an optimal policy. Our dual approach can then be used to generate an upper bound on the value of an optimal policy. If the difference between the expected value with this candidate policy and the upper bound on the optimal value is small, we may conclude that the candidate policy is “good enough” and not continue searching for a better policy. If the difference is large, it may be worthwhile to work harder to find a better policy and/or a tighter upper bound. In our inventory example, we will use the dual bounds to determine whether a simple myopic ordering policy is “good enough” or whether we need to consider more complex one- or two-period look-ahead policies. In the option-pricing example, we use the dual bounds to study the effectiveness of an exercise policy that ignores uncertainty about volatilities and interest rates. In both examples, we will also demonstrate how we can use the results of the dual problem to identify ways to improve these heuristic policies.

Our interest in this dual approach for DPs was motivated by the need to evaluate the quality of heuristic policies in applications, and inspired by Haugh and Kogan’s (2004) dual approach for placing bounds on the value of an American option; Rogers (2002) independently proposed a similar dual approach, also applied to option pricing. Both Haugh and Kogan (2004) and Rogers (2002) consider the use of what we call perfect information relaxations and establish their main results using martingale arguments. Haugh and Kogan propose a particular method for generating penalties or, in their terminology, “dual martingales” based on approximate value functions and demonstrate the use of this method in high-dimensional option-pricing problems. Andersen and Broadie (2004) propose an alternative method for generating dual martingales based on approximate policies. Glasserman (2004) provides a nice overview of this work.

We generalize the work of Haugh and Kogan (2004), Rogers (2002), and Andersen and Broadie (2004) in several ways. First, rather than focusing exclusively on option-pricing problems, we consider general stochastic DPs. Second, rather than focusing exclusively on perfect information relaxations, we consider general information relaxations. Finally, we present a general method for constructing good penalties that includes and extends the methods proposed by Haugh and Kogan and Andersen and Broadie. These generalizations expand the scope and flexibility of this dual approach.

The idea of relaxing the nonanticipativity constraints has also been studied in the stochastic programming literature (see, e.g., Rockafellar and Wets 1991, Shapiro and

Ruszczynski 2003, Shapiro et al. 2009). Rogers (2007) also recently (independently) proposed a dual approach for Markov decision processes. In short, though these alternative approaches have similarities with ours, our formulation is different and leads to results that we believe are both simpler and more general. The stochastic programming formulation requires the reward functions and set of feasible actions to be convex and the penalties considered are linear functions of the actions; they consider only perfect information relaxations. Rogers focuses on Markov decision processes and considers only perfect information relaxations and penalties that are a function of the state variable only; Rogers does not present any example applications. In contrast, our framework allows general reward functions and action spaces, allows general penalty functions, and considers imperfect as well as perfect information relaxations. Moreover, our duality proofs are quite simple and direct and do not rely on sophisticated convex duality or martingale arguments. Finally, our inventory control and option-pricing examples demonstrate the power of this dual approach in some complex problems of significant practical interest.

We begin in §2 by defining the basic framework and theory underlying the dual approach; the main results are analogous to the duality results of linear programming. We then illustrate the approach in the inventory control and option-pricing examples in §3–4. We offer a few concluding remarks in §5. The electronic companion provides supporting information: Appendix A contains most of the proofs; Appendix B compares our results to similar results in stochastic programming and develops the connections to linear programming more fully; and Appendix C provides some details of the adaptive inventory example. The electronic companion to this paper is available as part of the online version that can be found at <http://or.journal.informs.org/>.

## 2. The Basic Framework and Results

We begin by describing the general formulation of the primal stochastic DP in §2.1. We then present our main duality results in §2.2 and discuss an approach for generating good penalties in §2.3.

### 2.1. General Framework

Uncertainty in the DP is described by a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  where  $\Omega$  is the set of possible outcomes (with typical element  $\omega$ ),  $\mathcal{F}$  is a  $\sigma$ -algebra that describes the set of all possible events (an event is a subset of  $\Omega$ ), and  $\mathbb{P}$  is a probability measure describing the likelihoods of the various events.

Time is discrete and indexed by  $t = 0, \dots, T$ . The DM’s state of information evolves over time and is described by a filtration  $\mathbb{F} = (\mathcal{F}_0, \dots, \mathcal{F}_T)$  where the  $\sigma$ -algebra  $\mathcal{F}_t$  describes the DM’s state of information at the beginning of period  $t$ , i.e.,  $\mathcal{F}_t$  is the set of events that will be known to be true or false at time  $t$ . We will refer to  $\mathbb{F}$  as the *natural filtration*. We

require all filtrations to satisfy  $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$  for all  $t < T$  so the DM does not forget what she once knew. We will assume that  $\mathcal{F}_0 = \{\emptyset, \Omega\}$ , so the DM initially “knows nothing” about the outcome of the uncertainties. A function (or random variable)  $f$  defined on  $\Omega$  is *measurable* with respect to a  $\sigma$ -algebra  $\mathcal{F}_t$  (or  $\mathcal{F}_t$ -measurable) if for every Borel set  $R$  in the range of  $f$ , we have  $\{\omega: f(\omega) \in R\} \in \mathcal{F}_t$ ; we can interpret  $f$  being  $\mathcal{F}_t$ -measurable as meaning the result of  $f$  depends only on the information known in period  $t$ . A sequence of functions  $(f_0, \dots, f_T)$  is said to be *adapted* to a filtration  $\mathbb{F}$  (or  $\mathbb{F}$ -adapted) if each function  $f_t$  is measurable with respect to  $\mathcal{F}_t$ .

In the DP model, the DM will choose an action  $a_t$  in period  $t$  from the set  $A_t$ ; we let  $A \subseteq A_0 \times \dots \times A_T$  denote the set of all feasible action sequences  $a$ . The DM’s choice of actions is described by a *policy*  $\alpha$  that selects a sequence of actions  $a$  in  $A$  for each outcome  $\omega$  in  $\Omega$  (i.e.,  $\alpha: \Omega \rightarrow A$ ). We let  $\mathcal{A}$  denote the set of all policies. In the primal DP, we assume that the DM’s choices are *nonanticipative* in that the choice of action  $a_t$  in period  $t$  depends only on what is known at the beginning of period  $t$ . More formally, we require policies to be adapted to the natural filtration  $\mathbb{F}$  in that a policy’s selection of the first  $t+1$  actions  $(a_0, \dots, a_t)$  must be measurable with respect to  $\mathcal{F}_t$ . We let  $\mathcal{A}_{\mathbb{F}}$  be the set of all nonanticipative policies.<sup>1</sup>

The goal of the DP is to select a nonanticipative policy  $\alpha$  to maximize the expected total reward. The rewards are defined by a sequence of reward functions  $(r_0(a, \omega), \dots, r_T(a, \omega))$  where the reward in period  $t$  depends on the action sequence  $a$  selected and the outcome  $\omega$ . We let  $r(a, \omega) = \sum_{t=0}^T r_t(a, \omega)$  denote the total reward; discounting can be incorporated into the period reward function  $r_t$ . The primal DP is then:

$$\sup_{\alpha \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[r(\alpha)]. \quad (1)$$

Here  $\mathbb{E}[r(\alpha)]$  could be written more explicitly as  $\mathbb{E}[r(\alpha(\omega), \omega)]$ , where policy  $\alpha$  selects an action sequence that depends on the random outcome  $\omega$ , and the rewards  $r$  depend on the action sequence selected by  $\alpha$  and the outcome  $\omega$ . We will typically suppress the dependence on  $\omega$  and interpret  $r(\alpha)$  as a random variable representing the reward generated with policy  $\alpha$ .

It is instructive to write the primal DP (1) in the standard Bellman-style recursive form. First, we will assume that the period- $t$  rewards  $r_t$  are  $\mathcal{F}_t$ -measurable for each set of actions and depend only on the first  $t+1$  actions  $(a_0, \dots, a_t)$ ; we will write  $r_t(a)$  as  $r_t(a_0, \dots, a_t)$  with the understanding that  $(a_0, \dots, a_t)$  is selected from the full sequence of actions  $a$ . For  $t > 0$ , let  $A_t(a_0, \dots, a_{t-1})$  be the subset of period- $t$  actions  $A_t$  that are feasible given the prior choice of actions  $(a_0, \dots, a_{t-1})$ . We take the terminal value function  $V_{T+1}(a_0, \dots, a_T) = 0$  and, for  $t = 0, \dots, T$ , we define

$$V_t(a_0, \dots, a_{t-1}) = \sup_{a_t \in A_t(a_0, \dots, a_{t-1})} \{r_t(a_0, \dots, a_t) + \mathbb{E}[V_{t+1}(a_0, \dots, a_t) | \mathcal{F}_t]\}. \quad (2)$$

Here both sides are random variables (and therefore implicitly functions of the outcome  $\omega$ ) and we select an optimal action  $a_t$  for each outcome  $\omega$ . Because the rewards  $r_t$  are assumed to be  $\mathcal{F}_t$ -measurable and the expected continuation values are conditioned on  $\mathcal{F}_t$ , and thus  $\mathcal{F}_t$ -measurable, the objective function on the right is  $\mathcal{F}_t$ -measurable for each set of actions  $(a_0, \dots, a_t)$ . Thus, the supremum over actions  $a_t$  is also  $\mathcal{F}_t$ -measurable, which implies that  $V_t$  is  $\mathcal{F}_t$ -measurable. There is no loss in restricting the choice of actions  $a_t$  to be  $\mathcal{F}_t$ -measurable; therefore, if the suprema on the right side of (2) are attained, we can construct a nonanticipative optimal policy using this recursion. The final value  $V_0$  is equal to the optimal value of (1).

## 2.2. The Dual Approach

In our dual approach to the DP (1), we relax the requirement that the policies be nonanticipative and impose penalties that punish violations of the nonanticipativity constraints. We define relaxations of the nonanticipativity requirement by considering alternative information structures. We say that a filtration  $\mathbb{G} = (\mathcal{G}_0, \dots, \mathcal{G}_T)$  is a *relaxation* of the natural filtration  $\mathbb{F} = (\mathcal{F}_0, \dots, \mathcal{F}_T)$  if, for each  $t$ ,  $\mathcal{F}_t \subseteq \mathcal{G}_t \subseteq \mathcal{F}$ ; we abbreviate this by writing  $\mathbb{F} \subseteq \mathbb{G}$ .  $\mathbb{G}$  being a relaxation of  $\mathbb{F}$  means that the DM knows more in every period under  $\mathbb{G}$  than she knows under  $\mathbb{F}$ . The perfect information filtration  $\mathbb{I} = (\mathcal{I}_0, \dots, \mathcal{I}_T)$  is given by taking  $\mathcal{I}_t = \mathcal{F}$  for all  $t$ . We let  $\mathcal{A}_{\mathbb{G}}$  denote the set of policies that are adapted to  $\mathbb{G}$ . For any relaxation  $\mathbb{G}$  of  $\mathbb{F}$ , we have  $\mathcal{A}_{\mathbb{F}} \subseteq \mathcal{A}_{\mathbb{G}} \subseteq \mathcal{A}_{\mathbb{I}} = \mathcal{A}$ ; thus, as we relax the filtration, we expand the set of feasible policies.

The set of penalties  $\mathcal{Z}$  is the set of all functions  $z(a, \omega)$  that, like the total rewards, depend on the choice of action sequence  $a$  and the outcome  $\omega$ . As with rewards, we will typically write the penalties as an action-dependent random variable  $z(a)(=z(a, \omega))$  or a policy-dependent random variable  $z(\alpha)(=z(\alpha(\omega), \omega))$ , suppressing the dependence on the outcome  $\omega$ . We define the set  $\mathcal{Z}_{\mathbb{F}}$  of *dual feasible penalties* to be those penalties that do not penalize nonanticipative policies (in expectation), that is

$$\mathcal{Z}_{\mathbb{F}} = \{z \in \mathcal{Z}: \mathbb{E}[z(\alpha_F)] \leq 0 \text{ for all } \alpha_F \text{ in } \mathcal{A}_{\mathbb{F}}\}. \quad (3)$$

Policies that do not satisfy the nonanticipativity constraints (and thus are not feasible to implement) may have positive expected penalties.

We can place an upper bound on the expected reward associated with any nonanticipative policy by relaxing the nonanticipativity constraint on policies and imposing a dual feasible penalty. This simple result can be viewed as a version of the “weak duality lemma” for linear programming:

**LEMMA 2.1 (WEAK DUALITY).** *If  $\alpha_F$  and  $z$  are primal and dual feasible, respectively (i.e.,  $\alpha_F \in \mathcal{A}_{\mathbb{F}}$  and  $z \in \mathcal{Z}_{\mathbb{F}}$ ), and  $\mathbb{G}$  is a relaxation of  $\mathbb{F}$ , then*

$$\mathbb{E}[r(\alpha_F)] \leq \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[r(\alpha_G) - z(\alpha_G)]. \quad (4)$$

PROOF. With  $z$ ,  $\alpha_F$ , and  $\mathbb{G}$  as defined in the lemma, we have

$$\mathbb{E}[r(\alpha_F)] \leq \mathbb{E}[r(\alpha_F) - z(\alpha_F)] \leq \sup_{\alpha_G \in \mathcal{A}_G} \mathbb{E}[r(\alpha_G) - z(\alpha_G)].$$

The first inequality holds because  $z \in \mathcal{Z}_F$  (thus  $\mathbb{E}[z(\alpha_F)] \leq 0$ ) and the second because  $\alpha_F \in \mathcal{A}_F \subseteq \mathcal{A}_G$ .  $\square$

Thus, any information relaxation with any dual feasible penalty provides an upper bound on all DP solutions. With a fixed penalty  $z$ , weaker relaxations  $\mathbb{G}$  lead to larger sets of feasible policies  $\mathcal{A}_G$  and weaker bounds. For example, if we consider the perfect information relaxation  $\mathbb{I}$ , the set of relaxed policies  $\mathcal{A}_I$  is simply the set of all policies  $\mathcal{A}$  and all actions are selected with full knowledge of the outcome. Thus, the weak duality lemma implies that for any  $\alpha_F$  in  $\mathcal{A}_F$  and  $z$  in  $\mathcal{Z}_F$ ,

$$\mathbb{E}[r(\alpha_F)] \leq \sup_{\alpha \in \mathcal{A}} \mathbb{E}[r(\alpha) - z(\alpha)] = \mathbb{E}\left[\sup_{a \in A} \{r(a) - z(a)\}\right]. \quad (5)$$

If we take the penalty  $z = 0$ , this upper bound is the expected value with perfect information.

Note that the upper bound (5) is in a form that is convenient for Monte Carlo simulation: we can estimate the expected value on the right side of (5) by randomly generating outcomes  $\omega$  and solving a deterministic “inner problem” of choosing an action sequence  $a$  to maximize the penalized objective  $r(a, \omega) - z(a, \omega)$  for each  $\omega$ . For instance, in our inventory example, the perfect information relaxation assumes the DM has knowledge of all demands and costs before making any ordering decisions. We estimate the dual bound by randomly generating demand/cost scenarios in the “outer simulation,” and the inner problem is a simple deterministic DP that chooses optimal ordering quantities in each demand/cost scenario. With imperfect information relaxations, we can often still use Monte Carlo simulation to estimate the upper bounds. For instance, in our option-pricing example, we will randomly generate interest rates and volatilities in the outer simulation, and the inner problem is a one-dimensional DP that considers uncertainty in stock prices.

If we minimize over the dual feasible penalties in (4), we obtain the dual of the primal DP (1):

$$\inf_{z \in \mathcal{Z}_F} \left\{ \sup_{\alpha_G \in \mathcal{A}_G} \mathbb{E}[r(\alpha_G) - z(\alpha_G)] \right\}. \quad (6)$$

By the weak duality lemma, if we identify a policy  $\alpha_F$  and penalty  $z$  that are primal and dual feasible, respectively, such that equality holds in (4), then  $\alpha_F$  and  $z$  must be optimal for their respective problems. In such a case, there would be no gap between the values given by these primal and dual solutions. If the primal solution is bounded, there is always a dual feasible penalty that yields no gap. For example, consider the penalty  $z^*(a) = r(a) - v^*$  where  $v^*$  is the optimal value of the primal DP (1). This  $z^*$  is dual feasible (because  $\mathbb{E}[r(\alpha_F)] \leq v^*$  for all  $\alpha_F \in \mathcal{A}_F$ ) and trivially

optimal: no matter what policy is selected, the penalized objective function  $r(a) - z^*(a)$  is equal to  $v^*$ . The existence of this trivially optimal penalty is not helpful in practice because it requires knowing the optimal value  $v^*$  of the primal DP. It does, however, show that there is no gap between the solutions to the primal and dual problems and that, in principle, we could determine the maximal expected reward in the primal DP (1) by solving the dual problem (6). This result is analogous to the strong duality theorem of linear programming.

**THEOREM 2.1 (STRONG DUALITY).** *Let  $\mathbb{G}$  be a relaxation of  $\mathbb{F}$ . Then*

$$\sup_{\alpha_F \in \mathcal{A}_F} \mathbb{E}[r(\alpha_F)] = \inf_{z \in \mathcal{Z}_F} \left\{ \sup_{\alpha_G \in \mathcal{A}_G} \mathbb{E}[r(\alpha_G) - z(\alpha_G)] \right\}. \quad (7)$$

*Furthermore, if the primal problem on the left is bounded, the dual problem on the right has an optimal solution  $z^* \in \mathcal{Z}_F$  that achieves this bound.*

The “complementary slackness condition” further characterizes the relationship between the primal and dual problems, saying that for a primal-dual pair  $(\alpha_F^*, z^*)$  to be optimal, it is necessary and sufficient for  $\alpha_F^*$  to have zero expected penalty and for  $\alpha_F^*$  to “solve” the dual problem in the following sense.

**THEOREM 2.2 (COMPLEMENTARY SLACKNESS).** *Let  $\alpha_F^*$  and  $z^*$  be feasible solutions for the primal and dual problems respectively (i.e.,  $\alpha_F^* \in \mathcal{A}_F$  and  $z^* \in \mathcal{Z}_F$ ), with information relaxation  $\mathbb{G}$ . A necessary and sufficient condition for these to be optimal solutions for their respective problems is that  $\mathbb{E}[z^*(\alpha_F^*)] = 0$  and*

$$\mathbb{E}[r(\alpha_F^*) - z^*(\alpha_F^*)] = \sup_{\alpha_G \in \mathcal{A}_G} \mathbb{E}[r(\alpha_G) - z^*(\alpha_G)]. \quad (8)$$

Equation (8) can be interpreted as saying that with an optimal penalty, in dual problem the DM will be content to choose a policy that is nonanticipative even though she has the option of choosing a policy that is not. In applications, we will compare the heuristic policies  $\alpha_F$  used to compute a lower bound with the policies  $\alpha_G$  selected in the dual problem to see if we can identify some way to improve the heuristic policy.

Finally, we note a useful property of this dual approach: if we can simplify the primal problem by focusing on some subset of policies, we can restrict the dual problem to focus on policies in this same set. For example, if we know that the optimal policy for the primal problem is myopic or has a threshold structure, we can simplify the dual problem by considering only policies that have the same structure. This leads to dual bounds that are at least as tight and perhaps easier to compute than the dual bounds that do not include this constraint. We summarize this property as follows.

**PROPOSITION 2.1 (STRUCTURED POLICIES.)** *If for some  $\mathcal{S} \subseteq \mathcal{A}$  we have  $\sup_{\alpha_F \in \mathcal{A}_F} \mathbb{E}[r(\alpha_F)] = \sup_{\alpha_F \in \mathcal{S}_F} \mathbb{E}[r(\alpha_F)]$ , then, for any dual feasible  $z$ , we have*

$$\begin{aligned} \sup_{\alpha_F \in \mathcal{A}_F} \mathbb{E}[r(\alpha_F)] &\leq \sup_{\alpha_G \in \mathcal{S}_G} \mathbb{E}[r(\alpha_G) - z(\alpha_G)] \\ &\leq \sup_{\alpha_G \in \mathcal{A}_G} \mathbb{E}[r(\alpha_G) - z(\alpha_G)]. \end{aligned} \quad (9)$$

Moreover, the inequalities also hold for all  $z$  such that  $\mathbb{E}[z(\alpha_F)] \leq 0$  for all  $\alpha_F$  in  $\mathcal{S}_F$ .

For instance, in our option-pricing example, in the primal problem it is never optimal to exercise a call option prior to expiration, except possibly just before a dividend is paid. However, in the dual problem with a relaxed filtration, “early exercise” may be optimal. In our numerical experiments for this example, we will use this structural result and impose a “no early exercise” constraint in the dual problem for call options. The resulting bounds are both tighter and easier to compute than they would be without this constraint.

### 2.3. Good Penalties

In our discussion so far, we have considered the set of all dual feasible penalties. We now focus on identifying “good” penalties that are likely to be useful in practice. The main approach we will use to generate penalties is described in the following proposition. We will show shortly that we can, in principle, generate an optimal dual penalty using this approach, so that strong duality holds even when restricted to these “good” penalties.

**PROPOSITION 2.2 (CONSTRUCTING GOOD PENALTIES).** *Let  $\mathbb{G}$  be a relaxation of  $\mathbb{F}$  and let  $(w_0(a, \omega), \dots, w_T(a, \omega))$  be a sequence of generating functions defined on  $A \times \Omega$  where each  $w_t$  depends only on the first  $t + 1$  actions  $(a_0, \dots, a_t)$  of  $a$ . Define  $z_t(a) = \mathbb{E}[w_t(a) | \mathcal{G}_t] - \mathbb{E}[w_t(a) | \mathcal{F}_t]$  and  $z(a) = \sum_{t=0}^T z_t(a)$ . Then:*

- (i) *For all  $\alpha_F \in \mathcal{A}_F$ , we have  $\mathbb{E}[z_t(\alpha_F) | \mathcal{F}_t] = 0$  for all  $t$ , and  $\mathbb{E}[z(\alpha_F)] = 0$ ; and*
- (ii)  *$(z_0(a), \dots, z_T(a))$  is adapted to  $\mathbb{G}$  and  $z_t$  depends only on the first  $t + 1$  actions  $(a_0, \dots, a_t)$  of  $a$ .*

Property (i) of the proposition implies that the penalties  $z$  generated using the proposition will always be dual feasible in that  $\mathbb{E}[z(\alpha_F)] \leq 0$  for  $\alpha_F$  in  $\mathcal{A}_F$ , but is stronger in that it implies the inequality defining feasibility holds with equality. The complementary slackness condition (Theorem 2) shows that an optimal penalty  $z^*$  will assign zero expected penalty to an optimal primal policy  $\alpha^*$ . Penalties generated using Proposition 2.2 will assign zero expected penalty to all nonanticipative policies.

Property (ii) of the proposition implies that the penalized objective function can be decomposed into period- $t$  components  $r_t - z_t$  that depend only on what is known at period  $t$  under  $\mathbb{G}$  and the actions chosen in or before period  $t$ . This

means we can solve the dual problem using a DP recursion like that of Equation (2) using the penalized rewards and based on filtration  $\mathbb{G}$  rather than  $\mathbb{F}$ . Specifically, the terminal dual value function is  $V_{T+1}^{\mathbb{G}}(a_0, \dots, a_T) = 0$  and, for  $t = 0, \dots, T$ , we have

$$\begin{aligned} V_t^{\mathbb{G}}(a_0, \dots, a_{t-1}) &= \sup_{a_t \in A_t(a_0, \dots, a_{t-1})} \{r_t(a_0, \dots, a_t) - z_t(a_0, \dots, a_t) \\ &\quad + \mathbb{E}[V_{t+1}^{\mathbb{G}}(a_0, \dots, a_t) | \mathcal{G}_t]\} \\ &= \sup_{a_t \in A_t(a_0, \dots, a_{t-1})} \{r_t(a_0, \dots, a_t) - \mathbb{E}[w_t(a_0, \dots, a_t) | \mathcal{G}_t] \\ &\quad + \mathbb{E}[w_t(a_0, \dots, a_t) | \mathcal{F}_t] \\ &\quad + \mathbb{E}[V_{t+1}^{\mathbb{G}}(a_0, \dots, a_t) | \mathcal{G}_t]\}. \end{aligned} \quad (10)$$

The initial value,  $V_0^{\mathbb{G}}$ , provides an upper bound on the primal DP (1) or, equivalently, (2).

We can construct an optimal penalty using Proposition 2.2 by taking the generating functions to be based on the optimal DP value function given by (2). Specifically, if we take  $w_t(a) = V_{t+1}^{\mathbb{G}}(a_0, \dots, a_t)$ , we arrive at an optimal dual penalty  $z^*(a)$  that we will refer to as the “ideal” penalty. It is easy to show by induction that with this choice of generating function, the dual value functions are equal to the corresponding primal value functions, i.e.,  $V_t^{\mathbb{G}} = V_t$ . This is trivially true for the terminal values (both are zero). If we assume that  $V_{t+1}^{\mathbb{G}} = V_{t+1}$ , terms cancel and (10) reduces to the expression for  $V_t$  given in Equation (2). Thus, with this choice of generating function, we obtain an optimal penalty for any information relaxation  $\mathbb{G}$ . The following theorem summarizes this result and adds a bit more.

**THEOREM 2.3 (THE IDEAL PENALTY).** *Let  $\mathbb{G}$  be a relaxation of  $\mathbb{F}$  and let  $z^*$  be defined as in Proposition 2.2 by taking  $w_t(a) = V_{t+1}^{\mathbb{G}}(a_0, \dots, a_t)$ . Then  $z^*$  is dual feasible and optimal in that*

$$\sup_{\alpha_F \in \mathcal{A}_F} \mathbb{E}[r(\alpha_F)] = \sup_{\alpha_G \in \mathcal{A}_G} \mathbb{E}[r(\alpha_G) - z^*(\alpha_G)]. \quad (11)$$

Moreover, if  $\alpha_F^* \in \mathcal{A}_F$  achieves the supremum for the primal problem on the left side of (11), then  $\alpha_F^*$  is also optimal for the dual problem on the right. Finally, if  $\mathbb{G}$  is the perfect information relaxation and  $\alpha_F^* \in \mathcal{A}_F$  is an optimal policy, then  $r(\alpha_F^*) - z^*(\alpha_F^*) = \mathbb{E}[r(\alpha_F^*)]$  almost always.

Although the value functions will not be known in applications, the form of  $z^*$  illustrates the ideal that we would like to approximate with our choice of penalties. Intuitively, we would like to choose penalties that eliminate the benefit of choosing actions based on the information in  $\mathbb{G}$  rather than relying on the information in the natural filtration  $\mathbb{F}$ . That is, we want to choose a generating function  $w_t$  so that the differences  $\mathbb{E}[w_t(a) | \mathcal{G}_t] - \mathbb{E}[w_t(a) | \mathcal{F}_t]$  approximate the differences  $\mathbb{E}[V_{t+1}(a) | \mathcal{G}_t] - \mathbb{E}[V_{t+1}(a) | \mathcal{F}_t]$  and the

conditional expectations ( $\mathbb{E}[w_t(a) | \mathcal{G}_t]$  and  $\mathbb{E}[w_t(a) | \mathcal{F}_t]$ ) are not too difficult to compute.

In applications, we can approximate  $z^*$  in a variety of ways. Haugh and Kogan (2004) and Andersen and Broadie (2004) proposed methods for generating penalties (or dual martingales) in the option pricing context that can be generalized to our setting. Generalizing Haugh and Kogan's approach, we can approximate the ideal penalty  $z^*$  by using an approximate value function  $\hat{v}_t(a)$  in place of the true value function  $V_t(a)$ . This leads to period- $t$  penalties of the form  $z_t(a) = \mathbb{E}[\hat{v}_{t+1}(a) | \mathcal{G}_t] - \mathbb{E}[\hat{v}_{t+1}(a) | \mathcal{F}_t]$ . To use this approach, we must somehow estimate or calculate the conditional expectations  $\mathbb{E}[\hat{v}_{t+1}(a) | \mathcal{G}_t]$  and  $\mathbb{E}[\hat{v}_{t+1}(a) | \mathcal{F}_t]$ . Haugh and Kogan consider the perfect information relaxation ( $\mathcal{G}_t = \mathcal{F}$ ) so  $\mathbb{E}[\hat{v}_{t+1}(a) | \mathcal{G}_t] = \hat{v}_{t+1}(a)$  can be evaluated directly for any sample path. They estimate  $\mathbb{E}[\hat{v}_{t+1}(a) | \mathcal{F}_t]$  using a “nested simulation” procedure: for each sample path and each period  $t$ , they estimate  $\mathbb{E}[\hat{v}_{t+1}(a) | \mathcal{F}_t]$  by generating random successors to the period- $t$  state and averaging the next period values  $\hat{v}_{t+1}(a)$  in these successor states. The penalties generated using this approach will lead to valid bounds as long as the nested estimates of these conditional expectations are unbiased; see Proposition 2.3(iv) below.

Andersen and Broadie (2004) also consider a perfect information relaxation, but base their penalty on a given policy rather than an approximate value function. In our framework, their approach can be seen as approximating the value function  $V_t(a)$  with  $v_t^\alpha(a) = \mathbb{E}[r(\alpha_t(a)) | \mathcal{F}_t]$ , where  $\alpha_t(a)$  denotes a policy that takes the first  $t$  actions ( $a_0, \dots, a_{t-1}$ ) to match those of  $a$  and then continues according to some given rule. The penalty is then  $z_t(a) = \mathbb{E}[v_{t+1}^\alpha(a) | \mathcal{G}_t] - \mathbb{E}[v_{t+1}^\alpha(a) | \mathcal{F}_t]$ ; with a perfect information relaxation, this is equivalent to  $z_t(a) = \mathbb{E}[r(\alpha_{t+1}(a)) | \mathcal{F}_{t+1}] - \mathbb{E}[r(\alpha_{t+1}(a)) | \mathcal{F}_t]$ . Andersen and Broadie generate sample paths in the outer simulation and estimate the conditional expectations using nested simulation. Whereas the nested simulations in the Haugh-Kogan approach consider a single period, here each period's nested simulation follows the specified policy through the end of the horizon or until the policy calls for stopping. Because each future period is considered in each nested simulation, the work involved in the Andersen-Broadie approach potentially grows with  $T^2$  where  $T$  is the number of periods considered in the model. Again, these penalties will lead to valid bounds as long as the estimates of these conditional expectations are unbiased.

In practice, there will typically be a trade-off between the quality of the bound and the computational effort required to compute it. We can control this trade-off through our choice of information relaxation and penalty. The following proposition provides some properties of penalties and information relaxations that are useful in understanding these trade-offs.

**PROPOSITION 2.3 (PROPERTIES OF PENALTIES AND RELAXATIONS).** (i) Let  $\mathbb{G}^1$  and  $\mathbb{G}^2$  be filtrations satisfying

$\mathbb{F} \subseteq \mathbb{G}^1 \subseteq \mathbb{G}^2$  and let  $z^1$  and  $z^2$  be penalties constructed using Proposition 2.2 with relaxations  $\mathbb{G}^1$  and  $\mathbb{G}^2$  and a common sequence of generating functions  $(w_0, \dots, w_T)$ . Then

$$\sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}^1}} \mathbb{E}[r(\alpha_G) - z^1(\alpha_G)] \leq \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}^2}} \mathbb{E}[r(\alpha_G) - z^2(\alpha_G)]. \quad (12)$$

(ii) For any two dual feasible penalties  $z^1$  and  $z^2$  and information relaxation  $\mathbb{G}$ , we have

$$\begin{aligned} \inf_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[z^2(\alpha_G) - z^1(\alpha_G)] &\leq \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[r(\alpha_G) - z^1(\alpha_G)] \\ &\quad - \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[r(\alpha_G) - z^2(\alpha_G)] \\ &\leq \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[z^2(\alpha_G) - z^1(\alpha_G)]. \end{aligned} \quad (13)$$

(iii) Let  $\mathbb{F}'$  and  $\mathbb{G}$  be filtrations satisfying  $\mathbb{F} \subseteq \mathbb{F}' \subseteq \mathbb{G}$  and let  $(w_0, \dots, w_T)$  be a sequence of generating functions satisfying the conditions of Proposition 2.2. The penalty  $z$  given by  $z_t(a) = \mathbb{E}[w_t(a) | \mathcal{G}_t] - \mathbb{E}[w_t(a) | \mathcal{F}_t']$  and  $z(a) = \sum_{t=0}^T z_t(a)$  satisfies the results of Proposition 2.2.

(iv) Let  $\mathbb{G}$  be a relaxation of  $\mathbb{F}$  and  $z(a) = \sum_{t=0}^T z_t(a)$  be a dual feasible penalty such that  $z_t(a)$  is  $\mathcal{G}_t$ -measurable and depends only on the first  $t+1$  actions of  $a$ . Suppose  $\hat{z}(a) = \sum_{t=0}^T \hat{z}_t(a)$  where each  $\hat{z}_t(a)$  depends only on the first  $t+1$  actions of  $a$ , and further suppose that each  $\hat{z}_t(a)$  is an unbiased estimate of  $z_t(a)$  in that  $\mathbb{E}[\hat{z}_t(a) | \mathcal{G}_t] = z_t(a)$ . Let  $\hat{\mathbb{G}}$  be a relaxation of  $\mathbb{G}$  that assumes that in addition to what is known under  $\mathbb{G}$ , the values of  $\hat{z}_t(a)$  are revealed in period  $t$ . Then

$$\sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[r(\alpha_G) - z(\alpha_G)] \leq \sup_{\alpha_G \in \mathcal{A}_{\hat{\mathbb{G}}}} \mathbb{E}[r(\alpha_G) - \hat{z}(\alpha_G)]. \quad (14)$$

The first result of the proposition says that if we generate penalties with a common set of generating functions, looser relaxations lead to weaker bounds. For example, we may find that the bounds given by using a simple generating function (say,  $w_t = 0$ ) may be “good enough” with one information relaxation, but not “good enough” with a looser relaxation.

The second result of the proposition can be viewed as a continuity property: if the penalties  $z^1$  and  $z^2$  are close in that the difference  $\mathbb{E}[z^2(\alpha_G) - z^1(\alpha_G)]$  is small for all  $\alpha_G$ , then the bounds provided by the two penalties will also be close. For example, if  $z^2$  is the ideal penalty  $z^*$  and therefore yields the optimal upper bound, the bound given by some other penalty  $z^1$  will exceed the optimal bound by no more than  $\sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[z^*(\alpha_G) - z^1(\alpha_G)]$ . In this sense, penalties that are close to the ideal penalty will lead to bounds that are close to optimal.

The third result can be helpful for determining penalties when  $\mathbb{E}[w_t(a) | \mathcal{F}_t]$  is difficult to calculate. For instance in the option-pricing example, if we assume that under the natural filtration  $\mathbb{F}$  volatility is unobserved, we may be able

to simplify the computation of bounds by calculating penalties using a filtration  $\mathbb{F}'$  that assumes that the volatility is observed.

The final result of Proposition 2.3 concerns the effects of errors when penalties are estimated, for example, using nested simulations as in Haugh and Kogan (2004) and Andersen and Broadie (2004). Here we can imagine the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  as including the uncertainties associated with the estimation of penalties as well as the original model uncertainties. These estimation uncertainties are not revealed under filtrations  $\mathbb{F}$  or  $\mathbb{G}$  and do not affect the rewards or penalties and thus are irrelevant to the primal and “true” dual problem. The estimates are, however, revealed under  $\hat{\mathbb{G}}$  and actions are selected to maximize the estimated penalized reward  $r(a) - \hat{z}(a)$  rather than the true penalized reward  $r(a) - z(a)$ . Here we see that when these estimated penalties are unbiased, we obtain estimates of the bounds that are valid but weaker than the bounds given by using the penalty  $z$  itself. Glasserman (2004) provides some numerical results studying the quality of the bounds in an option-pricing example with varying numbers of trials in the nested simulations. His results (and others’) show the importance of estimating penalties accurately. Our results for the option example in §4.7 confirm this finding.

## 2.4. Summary of Approach

Before turning to our examples, it may be worthwhile to summarize the steps involved in our approach. Given a dynamic programming model:

- Identify a heuristic policy that can be used in a simulation study to estimate a lower bound on the optimal value (or upper bound on the optimal cost) for the problem.
- Choose an information relaxation that makes it “easy” to determine optimal decisions given the additional information in the relaxation. It is often natural to start by considering a perfect information relaxation, although in some problems there may be other natural starting points.
- Find a penalty that does not greatly complicate the calculation of optimal decisions with the chosen information relaxation. We can start with zero penalty, but this may lead to weak upper bounds.
- Estimate lower and upper bounds on the optimal value.

In our examples, we will typically estimate the upper and lower bounds simultaneously in a single simulation.

- If the gap between bounds is sufficiently small, we may conclude that the heuristic policy is “good enough” for use in practice, and we are done. If not, we can study the differences between the heuristic policies and the dual policies and see if these suggest some ideas for improving the heuristic policies, relaxations, or penalties.

In the next two sections, we will study two complex examples and discuss issues involved in choosing heuristic policies, information relaxations, and penalties in these applications.

## 3. Example: Adaptive Inventory Control

Our first example is an adaptive inventory control model where demand is nonstationary and partially observed, meaning the probability distribution for demand changes over time and the true demand distribution is not known. These kinds of models are of significant practical interest, but are quite difficult to solve. Treharne and Sox (2002) consider several heuristic policies and evaluate the performance of these policies in a set of five-period examples that they were able to solve exactly. We illustrate our dual bounding approach by evaluating some of these heuristic policies in larger versions of Treharne and Sox’s examples.

### 3.1. The Model

The goal is to find a policy for ordering goods over  $T$  periods ( $t = 0, \dots, T - 1$ ) to minimize the expected total costs. The inventory level at the beginning of period  $t$  is denoted by  $x_t$  and the amount ordered in period  $t$  is  $a_t$ . The demand in period  $t$  is uncertain and denoted by  $d_t$ . The inventory level evolves according to  $x_{t+1} = x_t + a_t - d_t = x_0 + \sum_{\tau=0}^t (a_\tau - d_\tau)$ , where  $x_0$  is the initial inventory level. This evolution equation assumes unmet demand is backordered and appears as a negative inventory level entering the next period; the equation also assumes there is no lead time required to fulfill the orders. The order quantities and demands are assumed to be nonnegative integers.

The period- $t$  demand  $d_t$  is drawn from a distribution  $\delta_t$  that changes stochastically, following a Markov process. The demand  $d_t$  is observed at the end of period  $t$ , but the distribution  $\delta_t$  is never observed. We begin with a prior distribution  $\pi_0$  on the initial demand distribution  $\delta_0$  and update this over time with the period- $(t + 1)$  distribution  $\pi_{t+1}(\pi_t, d_t)$ , taking into account the prior beliefs  $\pi_t$ , the observed demand  $d_t$ , and the possibility of the distribution  $\delta_t$  changing.

In each period, there are ordering costs as well as costs associated with holding inventory or failing to meet demand. The cost of ordering  $a_t$  units is  $c_t a_t$ , where  $c_t$  is the cost of ordering one item or unit. The cost of holding inventory  $x_{t+1}$  from period  $t$  into period  $t + 1$  is  $f_t(x_{t+1}) = h_t \max(0, x_{t+1}) + p_t \max(0, -x_{t+1})$ , where  $h_t$  is the per-unit cost of holding excess inventory in period  $t$  and  $p_t$  is the per-unit penalty associated with backordering unmet demand in period  $t$ . Treharne and Sox assume a terminal cost of  $-c_T x_T$  to capture the value (or cost) of holding inventory (or unmet demand) at the end of the planning horizon. We generalize Treharne and Sox’s model by allowing the ordering costs  $c_t$  to vary following a Markov chain that is independent of the demands  $d_t$  and demand distributions  $\delta_t$ ; we assume that the period- $t$  ordering cost  $c_t$  is known at the beginning of period  $t$ . This generalization will allow us to consider a broader range of information relaxations and makes the problem harder to solve.

Placing this model in the general framework of §2.1, the actions  $a_0, \dots, a_{T-1}$  are the order quantities for each

period and the action sequences  $a$  are drawn from the set  $A$  of  $T$ -vectors of nonnegative integers. An outcome  $\omega$  is a sample path that includes the demands, demand distributions, and ordering costs for each period and a terminal cost  $c_T$ ; that is, the outcomes are of the form  $\omega = ((d_0, \delta_0, c_0), \dots, (d_{T-1}, \delta_{T-1}, c_{T-1}), (c_T))$ . The natural filtration  $\mathbb{F}$  corresponds to knowing the demands  $(d_0, \dots, d_{t-1})$  and costs  $(c_0, \dots, c_t)$  at the beginning of period  $t$ . Because the goal here is to minimize costs, we can either rewrite the primal DP (1) as a minimization problem or else take the rewards in (1) to be the negative costs.

The structure of the adaptive inventory model is perhaps clearer if we view the problem as a partially observable Markov decision process and write it recursively. The period- $t$  state variable is  $(x_t, c_t, \pi_t)$ , where  $x_t$  is the inventory level at the beginning of period  $t$ ,  $c_t$  is the ordering cost in period  $t$ , and  $\pi_t$  is the probability distribution on the period- $t$  demand distribution  $\delta_t$ . In this recursive formulation, it is convenient to take the decision variables to be the order-up-to level  $y_t = x_t + a_t$  rather than the order quantity  $a_t$ . We can then write the period- $t$  cost-to-go function  $J_t$ , for  $t = 0, \dots, T-1$ , as

$$\begin{aligned} J_t(x_t; c_t, \pi_t) \\ = -c_t x_t + \min_{y_t \geq x_t} \{c_t y_t + \mathbb{E}[f_t(y_t - \tilde{d}_t) \\ + J_{t+1}(y_t - \tilde{d}_t; \tilde{c}_{t+1}, \pi_{t+1}(\pi_t, \tilde{d}_t)) \mid c_t, \pi_t]\}. \end{aligned} \quad (15)$$

Here  $\tilde{d}_t$  and  $\tilde{c}_{t+1}$  denote the random period- $t$  demand and period- $(t+1)$  costs, and the terminal cost function is  $J_T(x_T; c_T, \pi_T) = -c_T x_T$ .

What makes this problem difficult to solve is that each demand sequence  $(d_0, \dots, d_{t-1})$  leads to a different  $\pi_t$  and, consequently, the number of scenarios that must be considered grows exponentially in the number of periods considered. For instance, the problems that Treharne and Sox solved to optimality had 5 time periods, 19 possible demand levels, and one ordering cost level. To find an optimal policy, they had to solve the optimization problem (15) for approximately 138,000 different  $(c_t, \pi_t)$ -scenarios. In our numerical examples, we will consider 10 time periods, 19 demand levels, and three cost levels; we would have to solve approximately  $10^{12}$  such optimization problems to find an optimal policy.

### 3.2. Heuristic Policies

Because of the complexity of the primal problem, Treharne and Sox propose using simpler “limited-look-ahead policies” that choose an order quantity that is optimal for a truncated version of the model that looks only zero, one or two periods into the future. For  $t = 0, \dots, T-1$ , the  $L$ -period look-ahead cost-to-go function is defined as

$$\begin{aligned} J_t^L(x_t; c_t, \pi_t) \\ = -c_t x_t + \min_{y_t \geq x_t} \{c_t y_t + \mathbb{E}[f_t(y_t - \tilde{d}_t) \\ + J_{t+1}^{L-1}(y_t - \tilde{d}_t; \tilde{c}_{t+1}, \pi_{t+1}(\pi_t, \tilde{d}_t)) \mid \pi_t, c_t]\}. \end{aligned} \quad (16)$$

In the terminal cases with  $t = T$  or  $L = -1$ , we take  $J_t^L(x_t; c_t, \pi_t) = -c_t x_t$ . When simulating the inventory system using an  $L$ -period look-ahead policy, we determine the order quantity for a particular  $(c_t, \pi_t)$ -scenario by solving (16) for the optimal order-up-to level  $y_t$ . We then draw the random demand  $d_t$  and next period cost  $c_{t+1}$ , calculate the updated probability distribution  $\pi_{t+1}$ , and repeat the process by finding the order quantity for the next period using the  $L$ -period look-ahead value function starting at  $(c_{t+1}, \pi_{t+1})$ .

The complexity of these limited-look-ahead policies grows exponentially with the look-ahead horizon  $L$ . In our numerical examples, we take  $L = 0, 1$ , and  $2$  and we must solve 1, 58, and 1,141 scenario-specific optimization problems (respectively) to determine the recommended order quantity for each period. If we estimate the expected costs of these policies using a simulation with  $T$  periods and  $K$  trials, we must solve  $KT$ ,  $58KT$ , or  $1,141KT$  optimization problems for the 0-, 1-, and 2-period look-ahead policies, respectively.

### 3.3. Information Relaxations

We will study three different information relaxations in this example, each of which allows us to avoid considering the full tree of all possible cost/demand scenarios. First, we will consider the perfect information relaxation. In this case, in the outer simulation we randomly generate the full sequence of ordering costs  $(c_0, \dots, c_T)$ , demand distributions  $(\delta_0, \dots, \delta_{T-1})$ , and actual demands  $(d_0, \dots, d_{T-1})$ . In the inner problem, we determine optimal order quantities by solving a simple deterministic DP. With this relaxation, we will be selecting random samples from the large tree of possible cost/demand scenarios.

Second, we will consider a tighter, imperfect information relaxation that assumes the demand distributions  $(\delta_0, \dots, \delta_{T-1})$  and actual demands  $(d_0, \dots, d_{T-1})$  are known in advance, but assumes the ordering costs  $c_t$  are not known until period  $t$ . In this case, we randomly generate the demand distributions and demands in the outer simulation. In the inner problem, we solve a small stochastic DP that determines cost-dependent order quantities for each period.

The third relaxation is tighter than the first two: it assumes that the actual costs  $c_t$  and demands  $d_t$  are revealed as in the natural filtration (in period  $t$  and period  $t+1$ , respectively), but the demand distribution  $\delta_t$  is known in period  $t$ ; the natural filtration assumes  $\delta_t$  is never observed. In this case, if we assume zero penalty, the dual problem can be formulated as a Markov DP with state variable  $(x_t, c_t, \delta_t)$ ; the number of scenarios that must be considered no longer grows exponentially in  $T$ , and this DP is easy to solve.

### 3.4. Penalties

As discussed in §2.3, the ideal penalty takes the generating function  $w_t$  to be the optimal continuation value, i.e.,

the period- $(t + 1)$  value function  $V_{t+1}$ . Here we will take the generating function  $w_t^L$  for the “ $L$ -period look-ahead penalty” to be the  $L$ -period look-ahead cost-to-go functions defined by Equation (16),

$$w_t = J_{t+1}^{L-1}(y_t - d_t; c_{t+1}, \pi_{t+1}(d_t, \pi_t)). \quad (17)$$

For example, in the myopic case with  $L = 0$ , the generating function is simply  $-c_{t+1}(y_t - d_t)$ . Although we would not expect the  $(L - 1)$ -period look-ahead cost-to-go functions to provide a very good approximation of the actual cost-to-go functions  $J_{t+1}$  (they consider the costs over a small fraction of the total time frame), these limited-look-ahead cost functions may provide a reasonable approximation of the change in costs due to having the additional information provided by relaxation  $\mathbb{G}$  instead of the natural filtration  $\mathbb{F}$ .

In the perfect information relaxation, the full sequence of demands and costs are known in advance and generated in the outer simulation. Let  $(\hat{d}_0^k, \dots, \hat{d}_{T-1}^k)$  and  $(\hat{c}_0^k, \dots, \hat{c}_T^k)$  denote the sequences of these values generated in the  $k$ th trial of the simulation and let  $\hat{\pi}_t^k$  denote the period- $t$  probability distribution on  $\delta_t$  given by starting with the prior distribution  $\pi_0$  and updating based on seeing  $(\hat{d}_0^k, \dots, \hat{d}_{t-1}^k)$ . Following Equation (10), we can write the inner problem in the  $k$ th trial with the  $L$ -period-look-ahead penalty as

$$\begin{aligned} \underline{J}_t^{L,k}(x_t) &= -\hat{c}_t^k x_t + \min_{y_t \geq x_t} \{ \hat{c}_t^k y_t - J_{t+1}^{L-1}(y_t - \hat{d}_t^k; \hat{c}_{t+1}^k, \hat{\pi}_{t+1}^k) \\ &\quad + \underline{J}_{t+1}^{L,k}(y_t - \hat{d}_t^k) + \mathbb{E}[f_t(y_t - \tilde{d}_t)] \\ &\quad + J_{t+1}^{L-1}(y_t - \tilde{d}_t; \tilde{c}_{t+1}, \pi_{t+1}(\hat{\pi}_t^k, \tilde{d}_t)) \mid \hat{\pi}_t^k, \hat{c}_t^k \} \end{aligned} \quad (18)$$

with terminal value  $\underline{J}_T^{L,k}(x_T) = -\hat{c}_T^k x_T$ . Note that the limited-look-ahead cost-to-go function  $J_{t+1}^{L-1}$  and the expectation in (18) would be calculated when determining the limited-look-ahead order quantity for this trial (see Equation (16)). Consequently, when simulating to estimate the expected costs with an  $L$ -period look-ahead policy, it is not difficult to simultaneously estimate the corresponding dual bound: we need only solve one additional scenario-specific optimization problem for each period.

The dual bounds are also easy to calculate with the generating function of Equation (17) for the imperfect information relaxation that assumes that the demands  $(d_0, \dots, d_{T-1})$  and demand distributions  $(\delta_0, \dots, \delta_{T-1})$  are known in advance, but the ordering costs  $c_t$  are revealed over time as in the natural filtration. In this case, the inner problem is a stochastic DP that explicitly considers the uncertainty about the ordering costs; see Appendix A.7 for details. By Proposition 2.2(i), we know that the bounds given by using this imperfect information relaxation will be at least as good as those given by the perfect information relaxation.

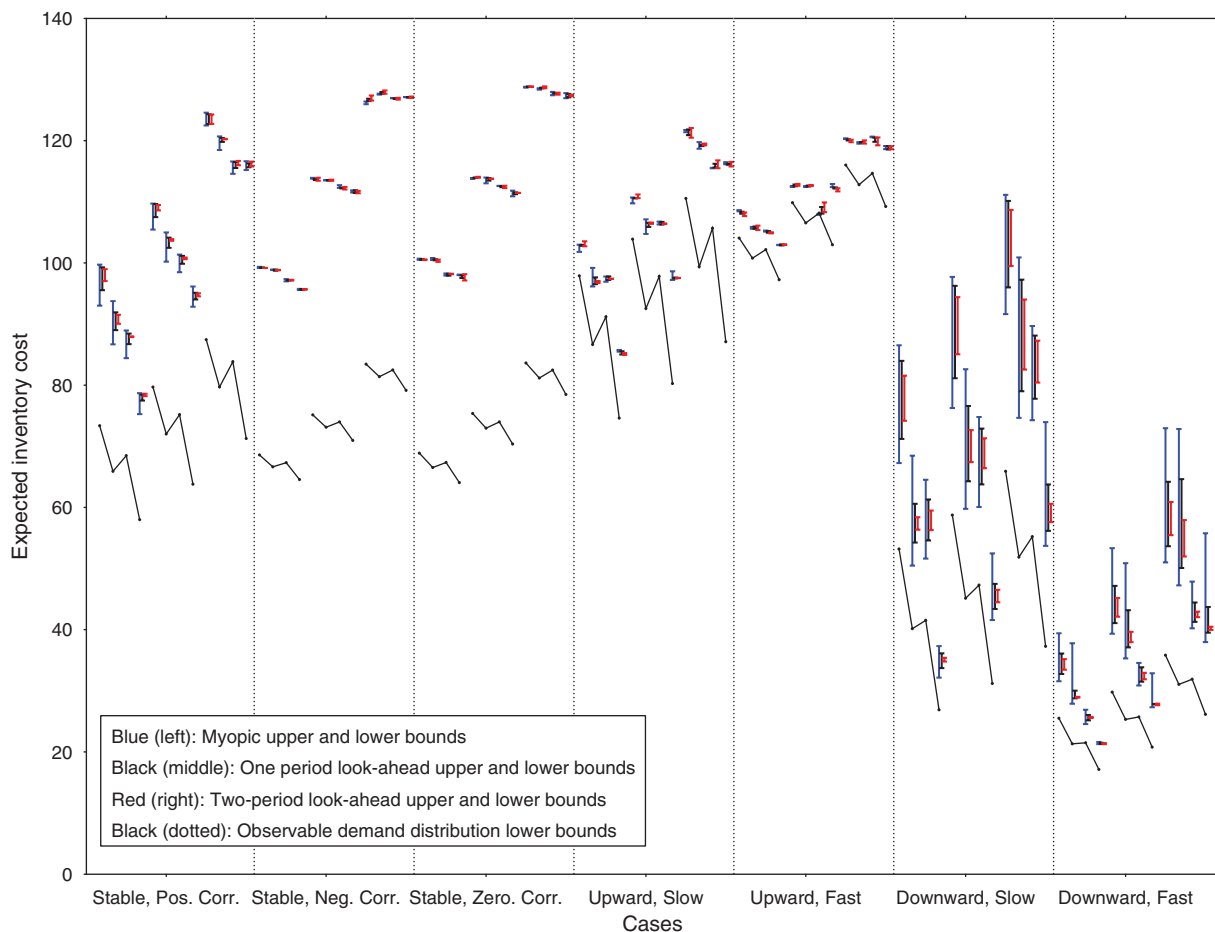
The third information relaxation we consider in this problem assumes the demand distributions  $\delta_t$  are observed in period  $t$ , but  $c_t$  and  $d_t$  are revealed over time as in the natural filtration. As discussed in §3.3, with zero penalty, this dual problem can be formulated as a Markov decision problem that is not difficult to solve. However, with this relaxation, the generating function of Equation (17) leads to an inner problem that is not easy to solve. The difficulty is that the generating functions depend on the probability distributions  $\pi_{t+1}$  that, in turn, depend on the whole history of demands  $(d_0, \dots, d_t)$ . This dependence destroys the Markovian structure that makes it easy to solve the inner problem with no penalty. Thus, the generating function (17) works well with the first two relaxations, but not with the third.

### 3.5. Numerical Results

In this section, we describe numerical results for the adaptive inventory control example. Our choice of parameters closely follows Treharne and Sox (2002). Specifically, following Treharne and Sox, we assume that there are three possible random demand distributions  $\delta_t$ , each of which is a truncated negative binomial distribution that ranges from 0 to 18 units. The three distributions are “low,” “medium,” and “high” and have means and standard deviations of (1, 1.01), (9, 3.01), and (16, 4.01), respectively, before truncation. We consider seven different transition probability matrices representing various trends for the demand distributions. The holding costs  $h_t$  are set to \$1.00 per unit and the backorder costs  $p_t$  are \$1.00, \$1.86, or \$4.00 per unit. Finally, we consider four different priors on the initial demand distribution  $\delta_0$ : the first, third, and fourth represent cases where the demands are most likely to be high, medium, or low, respectively; the second prior is a uniform distribution across the three different demand distributions. In total, there are 84 different combinations of parameters to consider (7 transition matrices  $\times$  3 backorder costs  $\times$  4 priors). In each case, we assume the initial ordering costs  $c_0$  are \$0.60 per unit and later costs take values \$0.00, \$0.60, or \$1.20 following a Markov chain. (These assumptions are described in detail in Appendix C.) Finally, we take the planning horizon  $T$  to be 10 periods and assume zero initial inventory.

In our numerical experiments, we calculate upper and lower bounds on the optimal expected costs using the zero-, one-, or two-step look-ahead policies and penalties. For each combination of model parameters, we estimate the bounds using a simulation of 1,000 trials. Figure 1 summarizes the results for the perfect information relaxation. Appendix C provides the numbers underlying this figure (estimated means and standard errors) as well as results for the imperfect information relaxation described in §3.3.

We call the plot of Figure 1 an “aquarium plot.” In the figure, there are 84 sets of bars, each appearing (if you have bad eyesight!) like a tropical fish. Each “fish” represents the results for a particular set of parameters and

**Figure 1.** Upper and lower bounds with the perfect information relaxation.

consists of three vertical bars with blue, black, and red colors and horizontal markers on each end. The blue bars on the left of each fish represent the myopic (or 0-period look-ahead) upper and lower bounds; the black bars in the middle represent the 1-period look-ahead upper and lower bounds; and the red bars on the right represent the two-period look-ahead upper and lower bounds. The different sets of parameters are grouped first according to the transition matrices (indicated at the bottom), then by backorder costs (with left to right representing high to low costs), and last by the initial priors.

In most cases, the gaps between bounds narrow as we increase the look-ahead horizon, albeit at varying rates. In many cases, the bounds are all quite narrow and the fish look like minnows; in these cases, we could probably assume that the myopic policies are “good enough” and not consider more complex policies.<sup>2</sup> In the cases with a “stable transition matrix, with positive correlation” (on the left of the figure), the fish have relatively wide tails on the left, but narrow quickly: here we may not be satisfied with the quality of the myopic policy, but may find the one- or two-period look-ahead policies to be “good enough.” There are, however, a few cases—with “downward, slow” and “downward, fast” transitions (on the right side of the

figure)—where the gaps remain relatively large even with a two-period look-ahead policy. We will return to these cases in §3.6 below.

Appendix C provides results for the imperfect information relaxation where all demands are assumed to be known in advance, but costs are revealed sequentially over time. The estimated bounds with imperfect information are quite similar to those with perfect information, but the imperfect information bounds are more precisely estimated. Across the 84 cases, the mean standard errors for the dual bounds with the imperfect information relaxation average \$0.216, \$0.172, and \$0.137 for the zero-, one-, and two-period look-ahead bounds, respectively. With the perfect information relaxation, the corresponding mean standard errors for the dual bounds average \$0.821, \$0.554, and \$0.374. Intuitively, the improved precision in the imperfect information bounds comes from eliminating random sampling variations associated with costs by explicitly enumerating the cost scenarios. In the imperfect information case, we also enumerate the cost scenarios when estimating the expected cost of the heuristic policy; this is somewhat more time consuming (for a fixed number of samples) but improves the precision of the estimated bounds.

**Table 1.** Computation times (seconds) for calculating bounds in the inventory example.

Look-ahead horizon ( $L$ )	Perfect information relaxation		Imperfect information relaxation	
	Heuristic policy	Dual bound	Heuristic policy	Dual bound
Zero periods	7.2	+0.2	9.3	+1.3
One period	46.1	+1.2	47.8	+3.7
Two periods	663	+1.2	667	+3.6

The run times required to calculate these bounds are shown in Table 1. We show the time required to evaluate the zero-, one-, or two-period look-ahead heuristic policies using 1,000 trials for one set of model parameters and the additional time required to calculate the dual bounds with these same 1,000 trials.<sup>3</sup> Here we see that once we have calculated the bounds associated with the heuristic policies (and the associated look-ahead value functions), it takes little additional time to compute the dual bounds. The myopic dual bounds are somewhat faster to compute than the one- and two-period look-ahead bounds because in the myopic case we know the objective function in Equation (18) is convex and can simplify the optimization problem. The imperfect information bounds take somewhat longer to compute than the perfect information bounds, because we must solve for dual optimal actions in each of the three possible cost states in each period rather than the one randomly chosen cost state that is considered in the perfect information case.

As discussed in Section §3.3, we can construct an alternative lower bound on expected costs by considering an information relaxation where the demands  $d_t$  and costs  $c_t$  are revealed over time according to the natural filtration but the demand distributions  $\delta_t$  are observed in period- $t$  (rather than never observed, as assumed in the natural filtration). If we take the penalty to be zero, this problem can be formulated as a Markov DP that takes approximately 0.08 seconds to solve. These “observable demand distribution” bounds are shown as connected dotted lines in Figure 1. These lines are well below the “fish” representing the limited-look-ahead bounds. Thus, in these examples, observing the demand distribution is quite valuable and, with no penalty, the corresponding bounds are quite weak.

### 3.6. Improving the Heuristic Policies and Bounds

We now consider the use of the dual results to identify better policies and bounds when the gaps are relatively large. We will focus on the cases with the “downward, slow” and “downward, fast” transition matrices. In these cases, demand may initially be high (with mean 16), but it may drop to medium (with mean 9) or low (with mean 1) this period, and when demand drops, it will not increase again. Comparing the order-up-to quantities (the  $y_t$ s) selected by

the myopic policy with those selected in the corresponding dual bound, we find that the dual problem takes advantage of the perfect information to reduce the order in the period when demand drops to the low demand state, thereby avoiding the cost of carrying excess inventory when the system enters the low state. It appears that the myopic policies order too much when the system is not in the low demand state and the dual penalties do not appropriately “punish” the DM in the dual problem for taking advantage of the perfect information about demand.

To understand why this is the case, note that the terminal value used in determining myopic policies and used as the generating function for the myopic dual bound,  $J_t^{-1}(x_t; c_t, \pi_t) = -c_t x_t$ , implicitly assumes that left-over inventory substitutes for future purchases. One way to perhaps improve the policies and bounds is to use the terminal values based on a model that assumes the demand distributions is observable. Specifically, we take the limited-look-ahead terminal value  $J_t^{-1}(x_t; c_t, \pi_t)$  to be  $\mathbb{E}[J_t^o(x_t, c_t, \tilde{\delta}_t) | \pi_t]$ , where  $J_t^o$  is the value function for a Markov DP that assumes the demand distribution  $\delta_t$  is observed in each period; this model was used to calculate the “observable demand distribution” bounds described in §3.5. As is evident in Figure 1, these observable demand value functions are not very good approximations of the true value functions (they greatly underestimate costs), but they are easy to compute and, unlike the original terminal values, they include the holding costs associated with having excess inventory in a low demand state.

This modification leads to dramatic improvements for the cases with the “downward, slow” and “downward, fast” transition matrices, with little additional work. For example, in the case with the “downward, slow” transition matrix, high backorder costs, and a high prior distribution, the myopic bounds with the modified terminal values were \$107.0 and \$107.5 as compared to \$92 and \$111 for the myopic bounds with the original terminal values; the run times were 7.7 and 7.4 seconds, respectively. (These results are for the perfect information relaxation and a simulation of 1,000 trials.) The myopic bounds for the other cases with “downward, slow” and “downward, fast” transition matrices are also much improved. In these cases, these modified myopic policies not only outperform the original myopic policies, they also outperform the significantly more complex one- and two-period look-ahead policies based on the original terminal values. (See Appendix C for detailed results for all cases.)

Although this modification of the myopic policies greatly improves the results for the cases with the “downward, slow” and “downward, fast” transition matrices, the modified myopic policies perform worse than the original myopic policies in some other cases, where the original myopic policies performed quite well. In all, comparing across the 84 different sets of parameters, we find that we can get within 2% of the optimal costs (and typically closer) using one of these two myopic policies. Thus, by

experimenting with the heuristic policies and comparing the expected costs and policies to the corresponding dual bounds and policies, we have identified simple heuristic policies that perform well in each of the 84 cases considered. Moreover, we *know* that we cannot do much better with more complex policies.

#### 4. Example: Option Pricing with Stochastic Volatilities and Interest Rates

An American call (put) option gives its owner the right to buy (sell) a stock at a specified strike price at any time before the option expires. To value an American option, we must use dynamic programming methods to determine an optimal policy for exercising the option. The original Black-Scholes-Merton model for valuing options on stocks assumes that the volatility of the stock price and the (risk-free) interest rate are both constant over time. In this setting, we can value an American option by solving a one-dimensional DP, typically represented as a binomial or trinomial lattice. In this example, we will consider the problem of valuing American options on a dividend-paying stock with stochastic volatilities and interest rates.

##### 4.1. The Model

We will consider fairly standard models of stock prices, volatilities, and interest rates and will not exploit any nonstandard properties of these models in our analysis. Specifically, our model of stock prices and volatilities follows Heston (1993) and our model of interest rates follows Medvedev and Scaillet's (2007) extension of Heston's model. With no dividends, the stock price  $s_\tau$  at time  $\tau$  has drift equal to the risk-free interest rate  $\gamma_\tau$  and instantaneous variance  $v_\tau$  ( $v_\tau$  is the square of the volatility);  $s_\tau$ ,  $\gamma_\tau$ , and  $v_\tau$  evolve according to the joint stochastic process

$$\begin{aligned} ds_\tau &= \gamma_\tau s_\tau d\tau + \sqrt{v_\tau} s_\tau dz_\tau^s \\ d\gamma_\tau &= -\kappa_\gamma (\gamma_\tau - \bar{\gamma}) d\tau + \sigma_\gamma \sqrt{\gamma_\tau} dz_\tau^\gamma, \\ dv_\tau &= -\kappa_v (v_\tau - \bar{v}) d\tau + \sigma_v \sqrt{v_\tau} dz_\tau^v, \end{aligned} \quad (19)$$

where  $\bar{\gamma}$  and  $\bar{v}$  are long-run average levels for  $\gamma_\tau$  and  $v_\tau$  (respectively);  $\kappa_\gamma$  and  $\kappa_v$  are the corresponding mean-reversion rates; and  $\sigma_\gamma$  and  $\sigma_v$  are the corresponding instantaneous volatilities. We will assume that the stochastic increments  $dz_\tau^s$  and  $dz_\tau^v$  for stock prices and volatilities have correlation  $\rho_{sv}$ . Following Medvedev and Scaillet, we will assume that  $dz_\tau^\gamma$  is uncorrelated with the other two factors; this simplifies our discussion somewhat but is not necessary for our approach. If the stock pays a dividend  $d_\tau$  at time  $\tau$ , the stock price drops instantaneously from  $s_{\tau-}$  (the price just before the dividend) to  $s_\tau = s_{\tau-} - d_\tau$ .

Following standard practice in option valuation, we assume that the stochastic differential Equations (19) are “risk-neutral” processes and the model parameters include

any required risk premiums. With this assumption, the value of any security whose value depends on  $s_\tau$  is given as the expected present value of its future payoffs, where expectations are calculated using the risk-neutral processes and payoffs are discounted at the risk-free rate  $\gamma_\tau$ . This implies that between dividends the discounted stock price follows a martingale, i.e., the expected present value of the stock at time  $\tau_2$ , discounted back to  $\tau_1$  values at the prevailing interest rate, is equal to the current price  $s_{\tau_1}$ .

This martingale property implies that there is no benefit to exercising a call option before expiration, except possibly immediately before a dividend is paid. Following Proposition 2.1, we will impose a constraint that enforces this “no early exercise” property when calculating bounds for call options. This “no early exercise” property does not hold for put options, and we must consider all possible exercise dates.

To place this problem in our discrete-time framework of §2, we will consider a discrete-time approximation of the diffusions (19) where the time until expiration is divided into  $T$  steps of length  $\delta$ . The outcomes  $\omega$  are sample paths that specify the stock price, volatility, and interest rates at each step:  $\omega = ((s_0, v_0, \gamma_0), \dots, (s_T, v_T, \gamma_T))$ . The actions  $a_t$  are to exercise the option or not (i.e.,  $a_t \in \{\text{exercise}, \text{do not exercise}\}$ ); the action space  $A$  includes the constraint that the option can be exercised at most once. Let  $\phi_t = \exp(-\delta(\sum_{i=0}^{t-1} \gamma_i))$  be a discount factor that converts the period- $t$  option payoff back to period-0 values using the time-varying risk-free interest rates  $\gamma_i$ . The period- $t$  reward for an option with strike price  $K$  is then  $\phi_t(s_t - K)$  for a call option ( $-\phi_t(s_t - K)$  for a put) if  $a_t = \text{exercise}$  and 0 otherwise.

The complexity of the primal DP depends on how we define the natural filtration  $\mathbb{F}$ . The simplest formulation is to assume that the stock price, volatility, and interest rates  $(s_t, v_t, \gamma_t)$  are observed in period  $t$  and that  $\mathcal{F}_t$  reflects knowledge of these processes up to time  $t$ . In this case, the primal DP can be formulated as a Markov decision process with three continuous state variables. In principle, this could be approximated using a three-dimensional grid to represent the state space. If we want good coverage of the state space, these grids may be quite large. For example, if we were to use a grid with 50 points for each dimension, we would need a total of  $50^3 = 125,000$  elements to represent the state space, and the probability transition matrix would have a total of  $(125,000)^2 \approx 1.6 \times 10^{10}$  elements. If we were to consider multifactor models of interest rates and/or volatilities, these DPs would be even more complex.

Alternatively and perhaps more realistically, we might consider a natural filtration  $\mathbb{F}$  that assumes that stock prices and interest rates are observed in each period, but recognizes the fact that the volatilities are never observed. The DP with this information structure could be formally modeled as a partially observed Markov decision problem with the state variable being  $(s_t, \gamma_t, \pi_t)$  where  $\pi_t$  is a probability distribution on  $v_t$ . Given the high dimensionality of  $\pi_t$ ,

the corresponding primal DP would be very difficult to formulate and solve.

#### 4.2. A Heuristic Policy

We can calculate a lower bound on the value of an option by simulating the option payoffs using any given exercise policy. We will generate lower bounds using an exercise policy that is optimal for a simplified model with constant volatilities and interest rates, set at their long-run means  $\bar{v}$  and  $\bar{\gamma}$ . This simplified option problem can be formulated as a DP with a one-dimensional state space and solved using standard lattice techniques. Our dual bounds will help us determine whether this simple exercise policy is “good enough” to value options in the more complex setting with stochastic volatilities and interest rates.

#### 4.3. Information Relaxations

Our primary focus will be on an imperfect information relaxation that assumes that the volatilities  $(v_0, \dots, v_T)$  and interest rates  $(\gamma_0, \dots, \gamma_T)$  are known in advance, but the stock price  $s_t$  is not known until period  $t$ . Thus, in the outer simulation we generate volatilities and interest rates and the inner problem is a one-dimensional option pricing problem that can be solved using a simple lattice. Although the stock prices remain uncertain, the information about volatilities and interest rates may be valuable. For example, if the volatilities are correlated with stock price movements (i.e.,  $\rho_{sv} \neq 0$ ), advance knowledge of the volatilities provides some information about future stock price movements. Even without such correlation, the volatilities affect the probability that an option will be “in the money,” and hence may have some bearing on the option values and exercise decisions.

We will also consider a perfect information relaxation that assumes that the volatilities  $(v_0, \dots, v_T)$ , interest rates  $(\gamma_0, \dots, \gamma_T)$ , and stock prices  $(s_0, \dots, s_T)$  are known in advance and generated in the outer simulation. In this case, the inner problem is a simple deterministic maximization problem where we choose the optimal exercise date or decline to exercise, with full knowledge of the penalized reward for exercising at each time. Note that these relaxations both assume the volatilities are known in advance and thus provide valid bounds whether we assume that the volatility is truly observed in period  $t$  or not.

#### 4.4. Penalties

We will focus on a simple penalty that approximately cancels the benefit of the information about stock prices provided by the information relaxation. As in the inventory example, our penalty will be derived from the model that is used to determine the heuristic policy. Here, the lower bound is given by simulating the complex model with stochastic volatilities and interest rates using a policy that is optimal for a simplified model with constant volatility and constant discount rates. In the simplified model, the

“delta” for the option,  $\Delta_t(s_t)$ , describes the sensitivity of the period- $t$  value of option to changes in the stock price  $s_t$  in period  $t$ . These deltas are straightforward to compute in the lattice used to determine these heuristic policies, and because  $\Delta_t(s_t)$  does not depend on the actual volatilities or interest rates, these deltas need only be calculated once when simulating to estimate the bounds.

We will use these deltas to approximate the impact of changing price expectations in our more complex model. Specifically, when the DM chooses to “wait” or hold the option (i.e., when  $a_t = \text{do not exercise}$  and  $t < T$ ), we take the generating function of Proposition 2.2 for period  $t$  to be

$$w_t^{\text{wait}} = \phi_t \Delta_t(s_t) (e^{-\gamma_t \delta} s_{t+1}) \quad (20)$$

where  $\phi_t$  is the previously defined discount factor that converts period- $t$  values to period-0 values. The period- $t$  penalty when the option is not exercised is then

$$\begin{aligned} z_t^{\text{wait}} &= \mathbb{E}[w_t^{\text{wait}} | \mathcal{G}_t] - \mathbb{E}[w_t^{\text{wait}} | \mathcal{F}_t] \\ &= \phi_t \Delta_t(s_t) (e^{-\gamma_t \delta} \mathbb{E}[\tilde{s}_{t+1} | \mathcal{G}_t] - e^{-\gamma_t \delta} \mathbb{E}[\tilde{s}_{t+1} | \mathcal{F}_t]) \\ &= \phi_t \Delta_t(s_t) (e^{-\gamma_t \delta} \mathbb{E}[\tilde{s}_{t+1} | \mathcal{G}_t] - s_t). \end{aligned} \quad (21)$$

The first two equalities follow from the definitions of  $z_t^{\text{wait}}$  and  $w_t^{\text{wait}}$  (respectively). In the third equality, we have used the martingale property for the stock prices ( $s_t = e^{-\gamma_t \delta} \mathbb{E}[\tilde{s}_{t+1} | \mathcal{F}_t]$ ). This penalty  $z_t^{\text{wait}}$  can be viewed as a crude first-order approximation of the change in the value of the option due to the extra information provided by  $\mathbb{G}$  in period  $t$ :  $\phi_t \Delta_t(s_t)$  approximates the sensitivity of the value of the option to changes in the period- $t$  stock price (discounted to present value terms) and  $(e^{-\gamma_t \delta} \mathbb{E}[\tilde{s}_{t+1} | \mathcal{G}_t] - s_t)$  represents the change in expected stock price in that period. When the option is exercised or allowed to expire without exercise, we take the generating function to be zero and the resulting penalty is also zero. Note that these penalties do not depend on how we define the natural filtration  $\mathbb{F}$ , as long as stock prices follow a martingale under the natural filtration. Thus, Equation (21) holds whether we assume that volatility is observed or not.

The penalties depend on the information relaxation  $\mathbb{G}$  through the  $\mathbb{E}[\tilde{s}_{t+1} | \mathcal{G}_t]$  term appearing in (21). If  $\mathbb{G}$  is the perfect information relaxation, then the stock price itself is known and  $\mathbb{E}[\tilde{s}_{t+1} | \mathcal{G}_t] = s_{t+1}$ . If  $\mathbb{G}$  assumes the volatilities  $(v_0, \dots, v_T)$  and interest rates  $(\gamma_0, \dots, \gamma_T)$  are known in advance but the stock price  $s_t$  is not observed until period  $t$ , then following the derivation in Appendix A.8, we can write

$$\mathbb{E}[\tilde{s}_{t+1} | \mathcal{G}_t] = e^{\gamma_t \delta} s_t \exp(\rho_{sv} \sqrt{v_t} (v_{t+1} - v_t) - \frac{1}{2} \rho_{sv}^2 v_t \delta). \quad (22)$$

Note that if  $\rho_{sv} = 0$ , then this expression simplifies and we have  $e^{-\gamma_t \delta} \mathbb{E}[\tilde{s}_{t+1} | \mathcal{G}_t] = s_t$ . Thus, with no correlation between stock prices and volatilities, the penalty given by (21) is identically zero.

With this form of penalty, we can solve the resulting inner problems efficiently using a recursive DP formulation. The terminal value of a call option is given by  $\bar{v}_{T+1} = 0$  and

$$\begin{aligned}\bar{v}_t(s_t) &= \max\{\phi_t(s_t - K), -z_t^{\text{wait}} + \mathbb{E}[\bar{v}_{t+1}(\tilde{s}_{t+1} | \mathcal{G}_t)]\} \\ &= \max\{\phi_t(s_t - K), \phi_t \Delta_t(s_t) s_t \\ &\quad + \mathbb{E}[-\phi_t \Delta_t(s_t) \tilde{s}_{t+1} + \bar{v}_{t+1}(\tilde{s}_{t+1}) | \mathcal{G}_t]\}. \quad (23)\end{aligned}$$

With the imperfect information relaxation, we will calculate bounds on the option value by randomly generating volatility and interest rates in the outer simulation and using a trinomial lattice to evaluate the recursion (23) in each scenario.

With the perfect information relaxation, stock prices are also generated in the outer simulation and we can rewrite the DP (23) for the inner problem as

$$\bar{v}_0 = \max\left\{\max_{t \in \{0, \dots, T\}} \{\phi_t(s_t - K) - \mu_t\}, -\mu_T\right\}, \quad (24)$$

where  $\mu_0 = 0$  and  $\mu_t = \sum_{i=0}^{t-1} z_i^{\text{wait}}$  for  $t > 0$ . Here,  $\mu_t$  is the accumulated penalty for waiting: if the DM exercises in period  $t$ , she “pays” the accumulated penalty for all of the periods she waited. Note that whenever we consider a generating function  $w_t$  that is zero when the option is exercised and equal to  $w_t^{\text{wait}}$  when the DM waits (as we have assumed here), then  $\mu_t = \sum_{i=0}^{t-1} z_i^{\text{wait}} = \sum_{i=0}^{t-1} (w_i^{\text{wait}} - \mathbb{E}[w_i^{\text{wait}} | \mathcal{F}_i])$  will be a martingale. This form of penalty can thus be interpreted as a “dual martingale,” as considered by Haugh and Kogan (2004), Rogers (2002), and Andersen and Broadie (2004). There is, however, a subtle difference in our formulations of the dual optimization problem that we discuss in Appendix A.9.

#### 4.5. Numerical Results

We will present numerical results for both put and call options, for a variety of model parameters. In all cases, we assume the options expire in one year ( $\tau_T = 1$ ), the initial stock price  $s_0$  is \$100, and the stock pays dividends equal to 1% of the stock price at times  $\tau = 0.25, 0.50, 0.75$ , and 1.00. The rest of our numerical assumptions are based on Medvedev and Scaillet (2007). The instantaneous variance process  $v_t$  is assumed to have mean-reversion rate  $\kappa_v = 1.58$ , volatility  $\sigma_v = 20\%$ , and initial value and long-run average  $v_0 = \bar{v} = 0.04$ . The interest rate process  $\gamma_t$  has  $\kappa_\gamma = .26$ ,  $\sigma_\gamma = 8\%$ , and  $\bar{\gamma} = \gamma_0 = 4\%$ . For both put and call options, we consider strike prices ( $K$ ) of \$90, \$100, and \$110 and correlations  $\rho_{sv}$  of  $-0.25, 0$ , and  $0.25$ .

In all cases, we will calculate bounds in a simulation with 5,000 trials. For the imperfect information bounds, we generate interest rates and volatilities in the outer simulation, and in each trial we solve the inner problem (23) using a trinomial lattice with 101 stages. In this lattice, the stock prices are fixed, but the probabilities are chosen to

match the mean and variance of the stock price process in each period, given the volatility and interest rate information. For the put options, we allow exercise at each stage in the lattice. For the call options, as discussed in §4.1, we limit exercise to expiration and the periods just before a dividend is paid. To ensure consistency across the different bounds, we use this same lattice to value the options using the heuristic policy and to calculate the perfect information bounds. To calculate the perfect information bounds, we randomly select a stock price path from the lattice in each outer simulation trial.

The results are summarized in Table 2; Table 3 shows the time required to compute these bounds for one set of model parameters. In Table 2 we see that, as expected based on Proposition 2.3(i), the imperfect information bounds are tighter than the corresponding perfect information bounds with penalties constructed using the same generating function. Of course, the imperfect information bounds take somewhat longer to compute. Including the penalties improves the performance of all of the bounds, except with the imperfect information relaxation when the correlation  $\rho_{sv}$  is zero: as discussed in §4.4, the penalties are identically zero in this case. The bounds with imperfect information and penalty are quite tight: the gap between the lower bound and this upper bound ranges from 0.2% to 0.8% of the value of the option. These gaps may be sufficiently small to conclude that the simple heuristic policy is good enough to use in practice. In addition, examining the mean standard errors (MSE) in Table 2, we see that the perfect information bounds are fairly precisely estimated.

#### 4.6. Improving the Heuristic Policy

We can also use the results of the dual problem to improve the heuristic policy in this example. Specifically, let us focus on the case with the largest duality gap, the put option with strike price of \$110 and a correlation  $\rho_{sv}$  of  $-0.25$ : the lower and upper bounds in this case are approximately \$13.42 and \$13.52. To see how the exercise policy might change with changing volatilities and interest rates, we fit a nonlinear regression model of the form  $E_t = \min(a_t, b_t \exp(\gamma_t \delta) + c_t) + d_t v_t$ , where  $E_t$  is the exercise threshold from the dual problem in period  $t$ :  $\gamma_t$  and  $v_t$  are the interest rate and instantaneous variance for that period, and  $a_t, b_t, c_t$ , and  $d_t$  are constants estimated in the regression. This choice of functional form was selected based on the graphical appearance of plots of  $E_t$  versus  $\gamma_t$  and  $v_t$ . If we assume that the volatility  $v_t$  is observed in period  $t$ , this regression model describes a nonanticipative exercise policy: exercise the option whenever  $s_t$  is less than the exercise threshold given by the regression equation. Simulating with this new policy, we found that the put option value increased by about \$0.06, reducing the duality gap by more than half. We could probe further and attempt to find a penalty that gives a tighter upper bound, but in this case the gap seems sufficiently small to suggest that further improvements would not be of much practical value.

**Table 2.** Bounds on option values.

Correlation ( $\rho_{sv}$ )	Put options						Call options					
	$\rho = -0.25$		$\rho = 0$		$\rho = +0.25$		$\rho = -0.25$		$\rho = 0$		$\rho = +0.25$	
	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Strike price = \$90												
Perfect info:												
No penalty:	5.899	0.117	5.508	0.109	5.140	0.100	19.596	0.202	19.755	0.216	20.037	0.227
With penalty:	3.422	0.018	3.331	0.019	3.221	0.018	14.042	0.018	13.963	0.018	13.865	0.018
Lower bound:	3.344	0.087	3.245	0.084	3.057	0.078	13.932	0.183	13.890	0.185	13.949	0.193
Imperfect info:												
No penalty:	3.764	0.034	3.270	0.015	3.380	0.005	14.807	0.028	13.899	0.015	14.957	0.060
With penalty:	3.383	0.014	3.270	0.015	3.163	0.015	14.004	0.014	13.899	0.015	13.803	0.015
Lower bound:	3.355	0.030	3.245	0.015	3.137	0.005	13.933	0.027	13.830	0.015	13.731	0.054
Strike price = \$100												
Perfect info:												
No penalty:	13.024	0.150	12.632	0.142	12.313	0.133	11.380	0.180	11.574	0.194	11.885	0.206
With penalty:	7.503	0.021	7.537	0.022	7.534	0.021	8.159	0.021	8.198	0.022	8.201	0.022
Lower bound:	7.391	0.128	7.376	0.125	7.300	0.121	8.056	0.160	8.101	0.166	8.144	0.173
Imperfect info:												
No penalty:	8.277	0.047	7.423	0.019	7.940	0.010	8.514	0.017	8.140	0.019	8.752	0.055
With penalty:	7.413	0.017	7.423	0.019	7.445	0.018	8.124	0.018	8.140	0.019	8.160	0.018
Lower bound:	7.354	0.044	7.370	0.018	7.391	0.011	8.088	0.018	8.107	0.019	8.121	0.052
Strike price = \$110												
Perfect info:												
No penalty:	22.839	0.148	22.444	0.141	22.121	0.132	5.689	0.140	5.975	0.155	6.314	0.168
With penalty:	13.676	0.020	13.806	0.021	13.902	0.020	4.311	0.021	4.455	0.022	4.568	0.022
Lower bound:	13.413	0.162	13.600	0.158	13.675	0.156	4.159	0.121	4.331	0.132	4.506	0.140
Imperfect info:												
No penalty:	15.143	0.052	13.655	0.017	14.673	0.019	4.430	0.009	4.425	0.018	4.824	0.044
With penalty:	13.520	0.016	13.655	0.017	13.796	0.017	4.274	0.017	4.425	0.018	4.567	0.017
Lower bound:	13.417	0.051	13.562	0.017	13.707	0.020	4.256	0.010	4.412	0.018	4.546	0.042

#### 4.7. Haugh-Kogan and Andersen-Broadie Style Penalties

We also experimented with Haugh-Kogan and Andersen-Broadie style bounds, based on the heuristic policy we use to calculate lower bounds. Recall from §2.3 that the Haugh-Kogan approach takes the generating function to be an approximate value function; here we use the value function from the simplified model with constant volatility and constant discount rates. In the Andersen-Broadie approach, we work directly with the heuristic policy. With both the Haugh-Kogan and Andersen-Broadie style bounds, we will consider the perfect information relaxation and use nested simulations to estimate the conditional expectations  $\mathbb{E}[w_t | \mathcal{F}_t]$  required to determine the penalties. In our nested simulations, we assume that the volatilities  $v_t$  are observed

in  $\mathcal{F}_t$ . Using Proposition 2.3(iii), this leads to a valid upper bound regardless of whether we assume volatilities are truly observed in the natural filtration.

If we let  $K$  be the number of trials in the outer simulation,  $N$  be the number of trials in each nested simulation, and  $T$  be the number of periods in the model, the Haugh-Kogan approach will require computational effort on the order of  $KNT$ ; each period requires a one-step nested simulation for each trial in the outer simulation. The nested simulations in the Andersen-Broadie approach continue until the option is exercised or expires and the approach therefore requires computational effort proportional to  $KNT^2$ . In our model we consider  $T = 100$  periods, so we should expect the Andersen-Broadie approach to be quite time consuming. Indeed, with  $K = 5,000$  trials in the outer simulation and  $N = 25$  trials in each nested simulation, it took approximately 20,000 seconds (5.6 hours) to estimate the Andersen-Broadie bounds. It took approximately 270 seconds to estimate the Haugh-Kogan bounds with the same number of trials. With  $N = 1,000$  trials in the inner simulation, these bounds take approximately 40 times longer to compute.

Table 4 shows the estimated Haugh-Kogan and Andersen-Broadie bounds for the case of a put option with strike price of \$110 with correlation  $\rho_{sv}$  equal to  $-0.25$ ,

**Table 3.** Run times for option bounds.

	Perfect information		Imperfect information	
	Puts	Calls	Puts	Calls
Lower bound:	7.2	4.8	21.1	12.1
Upper bound with no penalty:	+0.20	+0.17	+13.5	+11.0
Upper bound with penalty:	+0.24	+0.20	+20.8	+18.5

**Table 4.** Example Haugh-Kogan and Andersen-Broadie bounds.

	Trials in nested simulation ( $N$ )							
	10		25		100		1,000	
	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Haugh-Kogan bounds:	15.43	0.02	14.55	0.02	13.93	0.02	13.65	0.02
Andersen-Broadie bounds:	17.74	0.13	15.65	0.04	14.11	0.04	13.58	0.01

with varying number of trials ( $N$ ) in the nested simulation. All simulations involved  $K = 5,000$  trials in the outer simulation, except the  $N = 1,000$  case for the Andersen-Broadie bounds. The lower bound given by using the heuristic policy is approximately \$13.42 in this case (or \$13.48 with the improved policy discussed in §4.6). In these results, we see that sampling error in the nested simulations adversely affects the quality of the bounds, consistent with Proposition 2.3(iv). Although the Haugh-Kogan and Andersen-Broadie bounds with 1,000 trials in the nested simulation (\$13.65 and \$13.58, respectively) are better than the perfect information bound with our simple delta penalty (\$13.67), the Haugh-Kogan and Andersen-Broadie bounds were very time consuming to compute with this many nested trials. However, even with 1,000 nested trials, the Haugh-Kogan and Andersen-Broadie bounds were not as tight as the imperfect information bound with our simple delta penalty (\$13.52). Thus, in this example the imperfect information relaxation provides tighter bounds with this relatively easy-to-compute penalty.

## 5. Conclusions

We believe that the dual approach developed in this paper provides a powerful, general, and flexible approach for calculating upper bounds in DPs. In applications, the researcher can control the computational effort and the quality of the bound by choosing the penalties, information relaxations, and/or the number of simulations run. As discussed in the introduction, we see this dual approach as complementing approximate dynamic programming and the use of simulation methods with heuristic policies: given some candidate policy, we can use simulation to determine the value with this policy and use our dual approach to generate an upper bound on the value of an optimal policy. The gap between the lower and upper bounds gives an indication of how much better we could do with a more complex policy. In practice, we may find that we can identify policies that are “good enough” with relatively little work. We demonstrated these dual bounds in two complex applications (adaptive inventory control and option pricing) that are of significant practical interest, and the results appear to be promising. Lai et al. (2010), recently applied this dual

approach to evaluate heuristic policies used to manage a natural gas storage facility.

There is certainly an element of art in selecting penalties and information relaxations, just as there is art in selecting good heuristic policies and in selecting approximate value functions in approximate dynamic programming. The choice of information relaxation is particularly important because it determines which uncertainties are treated as stochastic in the inner problem and which are treated as deterministic. For instance, in the option-pricing example, it is straightforward to model stock price uncertainty with known but time-varying volatility and interest rates, but difficult to treat volatility, interest rates, and stock prices all as stochastic. In the adaptive inventory control problem, it is hard to consider the full tree of possible demand histories, but relatively easy to sample from this large tree. As discussed in §2.3, the key is to select a penalty that approximately cancels the benefit provided by the additional information. However, we must be mindful of the computational effort required to compute these penalties. In our examples, we considered simple penalties that were derived from the heuristic policy and were easy to compute.

There are a number of directions for possible future research on these dual methods. First, it would be interesting to consider continuous-time and/or infinite-horizon models as well as the discrete-time, finite-horizon DP models considered here. Second, we would like to study ways to optimize the dual bound through the use of a parameterized family of penalty functions. For example, we might allow the penalty function to be a weighted combination of candidate penalties and then optimize the weights when estimating the bound. More ambitiously, we might attempt to develop automatic methods for generating feasible (i.e., nonanticipative) policies or improvements on given policies using the results from the dual optimization problems. Finally, and perhaps most importantly, we need to build more experience through applying these techniques in examples. In so doing, we can perhaps develop a better understanding of what kinds of penalties and relaxations work well for what kinds of problems.

## 6. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at <http://or.journal.informs.org/>.

## Endnotes

1. Note that we assume that the set of possible action sequences does not depend on the outcome  $\omega$  and also that the probabilities associated with the outcomes do not depend on the selected actions. We could allow the set of possible action sequences to depend on  $\omega$  by restricting the set of policies to some subset of  $\mathcal{A}$ . The general formulation of the DP (1) would be unchanged, but the recursive formulation (2) would need to be modified to allow the

actions available in each period to depend (in a measurable way) on the outcome  $\omega$ . Problems with action-dependent probabilities can often be recast as equivalent problems with action-independent probabilities, sometimes quite naturally. For example, we could think of the inventory example of §3 as having random transitions from one inventory level to the next inventory level; the transition probabilities would then depend on the actions (the order quantities). Alternatively, we can formulate this problem (as we will) with demand as uncertain and independent of the actions. For a general problem, one could take the outcome  $\omega$  to be a series  $(U_0, \dots, U_T)$  of uniform random numbers where  $U_t$  revealed in period  $t$ ; we could then calculate the period- $t$  state from these random deviates and the chosen actions  $(a_0, \dots, a_{t-1})$ . There are a variety of ways one can reformulate a model to have action-independent probabilities. In applications we would want to exploit the specific structure of the problem under consideration.

2. Note that it is possible for the estimated duality gap to be negative: although our penalties have zero expected penalty for nonanticipative policies, there is no guarantee that these penalties will average exactly zero in a particular sample. When we saw negative gaps in our study, the estimated gaps were small compared to the associated standard errors.

3. All computations were performed using MATLAB on a Dell PC (with a 2.66 GHz Intel Core2 Quad CPU and 3.25 GB of RAM) running Microsoft Windows XP.

## Acknowledgments

The authors thank Xin Chen, Alan King, David Morton, Melvyn Sim, and Stathis Tompaidis for helpful comments on this paper. They are also grateful for the careful and detailed feedback provided by two anonymous referees.

## References

- Adelman, D., A. J. Mersereau. 2008. Relaxations of weakly coupled stochastic dynamic programs. *Oper. Res.* **56**(3) 712–727.
- Andersen, L., M. Broadie. 2004. Primal-dual simulation algorithm for pricing multidimensional American options. *Management Sci.* **50**(9) 1222–1234.
- Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.
- de Farias, D. P., B. Van Roy. 2003. The linear programming approach to approximate dynamic programming. *Oper. Res.* **51**(6) 850–865.
- Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York.
- Haugh, M. B., L. Kogan. 2004. Pricing American options: A duality approach. *Oper. Res.* **52**(2) 258–270.
- Heston, S. L. 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financial Stud.* **6**(2) 327–343.
- Lai, G., F. Margot, N. Secomandi. 2010. An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Oper. Res.*, ePub ahead of print February 4, <http://or.journal.informs.org/cgi/content/abstract/opre.1090.0768v1>.
- Medvedev, A., O. Scaillet. 2007. Pricing American options under stochastic volatility and stochastic interest rates. Working paper, Swiss Finance Institute, Geneva.
- Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, Hoboken, NJ.
- Rockafellar, R. T., R. J.-B. Wets. 1991. Scenarios and policy aggregation in optimization under uncertainty. *Math. Oper. Res.* **16**(1) 119–147.
- Rogers, L. C. G. 2002. Monte Carlo valuation of American options. *Math. Finance* **12** 271–286.
- Rogers, L. C. G. 2007. Pathwise stochastic optimal control. *SIAM J. Control Optim.* **46**(3) 1116–1132.
- Shapiro, A., A. Ruszczyński. 2003. Optimality and duality in stochastic programming. A. Ruszczyński, A. Shapiro, eds. *Stochastic Programming. Handbooks in Operations Research and Management Science*, Volume 10. Elsevier, Amsterdam.
- Shapiro, A., D. Dentcheva, A. Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics and the Mathematical Programming Society, Philadelphia.
- Trehan, J. T., C. R. Sox. 2002. Adaptive inventory control for non-stationary demand and partial information. *Management Sci.* **48**(5) 607–624.