Information Relaxations and Duality in Stochastic Dynamic Programs: A Review and Tutorial

David B. Brown[1] and James E. Smith[2]

[1]Fuqua School of Business
Duke University
dbbrown@duke.edu

[2]Tuck School of Business
Dartmouth College
jim.smith@dartmouth.edu

January 21, 2022

Abstract

In this paper, we provide an overview of the information relaxation approach for calculating performance bounds in stochastic dynamic programs (DPs). The technique involves (1) relaxing the temporal feasibility (or nonanticipativity) constraints so the decision-maker (DM) has additional information before making decisions and (2) incorporating a penalty that punishes the DM for violating the temporal feasibility constraints. The goal of this paper is to provide a self-contained overview of the key theoretical results of the information relaxation approach as well as a review of research that has successfully used these techniques in a broad range of applications. We illustrate the information relaxation approach on applications in inventory management, assortment planning, and portfolio optimization.

Subject classifications: Stochastic dynamic programs, information relaxations, approximate dynamic programming.

1. Introduction

In principle, dynamic programming (DP) provides a powerful framework for modeling complex decision problems where uncertainty is resolved and decisions are made over time. However, in practice, the "curse of dimensionality" – the fact that the size of the state space typically grows exponentially in the number of state variables considered – severely limits the complexity of problems that can be solved using DP methods. 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 reward (or upper bound on the expected cost) with an optimal policy, but Monte Carlo simulation typically does not provide a good way to identify an optimal policy or a provide a *performance bound*, i.e., an upper bound on the expected reward (or lower bound on expected cost) with an optimal policy. Consequently, researchers and practitioners using heuristic control policies often wonder how good a policy is and whether it is "good enough" to use in practice.

In this paper, we review the information relaxation approach for calculating performance bounds in stochastic DPs, following Brown, Smith and Sun (2010) (hereafter BSS (2010)) and related work. The information relaxation approach consists of two elements: (1) we relax the temporal feasibility (or 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 these relaxed constraints. Relaxing the temporal feasibility constraint allows the decision-maker (DM) to make decisions using more information than is truly available and thus leads to an upper bound on value. Without any penalty for using this additional information, the resulting performance bound is often quite weak. Informally, we say a penalty is dual feasible if it does not penalize temporally feasible policies. Though there exists a dual feasible penalty that provides a bound that is equal to the optimal value for the primal DP (i.e., strong duality holds), these ideal penalties are based on the optimal value function, which is typically not available in the applications of interest – if the value function were available, we would not need performance bounds. In practice, we typically use penalties based on approximate value functions to generate performance bounds.

By relaxing the temporal feasibility constraints, we can often greatly simplify the problem by reducing a complex stochastic DP to a series of scenario-specific deterministic optimization problems solved within a Monte Carlo simulation. To illustrate this idea, we will consider a dynamic assortment problem, where a retailer decides which products to offer for sale ("display") when

facing uncertain demand, drawn from a distribution with unknown parameters. Here a perfect information relaxation assumes the DM knows all demands and distribution parameters before deciding which products to display. With this information, the problem of choosing products to display is a deterministic optimization problem. The information relaxation performance bound can be estimated using Monte Carlo simulation by repeatedly drawing random demands and distributions and averaging the results. We can also consider imperfect information relaxations where, for example, the DM knows the demand distribution but not the realized demands.

1.1 Outline of the Paper

The goal of this paper is to provide a summary of the key ideas of information relaxation methods for stochastic DPs and demonstrate their use in several examples. The idea is to provide a "one-stop-shop" (or at least a "first stop") for researchers seeking to learn the key ideas and tools for using information relaxation methods.

Following a brief history and literature review in §1.2, in §§2-4, we describe the theory associated with the information relaxation approach. §2 establishes the basic framework and §3 presents the key theoretical results, both following BSS (2010). In §4, we study DPs with a convex structure and show how the use of "gradient" penalties leads to inner problems that are easy to solve; this section draws on Brown and Smith (2014*b*). Before considering specific examples in detail, in §5 we provide a summary of the information relaxation approach and advice on how to proceed in applications.

In sections §6-§8, we consider illustrative applications. §6 illustrates the basic results and methods in a simple inventory management example with and without uncertainty about the state of the world; this problem is simple enough that it can be solved to optimality, allowing us to compare the information relaxation performance bounds to the optimal value. In §7, we consider a more complex example based on the dynamic assortment problem studied in Caro and Gallien (2007); our discussion draws on Brown and Smith (2020). In §8, we illustrate the use of gradient penalties (introduced in §4) on dynamic portfolio optimization problems with transaction costs, building on the model and results of Brown and Smith (2011).

A reader eager to see examples could read §6 describing the inventory example and perhaps §7 on the dynamic assortment example in parallel with §2-§3 describing the general framework and main results. Similarly, one could read §8 describing the portfolio optimization example in parallel with §4 describing the theory for convex DPs.

In §9 and §10, we briefly review other work that has advanced information relaxation

methodology and successfully applied the information relaxation approach. §11 offers a few concluding remarks and suggestions for future research.

1.2 History and Literature Review

Our interest in information relaxation methods for DPs began with BSS (2010). As discussed in BSS (2010), we were motivated by the need to evaluate the quality of heuristic policies in applications. As an example of one such application, Lai et al. (2010) consider the problem of managing natural gas storage over time in the presence of stochastic price dynamics. In the model, the merchant may inject or withdraw natural gas in each period. This problem is naturally formulated as a stochastic DP but is challenging because the natural gas forward curve involves a high-dimensional model that leads to a very large state space for the stochastic DP. Lai et al. (2010) develop some policies based on approximations of the value function. Naturally, one might wonder how good these policies are: could one do better with other – perhaps more complex – policies or is the current one "good enough?" Such questions are common when studying complex dynamic models.

The information relaxation approach to calculating performance bounds for DPs in BSS (2010) was inspired by Haugh and Kogan (2004)'s "duality approach" for placing bounds on the value of an American option; Rogers (2002) independently proposed a similar 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 (2004) 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 (2003) provides a nice overview of this work. Subsequent work (e.g., Meinshausen and Hambly 2004; Schoenmakers 2012) in financial engineering extended these dual methods to multiple stopping problems, for example, derivatives with several exercise rights such as "swing options" in electricity markets or "chooser caps" in interest rate markets.

BSS (2010) generalizes Haugh and Kogan (2004), Rogers (2002), and Andersen and Broadie (2004) in several ways. First, rather than focusing exclusively on option pricing problems, it considers general stochastic DPs. Second, rather than focusing exclusively on perfect information relaxations, it considers general information relaxations. BSS (2010) also presents a general method for constructing good penalties that includes and extends the methods proposed by Haugh and Kogan (2004) and Andersen and Broadie (2004).

The idea of relaxing temporal feasibility (or nonanticipativity) constraints has also been studied in the stochastic programming literature (see, e.g., Rockafellar and Wets 1976; Shapiro et al. 2009). The stochastic programming formulation typically requires the reward functions and set of feasible actions to be convex and the penalties to be linear functions of the actions; they consider only perfect information relaxations. In contrast, the information relaxation approach described here allows general reward functions and action spaces, allows general penalty functions, and considers imperfect as well as perfect information relaxations. The connection between the stochastic programming formulation and the information relaxation approach is discussed in more detail in Appendix B of BSS (2010). That appendix also discusses connections between the information relaxation results and standard Lagrangian duality results for linear programs (LPs). In the LP formulation of the information relaxation problem, the decision variables are mixing weights on policies and the objectives and constraints (including the temporal feasibility constraints) are linear functions of these decision variables. In this LP formulation, the penalties of the information relaxation approach correspond to the Lagrange multipliers associated with the temporal feasibility constraints. However, as shown in §3 below, we can also use simple, direct arguments to establish the key information relaxation duality results without considering mixed policies or LP duality results.

We view this information relaxation approach as a complement to the use of simulation methods and approximate dynamic programming methods for studying DPs (see, e.g., Bertsekas and Tsitsiklis 1996; de Farias and Van Roy 2003; Powell 2007; Adelman and Mersereau 2008). As mentioned earlier, given a candidate policy (perhaps identified using a heuristic reasoning 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 reward with an optimal policy. The information relaxation performance bound can often be estimated with little additional effort in the same simulation and, as discussed, can help determine whether the proposed policy is "good enough" or if we should continue searching for a better policy, perhaps using more complex ADP techniques.

"Hindsight bounds" – perfect information bounds with no penalties – are popular in the theoretical computer science literature (see, e.g., Feldman et al. 2010). These bounds are used to establish theoretical guarantees, for example showing that an algorithm is guaranteed to produce a solution that is within, say, 50% of the optimal solution. As we will see in our numerical examples, perfect information bounds with no penalty are often quite weak. Balseiro and Brown (2019) show how one can incorporate penalties in such theoretical studies and improve the theoretical guarantees

to show, for example, that an algorithm or policy is asymptotically optimal in a given setting (see §9 below for more).

2. Basic Framework

We take a high-level and abstract view of a DP that emphasizes the role of information; this approach allows us to formalize information structures and relaxations and treat the information structure as a "variable" in our framework. Uncertainty in the DP is described by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the set of possible scenarios (with typical element ω), $\mathcal{F}$ is a σ-algebra that describes the set of all possible events (an event is a subset of Ω), and $\mathbb{P}$ is a probability measure describing the likelihoods of the various events.

Time is discrete and indexed by $t = 0, \ldots, T$. The DM's state of information evolves over time and is described by a filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ where the σ-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." A function (or random variable) f defined on Ω is *measurable* with respect to a σ-algebra $\mathcal{F}_t$ 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 value of f depends only on the information known in period t. A sequence of functions $(f_0, \ldots, f_T)$ is said to be *adapted* to a filtration $\mathbb{F}$ (or $\mathbb{F}$-adapted) if each function f_t is $\mathcal{F}_t$-measurable.

In the DP model, the DM will choose an action a_t in period t from the set A_t; we let $\boldsymbol{A}(\omega) \subseteq A_0 \times \cdots \times A_T$ denote the set of all *physically feasible* action sequences $\boldsymbol{a} = (a_0, \ldots a_T)$ in scenario ω. The DM's choice of actions is described by a *policy* α that selects a sequence of actions $\boldsymbol{a}$ in $\boldsymbol{A}$ for each scenario ω in Ω (i.e., $\alpha : \Omega \to \boldsymbol{A}$). To ensure the DM knows the feasible set when choosing actions in period t, we assume that the set of actions available in period t depends on the prior actions $\boldsymbol{a}_{t\text{-}1} = (a_0, \ldots, a_{t\text{-}1})$ and is $\mathcal{F}_t$-measurable for each set of prior actions. We let $\mathcal{A}$ denote the set of all *physically feasible* policies, i.e., those that ensure that $\alpha(\omega)$ is in $\boldsymbol{A}(\omega)$.

In the primal DP, we assume that the DM's choices are *temporally feasible* (or nonanticipative) in that the choice of action a_t in period t, $A_t(\boldsymbol{a}_{t\text{-}1})$, depends only on what is known at the beginning of period t; that is, we require policies to be adapted to the natural filtration $\mathbb{F}$. We let $\mathcal{A}_{\mathbb{F}}$ be the set of all temporally and physically feasible – or just *feasible* – policies.

The goal of the DP is to select a feasible policy α to maximize the expected total reward. The

rewards are defined by a sequence of reward functions $(r_0(\boldsymbol{a}_0, \omega), \ldots, r_T(\boldsymbol{a}_T, \omega))$ where the reward in period t depends on the action sequence $\boldsymbol{a}_t$ selected up to period t and the scenario ω. We let $r(\boldsymbol{a}, \omega) = \sum_{t=0}^{T} r_t(\boldsymbol{a}_t, \omega)$ denote the total reward; discounting can be incorporated into the period reward function r_t. The primal DP is then:

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

Here $\mathbb{E}[\, r(\alpha)\,]$ could be written more explicitly as $\mathbb{E}[\, r(\alpha(\omega), \omega)\,]$ where policy α selects an action sequence that depends on the random scenario ω and the rewards r depend on the action sequence selected by α and the scenario ω. We will typically suppress the dependence on ω and interpret $r(\alpha)$ as a random variable representing the reward generated with policy α. A policy α is *optimal* if it is feasible ($\alpha \in \mathcal{A}_{\mathbb{F}}$) and it attains the supremum in (1).

It is instructive to write the primal DP (1) in the form of a Bellman-style recursion. Let $A_t(\boldsymbol{a}_t)$ denote the subset of period-t actions A_t that are feasible given the prior choice of actions $\boldsymbol{a}_t$. We take the terminal value function $V_{T+1}^*(\boldsymbol{a}_T) = 0$ and, for $t = 0, \ldots, T$, we define

$$V_t^*(\boldsymbol{a}_{t-1}) = \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\big[\, r_t(\boldsymbol{a}_t) + V_{t+1}^*(\boldsymbol{a}_t) \,\big|\, \mathcal{F}_t \,\big]\,. \tag{2}$$

Here both sides are random variables (and therefore implicitly functions of the scenario ω) and we select an optimal action a_t for each scenario ω.[1] Since 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 sequence of actions $\boldsymbol{a}_t$. Given that the feasible actions $A_t(\boldsymbol{a}_{t\text{-}1})$ are assumed to be $\mathcal{F}_t$-measurable, the supremum over actions a_t is also $\mathcal{F}_t$-measurable which implies 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; so, if the suprema on the right side of (2) are attained, we can construct a temporally feasible optimal policy using this recursion. The final value V_0 is equal to the optimal value of (1).

Note that this formulation begins with an exogenous probability measure $\mathbb{P}$, implying the probabilities in the model are independent of the actions selected by the DM. Problems with action-dependent probabilities can be recast as equivalent problems with action-independent probabilities, sometimes quite naturally. For example, we could think of the dynamic assortment problem of §7

[1] In many problems modeled as DPs, it is standard to define a notion of a *state* and write the problem as a Markov decision process (MDP) as in, e.g., Bertsekas (2017). We deliberately avoid defining states in our framework: when we consider different information relaxations and penalties, the relevant state space needed to solve the relaxation as a DP may differ from that considered in the primal DP formulation. To avoid this complication, we write the DP using this random variable formulation instead.

as having state transition probabilities that depend on the display decisions. Alternatively, we can formulate this problem (as we will) with demand as uncertain and independent of the actions. More generally, one could take the scenario ω to be a series $(U_0, \ldots, U_T)$ of uniform random numbers where U_t is revealed in period t; using inverse transform sampling, we could then calculate the period-t state from these uniform random numbers and the chosen actions $\boldsymbol{a}_t$. There are a variety of ways one can formulate a DP model to have action-independent probabilities and, in principle, the assumption that probabilities are independent of the actions in the primal DP is without loss of generality. However, different formulations may lead to different information relaxations and performance bounds.

3. Main Results

In this section, we review the main results underlying the information relaxation approach. Our presentation follows BSS (2010) and the main results are stated in a format that mimics standard presentations of linear programming duality (see, e.g., Luenberger and Ye 2016; Bertsimas and Tsitsiklis 1997).

3.1 Duality Results

In the information relaxation approach to the DP (1), we relax the constraint that the policies must be temporally feasible and impose penalties that punish violations of these constraints. We define relaxations of the temporal feasibility requirement by considering alternative information structures: a filtration $\mathbb{G} = (\mathcal{G}_0, \ldots, \mathcal{G}_T)$ is a *relaxation* of another filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ if, for each t, $\mathcal{F}_t \subseteq \mathcal{G}_t \subseteq \mathcal{F}$; that is, the DM knows more in every period under $\mathbb{G}$ than is known under $\mathbb{F}$. We abbreviate this by writing $\mathbb{F} \subseteq \mathbb{G}$. The perfect information relaxation is $\mathbb{I} = (\mathcal{F}, \ldots, \mathcal{F})$, meaning the DM knows the scenario ω before making any decisions. 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 Π is the set of all functions $\pi(\boldsymbol{a}, \omega)$ that, like the total rewards, depend on the action sequence $\boldsymbol{a}$ and the scenario ω. As with rewards, we will typically write penalties as action-dependent random variables $\pi(\boldsymbol{a})$ $(= \pi(\boldsymbol{a}, \omega))$ or policy-dependent random variables $\pi(\alpha)$ $(= \pi(\alpha(\omega), \omega))$, suppressing the dependence on the scenario ω. We define the set $\Pi_{\mathbb{F}}$ of *dual feasible*

penalties to be those penalties that do not penalize (in expectation) temporally feasible policies:

$$\Pi_{\mathbb{F}} = \{\pi \in \Pi \ : \ \mathbb{E}[\,\pi(\alpha_{_F})\,] \leq 0 \ \ \forall \alpha_{_F} \in \mathcal{A}_{\mathbb{F}}\}\,.$$

Policies that are not temporally feasible may have positive expected penalties and we will use this to "punish" the DM for using information that would not be known in the natural filtration $\mathbb{F}$.

We can derive an upper bound on the expected reward associated with any feasible policy by relaxing the temporal feasibility constraints on policies and imposing a dual feasible penalty. This simple result is analogous to "weak duality" in linear programming and is the key result for applications. The proof follows directly from the definitions of information relaxations and dual feasible penalties.

Theorem 3.1 (Weak Duality, BSS (2010))**.** *If $\alpha_{_F}$ and π are primal and dual feasible respectively (i.e., $\alpha_{_F} \in \mathcal{A}_{\mathbb{F}}$ and $\pi \in \Pi_{\mathbb{F}}$) and $\mathbb{G}$ is a relaxation of $\mathbb{F}$, then*

$$\mathbb{E}[\,r(\alpha_{_F})\,] \leq \sup_{\alpha_{_G} \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\,r(\alpha_{_G}) - \pi(\alpha_{_G})\,]\,. \tag{3}$$

Proof. With π, $\alpha_{_F}$, and $\mathbb{G}$ as defined in the theorem statement, we have

$$\mathbb{E}[\,r(\alpha_{_F})\,] \leq \mathbb{E}[\,r(\alpha_{_F}) - \pi(\alpha_{_F})\,] \leq \sup_{\alpha_{_G} \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\,r(\alpha_{_G}) - \pi(\alpha_{_G})\,]\,.$$

The first inequality holds because $\pi \in \Pi_{\mathbb{F}}$ (thus $\mathbb{E}[\,\pi(\alpha_{_F})\,] \leq 0$ for any $\alpha_{_F} \in \mathcal{A}_{\mathbb{F}}$) and the second because $\alpha_{_F} \in \mathcal{A}_{\mathbb{F}} \subseteq \mathcal{A}_{\mathbb{G}}$. □

Thus any information relaxation with any dual feasible penalty will provide an upper bound on all feasible DP policies – including the optimal policy – thereby providing a performance bound.

With a fixed penalty π, weaker relaxations $\mathbb{G}$ lead to larger sets of feasible policies $\mathcal{A}_{\mathbb{G}}$ and weaker bounds:

Corollary 3.1 (Tighter Relaxations)**.** *If $\mathbb{F} \subseteq \mathbb{G} \subseteq \mathbb{G}'$, then*

$$\mathbb{E}[\,r(\alpha_{_F})\,] \leq \sup_{\alpha_{_G} \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\,r(\alpha_{_G}) - \pi(\alpha_{_G})\,] \leq \sup_{\alpha_{_G} \in \mathcal{A}_{\mathbb{G}'}} \mathbb{E}[\,r(\alpha_{_G}) - \pi(\alpha_{_G})\,]\ .$$

For example, as we will see in the "world driven" inventory control example of §6.4, the bounds given by one penalty may be "good enough" with one information relaxation but not "good enough" with a looser relaxation.

If we consider the perfect information relaxation $\mathbb{I}$, the set of relaxed policies $\mathcal{A}_{\mathbb{I}}$ is simply the set of all physically feasible policies $\mathcal{A}$ and all actions are selected knowing the scenario ω. Weak duality then implies that for any $\alpha_{_F}$ in $\mathcal{A}_{\mathbb{F}}$ and π in $\Pi_{\mathbb{F}}$,

$$\mathbb{E}[\, r(\alpha_{_F})\,] \leq \sup_{\alpha \in \mathcal{A}} \mathbb{E}[\, r(\alpha) - \pi(\alpha)\,] = \mathbb{E}\left[\sup_{\boldsymbol{a} \in \boldsymbol{A}(\omega)} \{r(\boldsymbol{a}, \omega) - \pi(\boldsymbol{a}, \omega)\} \right]. \tag{4}$$

The perfect information upper bound (4) is in a form that is convenient for Monte Carlo simulation: we can estimate the expected value on the right side of (4) by randomly generating scenarios ω and solving a deterministic "inner problem" where we choose a feasible action sequence $\boldsymbol{a} \in \boldsymbol{A}(\omega)$ to maximize the penalized objective $r(\boldsymbol{a}, \omega) - \pi(\boldsymbol{a}, \omega)$ for the given ω. For instance, in our portfolio optimization example in §8, the perfect information relaxation assumes the DM knows all asset returns before making any trading decisions. We estimate the information relaxation performance bound by randomly generating return scenarios and solving a deterministic inner problem that chooses optimal trading decisions in each return scenario given a particular form of penalty.

We can write the dual DP on the right side of (3) in a recursive form analogous to that for the primal DP (2). The terminal case is $V^{\mathbb{G}}_{T+1}(\boldsymbol{a}_T) = 0$, and, for $t = 0, \ldots, T$, we have

$$V^{\mathbb{G}}_t(\boldsymbol{a}_{t-1}) = \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\Big[\, r_t(\boldsymbol{a}_t) - \pi_t(\boldsymbol{a}_t) + V^{\mathbb{G}}_{t+1}(\boldsymbol{a}_t) \,\Big|\, \mathcal{G}_t \,\Big]. \tag{5}$$

As discussed following (2), this recursion leads to $\mathbb{G}$-adapted policies and value functions. The expectation of initial value, $\mathbb{E}\big[\, V^{\mathbb{G}}_0 \,\big]$, provides an upper bound on the primal DP (1) or (2).

With an imperfect information relaxation, the resulting "inner problems" may be stochastic DPs that ideally are easier to solve than the primal DP. For example, BSS (2010) considers an option-pricing problem where stock prices, interest rates, and volatilities are all uncertain and evolving over time. An imperfect information relaxation considered there treats stock prices as uncertain and interest rates and volatilities as known: i.e., $\mathcal{G}_t$ in (5) includes knowledge of all interest rates and volatilities through period T and stock prices up to period $t-1$. The resulting inner problem is a standard option-pricing problem which can be solved using, for example, a binomial or trinomial lattice to value an option with known, but time-varying interest rates and volatilities. The information relaxation bounds are estimated by Monte Carlo simulation where these inner problems are repeatedly solved for randomly generated sequences of interest rates and volatilities. In practice, the choice of information relaxation must be made with careful consideration of the complexity of the resulting inner problems; see §5 for more discussion on this point. The inventory management

and dynamic assortment problems of §6-§7 also consider imperfect information relaxations.

If we minimize over dual feasible penalties in (3), we obtain the dual of the primal DP (1):

$$\inf_{\pi \in \Pi_{\mathbb{F}}} \left\{ \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\, r(\alpha_G) - \pi(\alpha_G)\,] \right\}. \tag{6}$$

By weak duality, if we identify a policy α_F and penalty π that are primal and dual feasible, respectively, such that equality holds in (3), then α_F and π 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 $\pi^*(\boldsymbol{a}) = r(\boldsymbol{a}) - v^*$ where v^* is the optimal value of the primal DP (1). This π^* is dual feasible (since $\mathbb{E}[\, r(\alpha_F)\,] \leq v^*$ for all $\alpha_F \in \mathcal{A}_{\mathbb{F}}$) and trivially optimal: no matter what policy is selected, the penalized objective function $r(\boldsymbol{a}) - \pi^*(\boldsymbol{a})$ is equal to v^*. Of course, 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 3.2 (Strong Duality, BSS (2010))**.** *Let $\mathbb{G}$ be a relaxation of $\mathbb{F}$. Then*

$$\sup_{\alpha_F \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[\, r(\alpha_F)\,] = \inf_{\pi \in \Pi_{\mathbb{F}}} \left\{ \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\, r(\alpha_G) - \pi(\alpha_G)\,] \right\}. \tag{7}$$

Furthermore, if the primal problem on the left is bounded, the dual problem on the right has an optimal solution $\pi^* \in \Pi_{\mathbb{F}}$ that achieves this bound.

In §3.2 below, we will consider a penalty that we call the "ideal penalty" that is also optimal in (7) and gives more insight into penalties that are likely to perform well in practice.

Finally, as in linear programming, the complementary slackness condition characterizes the relationship between the primal and dual problems, saying that for a primal-dual pair (α_F^*, π^*) to be optimal, it is necessary and sufficient for α_F^* to have zero expected penalty with penalty π^* and for α_F^* to solve the dual problem in the following sense.

Theorem 3.3 (Complementary Slackness, BSS (2010))**.** *Let α_F^* and π^* be feasible solutions for the primal and dual problems respectively (i.e., $\alpha_F^* \in \mathcal{A}_{\mathbb{F}}$ and $\pi^* \in \Pi_{\mathbb{F}}$) with information relaxation $\mathbb{G}$. A necessary and sufficient condition for these to be optimal solutions for their respective problems*

is that $\mathbb{E}[\,\pi^*(\alpha_F^*)\,] = 0$ *and*

$$\mathbb{E}[\,r(\alpha_F^*) - \pi^*(\alpha_F^*)\,] = \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\,r(\alpha_G) - \pi^*(\alpha_G)\,]\,. \tag{8}$$

Proof. We first consider sufficiency. Consider any $\alpha_F^* \in \mathcal{A}_{\mathbb{F}}$ and $\pi^* \in \Pi_{\mathbb{F}}$ and suppose (8) holds and $\mathbb{E}[\,\pi^*(\alpha_F^*)\,] = 0$. Then we can rewrite the dual problem (on the right side of (3)) with this penalty as

$$\begin{aligned} \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\,r(\alpha) - \pi^*(\alpha)\,] &= \mathbb{E}[\,r(\alpha_F^*) - \pi^*(\alpha_F^*)\,] && \text{(using (8))} \\ &= \mathbb{E}[\,r(\alpha_F^*)\,] && \text{(since } \mathbb{E}[\,\pi^*(\alpha_F^*)\,] = 0\text{).} \end{aligned}$$

Then, by weak duality, α_F^* and π^* must be optimal.

To show necessity, first note that for any $\alpha_F^* \in \mathcal{A}_{\mathbb{F}}$ and $\pi^* \in \Pi_{\mathbb{F}}$, we have:

$$\begin{aligned} \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\,r(\alpha_G) - \pi^*(\alpha_G)\,] &\geq \sup_{\alpha_F \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[\,r(\alpha_F) - \pi^*(\alpha_F)\,] && \text{(because } \mathcal{A}_{\mathbb{F}} \subseteq \mathcal{A}_{\mathbb{G}}\text{)} \\ &\geq \mathbb{E}[\,r(\alpha_F^*) - \pi^*(\alpha_F^*)\,] && \text{(because } \alpha_F^* \in \mathcal{A}_{\mathbb{F}}\text{)} \\ &\geq \mathbb{E}[\,r(\alpha_F^*)\,] && \text{(because } \pi^* \in \Pi_{\mathbb{F}}\text{).} \end{aligned}$$

If $\alpha_F^* \in \mathcal{A}_{\mathbb{F}}$ and $\pi^* \in \Pi_{\mathbb{F}}$ are primal and dual optimal (respectively), then, by the strong duality theorem, the first and last terms above are equal which implies the intervening inequalities hold with equality and we have $\mathbb{E}[\,\pi^*(\alpha_F^*)\,] = 0$ and (8). □

Equation (8) can be interpreted as implying that with an optimal penalty, in the dual problem the DM will be content to choose a policy that is temporally feasible even though they have the option of choosing a policy that is not. In applications, we can study the differences between the action selected by the heuristic policies α_F used to compute a lower bound and the policies α_G selected in the dual problem to see if we can identify some way to improve the heuristic policy and/or dual bound. BSS (2010) demonstrates this idea in two examples. In an inventory management problem with a demand distribution that depends on an unobservable market state which changes stochastically over time, the myopic policies (α_F) tend to order too much given the possibility of decreased future demand whereas the dual policy (α_G) "cheats" and orders less before demand actually drops (something that would not be known in the natural filtration) and thereby avoids the cost of holding too much inventory in a low demand state. This observation suggested using an improved myopic value model that recognizes the possibility of reduced future demand (see §3.6 of BSS (2010)). Similar comparisons in an option pricing example lead to an improved exercise policy (see §4.6 of BSS (2010)).

3.2 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 method we will use to generate penalties is described in the following proposition.

Proposition 3.1 (Constructing Good Penalties, BSS (2010))**.** *Let* $(w_0(\boldsymbol{a}_0, \omega), \ldots, w_T(\boldsymbol{a}_T, \omega))$ *be a sequence of generating functions defined on* $A \times \Omega$ *where each* w_t *depends only on the first* $t+1$ *actions* $\boldsymbol{a}_t = (a_0, \ldots, a_t)$ *of* $\boldsymbol{a}$*. Similarly, let* $\boldsymbol{\alpha}_t$ *denote the first* $t+1$ *actions selected by policy* α*. Define*

$$\pi_t(\boldsymbol{a}_t) = w_t(\boldsymbol{a}_t) - \mathbb{E}[\, w_t(\boldsymbol{a}_t) \,|\, \mathcal{F}_t \,] \tag{9}$$

and $\pi(\boldsymbol{a}) = \sum_{t=0}^{T} \pi_t(\boldsymbol{a}_t)$*. Then, for all* α *in* $\mathcal{A}_{\mathbb{F}}$*, we have* $\mathbb{E}[\, \pi_t(\boldsymbol{\alpha}_t) \,|\, \mathcal{F}_t \,] = 0$ *(almost surely) for all* t *and* $\mathbb{E}[\, \pi(\alpha) \,] = 0$*. Thus* π *is dual feasible (i.e.,* $\pi \in \Pi_{\mathbb{F}}$*).*

This proposition implies that the penalties π generated using (9) will always be dual feasible in that $\mathbb{E}[\, \pi(\alpha_F) \,] \leq 0$ for α_F in $\mathcal{A}_{\mathbb{F}}$, but is stronger in that it implies the inequality defining feasibility holds with equality. The complementary slackness condition (Theorem 3.3) shows that an optimal penalty π^* will assign zero expected penalty to an optimal primal policy α^*. Penalties generated using Proposition 3.1 will assign zero expected penalty to *all* temporally feasible policies.[2]

The proof of Proposition 3.1 relies on the following lemma which we state without proof (for a proof, see Lemma A.1 of BSS (2010)).

Lemma 3.1. *Let* $z_t(a) = \mathbb{E}[\, w_t(a) \,|\, \mathcal{F}_t \,]$*. If* $w_t(a)$ *depends on the first* $t+1$ *actions in* a *and* α *is* $\mathbb{F}$*-adapted, then* $z_t(\alpha) = \mathbb{E}[\, w_t(\alpha) \,|\, \mathcal{F}_t \,]$*, almost surely.*

Note that the result of this lemma need not hold for policies that are not $\mathbb{F}$-adapted: In $z_t(\alpha)$ (with $z_t(a)$ as defined in the lemma), we calculate the "$\mathcal{F}_t$-average" in the conditional expectation and then select averaged values for actions selected according to policy α. In $\mathbb{E}[\, w_t(\alpha) | \mathcal{F}_t \,]$, we select values $w_t(a)$ for actions according to the policy α first and then calculate the $\mathcal{F}_t$-average. In these terms, the lemma says that if α is $\mathbb{F}$-adapted, $\mathcal{F}_t$-averaging and then selecting actions is equivalent to selecting actions and then $\mathcal{F}_t$-averaging. If the policy α is not $\mathbb{F}$-adapted, selecting then $\mathcal{F}_t$-averaging may not be the same as averaging then selecting.

[2]Note that BSS (2010) defines "good penalties" as having the form $\pi_t(\boldsymbol{a}_t) = \mathbb{E}[\, w_t(\boldsymbol{a}_t) \,|\, \mathcal{G}_t \,] - \mathbb{E}[\, w_t(\boldsymbol{a}_t) \,|\, \mathcal{F}_t \,]$ rather than the form of equation (9). Leaving out the $\mathbb{G}$-conditional expectations as done in (9) makes no difference in applications because we take $\mathbb{G}$-expectations of the penalties in the relaxed problem (e.g., in (5)), but it is simpler to define penalties independently of the information relaxation as we do here.

Proof of Proposition 3.1. Given an $\mathbb{F}$-adapted policy α, using Lemma 3.1 and the law of iterated expectations, we have

$$\mathbb{E}[\,\pi_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,] = \mathbb{E}[\,w_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,] - \mathbb{E}[\,\mathbb{E}[\,w_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,]\,|\,\mathcal{F}_t\,] = 0 \;\; \text{(almost surely)}.$$

Here the law of iterated expectations implies $\mathbb{E}[\,\mathbb{E}[\,w_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,]\,|\,\mathcal{F}_t\,] = \mathbb{E}[\,w_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,]$, almost surely. Summing π_t over time and using the law of iterated expectations again (to establish $\mathbb{E}[\,\mathbb{E}[\,\pi_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,]\,] = \mathbb{E}[\,\pi_t(\boldsymbol{\alpha}_t)\,]$), we have $\mathbb{E}[\,\pi(\alpha)\,] = 0$, as stated in the second claim of the proposition. □

We can construct an *ideal penalty* using Proposition 3.1 by taking the generating functions w_t to be based on the optimal DP value function (2) as

$$w_t(\boldsymbol{a}_t) = r_t(\boldsymbol{a}_t) + V_{t+1}^*(\boldsymbol{a}_t)\;. \tag{10}$$

Given a relaxation $\mathbb{G}$ of $\mathbb{F}$, the dual value function (5) with this penalty then becomes

$$V_t^{\mathbb{G}}(\boldsymbol{a}_{t-1}) = \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\Big[\,\mathbb{E}\big[\,r_t(\boldsymbol{a}_t) + V_{t+1}^*(\boldsymbol{a}_t)\,\big|\,\mathcal{F}_t\,\big] - V_{t+1}^*(\boldsymbol{a}_t) + V_{t+1}^{\mathbb{G}}(\boldsymbol{a}_t)\,\Big|\,\mathcal{G}_t\,\Big]\,. \tag{11}$$

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, noting that $\mathbb{E}[\,\mathbb{E}[\,\cdot\,|\,\mathcal{F}_t\,]\,|\,\mathcal{G}_t\,] = \mathbb{E}[\,\cdot\,|\,\mathcal{F}_t\,]$ (since $\mathbb{F} \subseteq \mathbb{G}$), the expression for $V_t^{\mathbb{G}}(\boldsymbol{a}_{t-1})$ above 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 3.4 (Ideal Penalties, BSS (2010))**.** *Let $\mathbb{G}$ be a relaxation of $\mathbb{F}$ and let π^* be defined as in Proposition 3.1 by taking $w_t(\boldsymbol{a}_t) = r(\boldsymbol{a}_t) + V_{t+1}^*(\boldsymbol{a}_t)$. Then π^* is dual feasible and optimal in that*

$$V_0^* = \sup_{\alpha_F \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[\,r(\alpha_F)\,] = \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\,r(\alpha_G) - \pi^*(\alpha_G)\,]\;. \tag{12}$$

Moreover, if $\alpha_F^* \in \mathcal{A}_{\mathbb{F}}$ achieves the supremum for the primal problem on the left side of (12) *(i.e., is optimal), then α_F^* is also optimal for the dual problem on the right. Finally, if $\alpha_G^* \in \mathcal{A}_{\mathbb{G}}$ is an optimal policy for the dual problem, then for almost all ω,*

$$r(\alpha_G^*, \omega) - \pi^*(\alpha_G^*, \omega) = V_0^*. \tag{13}$$

The last part of the result could alternatively be stated as $r(\alpha_G^*) - \pi^*(\alpha_G^*) = V_0^*$, almost surely, meaning the set of scenarios ω where this equality (or (13) does not hold) has probability zero. Informally, the "almost" here stems from the fact that optimal policies may be "suboptimal" on sets with probability zero and still be optimal and, similarly, versions of the conditional expectations $\mathbb{E}[\,\cdot\,|\,\mathcal{F}_t\,]$ (used in defining the ideal penalties) may differ on sets with probability zero.

Proof. The fact that π^* is dual feasible follows from Proposition 3.1 and the fact that it is optimal for the dual problem follows from the inductive argument in the text preceding the statement of the theorem. The fact that any $\alpha_F^* \in \mathcal{A}_{\mathbb{F}}$ that is optimal for the primal problem is also optimal for the dual problem then follows by the complementary slackness result, Theorem 3.3.

To establish the last part of the theorem, let us abuse notation a bit and write $V_t^*(\boldsymbol{a})$ in place of $V_t^*(\boldsymbol{a}_{t\text{-}1})$ with the understanding that the subsequence of actions $\boldsymbol{a}_t = (a_0, \ldots, a_t)$ is selected from the full sequence of actions $\boldsymbol{a}$; similarly with the rewards r_t and penalties π_t^*. If $\alpha_G^* \in \mathcal{A}_{\mathbb{G}}$ is optimal for the dual problem, we then have

$$\begin{aligned} r(\alpha_G^*) - \pi^*(\alpha_G^*) &= \sum_{t=0}^{T} r_t(\alpha_G^*) - \pi_t^\star(\alpha_G^*) \\ &= \sum_{t=0}^{T} \mathbb{E}\big[\, r_t(\alpha_G^*) + V_{t+1}^*(\alpha_G^*) \,\big|\, \mathcal{F}_t \,\big] - V_{t+1}^*(\alpha_G^*) \\ &= \sum_{t=0}^{T} V_t^*(\alpha_G^*) - V_{t+1}^*(\alpha_G^*) \text{ (almost surely)} \qquad (14) \\ &= V_0^*. \end{aligned}$$

The first and second equalities above follow from the definition of r and π^* and Lemma 3.1. With an ideal penalty, an optimal policy α_G^* for the dual problem almost surely satisfies (see equations (5) and (11))

$$\begin{aligned} V_t^{\mathbb{G}}(\boldsymbol{a}_{t-1}) &= \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\big[\, \mathbb{E}\big[\, r_t(\boldsymbol{a}_t) + V_{t+1}^*(\boldsymbol{a}_t) \,\big|\, \mathcal{F}_t \,\big] \,\big|\, \mathcal{G}_t \,\big] \\ &= \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\big[\, r_t(\boldsymbol{a}_t) + V_{t+1}^*(\boldsymbol{a}_t) \,\big|\, \mathcal{F}_t \,\big] \\ &= V_t^*(\boldsymbol{a}_{t-1}). \end{aligned}$$

Here we use the fact that $\mathbb{E}[\,\mathbb{E}[\,\cdot\,|\,\mathcal{F}_t\,]\,|\,\mathcal{G}_t\,] = \mathbb{E}[\,\cdot\,|\,\mathcal{F}_t\,]$ (since $\mathbb{F} \subseteq \mathbb{G}$) and the definition of V_t^* in equation (2). (Note this equality may fail on a set of measure zero for an optimal policy α_G^*.) This establishes (14) above. Continuing after (14), we find that adjacent terms in (14) cancel and (14) reduces to $V_0^*(\alpha_G^*) - V_{T+1}^*(\alpha_G^*)$. Here V_{T+1}^* was defined to be 0 and $V_0^*(\alpha_G^*)$ is equal to V_0^*. □

The last part of Theorem 3.4 notes that not only does the ideal penalty result in a dual problem (on the right side of (12)) whose expected value matches that of the primal DP (on the left side of (12)), the dual problem yields the optimal value of the primal DP in every scenario. For example, with a perfect information relaxation dual problem (4), if we were to estimate this bound with

Monte Carlo simulation using an ideal penalty, the simulation would return the optimal expected value for the primal DP in every sampled scenario and the estimate would have zero variance.

Note that if the period-t reward functions r_t are $\mathcal{F}_t$-measurable (as was assumed in BSS (2010)), then the reward function can be omitted from the definition of the ideal penalty, i.e., we can take the generating function to be

$$w_t(\boldsymbol{a}_t) = V^*_{t+1}(\boldsymbol{a}_t) \tag{15}$$

rather than (10) because $r_t(\boldsymbol{a}_t) = \mathbb{E}[\, r_t(\boldsymbol{a}_t) \,|\, \mathcal{F}_t \,]$ and the reward terms cancel in the definition of the penalty (9).

Of course, the optimal value functions will not be known in the applications of interest (if the value functions were known, we would not need performance bounds) and in such cases, the ideal penalty will not be available. However, the form of the ideal penalty π^* illustrates what we would like to approximate with our choice of penalties. In practice, we will typically take the generating functions in Proposition 3.1 to be based on approximate value functions $\hat{V}_{t+1}$ and consider penalties with period-t terms of the form:

$$\hat{\pi}_t(\boldsymbol{a}_t) = r_t(\boldsymbol{a}_t) + \hat{V}_{t+1}(\boldsymbol{a}_t) - \mathbb{E}\Big[\, r_t(\boldsymbol{a}_t) + \hat{V}_{t+1}(\boldsymbol{a}_t) \,\Big|\, \mathcal{F}_t \,\Big] \,. \tag{16}$$

Proposition 3.1 ensures the penalty $\hat{\pi} = \sum_{t=0}^{T} \hat{\pi}_t$ is dual feasible and thus leads to an upper bound on V_0. The key to obtaining a good bound from such an approximate value function is for the differences in (16) to provide a good approximation of the differences

$$r_t(\boldsymbol{a}_t) + V^*_{t+1}(\boldsymbol{a}_t) - \mathbb{E}\big[\, r_t(\boldsymbol{a}_t) + V^*_{t+1}(\boldsymbol{a}_t) \,\big|\, \mathcal{F}_t \,\big]$$

based on the true value function V^*_t. For example, penalties based on limited-lookahead approximate value functions may do well: though the limited-lookahead approximations do not approximate the value functions very well (because they include only a few periods of rewards), they may approximate the differences in true values well. For example, we will see that a myopic "smoothing penalty" performs well in the inventory example of §6.

These "good penalties" may also be helpful as control variates when estimating the expected reward associated with a heuristic policy, i.e., in estimating a primal lower bound. For example, given a heuristic policy $\hat{\alpha}$ that is feasible for the primal DP (1), we can write the expected total

reward as

$$\begin{aligned}\mathbb{E}[\, r(\hat{\alpha})\,] &= \mathbb{E}\left[\sum_{t=0}^{T} r_t(\hat{\boldsymbol{\alpha}}_t)\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{T} r_t(\hat{\boldsymbol{\alpha}}_t) - \left(r_t(\hat{\boldsymbol{\alpha}}_t) + \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) - \mathbb{E}\Big[\, r_t(\hat{\boldsymbol{\alpha}}_t) + \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big]\right)\right] , \end{aligned} \tag{17}$$

where the last expression incorporates a zero-mean penalty term (16) as a control variate. This control variate is of the form considered in the "approximating martingale-process method" for variance reduction developed in Henderson and Glynn (2002). If the value functions $\hat{V}_t$ are value functions corresponding to policy $\hat{\alpha}$ (so $\hat{V}_t(\hat{\boldsymbol{\alpha}}_{t\text{-}1}) = r_t(\hat{\boldsymbol{\alpha}}_t) + \mathbb{E}\Big[\, \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big]$) adjacent terms in (17) cancel and the expectations reduce to the expectation of a constant, $\mathbb{E}\Big[\, \hat{V}_0 \,\Big] = \hat{V}_0$. In this case, when estimating values by simulation, we would obtain a zero-variance estimate of the expected reward associated with policy $\hat{\alpha}$. If the functions $\hat{V}_t$ approximate the values given by the policy $\hat{\alpha}$ (or, more precisely, approximate the differences in values appearing in (17) well), we would expect to obtain low variance estimates of the value associated with a given policy.

3.3 Properties of Information Relaxation Bounds

Although the approximate value function $\hat{V}_{t+1}$ in (16) can be any function satisfying the conditions of Proposition 3.1, we can say more in the case where the approximate value function is an optimal value function for an approximating DP. Specifically, consider a DP defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and filtration $\mathbb{F}$ as in the original model (as described in §2), but with total rewards $\hat{r}$ instead of r and constraint set $\hat{\boldsymbol{A}}$ instead of $\boldsymbol{A}$. We say this approximate model is a (physical) *relaxation* of the original model if $r(\boldsymbol{a}, \omega) \leq \hat{r}(\boldsymbol{a}, \omega)$ holds for all $\boldsymbol{a}$ in $\boldsymbol{A}$ and for all ω (i.e., for all actions that are feasible for the original model) and $\boldsymbol{A}(\omega) \subseteq \hat{\boldsymbol{A}}(\omega)$ for all scenarios ω; we will abbreviate this by writing $r \leq \hat{r}$ and $\boldsymbol{A} \subseteq \hat{\boldsymbol{A}}$, respectively. Such relaxations arise naturally in many settings, e.g., the relaxation may come from relaxing physical constraints, such as a Lagrangian relaxation of a weakly coupled DP as in the dynamic assortment problem (see §7) or by considering a "frictionless" approximation that, for example, ignores transaction costs or taxes in portfolio optimization (see §8). Because the rewards and feasible sets are no smaller in the relaxed model, the relaxed model must be an upper bound on the optimal value in the original model, i.e.,

$$V_0 = \sup_{\alpha \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[\, r(\alpha)\,] \;\leq\; \hat{V}_0 = \sup_{\alpha \in \hat{\mathcal{A}}_{\mathbb{F}}} \mathbb{E}[\, \hat{r}(\alpha)\,] \,, \tag{18}$$

where $\hat{\mathcal{A}}_{\mathbb{F}}$ denotes the set of feasible policies for the relaxed problem. What is perhaps not obvious is that the information relaxation bound based on the penalty (16) from this relaxed value function $\hat{V}_t$ will be tighter than the bound (18) provided by the relaxed model itself. Part (ii) shows the same result holds with penalties generated by a supersolution to a DP (see, e.g., Puterman 1994, Proposition 5.3.1).

Proposition 3.2 (Improving Bounds)**.**

(i) (Brown and Smith 2014*b*) *Let $\hat{\pi}$ be the penalty given by* (16) *for approximate value functions $\hat{V}_t$. If the value functions $\hat{V}_t$ are the optimal value functions for a relaxed model with $\boldsymbol{A} \subseteq \hat{\boldsymbol{A}}$ and $r \leq \hat{r}$, then,*

$$\sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\, r(\alpha_G) - \hat{\pi}(\alpha_G)\,] \leq \hat{V}_0.$$

Moreover, for almost every scenario ω:

$$\sup_{\boldsymbol{a} \in \boldsymbol{A}(\omega)} \{r(\boldsymbol{a}, \omega) - \hat{\pi}(\boldsymbol{a}, \omega)\} \leq \hat{V}_0. \tag{19}$$

(ii) (Desai et al. 2011; Brown and Haugh 2017) *The conclusions of (i) also hold if the approximate value function $\hat{V}_t$ is a supersolution to the Bellman equation: that is, if $\hat{V}_t$ satisfies* (2) *with an inequality ($\geq$) rather than equality.*

Proof. (i) Using the ideal penalty result (Theorem 3.4) with the relaxed model, we know that, for almost every scenario ω,

$$\sup_{\boldsymbol{a} \in \hat{\boldsymbol{A}}(\omega)} \{\hat{r}(\boldsymbol{a}, \omega) - \hat{\pi}(\boldsymbol{a}, \omega)\} = \hat{V}_0.$$

Since $r(\boldsymbol{a}, \omega) \leq \hat{r}(\boldsymbol{a}, \omega)$ and $\boldsymbol{A}(\omega) \subseteq \hat{\boldsymbol{A}}(\omega)$, we have

$$\sup_{\boldsymbol{a} \in \boldsymbol{A}(\omega)} \{r(\boldsymbol{a}, \omega) - \hat{\pi}(\boldsymbol{a}, \omega)\} \leq \sup_{\boldsymbol{a} \in \hat{\boldsymbol{A}}(\omega)} \{\hat{r}(\boldsymbol{a}, \omega) - \hat{\pi}(\boldsymbol{a}, \omega)\} = \hat{V}_0\ .$$

(ii) This result follows from an induction argument similar to that used to establish the optimality of the ideal penalty (Theorem 3.4). Specifically, we show that $V_t^{\mathbb{G}}(\boldsymbol{a}_{t-1}) \leq \hat{V}_t(\boldsymbol{a}_{t-1})$ for almost every scenario. Taking $\hat{V}_{T+1}(\boldsymbol{a}_T) = 0$ establishes the base case. Now assume the result

holds for period $t+1$. Following (15) we have

$$\begin{aligned} V_t^{\mathbb{G}}(\boldsymbol{a}_{t-1}) &= \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\Big[\, \mathbb{E}\Big[\, r_t(\boldsymbol{a}_t) + \hat{V}_{t+1}(\boldsymbol{a}_t) \,\Big|\, \mathcal{F}_t \,\Big] - \hat{V}_{t+1}(\boldsymbol{a}_t) + V_{t+1}^{\mathbb{G}}(\boldsymbol{a}_t) \,\Big|\, \mathcal{G}_t \,\Big] \\ &\leq \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\Big[\, \mathbb{E}\Big[\, r_t(\boldsymbol{a}_t) + \hat{V}_{t+1}(\boldsymbol{a}_t) \,\Big|\, \mathcal{F}_t \,\Big] \,\Big|\, \mathcal{G}_t \,\Big] \\ &= \sup_{a_t \in A_t(\boldsymbol{a}_{t-1})} \mathbb{E}\Big[\, r_t(\boldsymbol{a}_t) + \hat{V}_{t+1}(\boldsymbol{a}_t) \,\Big|\, \mathcal{F}_t \,\Big] \\ &\leq \hat{V}_t(\boldsymbol{a}_{t-1})\,, \end{aligned}$$

where the first inequality follows from the induction assumption and the second inequality follows from the fact that $\hat{V}_t$ is a supersolution. □

The two results of the proposition are closely related: for example, the Lagrangian relaxation and frictionless model considered in §7 and §8 are physical relaxations that satisfy the conditions of the first result and generate approximate value functions that satisfy the supersolution condition of the second result. The fact that the bound (19) holds in every scenario is a useful diagnostic for checking results when estimating bounds using Monte Carlo simulation and suggests that information bounds generated by using penalties based on good physical relaxations (or supersolutions) will yield high-quality, low-variance performance bounds.

The information relaxation approach also allows us to exploit structural properties of the optimal policy for the primal problem: 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 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 such structural constraints. We summarize this property as follows.

Proposition 3.3 (Structured Policies, BSS (2010))**.** *If for some $\mathcal{S} \subseteq \mathcal{A}$ we have* $\sup_{\alpha_F \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[\, r(\alpha_F)\,] = \sup_{\alpha_F \in \mathcal{S}_{\mathbb{F}}} \mathbb{E}[\, r(\alpha_F)\,]$ *then, for any dual feasible π, we have*

$$\sup_{\alpha_F \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[\, r(\alpha_F)\,] \leq \sup_{\alpha_G \in \mathcal{S}_{\mathbb{G}}} \mathbb{E}[\, r(\alpha_G) - \pi(\alpha_G)\,] \leq \sup_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}[\, r(\alpha_G) - \pi(\alpha_G)\,]\ . \tag{20}$$

Moreover, the inequalities also hold for all π such that $\mathbb{E}[\,\pi(\alpha_F)\,] \leq 0$ for all α_F in $\mathcal{S}_{\mathbb{F}}$.

Proof. The first inequality in (20) follows from applying the weak duality result (Theorem 3.1) with the restricted policy space $\mathcal{S}$ in place of the full policy space $\mathcal{A}$. Note that, by definition, any dual feasible penalty π for the original problem with $\mathcal{A}$ satisfies $\mathbb{E}[\,\pi(\alpha_F)\,] \leq 0$ for all α_F in $\mathcal{A}_{\mathbb{F}}$. Since $\mathcal{S} \subseteq \mathcal{A}$, any such penalty will also be dual feasible with a restricted policy space, i.e,

$\mathbb{E}[\,\pi(\alpha_{_F})\,] \leq 0$ for all $\alpha_{_F}$ in $\mathcal{S}_{\mathbb{F}}$. This set of dual feasible penalties in the restricted policy space is larger than the set of dual feasible penalties in original space. Thus this first inequality holds on the larger set of penalties that are dual feasible with the restricted penalties.

The second inequality in (20) follows from the fact that $\mathcal{S}_{\mathbb{G}} \subseteq \mathcal{A}_{\mathbb{G}}$. □

We will illustrate the use of this result in the assortment planning example of §7 where the retailer chooses N_t products to display in period t. In that example, if products are *a priori* identical, then in the first period it doesn't matter which products are displayed. Thus there is no loss in optimality in imposing a restriction that the first N_0 products (in index order) are displayed in the first period. With a relaxed filtration, the products may no longer be identical in the first period, but we can obtain tighter bounds by imposing this restriction.

In practice, there will often be a trade-off between the quality of the bound and the computational effort required to compute it. As discussed earlier (see Corollary 3.1 and discussion after), we can control this trade-off through our choice of information relaxation $\mathbb{G}$ and penalty. We can also sometimes use the following result to simplify the calculation of "good penalties" when $\mathbb{E}[\,w_t(\boldsymbol{a}_t)\,|\,\mathcal{F}_t\,]$ is difficult to evaluate.

Proposition 3.4 (Simplifying Good Penalties, BSS (2010))**.** *Let $\mathbb{F}'$ be filtration satisfying $\mathbb{F} \subseteq \mathbb{F}'$ and let $(w_0, \ldots, w_T)$ be a sequence of generating functions satisfying the conditions of Proposition 3.1. The penalty π given by $\pi_t(\boldsymbol{a}_t) = w_t(\boldsymbol{a}_t) - \mathbb{E}[\,w_t(\boldsymbol{a}_t)\,|\,\mathcal{F}'_t\,]$ satisfies the conclusions of Proposition 3.1 and thus is dual feasible (i.e., $\pi \in \Pi_{\mathbb{F}}$).*

Proof. From Lemma 3.1 (since $\alpha_{_F}$ being $\mathbb{F}$-adapted implies $\alpha_{_F}$ is also $\mathbb{F}'$-adapted) and the law of iterated expectations (since $\mathcal{F}_t \subseteq \mathcal{F}'_t$), we have

$$\mathbb{E}[\,\pi_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,] = \mathbb{E}[\,w_t(\boldsymbol{\alpha}_t)\,|\,\mathcal{F}_t\,] - \mathbb{E}\big[\,\mathbb{E}\big[\,w_t(\boldsymbol{\alpha}_t)\,\big|\,\mathcal{F}'_t\,\big]\,\big|\,\mathcal{F}_t\,\big] = 0 \quad \text{(almost surely).}$$

The rest of the proof then proceeds as in the proof of Proposition 3.1. □

This result can be helpful if the natural filtration includes elements that are partially observed. For instance in the option pricing example with stochastic volatility in BSS (2010), the volatility is assumed to be not observed in the natural filtration. A correct calculation of $\mathbb{E}[\,w_t(\boldsymbol{a}_t)\,|\,\mathcal{F}_t\,]$ would require keeping track of a probability distribution on volatility which would be updated over time using Bayes rule, based on observed stock prices and interest rates. Using the result above, we can simplify the computation by calculating penalties using a filtration $\mathbb{F}'$ that assumes the volatility is observed. A similar situation arises in the inventory management example with uncertainty about the "state of the world" (see §6.4 below) if the state is not fully observed.

Proposition 4.3 of BSS (2010) has some additional results about information relaxations and penalties. One result can be interpreted as a continuity property (implying, for example, that penalties that are "close" to the ideal penalty, yield bounds that are close to the optimal value). Another result shows that if we estimate penalties using simulation methods (e.g., as in Haugh and Kogan 2004; Andersen and Broadie 2004), we obtain estimates of the bounds that are weaker than the bounds given by using the penalty itself.

4. Convex Dynamic Programs

Though the results of §3 hold for all DPs, our focus in this section will be on the case where the DP or its approximating model has a convex structure. The dynamic portfolio optimization example of §8 is an example of a convex DP as are many problems using linear systems and concave rewards (or convex costs), such as linear-quadratic control problems (e.g., Bertsekas 2017, §4.2). We will follow Brown and Smith (2014*b*) (hereafter BS (2014)) and restrict our attention to perfect information relaxations and assume that the actions in each period a_t are vectors of real numbers, i.e., in $\mathbb{R}^{n_t}$ for some finite n_t.

A *convex dynamic program* is a DP where the reward functions $r_t(\boldsymbol{a}_t, \omega)$ are concave functions of the actions $\boldsymbol{a}_t$ for each ω and the feasible set of actions $\boldsymbol{A}(\omega)$ is convex for each ω. With a convex DP, the primal DP (1) can be viewed as a large convex optimization problem with decision variables corresponding to choices of actions $\boldsymbol{a}$ for each scenario ω and a concave objective function, a convex set of constraints $\boldsymbol{A}(\omega)$ for each scenario, and a large set of equality constraints that link actions across scenarios and represent the temporal feasibility constraints. We can also show by induction that for a convex DP, the optimal value functions V_t given by the Bellman recursion (2) will be concave in actions; see BS (2014).

With convex DPs, though the rewards are concave and constraint sets are convex, with penalties like (16) based on approximate value functions (or the true value function), the penalized objective $r(\boldsymbol{a}) - \hat{\pi}(\boldsymbol{a})$ involves differences of concave functions, may not be concave in $\boldsymbol{a}$ and, consequently, the resulting inner problem in (4) may be difficult to solve. A natural way to address this issue is to replace the penalties with a first-order linear approximation so these differences are linear and the objective in the inner problems will be concave. To simplify the discussion here, we will focus on the case where the approximate value functions are differentiable. In practice, many approximate value functions are nondifferentiable (e.g., arising from piecewise linear approximations); see BS (2014) for the nondifferentiable case. We also focus on the case where the period-t reward functions r_t

are $\mathbb{F}$-adapted so the ideal penalties (and approximations thereof) can be defined using generating functions (15) that do not include the reward functions. This is for notational convenience and is true in the portfolio optimization example in §7.

Assuming the approximate value functions $\hat{V}_t$ are concave and differentiable in actions, we can take a first-order linear approximation around the nonanticipative (or $\mathbb{F}$-adapted) policy $\hat{\alpha}$:

$$\hat{V}_{t+1}(\boldsymbol{a}_t) \approx \nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t)^\intercal (\boldsymbol{a}_t - \hat{\boldsymbol{\alpha}}_t) + \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \ ,$$

where $\nabla \hat{V}_{t+1}(\boldsymbol{a}_t)$ denotes the gradient of $\hat{V}_{t+1}(\boldsymbol{a}_t)$ with respect to the first $t+1$ actions, evaluated at $\boldsymbol{a}_t$ and $\hat{\boldsymbol{\alpha}}_t$ denotes the first $t+1$ actions selected under policy $\hat{\alpha}$. Note that $\hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t)$ is a random variable (written more explicitly as $\hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t(\omega), \omega)$), the gradients are evaluated for each ω, and the resulting approximation is a random variable for each action sequence $\boldsymbol{a}_t$. We can then use this approximation as a generating function, taking

$$w_t(\boldsymbol{a}_t) = \nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t)^\intercal (\boldsymbol{a}_t - \hat{\boldsymbol{\alpha}}_t) + \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \tag{21}$$

in Proposition 3.1 to generate the *gradient penalty*:

$$\begin{aligned} \hat{\pi}_\nabla(\boldsymbol{a}) = \sum_{t=0}^{T} \Bigg(& \left(\nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) - \mathbb{E}\Big[\nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \Big] \right)^\intercal (\boldsymbol{a}_t - \hat{\boldsymbol{\alpha}}_t) \\ & + \left(\hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) - \mathbb{E}\Big[\hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \Big] \right) \Bigg) \ . \end{aligned} \tag{22}$$

(We use the assumption that $\hat{\boldsymbol{\alpha}}_t$ is $\mathcal{F}_t$-measurable to move $\hat{\boldsymbol{\alpha}}_t$ outside of the expectation.) This penalty is affine in actions $\boldsymbol{a}$ and, given a problem with concave rewards and convex action sets, the inner problem (4) with this penalty is a convex optimization problem. The final terms (inside the parentheses) are constant with respect to $\boldsymbol{a}$ and play the role of control variates, similar to the last terms in (17): they have zero mean and thus do not affect the expected value in the bound (4). However, these terms may be correlated with the reward terms in (4) and, as discussed in §3.2, including them in the penalty may help reduce the variance when estimating the bounds using Monte Carlo simulation.

What is striking about these gradient penalties is that the linear approximation, in principle, entails no loss in functionality when working with convex DPs. The gradient penalties are dual feasible (by construction) and hence generate valid bounds by weak duality (Theorem 3.1). Strong duality also holds: there exists a gradient penalty that generates a zero-variance, tight bound.

Moreover, when working with an approximate value function from a relaxed model that is a convex DP, the gradient penalty will improve on the bound given by the relaxed model in every scenario. We formalize these results for the differentiable case as follows.

Proposition 4.1 (Properties of Gradient Penalties, BS (2014))**.** *Suppose the approximate value functions $\hat{V}_t$ are concave in actions and differentiable. Let $\hat{\pi}_\nabla$ denote the gradient penalty defined by linearizing $\hat{V}_t$ around a $\mathbb{F}$-adapted policy $\hat{\alpha}$ as in* (22)*.*

(i) (Strong Duality) *If the original model is a convex DP and the approximate value functions $\hat{V}_t$ and policies $\hat{\alpha}$ are the optimal value functions and an optimal policy for this model, then, for almost every scenario ω,*

$$\max_{\boldsymbol{a} \in \boldsymbol{A}(\omega)} \{r(\boldsymbol{a}, \omega) - \hat{\pi}_\nabla(\boldsymbol{a}, \omega)\} = V_0 \ .$$

(ii) (Improving Bounds from Other Relaxations) *If the approximate value functions $\hat{V}_t$ are the optimal value functions for a relaxed model that is a convex DP with $\boldsymbol{A} \subseteq \hat{\boldsymbol{A}}$ and $r \leq \hat{r}$ and $\hat{\alpha}$ is an optimal policy for this relaxed model, then, for almost every scenario ω:*

$$\max_{\boldsymbol{a} \in \boldsymbol{A}(\omega)} \{r(\boldsymbol{a}, \omega) - \hat{\pi}_\nabla(\boldsymbol{a}, \omega)\} \leq \hat{V}_0 \ .$$

Note that the results above hold "pathwise" (i.e., for almost every scenario ω), which implies the dual bounds given by taking expectations over scenarios,

$$\mathbb{E}\left[\max_{\boldsymbol{a} \in \boldsymbol{A}(\omega)} \{r(\boldsymbol{a}, \omega) - \pi_\nabla(\boldsymbol{a}, \omega)\} \right] \ ,$$

will be equal to V_0 in part (i) and less than or equal to $\hat{V}_0$ in part (ii).

Proof. (i) To simplify the discussion, we will assume that the action choices are unconstrained; see BS (2014) for a full proof. Consider a gradient penalty $\hat{\pi}_\nabla$ defined by linearizing $\hat{V}_t$ around policy $\hat{\alpha}$, as in (22). If we omit the terms inside the parentheses that are constant in actions (which, as discussed earlier, serve as control variates), the inner problem for a given scenario reduces to

$$\begin{aligned} &\max_{\boldsymbol{a}} \left\{ \sum_{t=0}^{T} r_t(\boldsymbol{a}_t) - \left(\nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) - \mathbb{E}\Big[\nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \Big] \right)^\intercal (\boldsymbol{a}_t - \hat{\boldsymbol{\alpha}}_t) \right\} \qquad (23) \\ = &\max_{\boldsymbol{a}} \left\{ \sum_{t=0}^{T} r_t(\boldsymbol{a}_t) - \left(\begin{pmatrix} \nabla \hat{V}_t(\hat{\boldsymbol{\alpha}}_{t\text{-}1}) \\ \mathbf{0} \end{pmatrix} - \mathbb{E}\Big[\nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \Big] \right)^\intercal (\boldsymbol{a}_t - \hat{\boldsymbol{\alpha}}_t) \right\} . \end{aligned}$$

Here, in rearranging terms, we use the fact that $\hat{V}_{T+1} = 0$ and thus $\nabla \hat{V}_{T+1} = \mathbf{0}$. In this expression, $\nabla \hat{V}_t$ has dimension corresponding to $\boldsymbol{a}_{t\text{-}1}$ and, hence, its gradient needs to be padded with a $\mathbf{0}$ of the dimension of a_t to match the dimensionality of $\nabla \hat{V}_{t+1}$, which corresponds to $\boldsymbol{a}_t$.

Now, if $\hat{\alpha}$ is an optimal policy and $\hat{V}_t$ are the optimal value functions and the choices of actions are unconstrained, we know that

$$\hat{V}_t(\hat{\boldsymbol{\alpha}}_{t\text{-}1}) = r_t(\hat{\boldsymbol{\alpha}}_t) + \mathbb{E}\Big[\, \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big] = \max_{a_t} \Big\{ r_t(\hat{\boldsymbol{\alpha}}_{t\text{-}1}, a_t) + \mathbb{E}\Big[\, \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_{t\text{-}1}, a_t) \,\Big|\, \mathcal{F}_t \,\Big] \Big\} \ . \tag{24}$$

The "envelope theorem" and the first-order conditions for optimality then imply

$$\begin{pmatrix} \nabla \hat{V}_t(\hat{\boldsymbol{\alpha}}_{t\text{-}1}) \\ \mathbf{0} \end{pmatrix} = \nabla r_t(\hat{\boldsymbol{\alpha}}_t) + \mathbb{E}\Big[\, \nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big] \ . \tag{25}$$

Using this "consistency condition," we can rewrite the reduced inner problem (23) as

$$\max_{\boldsymbol{a}} \left\{ \sum_{t=0}^{T} r_t(\boldsymbol{a}_t) - \nabla r_t(\hat{\boldsymbol{\alpha}}_t)^{\intercal} (\boldsymbol{a}_t - \hat{\boldsymbol{\alpha}}_t) \right\} ,$$

which, given the concavity of r_t, is maximized by taking $\boldsymbol{a}_t = \hat{\boldsymbol{\alpha}}_t$ for all t. Thus, the reduced inner problem (23) yields an optimal value of $\sum_{t=0}^{T} r_t(\hat{\boldsymbol{\alpha}}_t)$. Using this and incorporating the control variate terms that were omitted in the reduced inner problem (23), the inner problem in this case is

$$\begin{aligned} \max_{\boldsymbol{a}} \{ r(\boldsymbol{a}) - \pi_{\nabla}(\boldsymbol{a}) \} &= \sum_{t=0}^{T} r_t(\hat{\boldsymbol{\alpha}}_t) - \Big(\hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) - \mathbb{E}\Big[\, \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big] \Big) \\ &= \hat{V}_0 + \sum_{t=0}^{T} r_t(\hat{\boldsymbol{\alpha}}_t) - \hat{V}_t(\hat{\boldsymbol{\alpha}}_{t\text{-}1}) + \mathbb{E}\Big[\, \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big] \\ &= \hat{V}_0 \ . \end{aligned}$$

Here, in the second equality, we use $\hat{V}_{T+1} = 0$ and rearrange terms. In the third equality, we use the fact that $\hat{\boldsymbol{\alpha}}_t$ is optimal (so the first equality in (24) holds). Thus using a gradient penalty based on the optimal value function will generate a zero-variance, tight bound.

(ii) The second result follows from the first result as in Proposition 3.2(i) above. □

In practice, with gradient penalties based on approximate value functions, the optimality conditions (24) and (25) in the proof of optimality above may be approximated and the quality of the resulting bounds will depend on the quality of the approximations. For a gradient penalty to provide good bounds, it is important that the linear approximations of the approximate value functions closely approximate the differences in the true value functions, i.e.,

$$\begin{aligned} &V_{t+1}(\boldsymbol{a}_t) - \mathbb{E}[\, V_{t+1}(\boldsymbol{a}_t) \,|\, \mathcal{F}_t \,] \\ &\approx \Big(\nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) - \mathbb{E}\Big[\, \nabla \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big] \Big)^{\intercal} (\boldsymbol{a}_t - \hat{\boldsymbol{\alpha}}_t) + \Big(\hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) - \mathbb{E}\Big[\, \hat{V}_{t+1}(\hat{\boldsymbol{\alpha}}_t) \,\Big|\, \mathcal{F}_t \,\Big] \Big) \ . \end{aligned}$$

In particular, it is important for the difference in gradients to approximate

$$\nabla V_{t+1}(\boldsymbol{\alpha}_t) - \mathbb{E}[\,\nabla V_{t+1}(\boldsymbol{\alpha}_t) \,|\, \mathcal{F}_t\,]$$

well. In this case, the optimal solutions in the inner problem will match or closely approximate those of the true optimal solutions. Errors in the constant terms $(V_{t+1}(\boldsymbol{\alpha}_t) - \mathbb{E}[\,V_{t+1}(\boldsymbol{\alpha}_t) \,|\, \mathcal{F}_t\,])$ are less important, as they will average zero when calculating the bounds.

The result of Proposition 4.1 generalizes directly to the setting of nondifferentiable value functions but there is an added complication of selecting appropriate gradients: with nondifferentiable value functions, there may be multiple (super)gradients. Any gradient selection will generate a valid bound. However, the equalities of Proposition 4.1 will hold for some gradient selection, but not all gradient selections. With nondifferentiable value functions, the choice of gradients can have a significant impact on the quality of the resulting information relaxation performance bounds. See BS (2014) for further discussion of the nondifferentiable case and example applications.

5. Summary of the Information Relaxation Approach

Before turning to examples, we informally summarize the steps involved in the information relaxation approach. Given a DP, we:

(i) 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.

(ii) 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 which leads to deterministic inner problems, though in some problems there may be other natural starting points.

(iii) Find a penalty that does not greatly complicate the calculation of optimal decisions with the chosen information relaxation.

  (a) We can start with zero penalty, but this often leads to weak upper bounds.

  (b) Identify an approximate value function that can be used as a generating function in Proposition 3.1. As discussed following that proposition, the key to obtaining a good bound is for the differences in (16) to provide a good approximation of the differences

$r_t(\boldsymbol{a}_t) + V^*_{t+1}(\boldsymbol{a}_t) - \mathbb{E}\big[\, r_t(\boldsymbol{a}_t) + V^*_{t+1}(\boldsymbol{a}_t) \,\big|\, \mathcal{F}_t \,\big]$ based on the true value function V^*_t. One can often use "myopic" approximations of the value function, such as the smoothing penalty in the inventory example of §6 and the dynamic assortment problem of §7. Simple linear approximations of value functions can often be used to generate penalties when the DP has a convex structure, as discussed in §4.

(c) If the simple (e.g., myopic) approximations do not perform well, consider approximations that are more "forward-looking," taking into account the expected rewards over a longer time horizon. This could be accomplished by adopting a longer time horizon in the approximation or by adding an approximation of these longer-term effects to the myopic approximation.

(d) Alternatively, if the value function approximations are based on an approximate DP where the value functions are approximated by linear combinations of basis functions, one might attempt to improve the approximation by considering a different set of basis functions.

(iv) Estimate lower and upper bounds on the optimal value. We will typically estimate the upper and lower bounds simultaneously in a single simulation, as many of the calculations involved in implementing a heuristic are also used for the bound.

(v) If the gap between the performance of the heuristic and information relaxation performance bound 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 (as discussed following Theorem 3.3) and see if these suggest some ideas for improving the heuristic policies, relaxations, or penalties. We may, for example, turn to better approximate value functions as discussed in step (iii) above.

Though we have described this as a sequential process, there are links between the steps in the process:

- The difficulty of solving the inner problem with a given information relaxation is often linked to the choice of penalty. For example, in the dynamic assortment problem of §7, the "censored demands" relaxation leads to efficient computations in the inner problem with zero penalty, but not with the other penalties we consider. In addition, we sometimes simplify the inner problems by introducing additional relaxations (e.g., relaxing the constraints, perhaps with Lagrangian techniques as we do with the dynamic assortment problem) to improve the

tractability of the inner problem.

- The choice of heuristic (in step (i)) and approximate value function for use in generating a penalty (in step (iii)) are often paired. For example, a heuristic that is a myopic policy may be paired with a myopic value function to generate a penalty; more generally we may use a heuristic that is "greedy" with respect to an approximate value function and use that approximate value function to form a penalty. We do this in the dynamic assortment example of §7, where a Lagrangian relaxation of the primal DP leads to an approximate value function used to generate a heuristic as well as a penalty.

There is clearly some "art" in applying information relaxation methods and, in the authors' experience, there is often some trial and error and iteration (and learning!) in this process. In the next three sections, we will study examples in detail and discuss issues involved in using information relaxation techniques in these examples.

6. Example: Inventory Management

In this section, we apply the information relaxation approach to a classic inventory management problem. We begin in §6.1 by considering a standard inventory model where the only uncertainties are the demands observed in each period; we focus on the perfect information relaxation. In §6.4, we consider a "world-driven model" where there is uncertainty about the costs and demand processes (as well as demands) and consider imperfect as well as perfect information relaxations. These examples are simple enough that the primal DP can be solved to optimality; thus we do not need the performance bounds in this problem. However, this simple setting provides an introduction to the use of information relaxation methods and allows us to compare the information relaxation performance bounds with optimal values to assess the quality of the bounds.[3]

6.1 Standard Model

The goal is to find a policy for ordering goods over $T+1$ periods to minimize the expected total costs. The inventory level in period t is denoted by x_t and the amount ordered in period t is a_t. The demand realized in period t is uncertain and denoted by d_t. We assume the order quantities and demands are nonnegative and, for convenience, assume $x_0 = 0$. The inventory level evolves

[3]This inventory management example was originally developed in the preparation of BSS (2010) but did not appear in the published version of that paper. The authors gratefully acknowledge the contributions of Peng Sun in the development of this example.

according to $x_{t+1} = x_t + a_t - d_t$; this evolution equation assumes that unmet demand is backordered and appears as a negative inventory level entering the next period. The total cost in period t is $c_t(a_t) + f_t(x_{t+1})$ where $c_t(a_t)$ is the cost of ordering a_t units of the good and $f_t(x_{t+1})$ is the cost of holding excess inventory (if x_{t+1} is positive) or having backordered demand (if x_{t+1} is negative) at the end-of-period t. In most of our discussion we will consider general cost functions c_t and f_t, but we will also consider the special case of "linear costs" where $c_t(a_t) = k_t a_t$ and $f_t(x_{t+1}) = \max(h_t x_{t+1}, -p_t x_{t+1})$.

Placing this model in the general framework of §2, the actions a_t are order quantities that may be restricted to be nonnegative integers or particular lot sizes. The scenarios ω correspond to realized demands $\boldsymbol{d} = (d_0, \ldots, d_T)$ and $\Omega \subseteq \mathbb{R}_+^{T+1}$. In the primal problem, we assume that at the beginning of period t the DM knows the prior demands $\boldsymbol{d}_{t-1} = (d_0, \ldots, d_{t-1})$; the period-$t$ demand d_t is revealed after the period-t ordering decision. We define the natural filtration $\mathbb{F}$ accordingly. Rather than maximizing expected rewards as in (1), here we minimize costs where the period-t cost is given by

$$r_t(\boldsymbol{a}_t, \boldsymbol{d}_t) = c_t(a_t) + f_t(x_{t+1}(\boldsymbol{a}_t, \boldsymbol{d}_t))$$

Note that the period costs here are $\mathcal{F}_{t+1}$-adapted but are not $\mathcal{F}_t$-adapted because of their dependence on the period-t demand. Also note that the inventory level x_{t+1} depends on all prior orders $\boldsymbol{a}_t$ and demands $\boldsymbol{d}_t$. The total cost is then $r(\boldsymbol{a}, \boldsymbol{d}) = \sum_{t=0}^{T} r_t(\boldsymbol{a}_t, \boldsymbol{d}_t)$ and the primal DP can be written as

$$\inf_{\alpha_F \in \mathcal{A}_{\mathbb{F}}} \mathbb{E}[\, r(\alpha_F)\,] \ .$$

6.2 Information Relaxations and Penalties

We first consider the relaxed problem given by taking $\mathbb{G}$ to be the perfect information relaxation $\mathcal{G}_t = \mathcal{F}$ for all t, corresponding to knowing all demands before making any ordering decisions. With a dual feasible penalty π, the performance bound (3) is

$$\mathbb{E}\left[\inf_{\boldsymbol{a} \in \boldsymbol{A}} \{ r(\boldsymbol{a}, \boldsymbol{d}) - \pi(\boldsymbol{a}, \boldsymbol{d}) \} \right] \tag{26}$$

where $\boldsymbol{A}$ denotes set of feasible orders $\boldsymbol{a} = (a_0, \ldots, a_T)$. Here, because we are minimizing costs rather than maximizing profits, dual feasible penalties must satisfy $\mathbb{E}[\, \pi(\alpha_F)\,] \geq 0$ (rather than ≤ 0) for all α_F in $\mathcal{A}_{\mathbb{F}}$ and the resulting performance bound is a lower bound on the minimal costs.

If we use the penalty $\pi = 0$, we obtain the perfect information lower bound on costs and the

inner problem in (26) corresponds to choosing an optimal ordering policy with full information about all demands. This inner problem is known as a dynamic lot-sizing problem with backorders (see, e.g., Florian et al. 1980) where we solve

$$\inf_{a \in A} \sum_{t=0}^{T} \{c_t(a_t) + f_t(x_{t+1}(\boldsymbol{a}_t, \boldsymbol{d}_t))\} \tag{27}$$

with a known demand sequence $\boldsymbol{d}$. This inner problem can be solved efficiently as a deterministic DP or using specialized algorithms for dynamic lot-sizing problems. We can estimate a lower bound on the minimal costs by repeatedly solving these dynamic lot-sizing problems with randomly generated demand sequences.

In the special case with linear costs, the perfect information bounds simplify further: the inner dynamic lot-sizing problem (27) can be formulated and solved as a shortest path problem (see, e.g., Ahuja et al. 1988). The solution leads to the determination of cost coefficients $\hat{k}_t$ that are independent of the specific demand levels. The total cost given a demand sequence $\boldsymbol{d}$ is then given by $\sum_{t=0}^{T} \hat{k}_t d_t$. Thus, in the linear case, given these cost coefficients $\hat{k}_t$, the perfect information lower bound can be written analytically as $\sum_{t=0}^{T} \hat{k}_t \mathbb{E}[\, d_t \,]$. If the ordering costs are constant over time, these cost coefficients $\hat{k}_t$ are simply the ordering costs k_t. However, if the ordering costs are time-varying, it may be cheaper to satisfy demand in a given period by ordering before (or after) that period and paying the holding costs (or backorder costs). The coefficients $\hat{k}_t$ represent the cheapest way to meet the known demand in period t.

Although these perfect information bounds are easy to compute, as we will see, the bounds are rather loose. To obtain tighter bounds, we consider a "smoothing" penalty that is straightforward to compute and partially cancels the benefits of knowing demands in advance. We define this penalty using the method of Proposition 3.1, taking the generating functions w_t to be the period rewards r_t. This leads to a period-t penalty

$$\pi_t(\boldsymbol{a}_t, \boldsymbol{d}_t) = r_t(\boldsymbol{a}_t, \boldsymbol{d}_t) - \mathbb{E}[\, r_t(\boldsymbol{a}_t, \boldsymbol{d}_t) \,|\, \mathcal{F}_t \,],$$

and the (total) penalty $\pi(\boldsymbol{a}, \boldsymbol{d}) = \sum_{t=0}^{T} \pi_t(\boldsymbol{a}, \boldsymbol{d})$. Here the conditional expectations $\mathbb{E}[\, \cdot \,|\, \mathcal{F}_t \,]$ are taken knowing the demands $\boldsymbol{d}_{t-1} = (d_0, \ldots, d_{t-1})$ in the first t periods. Intuitively, this penalty cancels the benefit of precise knowledge of the demand in a given period, as the exact period-t reward $r_t(\boldsymbol{a}_t, \boldsymbol{d}_t)$ is replaced by the expected reward $\mathbb{E}[\, r_t(\boldsymbol{a}_t, \boldsymbol{d}_t) \,|\, \mathcal{F}_t \,]$. However, this cancellation is "myopic" in that, unlike the ideal penalty constructed in Theorem 3.4, it cancels the effect of

perfect information on a single period's reward rather than the effect on the total rewards.

With this smoothing penalty, the inner dual problem in (26) becomes

$$\inf_{\boldsymbol{a}\in\boldsymbol{A}} \sum_{t=0}^{T} \{c_t(a_t) + \mathbb{E}[\, f_t(x_{t+1}(\boldsymbol{a}_t, \boldsymbol{d}_t)) \,|\, \mathcal{F}_t \,]\} \tag{28}$$

which can be written more explicitly as

$$\inf_{(a_0,\ldots,a_T)\in\boldsymbol{A}} \sum_{t=0}^{T} \left\{ c_t(a_t) + \mathbb{E}\left[f_t\left(\sum_{\tau=0}^{t-1} (a_\tau - d_\tau) + (a_t - \tilde{d}_t) \right) \middle| \boldsymbol{d}_{t-1} \right] \right\} .$$

In this case, the inner problem can still be viewed as a dynamic lot-sizing problem as in (27) but the inventory cost function $f_t(x_{t+1})$ is replaced by a smoothed version of it, $\hat{f}_t(x_{t+1}) = \mathbb{E}[\, f_t(x_{t+1}) \,|\, \mathcal{F}_t \,]$, that takes expectation over the uncertain demand $\tilde{d}_t$ given a particular earlier demand sequence $\boldsymbol{d}_{t-1}$. As with the perfect information bound, we can estimate a lower bound on the optimal costs by repeatedly solving these dynamic lot-sizing problems with randomly generated demand sequences. Note that if the demands are independent over time, these smoothed cost functions $\hat{f}_t$ need only be calculated once and stored. However, with dependence in the demands, we need to consider different smoothed cost functions $\hat{f}_t$ for different demand scenarios.

The smoothing bound can also be simplified if we have linear costs. If we drop the requirement that the order quantity a_t be nonnegative (i.e., relax this physical constraint), we obtain a weaker bound but the inner dual problem decomposes into a series of simple newsvendor problems; we call the resulting bound the newsvendor bounds. To describe this, rather than choosing order quantities a_t, we instead (equivalently) choose base stock levels $y_t = a_t + x_t$. The end-of-period inventory levels are then given by $x_{t+1} = y_t - d_t$. If we drop the nonnegativity constraint on the order quantity a_t, our choice of base stock levels is unconstrained and the inner problem in (28) can be rewritten as

$$\begin{aligned} &\min_{\boldsymbol{y}} \sum_{t=0}^{T} \left\{ k_t(y_t - y_{t-1} + d_{t-1}) + \mathbb{E}\Big[\, f_t(y_t - \tilde{d}_t) \,\Big|\, \mathcal{F}_t \,\Big] \right\} \\ &= \sum_{t=0}^{T} \left\{ k_t d_{t-1} + \min_{y_t} \left(y_t(k_t - k_{t+1}) + \mathbb{E}\Big[\, f_t(y_t - \tilde{d}_t) \,\Big|\, \boldsymbol{d}_{t-1} \,\Big] \right) \right\}, \end{aligned} \tag{29}$$

where, on the right side, we have gathered all terms involving y_t into a single summand; we take $d_{-1} = 0$ and $k_{T+1} = 0$. Examining (29), we see that the inner problem requires the solution of period-specific minimization problems that have the same form as the classical newsvendor problem:

		Time-Varying Demand and Constant Costs					Constant Demand and Time-Varying Costs				
Period	t	0	1	2	3	4	0	1	2	3	4
Ordering costs	k_t	2	2	2	2	2	7	8	3	4	1.5
Holding costs	h_t	1	1	1	1	-1	1	1	1	1	-1
Backorder costs	p_t	9	9	9	9	11	9	9	9	9	11
Mean demand	$\mathbb{E}[\,d_t\,]$	40	40	40	2	2	30	30	30	30	30

Table 1: Assumptions for the standard inventory model example

the optimal base-stock levels y_t in (29) are given by selecting an appropriate "critical fractile" as in the newsvendor model. We can then take expectations over (29) to obtain a lower bound on the total expected costs. If the demands are independent over time, we can calculate these critical fractiles and expectations once: no simulation is required to calculate the expectation of (29). If demands are dependent, we need to find the critical fractiles for period t conditional on the prior demands $\boldsymbol{d}_{t-1}$.[4]

6.3 Example Numerical Results

We illustrate these bounds by considering their performance in four examples from Zipkin (2000, pp. 380-381). The examples involve two different sets of assumptions – one with time-varying demand distributions and constant costs and the other with time-varying costs and constant demand distributions – and consider Poisson and geometric demand distributions. The examples all involve 5 periods and assume a linear cost model; the specific assumptions are shown in Table 1. With both sets of assumptions, the final period inventory costs can be viewed as including a salvage value where excess inventory is sold for \$2 per unit and unmet demand must be satisfied at \$2 per unit.

Table 2 shows the results for the various bounds as well as optimal values; since demands are independent and discrete in these examples, the primal DP recursion has a finite, one-dimensional state space (representing the current inventory level) and is easy to solve. The smoothing bounds were computed using Monte Carlo simulation with 10,000 scenarios; standard errors are reported for these results. The other bounds were computed numerically without simulation.

Examining the results in Table 2, we see that:

- The perfect information bounds, though easy to compute, are quite loose, as expected. With perfect information and no penalty, the DM simply orders to match the actual demand in

[4]With independent demands, this newsvendor bound is equivalent to a bound given in Zipkin (2000, p. 381).

each period, using the cheapest method available to do so (with costs given by $\hat{k}_t$ as discussed earlier).

- The newsvendor bounds are also easy to compute and are tight when demands have a Poisson distribution. With a geometric demand distribution, however, the newsvendor bounds are 11-12% below the optimal value. In the newsvendor bounds, the DM orders the newsvendor optimal quantities in each period, ignoring the fact that this is not possible when the leftover inventory exceeds the newsvendor-optimal quantity.

- The smoothing bound is somewhat harder to compute but does very well in all cases. The nonnegative order quantity constraints are respected in these inner problems and thus the bounds are better than the newsvendor bounds.

The differences in performance in the Poisson and geometric cases are due to the differences in variances: with a mean demand of 30 (as in the constant demand case), the standard deviations of the Poisson and geometric distributions are 5.48 and 30.50, respectively. The higher variance of the geometric distribution leads to higher inventory costs and leads the perfect information and newsvendor bounds to perform much worse. In this case, the constraint that is ignored in the newsvendor bound (that the order quantities must be nonnegative) is more likely to be binding and the newsvendor bounds are much worse than the smoothing bounds.

6.4 With Uncertainty about the "State of the World"

A useful feature of the information relaxation approach is that it can easily accommodate more complex inventory models where there are uncertainties about costs, demand processes, and/or dependence among the demands over time. Though these additional uncertainties may make the primal DP significantly more complicated to solve, we can easily incorporate these uncertainties into the relaxed problem. We illustrate this by considering a "world-driven demand" inventory model (as in Song and Zipkin 1993) where the model structure is the same as in §6.1 except now the cost and demand parameters depend on an underlying "state of the world" s_t which evolves stochastically over time.

As a specific example, we will consider an extension of the previous numerical example but with uncertainty about the ordering costs (k_t) and the mean demand ($\mathbb{E}[\,d_t\,]$). There are three possible cost levels and three possible mean demands and these evolve independently with the values and transition probabilities specified in Figure 1; we assume the costs and mean demands both start in period 0 in the middle state. Considering all possible combinations of costs and mean

Poisson Demand Distributions

	Time-Varying Demand and Constant Costs			Constant Demand and Time-varying Costs		
	Cost	Std. Err.	% of Opt.	Cost	Std. Err.	% of Opt.
Optimal value	293.94	-	-	752.45	-	-
Smoothing bound	292.58	0.20	100%	752.42	0.53	100%
Newsvendor bound	287.84	-	98%	752.37	-	100%
Perfect info. bound	248.00	-	84%	705.00	-	94%

Geometric Demand Distributions

	Time-Varying Demand and Constant Costs			Constant Demand and Time-Varying Costs		
	Cost	Std. Err.	% of Opt.	Cost	Std. Err.	% of Opt.
Optimal value	607.14	-	-	1092.10	-	-
Smoothing bound	599.85	1.22	99%	1068.53	2.56	98%
Newsvendor bound	538.99	-	89%	956.17	-	88%
Perfect info. bound	248.00	-	41%	705.00	-	65%

Table 2: Example bounds for the standard inventory model

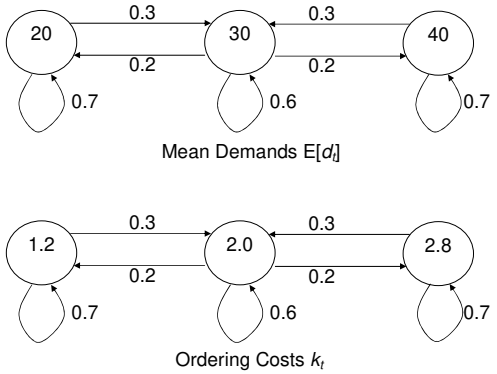

Figure 1: State transitions for the world-driven inventory model

demands gives a total of nine possible states of the world s_t. The holding and backorder costs (h_t and p_t) are the same as in the previous example (see Table 1) and, as before, we will consider Poisson and geometric demand distributions. Because the ordering costs k_t depend on the state s_t, the inventory rewards now depend on the state vector $\boldsymbol{s}_t = (s_0, \ldots, s_t)$ and can be written as $r_t(\boldsymbol{a}_t, \boldsymbol{d}_t, \boldsymbol{s}_t) = k_t(s_t)a_t + f_t(x_{t+1}(\boldsymbol{a}_t, \boldsymbol{d}_t))$. We assume that in period t (before placing that period's order) the DM knows the current state s_t as well as the prior demands $\boldsymbol{d}_{t-1}$ and states $\boldsymbol{s}_{t-1}$ and define the natural filtration $\mathbb{F}$ accordingly.

We will consider two relaxations of the natural filtration $\mathbb{F}$. First, as in §6.2, we consider

	Poisson Demands			Geometric Demands		
	Cost	Std. Err.	% of Opt.	Cost	Std. Err.	% of Opt.
Optimal value	348.69	-	-	644.45	-	-
Imperfect info. bound	346.15	0.75	99.3%	636.65	1.12	98.6%
Smoothing bound	347.23	0.79	99.6%	634.82	1.69	98.4%
Newsvendor bound	347.10	-	99.5%	628.43	-	97.4%
Perfect info. bound	300.00	-	86.0%	300.00	-	46.5%

Table 3: Numerical results for the world-driven demand example

the perfect information relaxation, which now means the DM has perfect knowledge of the actual demands $\boldsymbol{d}$ and states $\boldsymbol{s}$ before placing any orders. In this setting, the three bounds considered previously continue to apply in the same way. The perfect information bound given by taking $\pi = 0$ yields an inner problem that is a deterministic lot-sizing problem as in (27). The smoothing bound is given by taking the period-t penalty to be

$$\pi_t(\boldsymbol{a}_t, \boldsymbol{d}_t, \boldsymbol{s}_t) = r_t(\boldsymbol{a}_t, \boldsymbol{d}_t, \boldsymbol{s}_t) - \mathbb{E}[\, r_t(\boldsymbol{a}_t, \boldsymbol{d}_t, \boldsymbol{s}_t) \,|\, \mathcal{F}_t \,]$$

where the conditional expectations $\mathbb{E}[\, \cdot \,|\, \mathcal{F}_t \,]$ are taken knowing the period-t state and demand histories $\boldsymbol{s}_t$ and $\boldsymbol{d}_{t-1}$; this leads to an inner problem that is the same as (28). Finally, this smoothing bound can be simplified to yield a newsvendor bound with an inner problem like that of equation (29). In this case, however, because the mean demand and the ordering costs change in each state s_t, we take expectations over (29) to calculate the newsvendor bounds. The numerical results for these three bounds are shown in Table 3, along with the exact value for the primal problem. As with the standard inventory model, we find that the perfect information bounds with no penalty are rather loose, but the newsvendor and smoothing bounds are tight with the Poisson demand distributions and fairly tight with the geometric demand distribution.

Also shown in Table 3 are the results for an *imperfect information* bound that considers a relaxation $\mathbb{G}$ of the natural filtration $\mathbb{F}$ that assumes perfect information about all states $(s_0, \ldots, s_t)$ before any decisions are made, but assumes demand information is revealed sequentially as in the natural filtration $\mathbb{F}$. With no penalty, we have a dual problem that can be solved by repeatedly randomly generating a vector of states $\boldsymbol{s}$ and solving an inner problem that is standard stochastic inventory model with known costs and demand distributions. The inner problem is thus a stochastic DP, but it is simpler to solve than the primal DP that reflects uncertainty about the states s_t. The computational effort required to calculate these bounds will depend on how many scenarios one chooses to simulate, but the effort required for any given scenario is otherwise independent of the

number of possible states s_t of the world considered in the primal DP. As seen in Table 3, these bounds with zero penalty are tighter with the imperfect information relaxation than the perfect information relaxation, as one would expect based on Corollary 3.1. One could incorporate nonzero penalties into these imperfect information bounds, but these imperfect information bounds are quite tight and we have little incentive to attempt to improve them.

We can modify the primal DP for this world-driven inventory model by assuming that the cost component of the state is observed but the demand component of the state (representing the mean demand) is not observed before ordering in period t. In this case, the primal DP would be a partially observed Markov decision process: the DP state variable would include a probability distribution over the possible demand states and the DM would update his or her probability distribution on these states based on the observed demands. This modification would significantly complicate the primal DP and, with less information available to the DM, would lead to costs somewhat higher than the optimal value shown in Table 3. This modification is easily accommodated in the information relaxation dual problem. Specifically, we can use the result of Proposition 3.4 to simplify the calculation of penalties in the unobservable model ($\mathbb{F}$) by using a looser information relaxation ($\mathbb{F}'$) that assumes the demand components and the cost components are observed over time. This leads to the bounds we previously considered: though the optimal costs for the partially observed model would be higher than the costs for the fully observed model, the bounds we calculate in Table 3 would be unchanged and provide performance bounds for this partially observable model.

BSS (2010) present bounds for the adaptive inventory management model of Treharne and Sox (2002) where demand is nonstationary and partially observed, meaning the probability distribution for demand changes over time and the true demand distribution is not known. For the reasons discussed in the previous paragraph, the partially observed primal problem there is much harder than the simple inventory examples in §6.1-§6.4. BSS (2010) consider penalties based on limited-lookahead value functions, generalizing the myopic "smoothing" penalties considered in this simple inventory example.

7. Example: Dynamic Assortment Planning

In this section, we use information relaxation methods to study a dynamic assortment problem (DAP) with demand learning, using the model developed by Caro and Gallien (CG 2007) and the heuristics from Brown and Smith (BS 2020). This example is significantly more complicated than the previous one: it cannot be solved to optimality and the heuristics rely on Lagrangian

relaxations of the original DP, as does the inner DP considered in the information relaxation bounds. We describe the model in §7.1, a Lagrangian relaxation and associated heuristic in §7.2, and information relaxation bounds in §7.3. §7.4 provides illustrative numerical results. BS (2020) focus on Lagrangian relaxations and corresponding Lagrangian index policies, showing that the Lagrangian index policies are asymptotically optimal given many products. Here we focus on using information relaxations to generate performance bounds for a fixed number of products.

7.1 DAP: The Model

We consider a retailer who repeatedly chooses products to display from a set of S products available, subject to a shelf-space constraint that requires the number of products displayed in period t to be less than or equal to N_t. The demand rate for products is unknown and the DM updates beliefs about these rates over time using Bayes' rule. The demand for product s follows a Poisson distribution with an unknown rate γ_s. The demand rates are assumed to be independent across products and to have a gamma prior with shape parameter m_s and inverse scale parameter β_s ($m_s, \beta_s > 0$); this implies the mean of γ_s (and hence mean demand) is m_s/β_s.

The assumed distributions are convenient because they lead to nice forms for the demand distribution and Bayesian updating is easy. If a product is displayed, the observed demand in that period has a negative-binomial distribution (also known as the gamma-Poisson mixture). Given a gamma prior with parameters (m_s, β_s), after observing demand d_s for product s, the posterior distribution for the demand rate is a gamma distribution with parameters $(m_s + d_s, \beta_s + 1)$. If a product is not displayed, its state is unchanged.

The retailer's objective is to maximize the expected total profit earned. If a product is displayed, its reward for that period is assumed to be equal to the demand d_s; i.e., the unit margin is normalized to be one. If a product is not displayed, its reward is zero. We let a_s denote the decision variables in each period, where a_s equals 1 if product s is displayed and zero otherwise.

To streamline notation, we let $\boldsymbol{m} = (m_1, \ldots, m_S)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)$ denote the vectors of parameters describing DM's state of information about the S products, $\boldsymbol{a} = (a_1, \ldots, a_S)$ denote a vector of display decisions, and $\tilde{\boldsymbol{d}}_t = (\tilde{d}_{t,1}, \ldots, \tilde{d}_{t,S})$ denote a random vector of product demands in period t. After displaying the selected products, observing the demands for these products, and updating using Bayes' rule, the next period state of information is given by updated parameter vectors $\boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{a} \cdot \tilde{\boldsymbol{d}}_t$ (here "$\cdot$" denotes componentwise multiplication) and $\boldsymbol{\beta}' = \boldsymbol{\beta} + \boldsymbol{a}$. The

shelf-space constraint requires $\boldsymbol{a}$ to be in

$$\boldsymbol{A}_t \equiv \left\{ \boldsymbol{a} \in \{0,1\}^S \ : \ \mathbf{1}^\intercal \boldsymbol{a} \leq N_t \right\} \ . \tag{30}$$

where $\mathbf{1}$ is an S-vector of ones. Taking the terminal value $V^*_{T+1}(\boldsymbol{m}, \boldsymbol{\beta}) = 0$, we can write the optimal value function as

$$V^*_t(\boldsymbol{m}, \boldsymbol{\beta}) = \max_{\boldsymbol{a} \in \boldsymbol{A}_t} \mathbb{E}\Big[\boldsymbol{a}^\intercal \tilde{\boldsymbol{d}}_t + V^*_{t+1}\Big(\boldsymbol{m} + \boldsymbol{a} \cdot \tilde{\boldsymbol{d}}_t, \ \boldsymbol{\beta} + \boldsymbol{a} \Big) \Big| \boldsymbol{m}, \boldsymbol{\beta} \Big] \ . \tag{31}$$

Astute readers may notice that the natural filtration $\mathbb{F}$ here is action dependent: that is, decisions made in period t will be based on observing the history of demands for products that have been displayed, so the period-t state of information $\mathcal{F}_t$ will depend on the sequence of display decisions $(\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{t\text{-}1})$ in periods 0 to $t-1$. In this setting, a filtration $\mathbb{G}$ is an information relaxation of $\mathbb{F}$ if $\mathcal{F}_t \subseteq \mathcal{G}_t$ for any such sequence of actions.

7.2 DAP: Lagrangian Relaxations and Index Policies

The DP (31) is difficult to solve because the constraint (30) limiting the number of products displayed links decisions across products: the display decision for one product depends on the states of the other products. With more than a few products, the state space would be unmanageably large. In this subsection, we consider Lagrangian relaxations of (31) where we relax this linking constraint and decompose the value functions into computationally manageable subproblems. We then show how this Lagrangian relaxation can be used to generate a heuristic display policy as well as a bound on the performance with an optimal policy. The key results on Lagrangian relaxations of DPs (summarized in Proposition 7.1 below) are fairly standard in the literature on Lagrangian relaxations of DPs (e.g., Hawkins 2003; Adelman and Mersereau 2008); our discussion follows BS (2020).

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T) \geq \mathbf{0}$ denote a vector of Lagrange multipliers corresponding to the shelf-space constraints, requiring the DM to display at most N_t products in period t. Taking $L^{\boldsymbol{\lambda}}_{T+1}(\boldsymbol{m}, \boldsymbol{\beta}) = 0$, the Lagrangian DP has period-t value function given by

$$L^{\boldsymbol{\lambda}}_t(\boldsymbol{m}, \boldsymbol{\beta}) = \max_{\boldsymbol{a} \in \{0,1\}^S} \left\{ \mathbb{E}\Big[\boldsymbol{a}^\intercal \tilde{\boldsymbol{d}}_t + L^{\boldsymbol{\lambda}}_{t+1}\Big(\boldsymbol{m} + \boldsymbol{a} \cdot \tilde{\boldsymbol{d}}_t, \ \boldsymbol{\beta} + \boldsymbol{a} \Big) \Big| \boldsymbol{m}, \boldsymbol{\beta} \Big] + \lambda_t \left(N_t - \mathbf{1} \cdot \boldsymbol{a} \right) \right\} . \tag{32}$$

Compared to the DP (31), we have made two changes. First, we have incorporated the linking

constraint (30) into the objective by adding $\lambda_t(N_t - \mathbf{1}^\intercal \boldsymbol{a})$; with $\lambda_t \geq 0$, this term is nonnegative for all policies satisfying the linking constraint. Second, we have relaxed the linking constraint, allowing the DM to display as many products as desired (we require $\boldsymbol{a} \in \{0,1\}^S$ rather than $\boldsymbol{a} \in \boldsymbol{A}_t$). Both of these changes can only increase the optimal value so the Lagrangian value function provides an upper bound on the true value function.

The following proposition summarizes some of the key properties of this Lagrangian relaxation.

Proposition 7.1 (Properties of the Lagrangian, BS (2020))**.** *For all $\boldsymbol{m}, \boldsymbol{\beta}$, t, and $\boldsymbol{\lambda} \geq \mathbf{0}$,*

(i) (Decomposition) *The Lagrangian DP* (32) *can be decomposed into product-specific value functions*

$$L_t^{\boldsymbol{\lambda}}(\boldsymbol{m}, \boldsymbol{\beta}) = \sum_{\tau=t}^{T} \lambda_\tau N_\tau + \sum_{s=1}^{S} V_{t,s}^{\boldsymbol{\lambda}}(m_s, \beta_s) \tag{33}$$

where $V_{t,s}^{\boldsymbol{\lambda}}(m_s, \beta_s)$ is the value function for a product-specific DP: $V_{T+1,s}^{\boldsymbol{\lambda}}(m_s, \beta_s) = 0$ and

$$V_{t,s}^{\boldsymbol{\lambda}}(m_s, \beta_s) = \max_{a_s \in \{0,1\}} \mathbb{E}\Big[a_s \tilde{d}_{t,s} - \lambda_t + V_{t+1,s}^{\boldsymbol{\lambda}}\Big(m_s + a_s \tilde{d}_{t,s},\ \beta_s + a_s\Big) \,\Big|\, m_s, \beta_s \Big] \ . \tag{34}$$

(ii) (Weak Duality) $V_t^*(\boldsymbol{m}, \boldsymbol{\beta}) \leq L_t^{\boldsymbol{\lambda}}(\boldsymbol{m}, \boldsymbol{\beta})$.

(iii) (Convexity) *$L_t^{\boldsymbol{\lambda}}(\boldsymbol{m}, \boldsymbol{\beta})$ and $V_{t,s}^{\boldsymbol{\lambda}}(m_s, \beta_s)$ are piecewise linear and convex in $\boldsymbol{\lambda}$.*

The product-specific value functions (34) in the decomposed Lagrangian have a nice interpretation: intuitively, the period-t Lagrange multiplier λ_t can be interpreted as a charge for using the constrained resource in period t. Part (ii) of the Proposition says that $L_t^{\boldsymbol{\lambda}}(\boldsymbol{m}, \boldsymbol{\beta})$ can be used as a performance bound to assess the quality of a feasible policy. Part (iii) of the Proposition says the Lagrangian dual problem,

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \ L_1^{\boldsymbol{\lambda}}(\boldsymbol{m}, \boldsymbol{\beta}) \ , \tag{35}$$

is a convex optimization problem that can be solved using, for example, subgradient, linear programming, or cutting-plane methods. With an optimal set of Lagrange multipliers (i.e., solving (35)), the linking constraint (30) will hold "on average" (or in expectation) rather than in each state. See BS (2020) for further discussion of the Lagrangian dual problem, its solution, and properties of an optimal solution.

The optimal policies for the Lagrangian DP may not be physically feasible because they may violate the linking constraint (30). If, for example, many products' demands are higher than expected, more than N_t products may be displayed if we evaluate the products independently using

the policies from the product-specific DPs (34). However, we can use the Lagrangian relaxation to construct feasible policies that can be used as heuristics. The *Lagrangian index policy* is based on a priority index $i_{t,s}(m_s, \beta_s)$ that indicates the relative attractiveness of selecting product s in period t when the item is in state (m_s, β_s). This index uses the product-specific value function (34) to approximate the value added by selecting product s in period t when the item is in state (m_s, β_s),

$$i_{t,s}(m_s, \beta_s) = \mathbb{E}\Big[\, \tilde{d}_{t,s} + V^{\boldsymbol{\lambda}}_{t+1,s}\Big(m_s + \tilde{d}_{t,s}, \beta_s + 1\Big) \,\Big|\, m_s, \beta_s \,\Big] - V^{\boldsymbol{\lambda}}_{t+1,s}(m_s, \beta_s) \ . \tag{36}$$

Given priority indices for all products, the policies proceed as follows: (a) if there are more than N_t products with nonnegative indices, select the N_t products with the largest indices; (b) otherwise, select all products with nonnegative indices. The linking constraints will thus be satisfied and the Lagrangian index policies will be feasible. There may be cases where the index values are the same for some products (which may be in different states), leading to ambiguity in the choice of the "N_t products with the largest indices" in step (a). In the DAP example, it is fine to break such ties among products randomly but in other examples, it is important to use more sophisticated methods to break these ties; see BS (2020) for further discussion and examples.

In our numerical examples for the DAP, we will also consider a benchmark policy that in each period t selects the N_t products with the highest expected demand in the current state. This *myopic policy* is also an index policy where $V^{\boldsymbol{\lambda}}_{t+1,s}$ in (36) is replaced by 0.

Though we have described these policies as index policies, these policies may also be viewed as being greedy with respect to a value function approximation. For example, the Lagrangian index policy is equivalent to a policy that solves the original DP (31) with the Lagrangian (32) as an approximate continuation value:

$$\max_{\boldsymbol{a} \in \boldsymbol{A}_t} \mathbb{E}\Big[\, \boldsymbol{a}^\intercal \tilde{\boldsymbol{d}}_{t,s} + L^{\boldsymbol{\lambda}}_{t+1}\Big(\boldsymbol{m} + \boldsymbol{a} \cdot \tilde{\boldsymbol{d}}_t,\ \boldsymbol{\beta} + \boldsymbol{a}\Big) \,\Big|\, \boldsymbol{m}, \boldsymbol{\beta} \,\Big] \ . \tag{37}$$

Similarly, the myopic policy is greedy with respect to a value function approximation that replaces the Lagrangian $L^{\boldsymbol{\lambda}}_{t+1}$ in (37) with 0. The process of selecting the N_t largest indices is equivalent to solving these "greedy" optimization problems in each period.

7.3 DAP: Information Relaxation Bounds

Though the Lagrangian provides a performance bound (and optimal Lagrange multipliers provide the best such bound) as discussed in Proposition 3.2, we can improve on this Lagrangian bound

using information relaxations. In the DAP, the underlying uncertainties are the unknown Poisson demand rates γ_s for each product and the demand realizations $d_{t,s}$ for each product in each period. In the natural filtration $\mathbb{F}$, the demands $d_{t,s}$ are revealed for products after the products are displayed (if they are displayed); the demand rates γ_s are never revealed.

We will consider four different information relaxations:

(i) *Known demands* ($\mathbb{G}_d$): The DM knows all demands $d_{t,s}$ for all products in all periods, before making any display decisions (i.e., the DM knows what demand would be if a product were to be displayed); demands rates γ_s are never revealed.

(ii) *Known rates* ($\mathbb{G}_r$): The DM knows the demand rates γ_s for all products in advance, but product demands $d_{t,s}$ are revealed sequentially if/when products are displayed, as in the natural filtration.

(iii) *Perfect information* ($\mathbb{G}_p$): The DM knows both demands and demand rates in advance.

(iv) *Uncensored demands* ($\mathbb{G}_u$): Demands $d_{t,s}$ for products are revealed sequentially, whether the products are displayed or not; demand rates are never revealed.

These are all relaxations of the natural filtration and perfect information is the weakest of these four relaxations. The known-demands relaxation is weaker than the uncensored demands relaxation. The known-rates relaxation is neither weaker nor stronger than the known-demands and uncensored-demand relaxations. That is, $\mathbb{F} \subseteq \mathbb{G}_u \subseteq \mathbb{G}_d \subseteq \mathbb{G}_p$ and $\mathbb{F} \subseteq \mathbb{G}_r \subseteq \mathbb{G}_p$. As noted in Corollary 3.1, tighter relaxations will lead to tighter bounds for any given penalty.

We will consider three different penalties π_t given by different selections of generating functions w_t in Proposition 3.1:

(i) *Zero penalty*: $w_t = 0$, hence, $\pi_t = 0$.

(ii) *Smoothing penalty*: $w_t = \boldsymbol{a}^\mathsf{T}\boldsymbol{d}_t$ which leads to period-t penalty $\pi_t = \boldsymbol{a}^\mathsf{T}\left(\boldsymbol{d}_t - \mathbb{E}\left[\tilde{\boldsymbol{d}}_t \,\middle|\, \boldsymbol{m}, \boldsymbol{\beta}\right]\right)$ and penalized period-t reward function $\boldsymbol{a}^\mathsf{T}\mathbb{E}\left[\tilde{\boldsymbol{d}}_t \,\middle|\, \boldsymbol{m}, \boldsymbol{\beta}\right]$. Thus the penalty "smooths out" the benefit of knowing demand, as in the inventory example of §6.

(iii) *Lagrangian penalty*: Here we take $w_t = \boldsymbol{a}^\mathsf{T}\boldsymbol{d}_t + L^{\boldsymbol{\lambda}}_{t+1}(\boldsymbol{m}, \boldsymbol{\beta})$. In addition to "smoothing" demand as above, this penalty approximates the continuation value using the Lagrangian value function. Although we could use any $\boldsymbol{\lambda} \geq \boldsymbol{0}$, in our numerical examples we will take these to be optimal Lagrange multipliers $\boldsymbol{\lambda}^*$ given by solving the Lagrangian dual (35).

With these penalties, given knowledge of demands, the penalized rewards do not depend on the demand rates. Thus the perfect information relaxation $\mathbb{G}_p$ is equivalent to the known-demands

relaxation $\mathbb{G}_d$ with these penalties and will not be considered separately in our discussion of computations or numerical results below. However, $\mathbb{G}_p$ and $\mathbb{G}_d$ could lead to different results if the penalties used depend on the demand rates.

The different information relaxations and penalties require somewhat different computational approaches and we are not able to evaluate all penalties with all relaxations. As discussed following Theorem 3.1, in all cases the performance bound estimates are generated by Monte Carlo simulation where we draw a sample scenario representing a particular state of information for the DM and then we solve an inner problem given that state of information; we average across these scenarios to obtain an estimate of the performance bound. The nature of the inner problems varies with the information relaxation and penalty; see Table 4 for a summary. We discuss the inner problems with zero penalty first in §7.3.1 and then consider non-zero penalties in §7.3.2. We present numerical results in §7.4.

	Penalty		
Information Relaxation	Zero	Smoothing	Lagrangian
Known Demands ($\mathbb{G}_d$)	Pick N_t best demands	*Deterministic* inner DP with LR relaxation	
Known Rates ($\mathbb{G}_r$)	Pick N_t best rates	*Stochastic* inner DP with LR relaxation	
Uncensored Demands ($\mathbb{G}_u$)	Pick N_t best rates given past demands	Cannot be efficiently solved	

Table 4: Computational Methods for Inner Problems with Different Relaxations and Penalties

As noted in Proposition 3.3, if we can restrict attention to a subset of the available policies $\mathcal{A}_{\mathbb{F}}$ in the original problem without loss of optimality, we can impose these same restrictions on the policies $\mathcal{A}_{\mathbb{G}}$ for the relaxed model. In this example, if all items are initially identical we can restrict the policies to those that display the first (in label index order) N_0 products in the initial period (i.e., $s \leq N_0$) without loss of optimality. More generally, we can restrict the DM to policies that display products with $s \leq \sum_{\tau=0}^{t} N_\tau$ in period t. In our numerical examples, we impose these restrictions on display decisions. Imposing these restrictions can improve the information relaxation bound (i.e., lead to a lower value) because the information revealed in a particular scenario may suggest displaying some products (e.g., those with the highest demand in a given period) that are outside this restricted set.

7.3.1 Inner Problems with Zero Penalties

With zero penalties, the inner problems associated with the four relaxations described above are all easy to solve and the information-relaxation performance bounds are easy to estimate using Monte Carlo simulation. The key feature of these problems is that the (relaxed) state of information and rewards for any period are independent of all previous display decisions. Thus we can solve the inner problems by considering the display decisions in each period in isolation, without concern for the downstream effects of these display decisions.

- With the known-demands information relaxation $\mathbb{G}_d$, in the inner problem, the DM simply displays N_t largest demands $d_{t,s}$ in each period, in each scenario.

- With the known-rates relaxation $\mathbb{G}_r$, the DM displays the N_t products with the highest rates γ_s in each period, in each scenario. In any particular scenario, the expected reward in each period is the sum of the rates γ_s for the displayed products.

- With the uncensored demands relaxation $\mathbb{G}_u$, the DM selects the N_t products with the highest expected demands, conditional on the history of demands for the product. In given demand scenario, the expected reward in period t is the sum of $m_{t,s}/\beta_{t,s}$ for the selected products where $m_{t,s}$ and $\beta_{t,s}$ are the shape and scale parameters for the gamma distribution on rates in period t:

$$m_{t,s} = m_{0,s} + \sum_{\tau=0}^{t} d_{t,s} \tag{38}$$

$$\beta_{t,s} = \beta_{0,s} + t \tag{39}$$

  where $m_{0,s}$ and $\beta_{0,s}$ are the initial (period-0) shape and scale parameters for product s.

Imposing the restrictions on policies where the DM is restricted to displaying products with index $s \leq \sum_{\tau=0}^{t} N_\tau$ in period t (as discussed above) leads to some straightforward modifications in the simulation procedures just described (the choice of products must be from the restricted set) in the periods where the constraint may be binding. However, the zero-penalty performance bounds are still easy to estimate with these restrictions.

7.3.2 Inner Problems with Non-Zero Penalties

Though the inner problems with zero penalty are easy to solve, the inner problem with smoothing or Lagrangian penalties are difficult to solve because, like the original DP, the display constraint

(30) links decisions across products. We will use another Lagrangian relaxation to simplify these inner problems.

We first focus the known-demands information relaxation $\mathbb{G}_d$ with the Lagrangian penalty, i.e., taking the generating function in Proposition 3.1 to $w_t = \boldsymbol{a}^\intercal \boldsymbol{d}_t + L^{\boldsymbol{\lambda}}_{t+1}(\boldsymbol{m}, \boldsymbol{\beta})$. In this case, we can write the inner problem (3) for a given demand scenario as a deterministic DP. Given a demand scenario $\boldsymbol{d} = (\boldsymbol{d}_0, \ldots, \boldsymbol{d}_T)$ with $\boldsymbol{d}_t = (d_{0,t}, \ldots, d_{S,t})$, let $\hat{V}_{T+1}(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) = 0$ and, for earlier t, we recursively define

$$\hat{V}_t(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) = \max_{\boldsymbol{a} \in \boldsymbol{A}_t} \left\{ \boldsymbol{a}^\intercal \boldsymbol{d}_t - \pi_t(\boldsymbol{m}, \boldsymbol{\beta}, \boldsymbol{a}; \boldsymbol{d}_t) + \hat{V}_{t+1}(\boldsymbol{m} + \boldsymbol{a} \cdot \boldsymbol{d}_t, \boldsymbol{\beta} + \boldsymbol{a}; \boldsymbol{d}) \right\} \tag{40}$$

where

$$\pi_t(\boldsymbol{m}, \boldsymbol{\beta}, \boldsymbol{a}; \boldsymbol{d}_t) = \boldsymbol{a}^\intercal \boldsymbol{d}_t + L^{\boldsymbol{\lambda}}_{t+1}(\boldsymbol{m} + \boldsymbol{a} \cdot \boldsymbol{d}_t, \boldsymbol{\beta} + \boldsymbol{a}) - \mathbb{E}\Big[\boldsymbol{a}^\intercal \tilde{\boldsymbol{d}}_t + L^{\boldsymbol{\lambda}}_{t+1}\Big(\boldsymbol{m} + \boldsymbol{a} \cdot \tilde{\boldsymbol{d}}_t,\ \boldsymbol{\beta} + \boldsymbol{a}\Big) \,\Big|\, \boldsymbol{m}, \boldsymbol{\beta} \Big] \ . \tag{41}$$

Here the last term in (40) and the second term in (41) involve deterministic state transitions because the DM knows the demand for each product and each period. The expectation in (41), representing the $\mathcal{F}_t$-conditional expectation in (9), is calculated using the same state-dependent negative-binomial distributions for demand that were used in the original DP.

We now consider the inner DP (40) in more detail. First, note that even though the information relaxation $\mathbb{G}_d$ reveals all demands before making any decisions, we need to keep track of the DM's state of information (the parameters $(\boldsymbol{m}, \boldsymbol{\beta})$ of the demand distribution) over time; this is needed to calculate the expectations in the penalty (41). This inner DP is simpler than the original DP (31) because, for any given set of display decisions $\boldsymbol{a}$, we need only consider one possible next period state rather than taking expectations over many possible next period states. However, we still need to consider many possible $(\boldsymbol{m}, \boldsymbol{\beta})$ states in each period in these deterministic DPs as these states may be reached by some feasible sequences of display decisions.

Second, note that the penalty terms involving the Lagrangian $L^{\boldsymbol{\lambda}}_{t+1}$ decompose into the sum of product-specific values, as in (33), so the penalty π_t can be decomposed across products. However, the inner DP (40) does not decompose into product-specific subproblems because the constraint on the total number of products displayed ($\boldsymbol{a} \in \boldsymbol{A}_t$ where $\boldsymbol{A}_t$ is defined in (30)) links the decisions across items, as it did in the original DP (31). Thus, the inner DP – though deterministic – is still difficult to solve in problems with many items.

To decouple the inner DP (40), we relax the linking constraint in the same way that we

relaxed the original DP (31). Consider Lagrange multipliers $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T) \geq \mathbf{0}$ and let $\hat{L}^{\boldsymbol{\mu}}_{T+1}(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) = 0$. The period-$t$ *inner Lagrangian* with demand realization $\boldsymbol{d}$ is then given recursively as

$$\hat{L}^{\boldsymbol{\mu}}_t(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) = \max_{\boldsymbol{a} \in \{0,1\}^S} \left\{ \boldsymbol{a}^\intercal \boldsymbol{d}_t - \pi_t(\boldsymbol{m}, \boldsymbol{\beta}, \boldsymbol{a}; \boldsymbol{d}_t) + \hat{L}^{\boldsymbol{\mu}}_{t+1}(\boldsymbol{m} + \boldsymbol{a} \cdot \boldsymbol{d}_t, \boldsymbol{\beta} + \boldsymbol{a}; \boldsymbol{d}) + \mu_t \left(N_t - \mathbf{1}^\intercal \boldsymbol{a} \right) \right\} .$$

This can be decomposed into product-specific DPs as

$$\hat{L}^{\boldsymbol{\mu}}_t(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) = N_t \sum_{\tau=t}^{T} \mu_\tau + \sum_{s=1}^{S} \hat{V}^{\boldsymbol{\mu}}_{t,s}(x_s; \boldsymbol{d}_s)$$

where $\boldsymbol{d}_s = (d_{1,s}, \ldots, d_{T,s})$ is the demand sequence for product s and $\hat{V}^{\boldsymbol{\mu}}_{t,s}(x_s; \boldsymbol{d}_s)$ is an inner product-specific value function with $\hat{V}^{\boldsymbol{\mu}}_{T+1,s}(x_s; \boldsymbol{d}_s) = 0$ and

$$\begin{aligned} \hat{V}^{\boldsymbol{\mu}}_{t,s}(m_s, \beta_s; \boldsymbol{d}_s) = \max_{a_s \in \{0,1\}} \Bigg\{ & \left(m_s/\beta_s - \mu_t\right) a_s - V^{\boldsymbol{\lambda}}_{s,t+1}(m_s + a_s d_{t,s}, \beta_s + a_s) \\ & + \mathbb{E}\Big[V^{\boldsymbol{\lambda}}_{s,t+1}\Big(m_s + a_s \tilde{d}_{t,s}, \beta_s + a_s\Big) \,\Big|\, m_s, \beta_s \Big] + \hat{V}^{\boldsymbol{\mu}}_{t+1,s}(m_s + a_s d_{t,s}, \beta_s + a_s; \boldsymbol{d}_s) \Bigg\}, \end{aligned} \tag{42}$$

where $V^{\boldsymbol{\lambda}}_{t,s}$ is the value function for the product-specific DP (34). Note that we have used the fact that $\mathbb{E}\Big[\tilde{d}_t \,\Big|\, m_s, \beta_s \Big] = m_s/\beta_s$ in the expression above.

These inner product-specific DPs and the Lagrangian satisfy properties like those in Proposition 7.1. In particular, the Lagrangian is an upper bound on the inner DP: $\hat{V}_t(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) \leq \hat{L}^{\boldsymbol{\mu}}_t(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d})$ for all $\boldsymbol{m}, \boldsymbol{\beta}$, t, $\boldsymbol{d}$ and $\boldsymbol{\mu} \geq \mathbf{0}$. To ensure we have the best possible bound for a given $\boldsymbol{d}$ and initial state $(\boldsymbol{m}, \boldsymbol{\beta})$, we can solve the inner dual problem,

$$\min_{\boldsymbol{\mu} \geq \mathbf{0}} \ \hat{L}^{\boldsymbol{\mu}}_1(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) \ , \tag{43}$$

for an optimal $\boldsymbol{\mu}^*(\boldsymbol{m}, \boldsymbol{\beta}, \boldsymbol{d})$. This is a convex optimization problem and can be solved in a variety of ways (e.g., linear programming, subgradient methods, or the cutting-plane method discussed in BS (2020)). Moreover, if we take the inner Lagrange multipliers $\boldsymbol{\mu}$ to be equal to the "outer" Lagrange multipliers $\boldsymbol{\lambda}$ used to define the penalty, we can use an induction argument to show that $\hat{L}^{\boldsymbol{\lambda}}_t(\boldsymbol{m}, \boldsymbol{\beta}; \boldsymbol{d}) = L^{\boldsymbol{\lambda}}_t(\boldsymbol{m}, \boldsymbol{\beta})$ for all t and $\boldsymbol{d}$.[5] Thus, since $\boldsymbol{\lambda}$ is feasible but not necessarily optimal for

[5]Note that the $V^{\boldsymbol{\lambda}}_{s,t+1}(\cdot)$ and $\hat{V}^{\boldsymbol{\mu}}_{t+1,s}(\cdot)$ terms in (42) cancel if $\boldsymbol{\mu} = \boldsymbol{\lambda}$ and we have the induction hypothesis that $V^{\boldsymbol{\lambda}}_{s,t+1}(\cdot) = \hat{V}^{\boldsymbol{\lambda}}_{s,t+1}(\cdot)$. Then (42) reduces to the definition of $V^{\boldsymbol{\lambda}}_{s,t+1}(\cdot)$ in (34).

the inner Lagrangian dual problem (43), we have

$$\hat{V}_1(\boldsymbol{m},\boldsymbol{\beta};\boldsymbol{d}) \le \hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{m},\boldsymbol{\beta},\boldsymbol{d})}(\boldsymbol{m},\boldsymbol{\beta};\boldsymbol{d}) \le L_1^{\boldsymbol{\lambda}}(\boldsymbol{m},\boldsymbol{\beta}) \ .$$

We now briefly consider the other combinations of relaxations and non-zero penalties.

Known-demands relaxation $\mathbb{G}_d$ with smoothing penalty: With the smoothing penalty (generating function $w_t = \boldsymbol{a}^\intercal \boldsymbol{d}_t$) instead of the Lagrangian penalty ($w_t = \boldsymbol{a}^\intercal \boldsymbol{d}_t + L_{t+1}^{\boldsymbol{\lambda}}(\boldsymbol{m},\boldsymbol{\beta})$), we proceed exactly as above but without the Lagrangian terms $L_t^{\boldsymbol{\lambda}}$. We still need to keep track of the state $(\boldsymbol{m},\boldsymbol{\beta})$ to calculate expected rewards in the penalty (41) and still use the inner Lagrangian decomposition (the terms involving $\boldsymbol{\mu}$) to decouple the inner DP. These inner DPs with the smoothing penalty are thus about as difficult to solve as those with the Lagrangian penalty.

Known-rates relaxation $\mathbb{G}_r$: With the known-rates relaxation and the Lagrangian penalty, the approach is similar to that for the known-demands relaxation but the inner problems now require a stochastic rather than deterministic DP. Specifically, the known demands DP (40) now requires taking an expectation over demand, given the demand rates $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_S)$ and becomes:

$$\hat{V}_t(\boldsymbol{m},\boldsymbol{\beta};\boldsymbol{\gamma}) = \max_{\boldsymbol{a}\in\boldsymbol{A}_t} \mathbb{E}\Big[\, \boldsymbol{a}^\intercal \tilde{\boldsymbol{d}}_t - \pi_t(\boldsymbol{m},\boldsymbol{\beta},\boldsymbol{a};\tilde{\boldsymbol{d}}_t) + \hat{V}_{t+1}\Big(\boldsymbol{m} + \boldsymbol{a}\cdot\tilde{\boldsymbol{d}}_t, \boldsymbol{\beta}+\boldsymbol{a};\boldsymbol{d}\Big) \,\Big|\, \boldsymbol{\gamma}\,\Big] \ .$$

The decomposition proceeds as before and the inner product-specific value function (42) becomes

$$\begin{aligned}\hat{V}_{t,s}^{\boldsymbol{\mu}}(m_s,\beta_s;\gamma_s) = \max_{a_s\in\{0,1\}} \Bigg\{ & (m_s/\beta_s - \mu_t)\, a_s + \mathbb{E}\Big[\, V_{s,t+1}^{\boldsymbol{\lambda}}\Big(m_s + a_s\tilde{d}_{t,s}, \beta_s + a_s\Big) \,\Big|\, m_s,\beta_s\,\Big] \\ & - \mathbb{E}\Big[\, V_{s,t+1}^{\boldsymbol{\lambda}}\Big(m_s + a_s\tilde{d}_{t,s}, \beta_s + a_s\Big) + \hat{V}_{t+1,s}^{\boldsymbol{\mu}}\Big(m_s + a_s\tilde{d}_{t,s}, \beta_s + a_s;\gamma_s\Big) \,\Big|\, \gamma_s\,\Big] \Bigg\},\end{aligned}$$

Here the first expectations over demand are the $\mathbb{F}$-conditional expectations in the penalty and are calculated using the negative-binomial distribution with parameters (m_s,β_s), as before. The second expectations over demand are the $\mathbb{G}_r$-conditional expectations in the relaxed DP recursion (5) and are calculated using the Poisson distribution with the known rate γ_s. These inner DPs are significantly more complicated than those for the known-rates relaxation: not only are the inner DPs stochastic rather than deterministic, but one must also consider a broader range of (m_s,β_s) states in each period as more states may be reached given uncertain demand realizations and feasible sequences of display decisions. Using the smoothing penalty instead of the Lagrangian penalty in the known-rates relaxations leads to some simplifications (as in the known-demands relaxation), but results in a stochastic DP that is about as difficult to solve as with the Lagrangian penalty.

Uncensored demands relaxation $\mathbb{G}_u$: As in the known-rates relaxation $\mathbb{G}_r$ just discussed, the uncensored demand relaxation also results in a stochastic inner DP. This DP can also be decomposed using an inner Lagrangian relaxation. However, with both the smoothing and Lagrangian penalties, the resulting inner product-specific DPs are too complex to be efficiently solved. The issue is that in these inner product-specific DPs, not only one must keep track of the reachable (m_s, β_s)-states to calculate the $\mathbb{F}$-conditional expectations in the penalties, one must also keep track of a different set of (m_s, β_s)-states (given by (38) and (39)) to represent the DM's stochastically evolving state of information in the $\mathbb{G}_u$-relaxation. The DP thus has a four-dimensional state space that is too large to be efficiently solved.

7.4 DAP: Numerical Experiments

In our numerical examples, we will consider parameters similar to those in CG (2007) and BS (2020). We consider a horizon of $T = 12$. We assume that all products are *a priori* identical with gamma distribution parameters $(m_s, \beta_s) = (1.0, 0.1)$ (so the mean and standard deviation for the demand rate are both 10).[6] We assume that the DM can display 20% of the products available in each period, i.e., $N_t = 0.20S$, and consider the case when the total number of products S equals 5, 20, and 50.

We will consider the myopic and Lagrangian index policies (described in §7.2) and the information relaxations and penalties described in §7.3 (summarized in Table 4). We use Monte Carlo simulation to estimate the performance of the heuristics and bounds with 1,000 scenarios each, except for the known-rates bounds with smoothing and Lagrangian penalties; these are more time-consuming to compute (for the largest problems with $S = 50$, computing these bounds took about about two minutes per scenario versus 0.1 seconds per scenario for the known-demands bounds with smoothing and Lagrangian penalties) and hence we use 100 scenarios instead. We use the cutting plane method described in BS (2020)) to find the optimal Lagrange multipliers for the Lagrangian dual problems that arise. All calculations were done on a desktop PC using Matlab with the MOSEK Optimization Toolbox.

Table 5 provides a summary of the results. In terms of the policies, the Lagrangian index policy performs much better than the myopic index policy; this is not surprising, since the myopic policy ignores the downstream benefits of learning about demands. In terms of the information

[6] In our numerical examples, we truncate the demand distributions at $\bar{d} = 150$ (thereby including 99.9999% of the possible demand scenarios). In period t, there are $\sum_{\tau=0}^{t-1} \left((\tau - 1)\bar{d} + 1\right)$ possible states, representing the values of (m, β) that could be obtained under some policy.

$S = 5$ products						
Policies						
Myopic	209.93	(0.85)				
Lagrangian	**224.27**	(0.67)				
Performance Bounds						
Lagrangian relaxation	239.43					
Information relaxation	Zero Penalty		Smoothing Penalty		Lagrangian Penalty	
Known demands $\mathbb{G}_d$	277.02	(4.44)	236.93	(3.35)	227.95	(0.59)
Known rates $\mathbb{G}_r$	270.74	(4.51)	260.63	(14.09)	**225.71**	(2.41)
Uncensored $\mathbb{G}_u$	255.18	(4.04)	NA		NA	
$S = 20$ products						
Policies						
Myopic	869.44	(1.42)				
Lagrangian	**940.81**	(0.79)				
Performance Bounds						
Lagrangian relaxation	957.73					
Information relaxation	Zero Penalty		Smoothing Penalty		Lagrangian Penalty	
Known demands $\mathbb{G}_d$	1257.41	(10.21)	1049.41	(7.76)	951.90	(0.48)
Known rates $\mathbb{G}_r$	1221.35	(10.31)	1116.56	(25.30)	**949.51**	(1.62)
Uncensored $\mathbb{G}_u$	1151.25	(9.32)	NA		NA	
$S = 50$ products						
Policies						
Myopic	2180.05	(2.36)				
Lagrangian	**2373.04**	(1.06)				
Performance Bounds						
Lagrangian relaxation	2394.31					
Information relaxation	Zero Penalty		Smoothing Penalty		Lagrangian Penalty	
Known demands $\mathbb{G}_d$	3165.66	(16.32)	2634.31	(12.28)	2390.75	(0.38)
Known rates $\mathbb{G}_r$	3074.18	(16.48)	2851.35	(49.25)	**2386.22**	(2.84)
Uncensored $\mathbb{G}_u$	2894.86	(14.76)	NA		NA	

Table 5: Numerical results for dynamic assortment examples. For simulated results, the table shows estimated mean values with mean standard errors of simulated estimates in parentheses. For each S, the best policy and performance bound is in bold.

relaxation bounds, with zero penalty we see that the uncensored relaxation provides the best performance bound in each case but that these zero penalty performance bounds are all worse than the performance bound from the Lagrangian relaxation. The smoothing penalty improves upon zero penalty in each case. The Lagrangian penalty also improves upon the smoothing penalty in each case and, as stated in Propositions 3.2, the corresponding information relaxation bounds are tighter than the Lagrangian relaxation performance bound.

Because the information relaxations considered are ordered (recall from §7.3 that $\mathbb{F} \subseteq \mathbb{G}_u \subseteq \mathbb{G}_d \subseteq \mathbb{G}_p$ and $\mathbb{F} \subseteq \mathbb{G}_r \subseteq \mathbb{G}_p$), we might expect the performance bounds given by the same penalties to reflect this ordering with weaker relaxations leading to weaker penalties, as noted in Corollary 3.1. This ordering is reflected in the numerical results of Table 5 for the zero penalty case. However, this order need not hold for our numerical results for the smoothing and Lagrangian penalties because we have used a Lagrangian relaxation to simplify the solution of the inner problems. Our numerical results are thus upper bounds on the performance bounds that we would find if we were able to solve these inner problems exactly. These upper bounds on the performance bounds need not be ordered in the way the exact performance bounds would be. Indeed, in the numerical results in Table 5, we see that the reported bounds with the smoothing penalty do not reflect the ordering of the information relaxations. Specifically, the known rates $\mathbb{G}_r$ bounds are worse than the perfect information $\mathbb{G}_p$ bounds (which are equal to the known demand $\mathbb{G}_d$), even though $\mathbb{G}_r \subseteq \mathbb{G}_p$. However, the bounds with the Lagrangian penalty do reflect the information relaxation ordering.

The results with the Lagrangian penalty demonstrate the value of the information relaxation bounds, especially for small values of S: with $S = 5$, using the Lagrangian relaxation bound, we can conclude the Lagrangian index policy is within $(239.43/224.27-1) = 6.76\%$ of optimal, whereas the known-rates relaxation shows the Lagrangian index policy is in fact within $(225.71/224.27 - 1) = 0.64\%$ of optimal. The relative improvement provided by the information relaxations is somewhat less in the case of larger S; this reflects the fact that the Lagrangian index policy and Lagrangian bound are asymptotically optimal, with the relative gap between the two approaching zero as S grows large (see §6 of BS (2020)). However, the information relaxation bound reduces the gap between the policy and bound by approximately one-half when $S = 20$ and $S = 50$. For example, with S=50, the Lagrangian performance bound shows that the Lagrangian index policy is within $(2394.31/2373.04-1) = 0.90\%$ of optimal, whereas the known-rates relaxation shows the Lagrangian index policy is within $(2386.22/2373.04 - 1) = 0.56\%$ of optimal.

8. Example: Portfolio Optimization with Transaction Costs

As an example of a convex DP, in this section, we study a dynamic portfolio optimization problem with transaction costs, drawing on Brown and Smith (2011) (BS (2011)). Here, the approximate model is a "frictionless" portfolio optimization model that ignores transactions costs; this is a (physical) relaxation of the original model and is not difficult to solve to optimality. These frictionless value functions are differentiable and hence the results of Proposition 4.1 apply and, in particular, by part (ii) of that proposition, we can calculate information relaxation bounds that improve on the bound provided by the frictionless model.

We begin by describing the portfolio optimization model in §8.1, the information relaxation bounds in §8.2, and provide some numerical examples in §8.3. We note that the construction of the penalties in BS (2011) is different than our construction here; as discussed in Brown and Smith (2014*a*), we could have done somewhat better applying the approach of Proposition 4.1 (as we do here) instead.

8.1 Portfolio Optimization Model

There are n risky assets and a risk-free asset (cash). The risk-free rate ρ_f is assumed to be known and constant over time. The returns of the risky assets are stochastic and denoted by $\boldsymbol{\rho}_t = (\rho_{t,1}, \ldots, \rho_{t,n})$ where $\rho_{t,i} \geq 0$ is the (gross) return on asset i from period $t-1$ to period t.

The monetary values of the risky asset holdings at the beginning of period t are described by the vector $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,n})$; the cash position in period t is denoted c_t. We let the trade vector $\boldsymbol{a}_t = (a_{t,1}, \ldots, a_{t,n})$ denote the amounts (also in monetary values) of risky assets bought (if $a_{t,n} > 0$) or sold (if $a_{t,n} < 0$) in period t. The transaction costs associated with trade vector $\boldsymbol{a}_t$ are given by $\kappa(\boldsymbol{a}_t)$. In our general analysis and approach, we will assume that $\kappa(\boldsymbol{a}_t)$ is a nonnegative and convex function of the trades $\boldsymbol{a}_t$ with $\kappa(\boldsymbol{0}) = 0$. In our numerical experiments, we will focus on the special case of proportional transaction costs with

$$\kappa(\boldsymbol{a}_t) = \sum_{i=1}^{n} \left(\delta_i^+ a_{t,i}^+ - \delta_i^- a_{t,i}^- \right) \ , \tag{44}$$

where $a_{t,i}^+ = \max(a_{t,i}, 0)$ and $a_{t,i}^- = \min(a_{t,i}, 0)$ denote the positive and negative components of the trades and $\delta_i^+, \delta_i^- \geq 0$ are the proportional costs for buying and selling (respectively) asset i. Alternatively, we could use a quadratic function for transaction costs to capture a "linear price impact," where trades lead to temporary linear changes in prices. Many other forms are possible.

Taking transaction costs into account, the asset holdings and cash position evolve according to:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{\rho}_{t+1} \cdot (\boldsymbol{x}_t + \boldsymbol{a}_t), \\ c_{t+1} &= r_f(c_t - \mathbf{1}^\intercal \boldsymbol{a}_t - \kappa(\boldsymbol{a}_t)). \end{aligned}$$

Here $\cdot$ denotes the componentwise product of two vectors (so $x_{t+1,i} = \rho_{t+1,i}(x_{t,i} + a_{t,i})$) and $\mathbf{1}$ is an n-vector of ones. The investor's wealth w_t in period t is the sum of the risky asset and cash positions, i.e.,

$$w_t = \mathbf{1}^\intercal \boldsymbol{x}_t + c_t \ .$$

The investor's goal is to maximize the expected utility of terminal wealth, $\mathbb{E}[\,U(w_T)\,]$, where U is a nondecreasing and concave utility function. Note that in this formulation, we define wealth in terms of the market value of the portfolio. We could have instead defined wealth in terms of the liquidation value of the portfolio, including the transaction costs associated with liquidation. In this case, we would take $w_t = \mathbf{1}^\intercal \boldsymbol{x}_t - \kappa(-\boldsymbol{x}_t) + c_t$. Our general approach works in either case, though the numerical results would be somewhat different.

We will assume that the investor's trades $\boldsymbol{a}_t$ in period t are restricted to a closed, convex set $A_t(\boldsymbol{x}_t, c_t)$. In our numerical experiments, we will focus on the case where the investor is not allowed to have short positions in risky assets or cash, so given an asset position $(\boldsymbol{x}_t, c_t)$, the allowed trades are

$$A_t(\boldsymbol{x}_t, c_t) = \left\{ \boldsymbol{a}_t \in \mathbb{R}^n \ : \ \boldsymbol{x}_t + \boldsymbol{a}_t \geq 0, \ c_t - \mathbf{1}^\intercal \boldsymbol{a}_t - \kappa(\boldsymbol{a}_t) \geq 0 \right\}. \tag{45}$$

BS (2011) also consider numerical results for the case where short positions are allowed, but there is a margin requirement that limits the total (long or short) position in risky assets. In general, we consider sets of allowed trades $A_t(\boldsymbol{x}_t, c_t)$ defined in terms of a set H_t of allowed final positions (or holdings): $\boldsymbol{a}_t \in A_t(\boldsymbol{x}_t, c_t)$ if and only if $(\boldsymbol{x}_t + \boldsymbol{a}_t, c_t - \mathbf{1}^\intercal \boldsymbol{a}_t - \kappa(\boldsymbol{a}_t)) \in H_t$. We assume that the allowed set of final positions H_t is closed, convex, and nondecreasing in c_t (if $(\boldsymbol{x}_t, c_t') \in H_t$ and $c_t' \leq c_t''$ then $(\boldsymbol{x}_t, c_t'') \in H_t$). This implies that $A_t(\boldsymbol{x}_t, c_t)$ is convex for each $(\boldsymbol{x}_t, c_t)$.

We will allow the possibility that returns exhibit some degree of predictability. To model this, we let $\boldsymbol{z}_t$ denote a vector of observable market state variables that provides information about the returns $\boldsymbol{\rho}_{t+1}$ of the risky assets. We will assume that $\boldsymbol{z}_t$ follows a Markov process. The returns $\boldsymbol{\rho}_{t+1}$ may depend on $\boldsymbol{z}_t$ but, given $\boldsymbol{z}_t$, the returns are assumed to be conditionally independent of prior returns and earlier values of the market state variable.

This portfolio optimization problem can be formulated as a stochastic dynamic program with

state variables consisting of the current positions in risky assets and cash $(\boldsymbol{x}_t, c_t)$ and the market state variable $(\boldsymbol{z}_t)$. We take the terminal value function to be the utility of terminal wealth, $V_T^*(\boldsymbol{x}_T, c_T, \boldsymbol{z}_T) = U(\mathbf{1}^\intercal \boldsymbol{x}_T + c_T)$, and earlier value functions V_t^* are given recursively as

$$V_t^*(\boldsymbol{x}_t, c_t, \boldsymbol{z}_t) = \max_{\boldsymbol{a}_t \in A_t(\boldsymbol{x}_t, c_t)} W_t(\boldsymbol{a}_t, \boldsymbol{x}_t, c_t, \boldsymbol{z}_t) \tag{46}$$

$$W_t(\boldsymbol{a}_t, \boldsymbol{x}_t, c_t, \boldsymbol{z}_t) = \mathbb{E}\big[\, V_{t+1}^*\big(\tilde{\boldsymbol{\rho}}_{t+1} \cdot (\boldsymbol{x}_t + \boldsymbol{a}_t),\ \rho_f(c_t - \mathbf{1}^\intercal \boldsymbol{a}_t - \kappa(\boldsymbol{a}_t)),\ \tilde{\boldsymbol{z}}_{t+1}\big) \,\big|\, \boldsymbol{z}_t \,\big]. \tag{47}$$

In (47), expectations are taken over the random asset returns $\tilde{\boldsymbol{\rho}}_{t+1}$ and the next-period market state $\tilde{\boldsymbol{z}}_{t+1}$. We will assume that these expectations are well defined and that maxima in (46) are attained by some set of trades.

The following proposition states some key properties of this portfolio optimization model.

Proposition 8.1 (Properties of the Portfolio Optimization Model, BS (2011))**.**

(i) *For any market state $\boldsymbol{z}_t$, $V_t^*(\boldsymbol{x}_t, c_t, \boldsymbol{z}_t)$ is nondecreasing in cash c_t and jointly concave in the asset position $(\boldsymbol{x}_t, c_t)$.*

(ii) *For any market state $\boldsymbol{z}_t$, $W_t(\boldsymbol{a}_t, \boldsymbol{x}_t, c_t, \boldsymbol{z}_t)$ is jointly concave in the trades $\boldsymbol{a}_t$ and asset position $(\boldsymbol{x}_t, c_t)$.*

Thus, for any given market state $\boldsymbol{z}_t$ and asset and cash position $(\boldsymbol{x}_t, c_t)$, the optimization problem (46) is convex: we are maximizing a concave function over a convex set. Unfortunately, the dimension of the state space makes the portfolio optimization problem very difficult to solve, even with just a few risky assets. For example, suppose the market state variable $\boldsymbol{z}_t$ is one-dimensional. If we approximated the state space using a grid with 20 points for this market state variable and 100 points for each of the $n+1$ asset positions, the state space would consist of $20 \times 100^{n+1}$ states. To determine the value function $V_t^*(\boldsymbol{x}_t, c_t, \boldsymbol{z}_t)$ on this grid, we would have to solve the optimization problem (46) for each of these $20 \times 100^{n+1}$ states in each period. In our numerical examples with $n = 3$ risky assets and predictability, the state space would include $20 \times 100^4 = 2$ billion elements. With $n = 10$ risky assets and no predictability, the state space would include $100^{11} = 10^{22}$ elements. Moreover, each of these optimization problems involves expectations (47) over the $(n+1)$-dimensional space of $(\boldsymbol{\rho}_{t+1}, \boldsymbol{z}_{t+1})$ scenarios and we would have to somehow interpolate between grid points when solving for the optimal trades.

If there are no transaction costs ($\kappa = 0$), the portfolio optimization problem can be greatly simplified by taking the dynamic programming state variables to be the current wealth (w_t) and

market state variable ($\boldsymbol{z}_t$); we no longer need to consider the specific asset positions ($\boldsymbol{x}_t, c_t$). In this simpler dynamic program, the decision variables are the post-trade positions in risky assets $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t + \boldsymbol{a}_t$. Let $X_t(w_t)$ denote the set of possible post-trade positions in risky assets given initial wealth w_t; that is $X_t(w_t) = \left\{\hat{\boldsymbol{x}}_t : (\hat{\boldsymbol{x}}_t, w_t - \mathbf{1}^\intercal \hat{\boldsymbol{x}}_t) \in H_t\right\}$. For example with no transaction costs, the case described by (45) where the investor is not allowed to have short positions corresponds to a feasible set of post-trade asset positions of the form

$$X_t(w_t) = \left\{\hat{\boldsymbol{x}}_t \in \mathbb{R}^n \ : \ \hat{\boldsymbol{x}}_t \geq 0, \ \mathbf{1}^\intercal \hat{\boldsymbol{x}}_t \leq w_t\right\}.$$

We can then write the recursion for this "frictionless model" as follows: The terminal value function is $V_T^f(w_T, \boldsymbol{z}_T) = U(w_T)$ and earlier value functions are

$$\begin{aligned} V_t^f(w_t, \boldsymbol{z}_t) &= \max_{\hat{\boldsymbol{x}}_t \in X_t(w_t)} W_t^f(\hat{\boldsymbol{x}}_t, w_t, \boldsymbol{z}_t) \ , \\ W_t^f(\hat{\boldsymbol{x}}_t, w_t, \boldsymbol{z}_t) &= \mathbb{E}\Big[V_{t+1}^f\big(\tilde{\boldsymbol{\rho}}_{t+1}^\intercal \hat{\boldsymbol{x}}_t + \rho_f(w_t - \mathbf{1}^\intercal \hat{\boldsymbol{x}}_t), \ \tilde{\boldsymbol{z}}_{t+1}\big) \,\Big|\, \boldsymbol{z}_t \Big] \ . \end{aligned} \tag{48}$$

This frictionless model also has a convex structure and its results can be related to those of the more complicated model with transaction costs.

Proposition 8.2 (Properties of the Frictionless Portfolio Optimization Model, BS (2011))**.**

(i) *For any market state $\boldsymbol{z}_t$, $V_t^f(w_t, \boldsymbol{z}_t)$ is nondecreasing and concave in wealth w_t.*

(ii) *For any market state $\boldsymbol{z}_t$, $W_t^f(\hat{\boldsymbol{x}}_t, w_t, \boldsymbol{z}_t)$ is jointly concave in the post-trade asset positions $\hat{\boldsymbol{x}}_t$ and wealth w_t.*

(iii) *For any market state $\boldsymbol{z}_t$ and asset position $(\boldsymbol{x}_t, c_t)$, $V_t^*(\boldsymbol{x}_t, c_t, \boldsymbol{z}_t) \leq V_t^f\big(\mathbf{1}^\intercal \boldsymbol{x}_t + c_t, \boldsymbol{z}_t\big)$.*

Thus, to solve the frictionless model, we need to solve a convex optimization problem for each market state $\boldsymbol{z}_t$ and wealth w_t. For example, if the market state variable $\boldsymbol{z}_t$ is one-dimensional, we could solve this dynamic program on a two-dimensional grid involving $\boldsymbol{z}_t$ and w_t. The expectations over $(\tilde{\boldsymbol{\rho}}_{t+1}, \tilde{\boldsymbol{z}}_{t+1})$ in (48) will still be high-dimensional if we have many assets but can be evaluated using various methods. In our numerical experiments, we will approximate these expectations using discrete approximations of the underlying distributions (see §5.1 of BS (2011) for details).

If the investor has a power utility function, the frictionless model simplifies further. Specifically, suppose

$$U(w_T) = \frac{1}{1-\gamma} \ w_T^{1-\gamma},$$

where $\gamma > 0$ is the coefficient of relative risk aversion; in the case where $\gamma = 1$, $U(w_T) = \ln(w_T)$. We can then write the value function as

$$V_t^f(w_t, \boldsymbol{z}_t) = \frac{1}{1-\gamma}\ w_t^{1-\gamma}\ \phi_t(\boldsymbol{z}_t)\ , \tag{49}$$

where $\phi_t(\boldsymbol{z}_t)$ is defined recursively with $\phi_T(\boldsymbol{z}_T) = 1$ and

$$\frac{1}{1-\gamma}\phi_t(\boldsymbol{z}_t) = \max_{\hat{\boldsymbol{\theta}}_t \in X_t(1)} \mathbb{E}\left[\frac{1}{1-\gamma}(\tilde{\boldsymbol{\rho}}_{t+1}^\intercal \hat{\boldsymbol{\theta}}_t + \rho_f(1 - \mathbf{1}^\intercal \hat{\boldsymbol{\theta}}_t))^{1-\gamma}\ \phi_{t+1}(\tilde{\boldsymbol{z}}_{t+1}) \,\middle|\, \boldsymbol{z}_t \right]. \tag{50}$$

Here $\hat{\boldsymbol{\theta}}_t = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ are the post-trade fractions of wealth w_t invested in the risky assets. In this case, the dimension of the state space is equal to the dimension of the market state variable $\boldsymbol{z}_t$.

8.2 Information Relaxation Bounds

Following the approach described in §4, we can use the frictionless model as an approximate value function to construct a gradient penalty of the form of (22). Let $\boldsymbol{a} = (\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{T-1})$ denote the vector of trades made over all T periods and $\hat{\boldsymbol{\alpha}}^* = (\hat{\boldsymbol{\alpha}}_0^*, \ldots, \hat{\boldsymbol{\alpha}}_{T-1}^*)$ denote the vector of trades over all T periods following an optimal policy in the frictionless model. Given a return $\boldsymbol{\rho} = (\boldsymbol{\rho}_0, \ldots, \boldsymbol{\rho}_T)$ and market state scenario $\boldsymbol{z} = (\boldsymbol{z}_0, \ldots, \boldsymbol{z}_T)$, the period-$t$ generating function for the gradient penalty (21) is then

$$\nabla_{\boldsymbol{a}} V_{t+1}^f(w_{t+1}^f(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho}), \boldsymbol{z}_{t+1})^\intercal\, (\boldsymbol{a} - \hat{\boldsymbol{\alpha}}^*) + V_{t+1}^f(w_{t+1}^f(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho}), \boldsymbol{z}_{t+1}) \tag{51}$$

where $w_t^f(\boldsymbol{a}, \boldsymbol{\rho})$ denotes the wealth in the frictionless model at time t given trades $\boldsymbol{a}$ and returns $\boldsymbol{\rho}$. Given a power utility function, using (49) and the chain rule, the gradient in (51) can be written as

$$\nabla_{\boldsymbol{a}} V_{t+1}^f(w_{t+1}^f(\boldsymbol{a}, \boldsymbol{\rho}), \boldsymbol{z}_{t+1}) = \left(w_{t+1}^f(\boldsymbol{a}, \boldsymbol{\rho})\right)^{-\gamma} \phi_{t+1}(\boldsymbol{z}_{t+1}) \cdot \begin{bmatrix} \nabla_{\boldsymbol{a}_0} w_0^f(\boldsymbol{a}, \boldsymbol{\rho}) \\ \vdots \\ \nabla_{\boldsymbol{a}_{t+1}} w_{t+1}^f(\boldsymbol{a}, \boldsymbol{\rho}) \end{bmatrix}\ , \tag{52}$$

where

$$\nabla_{\boldsymbol{a}_\tau} w_t^f(\boldsymbol{a}, \boldsymbol{\rho}) = \prod_{\tau'=\tau+1}^{t+1} (\boldsymbol{\rho}_{\tau'} - \rho_f \cdot \mathbf{1})\ . \tag{53}$$

These gradients require some "bookkeeping" to keep track of the compounding effects of earlier trades on period-t wealth, captured through (52) and (53). However, the other terms involved, namely $\phi_{t+1}(\boldsymbol{z}_{t+1})$ and $w^f_{t+1}(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho})$, are calculated when solving the frictionless model (50), with wealth $w^f_{t+1}(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho})$ being derived from the optimal trading weights $\hat{\boldsymbol{\theta}}_t$ found for the frictionless model. The gradient penalty (22) is then

$$\hat{\pi}_\nabla(\boldsymbol{a}, \boldsymbol{\rho}, \boldsymbol{z}) = \sum_{t=0}^{T-1} \Bigg(\Big(\nabla_{\boldsymbol{a}} V^f_{t+1}(w^f_{t+1}(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho}), \boldsymbol{z}_{t+1}) - \mathbb{E}\Big[\nabla_{\boldsymbol{a}} V^f_{t+1}(w^f_{t+1}(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho}), \boldsymbol{z}_{t+1}) \,\Big|\, \mathcal{F}_t \Big] \Big)^\intercal (\boldsymbol{a} - \hat{\boldsymbol{\alpha}}^*) \\ + \Big(V^f_{t+1}(w^f_{t+1}(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho}), \boldsymbol{z}_{t+1}) - \mathbb{E}\Big[V^f_{t+1}(w^f_{t+1}(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\rho}), \boldsymbol{z}_{t+1}) \,\Big|\, \mathcal{F}_t \Big] \Big) \Bigg) \,. \tag{54}$$

With this gradient penalty based on the frictionless model, for a given scenario (described by returns $\boldsymbol{\rho}$ and market states $\boldsymbol{z}$), the perfect information inner problem (4) is

$$\max_{\boldsymbol{a} \in \boldsymbol{A}(\boldsymbol{\rho})} \{ U(w_T(\boldsymbol{a}, \boldsymbol{\rho})) - \pi_\nabla(\boldsymbol{a}, \boldsymbol{\rho}, \boldsymbol{z}) \} \,, \tag{55}$$

where $\boldsymbol{A}(\boldsymbol{\rho})$ denotes the set of feasible trades and $w_T(\boldsymbol{a}, \boldsymbol{\rho})$ the terminal wealth given returns $\boldsymbol{\rho}$, *both taking transaction costs into account*. Here, in the inner problem, the DM "knows" all future returns $\boldsymbol{\rho}$ and market states $\boldsymbol{z}$ in advance and chooses trades $\boldsymbol{a}$ to maximize the utility of end-of-horizon wealth $U(w_T(\boldsymbol{a}, \boldsymbol{\rho}))$, paying penalty $\pi_\nabla(\boldsymbol{a}, \boldsymbol{\rho}, \boldsymbol{z})$. Since adding a linear penalty to a concave objective preserves concavity, the inner problem (55) is a convex optimization problem with $n \times T$ decision variables, corresponding to the trades in each of n assets in each of T periods. From Proposition 4.1(ii), using the gradient penalty based on the frictionless model as in (54), the value of the inner problem will be smaller than the value of the frictionless model in every scenario.

In our numerical examples, we will assume proportional transaction costs given by (44) and a power utility function. In this case, we can simplify the optimization problem by decomposing the trades $\boldsymbol{a}$ into positive and negative components $\boldsymbol{a} = \boldsymbol{a}^+ - \boldsymbol{a}^-$ where $\boldsymbol{a}^+, \boldsymbol{a}^- \geq 0$. With proportional transaction costs, the terminal wealth $w_T(\boldsymbol{a}^+ - \boldsymbol{a}^-, \boldsymbol{\rho})$ is then linear in $(\boldsymbol{a}^+, \boldsymbol{a}^-)$ and the inner problem has a "smooth" concave objective function with $2 \times n \times T$ decision variables corresponding to the positive and negative components of each trade in each period.[7]

[7] As mentioned earlier, this is not the penalty construction used in BS (2011). The gradient penalties in BS (2011) differ from the gradient penalties (54) in that the penalties in BS (2011) consider only terminal wealth effects and do not include the term involving the expectation under $\mathcal{F}_t$. Although such gradient penalties are dual feasible, it can be shown that the gradient penalties (54) lead to tighter bounds than the gradient penalties in BS (2011) in every scenario; see Brown and Smith (2014*a*) for a proof.

8.3 Numerical Examples

Table 6 shows results for the dynamic portfolio optimization model on a set of examples using parameters from BS (2011). In these examples, we take $T = 12$ and use $n = 3$ risky assets with a return predictability model calibrated as in BS (2011). We use power utility with risk aversion coefficient γ and proportional transaction costs with rate δ, varying γ and δ each across three values, as in BS (2011). For feasible trading policies, we consider the policies in BS (2011); these policies use various forms of limited-lookahead approximations in selecting trades in each period.

The reported results are in terms of certainty equivalent returns: given a time horizon of T months and a mean utility calculated in a simulation of $\hat{\mu}$, the annualized certainty equivalent return is defined as the constant annual return $\hat{\rho}$ that yields utility $\hat{\mu}$, i.e., the $\hat{\rho}$ that solves:

$$\hat{\mu} = U\Big(w_0 \cdot \hat{\rho}^{T/12}\Big) \tag{56}$$

where w_0 is the initial wealth and U the power utility function. We estimate mean standard errors for these certainty equivalent returns and the duality gaps (the differences between upper and lower bounds on optimal returns) using the "delta method" (see, e.g., Casella and Berger 2002, p. 240) based on a first-order Taylor series expansion of the certainty equivalent formula (i.e., the inverse of equation (56)).

In Table 6, we see that the penalized information relaxation bounds are quite tight, with the average gap between heuristic and bound ranging from 13 basis points to 69 basis points, with higher transaction cost rates leading to somewhat larger gaps. As required by Proposition 4.1(i), the penalized perfect information upper bounds are tighter than the upper bounds from the frictionless model. The perfect information bounds reduce the suboptimality gap implied by the frictionless model by 34% to over 80%; the improvement from the frictionless model is most pronounced in examples with lower risk aversion (e.g., $\gamma = 1.5$ here) which have more rebalancing in the frictionless model and more transaction costs that are not reflected in the frictionless bound. The information relaxation bounds involve solving a convex optimization problem in every scenario; evaluating 1,000 scenarios took less than 10 seconds total on a desktop computer (see also Table 1 of BS (2011)). Using the penalized perfect information bounds, it is clear that the best-performing feasible policies are nearly optimal in each case, a conclusion that is less clear from the frictionless bounds alone.

				Performance Bounds			Suboptimality Gaps	
Risk Aversion (γ)	Trans. Costs (δ)	Best Policy		Information Relaxation		Frictionless Model	Information Relaxation	Frictionless Model
1.5	0.5%	6.57	(0.13)	6.70	(0.00)	7.25	0.13	0.68
1.5	1.0%	6.01	(0.15)	6.25	(0.01)	7.25	0.24	1.24
1.5	2.0%	5.03	(0.19)	5.51	(0.02)	7.25	0.48	2.22
3	0.5%	3.50	(0.07)	3.81	(0.00)	3.99	0.31	0.48
3	1.0%	3.21	(0.10)	3.65	(0.01)	3.99	0.44	0.77
3	2.0%	2.69	(0.12)	3.38	(0.01)	3.99	0.69	1.30
8	0.5%	1.60	(0.03)	1.73	(0.00)	1.79	0.13	0.19
8	1.0%	1.50	(0.04)	1.67	(0.00)	1.79	0.17	0.30
8	2.0%	1.31	(0.05)	1.57	(0.01)	1.79	0.26	0.48

Table 6: Numerical results for dynamic portfolio optimization examples. The policy performances, performance bounds, and gaps are all reported as certainty equivalent returns in %. Mean standard errors for estimated values are in parentheses.

9. Advances in Methodology

In this section, we will briefly discuss some methodological advances in information relaxation techniques. In the next section, we describe applications that also frequently involve methodological advances but tend to be more focused on a specific application.

9.1 Pathwise Optimization

Desai et al. (2012) study the use of perfect information relaxations on high-dimensional optimal stopping problems and develop a *pathwise optimization* approach that optimizes the penalty function. Specifically, they consider a linear-programming-based approximate value function (as in, e.g., de Farias and Van Roy 2003) where the value function is approximated as a weighted combination of a pre-specified set of basis functions. They then use this approximation architecture to construct penalties as in Proposition 3.1, choosing the weights to minimize the penalized perfect information bound. They also use the optimized penalties to suggest good heuristic (feasible) policies. Desai et al. (2012) show that the pathwise optimization problem is a convex optimization problem that can be solved by sample-based approximations and stochastic subgradient methods. In a sense, this pathwise optimization process formalizes some of the trial and error in finding good heuristics and penalties discussed in §5, though there are still significant opportunities for trial and error and iteration in, for example, selecting the set of basis functions.

In addition to theoretical results on the quality of these approximations, Desai et al. (2012)

show that the pathwise optimization approach provides tighter performance bounds than those produced by approximate linear programming alone (see Proposition 3.2(ii) above). They present numerical results for some examples of high-dimensional Bermudan options and show that the pathwise problem can be solved in a few seconds and produces high-quality policies and performance bounds.

Desai et al. (2011) generalize the pathwise optimization approach to Markov decision processes (MDPs) and show that for linear-convex control problems the pathwise optimization approach also leads to convex optimization problems that can be efficiently solved. Recently, Yang et al. (2020) extended the pathwise optimization approach to merchant energy production problems and have developed preconditioning methods that significantly improve the efficiency of the approach.

9.2 Infinite-Horizon Problems

Although the framework and theory described in §2 and §3 focus on finite-horizon DPs, in principle we can apply information relaxation methods in infinite-horizon problems as well. From a practical standpoint, this creates two related challenges. First, with a perfect information relaxation, we need to generate finite scenarios for the inner problems because we cannot perform computations with an infinitely long, randomly generated series. We can use a finite-horizon approximation in these cases but in many problems (e.g., average reward problems or discounted problems with a discount factor close to one), a long horizon may be necessary to obtain a good approximation. Second, even with a perfect information relaxation, the resulting deterministic inner problems may still be difficult to solve, particularly with long time horizons.

Motivated by these challenges, Brown and Haugh (2017) study information relaxation methods for infinite horizon MDPs with discounting. Building upon and generalizing an idea from Rogers (2007), Brown and Haugh (2017) develop a general class of *reformulations* of the primal MDP that involve changing the state transition function and correcting for the change in transition probabilities by multiplying rewards by likelihood ratio factors. The goal of these reformulations is to simplify the resulting information relaxation inner problems and lead to finite-horizon inner problems; the reformulations generalize the idea of using a random stopping time to convert a discounted MDP to a finite-horizon DP (e.g., Puterman 1994, Proposition 5.3.1).

Brown and Haugh (2017) show that weak and strong duality (as in Theorem 3.1 and Theorem 3.4) continues to hold when information relaxations are applied to these reformulated MDPs. They also show that when the penalty is generated as in Proposition 3.1 with a generating function that is a "supersolution" of the MDP, the information relaxation upper bound is tighter

than the performance bound from the supersolution itself (see Proposition 3.2(ii) above). This result is similar to Theorem 4 of Desai et al. (2011), though Brown and Haugh (2017) show the result in a framework involving the reformulations they study.

9.3 Hindsight Analysis

Perfect information bounds without penalties have long been used in the analysis of algorithms in the theoretical computer science and operations research communities and are often referred to as "hindsight bounds" or "offline optimal bounds." In this work, the performance of a particular algorithm is compared against that of a clairvoyant who makes decisions with advance knowledge of all uncertainties. Typically the goal is to prove the algorithm performs well either in terms of a constant factor guarantee or in some asymptotic regime of interest. As a classic example, Talluri and van Ryzin (1998) show that static bid-price policies are asymptotically optimal in network revenue management problems as the initial capacities and time horizon grow large; they also show that perfect information bounds are asymptotically optimal in this setting. Although hindsight (perfect information) bounds have been used successfully in many other problems, as we saw in our numerical examples, with no penalty, the resulting bounds are often quite weak. Balseiro and Brown (2019) consider the use of a penalty within the approach to strengthen the bounds in the theoretical analysis of algorithms. Specifically, they consider a general framework involving finite-horizon MDPs and use an approximate value function (a "Q-factor") in the construction of Proposition 3.1 to generate both a feasible policy as well as a penalized perfect information bound.

Balseiro and Brown (2019) demonstrate the technique on stochastic knapsack problems (Dean et al. 2008) where the size of each item is random and not revealed until an item is selected. They focus on an approximate value function corresponding to a feasible "greedy" policy that ranks items in order of their ratio of value over expected size. Using this approximate value function to generate a penalty in the perfect information problem, they show that the optimal solution in the inner problem is close to the greedy policy in every scenario, which implies the greedy policy is asymptotically optimal as the number of items and capacity grow large. Balseiro and Brown (2019) also apply information relaxations with penalties to show the asymptotic optimality of some simple policies in stochastic scheduling with parallel machines and optimal sequential search problems. Similarly, Balseiro et al. (2018) show how to use information relaxations to analyze the performance of static routing policies in stochastic scheduling problems on unrelated machines. In all of these examples, the penalty is essential in the analysis: the perfect information relaxation bound with zero penalty is not tight in the asymptotic regime of interest.

10. Applications

In this section, we briefly discuss several application areas where researchers have successfully applied information relaxation methods. In each application area, we highlight one or two papers: we briefly describe the model and the use of information relaxations in the problem and discuss some of the issues and challenges in the application. We also typically list several related applications in the area without providing details. This is a rapidly growing research area and, in this review, we aim to be representative rather than exhaustive.

10.1 Energy and Commodity Applications

Many researchers have applied information relaxation methods to problems related to managing commodities, especially energy commodities. As discussed in §1.2, Lai et al. (2010) consider the problem of a merchant managing natural gas storage over time in the presence of stochastic price dynamics where they may inject or withdraw a certain amount of natural gas in each period. The problem is challenging because the natural gas forward curve is represented using a high-dimensional model which leads to a very large state space for the stochastic DP. Lai et al. (2010) develop some policies based on approximations of the value function and use information relaxation methods to generate upper bounds on the optimal value using penalties based on these approximate value functions. In extensive numerical experiments, they find that the performance bound using the penalized information relaxation is typically quite close (within a few percent) of the performance of their policy. In addition, they find that the penalty is essential: with no penalty, the performance bounds are typically quite weak, sometimes a factor of two or more larger than the penalized information relaxation bound.

Information relaxation methods have also been applied to related problems, including other problems in natural gas storage (Secomandi 2015; Nadarajah et al. 2015, 2017; Nadarajah and Secomandi 2018), optimal procurement, processing, and trade of commodities (Devalkar et al. 2011), electricity generation and storage problems (Hinz and Yee 2018; Lin et al. 2020), managing renewable power purchase agreements (Trivella et al. 2018), and merchant energy production (Yang et al. 2020; Trivella et al. 2021).

10.2 Sequential Exploration Problems

Though the applications above relate to the downstream oil and gas and energy markets, Brown and Smith (2013) study an upstream application in oil and gas exploration. In this model, there is

a set of "target" sites in a given geographical area. Each target may contain oil, gas, or be "dry," and there is a joint probability distribution described by a Bayesian network that describes the probability of all possible outcomes at each target and the dependence among them. Brown and Smith (2013) decompose the network into smaller manageable "clusters" and consider an imperfect information relaxation in which each cluster has perfect information about the outcomes for all other clusters. The resulting inner problems are *bandit superprocesses* (Whittle 1980) which are still challenging to solve; Brown and Smith (2013) develop an easy-to-compute upper bound of these information relaxation inner problems. In numerical examples based on a model from a Norwegian oil company, they find that the resulting performance bounds are quite close to the performance of a given feasible policy, also derived from the bandit superprocess models.

10.3 Portfolio Optimization

Researchers have used information relaxation methods to evaluate the performance of heuristic policies in other complex dynamic portfolio optimization problems that are (at a high level) similar to the example considered in §8. For example, Haugh et al. (2016) consider dynamic portfolio optimization involving capital gains taxes. Computing an optimal policy for such problems is very difficult because the prices at which assets are bought and sold in each period must be tracked as part of the state space. Haugh et al. (2016) develop a set of heuristic policies and show how to apply information relaxations to assess the performance of these policies. Specifically, they use gradient penalties as discussed in §4 based on a frictionless model without taxes. The resulting perfect information inner problems are still difficult to solve (they involve the maximization of a nonconcave objective) but they show how to obtain good upper bounds on these inner problems. On a large set of numerical examples, they find that their heuristic policies are typically within a few basis points of the information relaxation bounds, thereby showing the policies are nearly optimal for this difficult problem.

Other applications of information relaxation methods to portfolio optimization include models with transaction costs (Brown and Smith 2011, as discussed in §8; Broadie and Shen 2016; and Mei and Nogales 2018), models with high-dimensional market states (Broadie and Shen 2017), continuous-time models (Ye and Zhou 2015), portfolio execution problems (Haugh and Wang 2014*a*), multiple stopping problems (Chandramouli and Haugh 2012), stochastic regime-switching models (Hinz and Yee 2017), and bilateral counter-party risk models (Bender et al. 2018).

There have also been numerous applications of information relaxation methods in pricing complex options, including the early work of Haugh and Kogan (2004), Rogers (2002), and Andersen

and Broadie (2004) discussed in §1.2 (and reviewed in Glasserman (2003)), in BSS (2010), as well as Desai et al. (2011) which was discussed in §9.1. There are many applications in option pricing that build on this early work.

10.4 Inventory Management

There have also been numerous applications of information relaxation methods in inventory settings. We highlight Bernstein et al. (2016) which considers the problem of a retailer jointly managing inventory levels and prices with a price-sensitive demand model and lead times. With a lead time of L periods, the resulting stochastic DP involves an $L+1$-dimensional state space reflecting supply due in each of the L next periods as well as the inventory level. With significant lead times, the state space can be quite large. The authors develop a myopic heuristic policy for this joint inventory-pricing problem and use information relaxation methods to provide performance bounds. The penalties they use are like the gradient penalties of §4 and are based on a linear approximation of a myopic value function used in their heuristic policy. Their numerical experiments show that the policies and information relaxation bounds are typically within a few percent of each other, with the gaps growing with longer lead times.

As discussed at the end of §6, BSS (2010) consider information relaxation performance bounds for the adaptive inventory model of Treharne and Sox (2002) with uncertainty about the underlying demand distribution. Brown and Smith (2014*b*) consider an inventory problem with lost sales and lead times (see, e.g., Zipkin 2008) and use the gradient penalty construction described in §4 to generate penalties that are linear in actions. In this setting, however, nondifferentiability (induced by the lost sales) plays an important role and gradients must be selected carefully to obtain good bounds. Other applications of information relaxation methods to inventory management problems include stochastic lot-sizing problems (Federgruen et al. 2015), Bayesian inventory management with demand change-points (Wang and Mersereau 2017), inventory management with autoregressive demand distributions (Brown and Haugh 2017), perishable inventory models (Lin et al. 2020), and managing inventory for multi-dose vaccines such as many COVID-19 vaccines (Shumsky et al. 2021).

10.5 Reinforcement Learning

Several researchers have recently applied information relaxations to problems and algorithms that arise in the area of reinforcement learning. We highlight Min et al. (2019) which considers a finite-

horizon, Bayesian multi-armed bandit problem where a decision-maker has prior beliefs about an unobservable parameter on each arm. In each period, the decision-maker selects an arm to pull, collects a random reward, and updates their beliefs about the arm's parameters. Although this problem has a known optimal solution in the infinite-horizon setting with discounting (in each period, pull the arm with the highest Gittens index), there is no known simple form for the optimal policy in the finite-horizon setting. A classical approach to such problems – Thompson sampling (TS) (Thompson 1933) – involves information relaxations: in each period, each arm's parameter is fictitiously "sampled" from the current priors and the arm with the highest resulting expected reward is selected in that period.

Min et al. (2019) show that the TS policy and the TS clairvoyant benchmark fits into an information relaxation framework with a particular penalty function; this penalty function effectively replaces the realized rewards associated with each arm with their expected value given the parameter (similar to the "smoothing" penalties considered in the examples of §6 and §7). The authors also develop three other penalty functions that can be used to generate feasible policies and also provide upper bounds. Min et al. (2019) (a) show that the upper bounds using these three penalties are tighter than the upper bound using Thompson sampling; (b) prove that the feasible policies generated by two of the three penalties lead to similar regret as Thompson sampling; and (c) in extensive numerical experiments demonstrate strong empirical performance of these policies compared to TS.

As a second application area to highlight, we consider Monte Carlo tree search (MCTS), MCTS is a widely studied algorithm used to solve decision problems (including games) in the artificial intelligence literature. Roughly speaking, the MCTS algorithm works by maintaining an approximation of a full decision tree (representing a finite-state and finite-action Markov decision process) then progressively expanding new nodes of the approximation. Although MCTS converges to an optimal action in the current state, in the worst case this requires a construction of the full decision tree. Jiang et al. (2020) develop a primal-dual form of MCTS with improved performance that uses information relaxation methods. Their method relies on generating information relaxation bounds in the tree expansion steps to "prune" parts of the tree that should not be visited by an optimal policy. Jiang et al. (2020) show that this approach retains the theoretical convergence behavior of MCTS and leads to improved performance empirically on a challenging application in ride-sharing. In a similar vein, El Shar and Jiang (2020) use information relaxation bounds to improve the performance of Q-learning algorithms.

10.6 Other Applications

Information relaxation methods have also been applied to other problems related to operations management, including vehicle routing (Goodson et al. 2013; Kullman et al. 2021), network revenue management (Brown and Smith 2014*b*), queueing (Brown and Haugh 2017; Farahani et al. 2020), ambulance dispatch (Marla and Bassamboo 2020), and repositioning of shipping containers (Lu et al. 2020). Other more methodologically oriented application areas include control theory (Desai et al. 2011; Haugh and Lim 2012), partially observable Markov decision processes (Haugh and Lacedelli 2019), and game theory (Haugh and Wang 2014*b*; Kogan and Mitra 2019).

11. Conclusions

In this paper, we have reviewed the information relaxation approach for obtaining performance bounds in stochastic DPs, describing the fundamental ideas underlying the approach and how one can apply it. We hope that this paper will be useful to researchers looking to learn about information relaxation techniques with the goal of advancing the methodology or applying it to new applications or theoretical questions of interest.

Looking forward, there are many interesting directions for future research related to information relaxation methods, including the following:

(i) Methodological challenges remain in applying information relaxation techniques in some problems, particularly when the resulting inner problems have large state spaces. For example, in the dynamic assortment examples in §7, the inner problems were linked across products (by the linking constraint (30)) and we used Lagrangian relaxations to relax this constraint and decompose the problem across products. However, if we had rewards based on choice models that capture substitution effects between products (e.g., multinomial logit models), the inner problem could not be decomposed in this way. However, one could perhaps use the reformulation techniques described in Rogers (2007) and Brown and Haugh (2017) to reduce the number of states that need to be considered in these inner problems. We believe that the design of reformulations to obtain good bounds is an underexplored research area.

(ii) The use of information relaxation techniques as a method for refining "hindsight bounds" in theoretical analysis of algorithms represents another exciting direction for future research. Balseiro and Brown (2019) develop some results along these lines and illustrate these ideas on three examples. We believe similar techniques can be applied in many other problems.

(iii) It would be interesting to develop methods that systematically generate feasible policies from information relaxations. As discussed in §5, we have used information relaxations in applications to improve feasible policies but typically this process requires some trial and error. Could this be automated in some way? For example, can information relaxation methods be used as part of a "primal-dual" algorithm that progressively updates the policy (or value function approximation) and penalty? The recent work of Min et al. (2019), Trivella et al. (2018), and Chen et al. (2020) provides some ideas along these lines.

(iv) Building upon the previous point, there are a number of interesting connections to methods in reinforcement learning. The papers described in §10.5 may serve as starting points for further developments in this area.

Although the list of research areas above focuses on methodological questions related to the information relaxation approach, we would also like to continue to see applications of information relaxation techniques in a wide variety of settings. Such applications will no doubt lead to further methodological challenges and ideas for future research.

References

Adelman, D. and Mersereau, A. J. (2008), 'Relaxations of weakly coupled stochastic dynamic programs', *Operations Research* **56**(3), 712–727.

Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1988), *Network flows*, Cambridge, Mass.: Alfred P. Sloan School of Management.

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

Balseiro, S. R. and Brown, D. B. (2019), 'Approximations to stochastic dynamic programs via information relaxation duality', *Operations Research* **67**(2), 577–597.

Balseiro, S. R., Brown, D. B. and Chen, C. (2018), 'Static routing in stochastic scheduling: Performance guarantees and asymptotic optimality', *Operations Research* **66**(6), 1641–1660.

Bender, C., Gärtner, C. and Schweizer, N. (2018), 'Pathwise dynamic programming', *Mathematics of Operations Research* **43**(3), 965–995.

Bernstein, F., Li, Y. and Shang, K. (2016), 'A simple heuristic for joint inventory and pricing models with lead time and backorders', *Management Science* **62**(8), 2358–2373.

Bertsekas, D. P. (2017), *Dynamic Programming and Optimal Control*, Vol. 1, 4th edn, Athena Scientific.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996), *Neuro-dynamic programming*, Athena Scientific.

Bertsimas, D. and Tsitsiklis, J. N. (1997), *Introduction to Linear Optimization*, Athena Scientific Series in Optimization and Neural Computation, 6, Athena Scientific.

Broadie, M. and Shen, W. (2016), 'High-dimensional portfolio optimization with transaction costs', *International Journal of Theoretical and Applied Finance* **19**(04).

Broadie, M. and Shen, W. (2017), 'Numerical solutions to dynamic portfolio problems with upper bounds', *Computational Management Science* **14**(2), 215–227.

Brown, D. B. and Haugh, M. B. (2017), 'Information relaxation bounds for infinite horizon markov decision processes', *Operations Research* **65**(5), 1355–1379.

Brown, D. B. and Smith, J. E. (2011), 'Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds', *Management Science* **57**(10), 1752–1770.

Brown, D. B. and Smith, J. E. (2013), 'Optimal sequential exploration: Bandits, clairvoyants, and wildcats', *Operations research* **61**(3), 644–665.

Brown, D. B. and Smith, J. E. (2014*a*), 'Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds (addendum on gradient penalties)'.

Brown, D. B. and Smith, J. E. (2014*b*), 'Information relaxations, duality, and convex stochastic dynamic programs', *Operations Research* **62**(6), 1394–1415.

Brown, D. B. and Smith, J. E. (2020), 'Index policies and performance bounds for dynamic selection problems', *Management Science* **66**(7), 3029–3050.

Brown, D. B., Smith, J. E. and Sun, P. (2010), 'Information relaxations and duality in stochastic dynamic programs', *Operations Research* **58**(4-part-1), 785–801.

Caro, F. and Gallien, J. (2007), 'Dynamic assortment with demand learning for seasonal consumer goods', *Management Science* **53**(2), 276–292.

Chandramouli, S. S. and Haugh, M. B. (2012), 'A unified approach to multiple stopping and duality', *Operations Research Letters* **40**(4), 258–264.

Chen, N., Ma, X., Liu, Y. and Yu, W. (2020), 'Information relaxation and a duality-driven algorithm for stochastic dynamic programs', *arXiv preprint arXiv:2007.14295* .

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

Dean, B. C., Goemans, M. X. and Vondrák, J. (2008), 'Approximating the stochastic knapsack problem: The benefit of adaptivity', *Mathematics of Operations Research* **33**(4), 945–964.

Desai, V., Farias, V. F. and Moallemi, C. C. (2012), 'Pathwise optimization for optimal stopping problems', *Management Science* **58**, 2292–2308.

Desai, V., Farias, V. and Moallemi, C. (2011), 'Bounds for markov decision processes', *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control,(FL Lewis, D. Liu, eds.)* pp. 452–473.

Devalkar, S., Anupindi, R. and Sinha, A. (2011), 'Integrated optimization of procurement, processing, and trade of commodities', *Operations Research* **59**, 1369–1381.

El Shar, I. and Jiang, D. (2020), Lookahead-bounded q-learning, *in* 'International Conference on Machine Learning', PMLR, pp. 8665–8675.

Farahani, M. H., Dawande, M. and Janakiraman, G. (2020), 'Order now, pickup in 30 minutes: Managing queues with static delivery guarantees'. Forthcoming in *Operations Research*, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3557605`.

Federgruen, A., Guetta, D. and Iyengar, G. (2015), Information relaxation-based lower bounds for the stochastic lot sizing problem with advanced demand information, Technical report, Working paper, Columbia University.

Feldman, J., Henzinger, M., Korula, N., Mirrokni, V. S. and Stein, C. (2010), Online stochastic packing applied to display ad allocation, *in* 'Proceedings of the 18th annual European conference on Algorithms: Part I', ESA'10, Springer-Verlag, pp. 182–194.

Florian, M., Lenstra, J. K. and Rinnooy Kan, A. (1980), 'Deterministic production planning: Algorithms and complexity', *Management Science* **26**(7), 669–679.

Glasserman, P. (2003), *Monte Carlo methods in financial engineering*, Vol. 53, Springer Science & Business Media.

Goodson, J. C., Ohlmann, J. W. and Thomas, B. W. (2013), 'Rollout policies for dynamic solutions to the multivehicle routing problem with stochastic demand and duration limits', *Operations Research* **61**(1), 138–154.

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

Haugh, M. B. and Lacedelli, O. R. (2019), 'Information relaxation bounds for partially observed markov decision processes', *IEEE Transactions on Automatic Control* **65**(8), 3256–3271.

Haugh, M. B. and Lim, A. E. (2012), 'Linear-quadratic control and information relaxations', *Operations Research Letters* **40**(6), 521–528.

Haugh, M., Iyengar, G. and Wang, C. (2016), 'Tax-aware dynamic asset allocation', *Operations Research* **64**(4), 849–866.

Haugh, M. and Wang, C. (2014*a*), 'Dynamic portfolio execution and information relaxations', *SIAM Journal on Financial Mathematics* **5**(1), 316–359.

Haugh, M. and Wang, C. (2014*b*), 'Information relaxations and dynamic zero-sum games', *arXiv preprint arXiv:1405.4347* .

Hawkins, J. T. (2003), A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications, PhD thesis, Massachusetts Institute of Technology.

Henderson, S. G. and Glynn, P. W. (2002), 'Approximating martingales for variance reduction in markov process simulation', *Mathematics of Operations Research* **27**(2), 253–271.

Hinz, J. and Yee, J. (2017), 'Stochastic switching for partially observable dynamics and optimal asset allocation', *International Journal of Control* **90**(3), 553–565.

Hinz, J. and Yee, J. (2018), 'Optimal forward trading and battery control under renewable electricity generation', *Journal of Banking & Finance* **95**, 244–254.

Jiang, D. R., Al-Kanj, L. and Powell, W. B. (2020), 'Optimistic monte carlo tree search with sampled information relaxation dual bounds', *Operations Research* **68**(6), 1678–1697.

Kogan, L. and Mitra, I. (2019), 'Near-rational equilibria in heterogeneous-agent models: A verification method'. Working paper, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3465120`.

Kullman, N. D., Goodson, J. C. and Mendoza, J. E. (2021), 'Electric vehicle routing with public charging stations', *Transportation Science* **55**(3), 637–659.

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

Lin, Q., Nadarajah, S. and Soheili, N. (2020), 'Revisiting approximate linear programming: Constraint-violation learning with applications to inventory control and energy storage', *Management Science* **66**(4), 1544–1562.

Lu, T., Lee, C.-Y. and Lee, L.-H. (2020), 'Coordinating pricing and empty container repositioning in two-depot shipping systems', *Transportation Science* **54**(6), 1697–1713.

Luenberger, D. G. and Ye, Y. (2016), *Linear and Nonlinear Programming: Fourth edition*, Springer.

Marla, L. and Bassamboo, A. (2020), 'Information relaxation bounds for evaluating ambulance dispatch and allocation policies'. Working paper, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3519306`.

Mei, X. and Nogales, F. J. (2018), 'Portfolio selection with proportional transaction costs and predictability', *Journal of Banking & Finance* **94**, 131–151.

Meinshausen, N. and Hambly, B. M. (2004), 'Monte carlo methods for the valuation of multiple-exercise options', *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* **14**(4), 557–583.

Min, S., Maglaras, C. and Moallemi, C. C. (2019), Thompson sampling with information relaxation penalties, *in* 'Advances in Neural Information Processing Systems', pp. 3549–3558.

Nadarajah, S., Margot, F. and Secomandi, N. (2015), 'Relaxations of approximate linear programs for the real option management of commodity storage', *Management Science* **61**(12), 3054–3076.

Nadarajah, S., Margot, F. and Secomandi, N. (2017), 'Comparison of least squares monte carlo methods with applications to energy real options', *European Journal of Operational Research* **256**(1), 196–204.

Nadarajah, S. and Secomandi, N. (2018), 'Merchant energy trading in a network', *Operations Research* **66**(5), 1304–1320.

Powell, W. B. (2007), *Approximate Dynamic Programming: Solving the curses of dimensionality*, Vol. 703, John Wiley & Sons.

Puterman, M. L. (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st edn, John Wiley & Sons, Inc., USA.

Rockafellar, R. T. and Wets, R.-B. (1976), Nonanticipativity and l 1-martingales in stochastic optimization problems, *in* 'Stochastic Systems: Modeling, Identification and Optimization, II', Springer, pp. 170–187.

Rogers, L. (2002), 'Monte carlo valuation of American options', *Mathematical Finance* **12**, 271–286.

Rogers, L. (2007), 'Pathwise stochastic optimal control', *SIAM Journal on Control and Optimization* **46**, 1116–1132.

Schoenmakers, J. (2012), 'A pure martingale dual for multiple stopping', *Finance and Stochastics* **16**(2), 319–334.

Secomandi, N. (2015), 'Merchant commodity storage practice revisited', *Operations Research* **63**(5), 1131–1143.

Shapiro, A., Dentcheva, D. and Ruszczyński, A. (2009), *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Series on Optimization 9, Society for Industrial and Applied Mathematics (SIAM).

Shumsky, R. A., Smith, J. E., Hoen, A. G. and Gilbert, M. (2021), 'Allocating covid-19 vaccines: Save one for the second dose?'. Working paper, Dartmouth College.

Song, J.-S. and Zipkin, P. (1993), 'Inventory control in a fluctuating demand environment', *Operations Research* **41**(2), 351–370.

Talluri, K. and van Ryzin, G. (1998), 'An analysis of bid-price controls for network revenue management', *Management Science* **44**(11), 1577–1593.

Thompson, W. R. (1933), 'On the likelihood that one unknown probability exceeds another in view of the evidence of two samples', *Biometrika* **25**(3/4), 285–294.

Treharne, J. T. and Sox, C. R. (2002), 'Adaptive inventory control for nonstationary demand and partial information', *Management Science* **48**(5), 607–624.

Trivella, A., Mohseni Taheri, D. and Nadarajah, S. (2018), 'Meeting corporate renewable power targets'. Working paper, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3294724`.

Trivella, A., Nadarajah, S., Fleten, S.-E., Mazieres, D. and Pisinger, D. (2021), 'Managing shutdown decisions in merchant commodity and energy production: A social commerce perspective', *Manufacturing & Service Operations Management* **23**(2), 311–330.

Wang, Z. and Mersereau, A. J. (2017), 'Bayesian inventory management with potential change-points in demand', *Production and Operations Management* **26**(2), 341–359.

Whittle, P. (1980), 'Multi-armed bandits and the gittins index', *Journal of the Royal Statistical Society: Series B (Methodological)* **42**(2), 143–149.

Yang, B., Nadarajah, S. and Secomandi, N. (2020), 'Pathwise optimization for merchant energy production'. Working paper, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3510676`.

Ye, F. and Zhou, E. (2015), 'Information relaxation and dual formulation of controlled markov diffusions', *IEEE Transactions on Automatic Control* **60**(10), 2676–2691.

Zipkin, P. (2008), 'On the structure of lost-sales inventory models', *Operations Research* **56**(4), 937–944.

Zipkin, P. H. (2000), *Foundations of inventory management*, Boston: McGraw Hill.