Information Relaxation Bounds for Infinite Horizon Markov Decision Processes

David B. Brown
Fuqua School of Business
Duke University
dbbrown@duke.edu

Martin B. Haugh
Department of IE&OR
Columbia University
mh2078@columbia.edu

This version: April 1, 2016
First version: September 6, 2014

Abstract

We consider the information relaxation approach for calculating performance bounds for stochastic dynamic programs (DPs), following Brown, Smith, and Sun (2010). This approach generates performance bounds by solving problems with relaxed nonanticipativity constraints and a penalty that punishes violations of these constraints. In this paper, we study infinite horizon DPs with discounted costs and consider applying information relaxations to reformulations of the DP. These reformulations use different state transition functions and correct for the change in state transition probabilities by multiplying by likelihood ratio factors. These reformulations can greatly simplify solution of the information relaxations, both in leading to finite horizon subproblems and by reducing the number of states that need to be considered in these subproblems. We show that any reformulation leads to a lower bound on the optimal cost of the DP when used with an information relaxation and a penalty built from a broad class of approximate value functions. We refer to this class of approximate value functions as *subsolutions*, and this includes approximate value functions based on Lagrangian relaxations as well as those based on approximate linear programs. We show that the method, in theory, recovers a tight lower bound using any reformulation, and is guaranteed to improve upon the lower bounds from subsolutions. Finally, we apply the method to an inventory control application with an autoregressive demand process, as well as dynamic service allocation in a multiclass queue. In our examples, we find that the information relaxation lower bounds are easy to calculate and are very close to the expected cost using simple heuristic policies, thereby showing that these heuristic policies are nearly optimal.

Keywords: infinite horizon dynamic programs, information relaxations, Lagrangian relaxations, inventory control, multiclass queues.

1. Introduction

Dynamic programming provides a powerful framework for analyzing sequential decision making in stochastic systems. The resulting problems are often very difficult to solve, however, especially for systems that can evolve over many states. When dynamic programs (DPs) are too difficult to solve, we must resort to heuristic policies that are generally suboptimal. In evaluating these heuristic policies, bounds on suboptimality can be very helpful: given a good bound on the performance of the (unknown) optimal policy, we may find that a heuristic policy is nearly optimal, and thus conclude that efforts to improve the heuristic are not worthwhile.

In this paper, we consider the information relaxation approach for calculating performance bounds for stochastic dynamic programs (DPs), following Brown, Smith, and Sun (2010; hereafter BSS). In BSS, bounds are generated by (i) relaxing the nonanticipativity constraints that require the decision maker to make decisions based only on the information available at the time the decision is made and (ii) incorporating a penalty that punishes violations of these nonanticipativity constraints. For a general class of finite horizon DPs, BSS show how to construct penalties that are dual feasible and lead to lower bounds using this approach, and refer to this as weak duality. BSS also show strong duality in that there exists an "ideal" penalty for which the resulting lower bound equals the optimal cost. As an illustrative example, later in this paper we study a multiclass queueing application in which a server allocates service to different customer classes in a queue. A perfect information relaxation in this problem involves making service decisions with advance knowledge of all future arrivals and service times; by repeatedly sampling arrival and service scenarios and solving a perfect information "inner problem" in each scenario, we can obtain a lower bound on the cost associated with an optimal service policy.

Our first goal in this paper is to study the use of information relaxations for a class of infinite horizon DPs with discounted costs. We show that weak and strong duality also go through in this setting. This is conceptually straightforward: for example, a perfect information relaxation corresponds to revealing an infinitely long sample path of all future uncertainties. In practice, however, we need to account for the fact that we cannot simulate infinitely long sequences. One resolution to this is to consider applying information relaxations to a standard, equivalent formulation (see, e.g., Puterman 1994, Ch. 5) in which there is no discounting, but rather an absorbing, costless state that is reached with probability $1 - \delta$ in each period (δ is the discount factor). With perfect information, the resulting inner problems are finite horizon problems.

We illustrate the approach on an infinite horizon inventory control problem where the demand distribution in each period depends on past demands over several periods (e.g., when demands follow autoregressive processes). The resulting DP may have many states due to the need to include past demand realizations over several periods as part of the state. In a perfect information relaxation, all demands are revealed (up to the absorbing time), and the problem is much easier to solve, as only the inventory level needs to be tracked as an explicit state in every period. In our examples, we evaluate the performance of a myopic policy that only considers costs one period ahead. With a perfect information relaxation and a penalty based on the myopic approximate value function, we find in our examples that the myopic policy is nearly optimal.

The inventory control application represents an example where direct application of information relaxations can work well, as much of the complexity in the problem results from a high-dimensional stochastic process (demands) that does not depend on actions. When demands are revealed, this greatly reduces the number of states. In other problems, however, a direct application of information relaxations can be challenging, e.g., when a high-dimensional component of the state space is endogenously affected by actions. This is the case in the multiclass queueing application we later study. We address this challenge by working with *reformulations* of the primal DP. Specifically, we allow for general reformulations of the state transition function with changes in state transition probabilities corrected through multiplication by likelihood ratio terms. This idea is motivated by the work of Rogers (2007), who studies reformulations in which state transitions do not depend on actions. We generalize this idea by allowing, for instance, reformulations in which states have a partial dependence on actions.

These reformulations can greatly simplify solving the information relaxation inner problems while still leading to high-quality lower bounds. Many reformulations are possible, and we show that even reformulations that are not equivalent to the original DP (in that the expected costs for some policies may be different under the reformulation) lead to lower bounds on the optimal cost of the original DP. Specifically, weak duality holds for any reformulation of the state transition function, provided we generate penalties from approximate value functions that are *subsolutions* to the optimal value function (a subsolution is a function that satisfies the Bellman optimality equation with an inequality in place of equality). It is well-known that subsolutions provide lower bounds on the optimal value function in every state, and we show that by using the information relaxation approach with a penalty built from a subsolution, we improve the lower bound from the subsolution. We also show that strong duality continues to hold for arbitrary reformulations.

Finally, we apply the approach to the problem of service allocation in a multiclass queue. Such models are well-studied and of significant practical interest. The particular model we study is complicated due to the fact that delay costs are assumed to be convex: good service policies need to judiciously balance serving customers with shorter service times against serving those that are more congested in the system. Our examples are far too large to solve exactly. We consider both *uncontrolled* and *partially controlled formulations* of this problem; the information relaxations of these formulations have far fewer states to consider than the original DP. In our examples, we obtain lower bounds that are very close to the expected cost of the heuristic policies we study, thereby showing that the heuristic policies are nearly optimal.

1.1. Literature Review and Outline

BSS follows Haugh and Kogan (2004) and Rogers (2002), who independently developed methods for calculating performance bounds for the valuation of American options. BSS extends these ideas to general, finite horizon DPs with general (i.e., imperfect) information relaxations. Rogers (2007) develops perfect information relaxations for Markov decision processes and he uses a change of measure approach similar to the "uncontrolled formulation" case in our approach. Desai, Farias, and Moallemi (2011) show how to improve on bounds from approximate linear programming with perfect information relaxations, and Brown and Smith (2011, 2014) show that using information relaxations with "gradient penalties" improves bounds from relaxed DP models when the DP has a convex structure. In independent work, Ye, Zhu, and Zhou (2014) study weakly coupled DPs and show that improved bounds can be obtained by combining perfect information relaxations with Lagrangian relaxations; they do not consider reformulations. We also use Lagrangian relaxations in our multiclass queueing examples and find the reformulations to be quite useful in that application.

Information relaxations have been used in a variety of applications, including valuation of American options in Rogers (2002), Haugh and Kogan (2004), Andersen and Broadie (2004), BSS, Chen and Glasserman (2007), and Desai, Farias, and Moallemi (2012). Information relaxations are applied to inventory management problems in BSS. Lai, Margot, and Secomandi (2010) use information relaxations in studying valuations for natural gas storage, as do Devalkar, Anupindi, and Sinha (2011) in an integrated model of procurement, processing, and commodity trading. Information relaxations are used in Brown and Smith (2011) for dynamic portfolio optimization problems with transaction costs, as well as in Haugh and Wang (2014a) for dynamic portfolio execution problems and Haugh, Iyengar, and Wang (2014) for dynamic portfolio optimization with taxation. Brown and Smith (2014) apply information relaxations to network revenue management problems and inventory management with lost sales and lead times. Kim and Lim (2016) develop a weak duality result for robust multi-armed bandits. Kogan and Mitra (2013) use information relaxations to evaluate the accuracy of numerical solutions to general equilibrium problems in an infinite horizon setting; in their examples they use finite horizon approximations. Haugh and Wang (2014b) develop information relaxations for dynamic zero-sum games. A recurring theme in these papers is that relatively easy-to-compute policies are often nearly optimal; the bounds from information relaxations are essential in showing that.

In Section 2, we formulate the general class of primal DPs we study and develop the basic duality results, analogous to the results in BSS. We illustrate the method on the inventory control application in Section 3. In Section 4, we describe the reformulation framework, then describe the duality results applied to the reformulations. Section 5 presents the multiclass queueing application. Section 6 discusses directions for future research. Most proofs and some detailed derivations are presented in Appendix A.

2. Problem Formulation and Basic Duality Results

In this section, we first describe the class of infinite horizon stochastic dynamic programs that we will study, and then describe the basic theory of information relaxations and duality applied to these problems. Section 2.2 essentially reviews the setup in BSS and provides infinite horizon analogues of the duality results in BSS.

2.1. The Primal Dynamic Program

We will work with a Markov decision process (MDP) formulation. Time is discrete and indexed by t, starting with $t = 0$. We let x_t and a_t denote the state and action, respectively, at time t, and the initial state x_0 is known. We denote the state space by $\mathbb{X}$ and assume this is a countable set. We denote the action space by $\mathbb{A}$ and the feasible action set in state x_t by $A(x_t)$ and assume $A(x_t)$ is finite for every $x_t \in \mathbb{X}$. We let $p(x_{t+1}|x_t, a_t)$ denote the probability of a state transition to $x_{t+1} \in \mathbb{X}$ when the state is $x_t \in \mathbb{X}$ and $a \in A(x_t)$ is the action. Costs depend on states and actions and are denoted by $c(x_t, a_t)$, which we assume to be uniformly bounded. The objective is to minimize the total expected discounted costs over an infinite horizon, where $\delta \in (0, 1)$ is a known discount factor. We refer to this problem as the *primal DP*.

It will be useful for us to equivalently write the problem in terms of an exogenous stochastic process and a state transition function. We let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space, representing the full product space for a sequence $\{w_t\}_{t\geq 1}$ of IID random variables, each supported on a set $\mathbb{W} \subseteq \mathbb{R}$. We refer to realizations of $\{w_t\}_{t\geq 1}$ as *sample paths.* We assume that the values $w_1, \ldots, w_t$ are known at time t, and we let $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$, where $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$ for all $t \geq 0$, denote the corresponding *natural filtration* describing this information at time t. We take $\mathcal{F}_0 = \{\varnothing, \Omega\}$, i.e., initially "nothing is known" about the sequence of w_t's. States evolve as $x_{t+1} = f(x_t, a_t, w_{t+1})$ for a given state transition function $f : \mathbb{X} \times \mathbb{A} \times \mathbb{W} \mapsto \mathbb{R}$, where, for all t, $x_{t+1} \in \mathbb{X}$, $x_t \in \mathbb{X}$, and $a_t \in A(x_t)$, and $\mathbb{P}(\{w_{t+1} \in \mathbb{W} : f(x_t, a_t, w_{t+1}) = x_{t+1}\}|\mathcal{F}_t) = p(x_{t+1}|x_t, a_t)$ holds.[1] Throughout the paper, $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$. Thus, for any $v : \mathbb{X} \mapsto \mathbb{R}$, and any $x_t \in \mathbb{X}$ and $a \in A(x_t)$, we have $\mathbb{E}[v(f(x_t, a_t, w_{t+1}))] = \sum_{x_{t+1}\in\mathbb{X}} p(x_{t+1}|x_t, a_t)v(x_{t+1})$.

A *policy* $\alpha := \{\alpha_t\}_{t\geq 0}$ is a sequence of functions, each mapping from $\{w_t\}_{t\geq 1}$ to feasible actions; we will focus on non-randomized policies. We let $\mathcal{A}$ denote the set of all policies. A policy is *primal feasible* if each α_t is $\mathcal{F}_t$-measurable, i.e., α is $\mathbb{F}$-adapted, and we let $\mathcal{A}_{\mathbb{F}}$ denote this set of policies. We will sometimes restrict (without loss of optimality in the primal DP) to *stationary* policies, where α_t only depends on x_t and α_t is constant over t. We let $\mathcal{A}_{\mathrm{S}}$ denote the set of stationary policies. We denote the expected cost of a feasible policy $\alpha \in \mathcal{A}_{\mathbb{F}}$ starting in state x_0 by $v_\alpha(x_0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \delta^t c(x_t, \alpha_t)\right]$. Finally, $v^\star(x)$ denotes the optimal expected cost from state x among all $\mathbb{F}$-adapted policies.

The following results are standard (e.g., Puterman 1994, Thms. 6.1.1 and 6.2.10):

(a) For any $\alpha \in \mathcal{A}_{\mathrm{S}}$, the associated value function $v_\alpha(x)$ satisfies, for every $x \in \mathbb{X}$,

$$v_\alpha(x) \quad = \quad c(x, \alpha(x)) + \delta\mathbb{E}[v_\alpha(f(x, \alpha(x), w))].$$

(b) There exists an optimal stationary policy. Moreover, the optimal value function $v^\star$ is bounded and satisfies, for every $x \in \mathbb{X}$,

$$v^\star(x) \quad = \quad \min_{a\in A(x)} \{c(x, a) + \delta\mathbb{E}[v^\star(f(x, a, w))]\}. \tag{1}$$

We assume a countable state space and finite action sets primarily to simplify exposition and to avoid

[1] This can always be done, for instance, by treating states as natural numbers, taking each w_t to be uniformly distributed on $\mathbb{W} = [0, 1]$, and setting $f(x, a, w)$ to be the generalized inverse distribution function corresponding to the state transition probabilities $p(\cdot|x, a)$ for all $x \in \mathbb{X}$, $a \in A(x)$.

technical issues related to measurability that are not our focus. All of our examples satisfy these assumptions. The basic theory of information relaxations does not require such assumptions. In moving to more general state spaces, care needs to be taken to ensure the existence of optimal policies as well as appropriate measurability properties of value functions; see, e.g., Bertsekas and Shreve (1996) or, for a more recent overview of key results, Feinberg (2011).

2.2. Information Relaxations and Duality

In the primal DP, feasible policies must be $\mathbb{F}$-adapted, or *nonanticipative* in that they cannot depend on the realizations of future uncertainties. We consider relaxations of the primal DP that relax the requirement that policies be nonanticipative and impose penalties that punish violations of these constraints. We define relaxations of the nonanticipativity constraints by considering alternative information structures.

Formally, we say that another filtration $\mathbb{G} = \{\mathcal{G}_t\}_{t\geq 0}$ is a *relaxation* of the natural filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ if, for each t, $\mathcal{F}_t \subseteq \mathcal{G}_t$. In words, $\mathbb{G}$ being a relaxation of $\mathbb{F}$ means that the decision maker knows at least as much (and possibly more) about uncertainties in every period under $\mathbb{G}$ than under $\mathbb{F}$. For example, we will often focus on the *perfect information relaxation*, which is given by taking $\mathcal{G}_t = \mathcal{F}$ for all t. We let $\mathcal{A}_\mathbb{G}$ denote the set of policies that are adapted to $\mathbb{G}$. For any relaxation $\mathbb{G}$ of $\mathbb{F}$, we have $\mathcal{A}_\mathbb{F} \subseteq \mathcal{A}_\mathbb{G}$; thus, as we relax the filtration, we expand the set of feasible policies.

Penalties, like costs, depend on states and actions and are incurred in each period. In their basic duality results, BSS consider a general set of *dual feasible* penalties, which are functions added to the cost that do not have a positive expected value for any primal feasible policy $\alpha \in \mathcal{A}_\mathbb{F}$. BSS show how to construct dual feasible penalties through differences in conditional expectations of any "generating function." We will essentially follow this construction.

Specifically, we generate penalties by adding the terms $\pi_t := \delta^{t+1}(\mathbb{E}[v(f(x_t, a_t, w))] - v(x_{t+1}))$ to the costs in each period, where $v : \mathbb{X} \mapsto \mathbb{R}$ is a bounded function. For any $\alpha \in \mathcal{A}_\mathbb{F}$, the sequence $\{\pi_t\}_{t\geq 0}$ are martingale differences under $\mathbb{F}$, and therefore for any finite T, the sum $\Pi_T := \sum_{t=0}^{T} \pi_t$ is a martingale under $\mathbb{F}$. Since v is bounded and $\delta \in (0,1)$, Π_T is bounded, and thus by the bounded convergence theorem, we conclude, for any $\alpha \in \mathcal{A}_\mathbb{F}$, that

$$\mathbb{E}[\, \textstyle\sum_{t=0}^{\infty} \delta^{t+1} \left(\mathbb{E}[v(f(x_t, \alpha_t, w))] - v(x_{t+1})\right)] \quad = \quad 0. \tag{2}$$

These terms may, however, have positive expected value for policies that violate the nonanticipativity constraints, and thus can serve as penalties in the information relaxations.

We can obtain a lower bound on the expected discounted cost associated with any primal feasible policy by relaxing the nonanticipativity constraints on policies and imposing a penalty constructed in this way. This is stated in the following "weak duality" result, which is analogous to Lemma 2.1 in BSS (2010). In what follows, we let $\mathcal{V}$ denote the set of bounded, real-valued functions on $\mathbb{X}$.

Lemma 2.1 (Weak Duality). *For any $\alpha \in \mathscr{A}_{\mathbb{F}}$, any relaxation $\mathbb{G}$ of $\mathbb{F}$, and any $v \in \mathscr{V}$, then*

$$v_\alpha(x_0) \;\geq\; \inf_{\alpha_G \in \mathscr{A}_{\mathbb{G}}} \mathbb{E}\left[\textstyle\sum_{t=0}^{\infty} \delta^t \left(c(x_t, \alpha_{G,t}) + \delta \mathbb{E}[v(f(x_t, \alpha_{G,t}, w))] - \delta v(x_{t+1})\right)\right] .$$

Proof. We have

$$\begin{aligned} v_\alpha(x_0) &= \mathbb{E}\left[\textstyle\sum_{t=0}^{\infty} \delta^t c(x_t, \alpha_t)\right] \\ &= \mathbb{E}\left[\textstyle\sum_{t=0}^{\infty} \delta^t \left(c(x_t, \alpha_t) + \delta \mathbb{E}[v(f(x_t, \alpha_t, w))] - \delta v(x_{t+1})\right)\right] \\ &\geq \inf_{\alpha_G \in \mathscr{A}_{\mathbb{G}}} \mathbb{E}\left[\textstyle\sum_{t=0}^{\infty} \delta^t \left(c(x_t, \alpha_{G,t}) + \delta \mathbb{E}[v(f(x_t, \alpha_{G,t}, w))] - \delta v(x_{t+1})\right)\right] . \end{aligned}$$

The first equality is the definition of v_α. The second equality follows from (2). The inequality follows since $\alpha \in \mathscr{A}_{\mathbb{F}}$ and $\mathscr{A}_{\mathbb{F}} \subseteq \mathscr{A}_{\mathbb{G}}$. □

Thus any information relaxation with any penalty constructed this way provides a lower bound on the expected discounted cost generated by any primal feasible policy, and in particular, the optimal policy.

With $\mathbb{G}$ as the perfect information relaxation, for example, the set of relaxed policies $\mathscr{A}_{\mathbb{G}}$ is the set of all policies $\mathscr{A}$, where feasible actions are selected with full knowledge of the (infinite) sample path $\boldsymbol{w} := \{w_t\}_{t\geq 1}$. In this case, the weak duality lemma implies that for any $v \in \mathscr{V}$,

$$\begin{aligned} v^\star(x_0) &\geq \inf_{\alpha_G \in \mathscr{A}} \mathbb{E}\left[\textstyle\sum_{t=0}^{\infty} \delta^t \left(c(x_t, \alpha_{G,t}) + \delta \mathbb{E}[v(f(x_t, \alpha_{G,t}, w))] - \delta v(x_{t+1})\right)\right] \\ &= \mathbb{E}\Big[\min_{\boldsymbol{a} \in \mathbf{A}(\boldsymbol{w})} \textstyle\sum_{t=0}^{\infty} \delta^t \left(c(x_t, a_t) + \delta \mathbb{E}[v(f(x_t, a_t, w))] - \delta v(x_{t+1})\right)\Big], \end{aligned} \tag{3}$$

where $\boldsymbol{a} := \{a_t\}_{t\geq 0}$ and $\mathbf{A}(\boldsymbol{w})$ denotes the set of feasible actions given the sample path $\boldsymbol{w}$. If we take $v = 0$, this lower bound is the expected value of the optimal policy with perfect information.

Note that the lower bound (3) is in a form that appears convenient for Monte Carlo simulation: we can estimate the expected value on the right side of (3) by randomly generating sample paths $\boldsymbol{w}$ and for each such $\boldsymbol{w}$ solving the deterministic *inner problem* of choosing a feasible action sequence $\boldsymbol{a} \in \mathbf{A}(\boldsymbol{w})$ to minimize the penalized costs:

$$\min_{\boldsymbol{a} \in \mathbf{A}(\boldsymbol{w})} \textstyle\sum_{t=0}^{\infty} \delta^t \left(c(x_t, a_t) + \delta \mathbb{E}[v(f(x_t, a_t, w))] - \delta v(x_{t+1})\right) . \tag{4}$$

Here, unlike the primal DP, we need only consider actions for a particular sample path $\boldsymbol{w}$ and we need not consider the nonanticipativity constraints that link actions across sample paths in the primal DP. The inner problem (4) is an infinite horizon dynamic program with deterministic state transitions given by $x_{t+1} = f(x_t, a_t, w_{t+1})$ for every feasible action $a_t \in A(x_t)$. A lower bound on the optimal value of the primal DP would then be obtained by averaging the optimal solution to (4) over simulated sample paths $\boldsymbol{w}$.

The intuition for the "generating function" v stems from approximating the optimal value function. If v is a good approximation to $v^\star$, then the penalty terms in (4) should nearly compensate for the additional information in $\mathbb{G}$ by deducting the continuation value in the next-period state and adding back in the expected continuation value under the natural filtration, $\mathbb{F}$. When $v = v^\star$, we obtain an “ideal penalty” that recovers a tight bound. This following result, analogous to Thm. 2.1 in BSS, formalizes this.

Proposition 2.1 (Strong Duality). *For any relaxation $\mathbb{G}$ of $\mathbb{F}$,*

$$\sup_{v \in \mathcal{V}} \inf_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \delta^t \big(c(x_t, \alpha_{G,t}) + \delta \mathbb{E}[v(f(x_t, \alpha_{G,t}, w))] - \delta v(x_{t+1})\big)\big] \quad = \quad v^\star(x_0) \ .$$

Of course, we are working under the assumption that $v^\star$ is difficult to calculate, so in such cases we would not be able to use this ideal penalty. We may, however, have an approximation of $v^\star$ given by an approximate value function v. We could then use v to generate the penalty and, by Lemma 2.1, we obtain a lower bound on $v^\star(x_0)$ using any information relaxation. Proposition 2.1 states that we can in principle get a tight lower bound with this approach, and if v is a good approximation to $v^\star$, intuitively we would expect this lower bound to be close to $v^\star(x_0)$.

2.3. Duality Results with Absorption Time Formulation

The perfect information relaxation described above involves simulating the infinitely long sequences $\{w_t\}_{t \geq 1}$, and in practice we cannot do this. A simple way around this issue would be to apply information relaxations to a finite horizon approximation of the primal DP, where the horizon is chosen to be sufficiently long. To ensure this approach provides a lower bound to the infinite horizon problem, we would then need to correct for any omitted tail costs (e.g., by adding $\delta^T \underline{c}/(1-\delta)$ to the total costs, where $\underline{c}$ is a lower bound on $c(x,a)$ over all feasible state-action pairs, and T is the approximating horizon length). We will discuss finite horizon approximations again in Section 4.4.

Another approach is to apply information relaxations to a well-known, equivalent formulation of the primal DP in which there is no discounting but instead $\mathbb{X}$ includes a costless, absorbing state x^a that is reached with probability $1-\delta$ from every state and for any feasible action. Otherwise, conditional on not absorbing, the distribution of state transitions is as before. In this formulation, we can equivalently express the expected cost of any policy $\alpha \in \mathcal{A}_{\mathbb{F}}$ as

$$v_\alpha(x_0) = \mathbb{E}\big[\textstyle\sum_{t=0}^{\tau} c(x_t, \alpha_t)\big], \tag{5}$$

where $\tau := \inf\{t : x_t = x^a\}$ is the absorption time for reaching x^a and is geometric with parameter $1-\delta$ and supported on $\mathbb{N}^+$. The expectations in (5) are over the same probability space as before, but states evolve according to a different state transition function, which we denote by s, that includes a probability $1-\delta$ of absorption for every feasible state-action pair, but conditional on not absorbing, has the same state transition probabilities as under f. We refer to this problem as the *absorption time formulation* and the equivalence to the discounted formulation above is standard (e.g., Puterman 1994, Proposition 5.3.1).

We now state the main duality results using the absorption time formulation. These results are analogous to the duality results for the discounted formulation above and follow as a special case of a more general result that we will later show (Thm. 4.1). In everything that follows, we assume all $v \in \mathcal{V}$ satisfy $v(x^a) = 0$.

Proposition 2.2 (Duality Results, Absorption Time Formulation). *For any relaxation $\mathbb{G}$ of $\mathbb{F}$, the following hold:*

(i) Weak duality. For any $\alpha \in \mathcal{A}_{\mathrm{S}}$ and any $v \in \mathcal{V}$,

$$v_\alpha(x_0) \;\geq\; \inf_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}\left[\, \textstyle\sum_{t=0}^{\tau} c(x_t, \alpha_{G,t}) + \mathbb{E}[v(s(x_t, \alpha_{G,t}, w))] - v(x_{t+1}) \right]. \tag{6}$$

(ii) Strong duality.

$$\sup_{v \in \mathcal{V}} \inf_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}\left[\, \textstyle\sum_{t=0}^{\tau} c(x_t, \alpha_{G,t}) + \mathbb{E}[v(s(x_t, \alpha_{G,t}, w))] - v(x_{t+1}) \right] \;=\; v^\star(x_0)\ .$$

In addition, with $v = v^\star$ we attain the supremum almost surely.

Using the absorption time formulation with $\mathbb{G}$ as the perfect information relaxation, the absorption time τ is revealed along with the sample path $\boldsymbol{w} = \{w_1, \ldots, w_\tau\}$, and the inner problem then is

$$\min_{\boldsymbol{a} \in \mathbf{A}(\boldsymbol{w})} \textstyle\sum_{t=0}^{\tau} \left(c(x_t, a_t) + \mathbb{E}[v(s(x_t, a_t, w))] - v(x_{t+1}) \right)\ . \tag{7}$$

In (7), we need to solve a deterministic but finite horizon dynamic program, with τ time periods and no discounting. The weak duality result in Prop. 2.2(i) then tells us that by simulating sample paths and averaging the optimal costs in (7), we obtain a lower bound on the optimal expected cost, $v^\star(x_0)$. In this case, the inner problem (7) for a given sample path $\{w_1, \ldots, w_\tau\}$ can be solved by the recursion

$$v_t^{\mathbb{G}}(x_t) \;=\; \min_{a \in A(x_t)} \Big\{ c(x_t, a) + \mathbb{E}\left[v(s(x_t, a, w))\right] - v(s(x_t, a, w_{t+1})) + v_{t+1}^{\mathbb{G}}(s(x_t, a, w_{t+1})) \Big\}, \tag{8}$$

for $t = 0, \ldots, \tau - 1$, with the boundary condition $v_\tau^{\mathbb{G}} = 0$, which follows from the fact that $x_\tau = x^a$, $c(x^a, a) = 0$, and $v(x^a) = 0$. In expressing (7) as this recursion, we are using the fact that $x_{t+1} = s(x_t, a_t, w_{t+1})$; in particular note that $v_t^{\mathbb{G}}$ depends on the sample path, and we have suppressed this dependence to simplify notation. The optimal value $v_0^{\mathbb{G}}(x_0)$ equals the optimal value of the inner problem (7). Thus by simulating sample paths $\{w_1, \ldots, w_\tau\}$, solving (8), and averaging, we obtain a lower bound on the optimal value: $\mathbb{E}\left[v_0^{\mathbb{G}}(x_0)\right] \leq v^\star(x_0)$.

The recursion (8) also provides intuition for the strong duality result (Prop. 2.2(ii)). When $v = v^\star$, we can argue by induction that $v_t^{\mathbb{G}}(x_t) = v^\star(x_t)$ in each time period and for possible every state. To see this, note that $v_\tau^{\mathbb{G}} = v^\star(x^a) = 0$. The induction hypothesis is that $v_{t+1}^{\mathbb{G}}(x_{t+1}) = v^\star(x_{t+1})$ for all possible next-period states at time t. Using this and $v = v^\star$, we see that the final two terms in (8) cancel, and we have

$$v_t^{\mathbb{G}}(x_t) \;=\; \min_{a \in A(x_t)} \left\{ c(x_t, a) + \mathbb{E}\left[v^\star(s(x_t, a, w))\right] \right\} \;=\; \min_{a \in A(x_t)} \left\{ c(x_t, a) + \delta \mathbb{E}\left[v^\star(f(x_t, a, w))\right] \right\} \;=\; v^\star(x_t),$$

where the second equality follows by the fact that $v^\star(x^a) = 0$ and the fact that s and f have the same transition probabilities, conditional on absorption not occurring; the second equality follows from (1). This verifies the induction hypothesis. Note that in this case, we get $v_0^{\mathbb{G}}(x_0) = v^\star(x_0)$ for every sample path $\{w_1, \ldots, w_\tau\}$: we recover a tight bound that holds almost surely and could be estimated with zero variance.

To solve (8), the number of states that we need to consider will depend on the problem structure and, in particular, in the manner in which actions influence state transitions. In problems for which the primal DP has many states due to exogenous uncertainties that are not affected by actions, but a manageable number of states to consider through actions, then (8) will be easy to solve. This is the case in the inventory control examples we discuss in Section 3: here the demand process is potentially high dimensional, but in the perfect information relaxation, demands are revealed and we need only track inventory levels as a state variable. In other problems, however, states may be affected through actions in such a way that recursion (8), though deterministic, still requires us to consider many possible states. This is the case in the multiclass queueing application we discuss in Section 5. We will discuss potential remedies for this issue in Section 4.

2.3.1. Suboptimality Gap Estimates

A useful feature of the absorption time formulation is that we can simultaneously estimate the expected costs of primal feasible policies and lower bounds from the information relaxations via simulations using common samples. In particular, when estimating the cost of a stationary policy $\alpha \in \mathscr{A}_{\mathrm{S}}$, we can include the penalty terms in each period, as these have zero mean for any $\alpha \in \mathscr{A}_{\mathbb{F}}$ (see Prop. A.1). For a given sample path, evaluating the cost of α with these terms included is equivalent to (8) with actions fixed as those under α:

$$\hat{v}_t(x_t) \quad = \quad c(x_t, \alpha(x_t)) + \mathbb{E}\left[v(s(x_t, \alpha(x_t), w))\right] - v(s(x_t, \alpha(x_t), w_{t+1})) + \hat{v}_{t+1}(s(x_t, \alpha(x_t), w_{t+1}))\Big\} \;, \quad (9)$$

for $t = 0, \ldots, \tau-1$, where $\hat{v}_\tau = 0$. Since $\mathbb{E}\left[v(s(x_t, \alpha(x_t), w))\right] - v(s(x_t, \alpha(x_t), w_{t+1}))$ has zero mean, $\mathbb{E}[\hat{v}_0(x_0)] = v_\alpha(x_0)$; these terms from the penalty do, however, serve as control variates that may reduce the variance associated with estimating the expected cost of the policy. Moreover, since the actions selected by α are feasible in (8), $\hat{v}_0(x_0) - v_0^{\mathbb{G}}(x_0) \geq 0$ holds in every sample path.

Thus we can view the random variable $\hat{v}_0(x_0) - v_0^{\mathbb{G}}(x_0)$ as an estimate of the suboptimality "gap" of the policy α: this gap value is almost surely nonnegative, and, by weak duality, in expectation provides an upper bound on the suboptimality of α, i.e., an upper bound on $v_\alpha(x_0) - v^\star(x_0)$. When v is a good approximation of $v^\star$ and α is a good approximation of a policy that is optimal to the primal DP, the values $\hat{v}(x_0)$ and $v_0^{\mathbb{G}}(x_0)$ will be highly correlated and relatively few samples will lead to precise estimates of the suboptimality gap of α. When $v = v^\star$ and α is an optimal policy, by Prop. 2.2(ii), $\hat{v}_0(x_0) - v_0^{\mathbb{G}}(x_0) = 0$ holds almost surely.

3. Application to Inventory Control

We illustrate the information relaxation approach on an infinite horizon inventory control problem in which the distribution of demands may depend on past realizations of demand. This kind of dependence can arise in a variety of inventory management problems, such as Bayesian demand models or autoregressive moving average (ARMA) demand processes. The resulting DPs may be quite complex and difficult to solve, and many researchers have studied the performance of simple heuristic policies for a variety of demand models. A classic result due to Veinott (1965) (see Thm. 6.1 of that paper) provides sufficient conditions for myopic policies to be optimal in a general class of inventory control problems with demand distributions that may depend on the entire past history of demand realizations. The result in Veinott (1965) states that if the inventory position that is myopically optimal is reachable in every state, then the myopic policy is optimal for the full DP. This result has sparked much interest in the performance of myopic policies in related models. For example, Johnson and Thompson (1975) show myopic policies are optimal for an inventory model with ARMA demand and no backlogging; their result also requires a specific truncation of the demand shocks. Graves (1999) proposes the use of myopic policies in a problem with ARMA demands and lead times, although he does not claim a myopic policy is optimal for that problem. Lu, Song, and Regan (2006) study a finite horizon inventory model with forecasts of all future demands and show that a condition analogous to that in Veinott (1965) is in fact necessary for myopic policies to be optimal. Here we study an inventory model in which the demand distribution may depend on past realizations of demand and required conditions for the myopic policy to be optimal need not hold. We will use information relaxations to assess the suboptimality of myopic policies. As mentioned in Section 1, other researchers have used information relaxations in inventory control, albeit for finite horizon models with different demand processes.

3.1. The Model

We consider an infinite horizon, discrete-time, single-item inventory control problem with discounted costs, zero lead times, and backorders. At each point in time, $y_t \in \mathbb{Z}$ denotes the inventory level at the start of the period. We let $Y = \{-\underline{y}, \dots, 0, \dots, \overline{y}\}$, where $\underline{y} \in \mathbb{Z}_+$, $\overline{y} \in \mathbb{Z}_+$ denote the set of possible inventory levels, and the limits represent capacity limitations. In each period, the decision maker may order a_t units, where the constraint set is $A(y_t) = \{a_t \in \mathbb{Z}_+ : y_t + a_t \leq \overline{y}\}$. The costs in each period are given by $c(y_t, a_t) = c_o a_t + c_h y_t^+ + c_b y_t^-$, where $y^+ = \max\{0, y\}$ and $y^- = \max\{0, -y\}$, and $c_o \geq 0$, $c_h \geq 0$, and $c_b \geq 0$ represent the marginal cost associated with orders, held inventory, and backorders, respectively. Immediately after an order is placed, the order arrives, then a random demand $d_{t+1} \in \mathbb{Z}_+$ is realized, and the next-period inventory level y_{t+1} evolves as $y_{t+1} = g(y_t + a_t - d_{t+1})$, where $g(y) = \max(y, -\underline{y})$. In this setup, we charge holding and backorder costs at the start of each period based on the incoming inventory level; this is equivalent to a model that charges holding and backorder costs at the end of each period but with holding and backorder costs scaled as δc_h and δc_p.

We assume the demand distribution in each period depends on past realizations of demands. Specifically, we assume the distribution of the demand d_{t+1} depends on the previous $k \geq 1$ demand realizations $d_t, d_{t-1}, \dots, d_{t-k+1}$. In our specific examples below, we assume the demand distribution has a known shape (e.g., Poisson or geometric), and the expected demand in each period follows the stochastic process

$$\mathbb{E}[d_{t+1} | d_t, \dots, d_{t-k+1}] \quad = \quad \beta_0 + \beta_1 d_t + \dots + \beta_k d_{t-k+1}, \tag{10}$$

for some known coefficients $\beta_0, \ldots, \beta_k$. The framework we discuss here allows for more general dependence structures between the current demand distribution and past demands.[2]

The DP states in this problem are given by $x_t = (y_t, \mathcal{D}_t)$, where $\mathcal{D}_t := (d_t, \ldots, d_{t-k+1})$, and the state transition function f is given by $f(x_t, a_t, d_{t+1}) = (g(y_t + a_t - d_{t+1}), \mathcal{D}_t^+(d_{t+1}, \mathcal{D}_t))$, where the past demand states evolve as $\mathcal{D}_t^+(d_{t+1}, \mathcal{D}_t) = (d_{t+1}, d_t, \ldots, d_{t-k+2})$. We assume y_0 and $\mathcal{D}_0$ are known; the resulting DP is:

$$v^\star(y_t, \mathcal{D}_t) \quad = \quad \min_{a_t \in A(y_t)} \{c(y_t, a_t) + \delta \mathbb{E}[v^\star(g(y_t + a_t - d_{t+1}), \mathcal{D}_t^+(d_{t+1}, \mathcal{D}_t))|\mathcal{D}_t]\} \ , \tag{11}$$

where $\mathbb{E}[\cdot|\mathcal{D}_t]$ denotes the expectation over next-period demands given the past demand set $\mathcal{D}_t$. Since solving (11) may be difficult in general - the state space is $k+1$ dimensional - we will consider a suboptimal, heuristic policy. Specifically, we consider the *myopic policy*, which chooses order quantities a_t^{m} in each state $(y_t, \mathcal{D}_t)$ according to

$$a_t^{\mathrm{m}} \quad = \quad \arg \min_{a_t \in A(y_t)} \{c_o a_t + \delta \mathbb{E}[v^{\mathrm{m}}(g(y_t + a_t - d))|\mathcal{D}_t]\} \ , \tag{12}$$

where $v^{\mathrm{m}}(y_t) := -c_o y_t + c_h y_t^+ + c_b y_t^-$. With this approximation, the second and third term are the next-period holding and backorder costs, and the first term reflects the fact that in the next period any held inventory substitutes for orders. Intuitively, the myopic policy considers the impact that the current demand state $\mathcal{D}_t$ has on the next-period costs, but ignores the downstream impact that $\mathcal{D}_t$ has on costs in later periods, as well as the future evolution of inventory positions. Solving (12) in a given state is easy and reduces to a critical fractile calculation based on the distribution of d_{t+1} given $\mathcal{D}_t$, and ordering up to this level.

Note that in this model we use the standard assumption that negative order quantities are not permitted. Thus, noting the “reachability” result from Veinott (1965), we would expect the myopic policy to perform well provided demand does not frequently drop substantially from one period to the next, which would lead to incoming inventory positions that are overstocked relative to the myopically optimal position. In our examples, these types of drops in demand are possible, so the myopic policy need not be optimal.

3.2. Perfect Information Relaxation and Penalties

To obtain lower bounds on the optimal cost, we will use the perfect information relaxation with the absorption time formulation. In this relaxation, each sample path corresponds to a realization of the absorption time τ, drawn from a geometric distribution with parameter $1 - \delta$, along with a sample path of demands $d_1, \ldots, d_\tau$, generated by the stochastic process described above. For a given sample path, since demands are fixed and known, we need only keep track of the inventory level in solving the deterministic inner problems.

[2] Note that although demands are dependent over time, we could equivalently write this in terms of IID uncertainties as in the general setup, e.g., with IID uniform $[0,1]$ random variables each being mapped to demand realizations using the generalized inverse distribution function corresponding to the current demand state $\mathcal{D}_t$.

We consider two penalties. First, the case with zero penalty, i.e., $v = 0$. In this case, (8) takes the form

$$v_t^{\mathbb{G}}(y_t) \quad = \quad \min_{a_t \in A(y_t)} \left\{ c(y_t, a_t) + v_{t+1}^{\mathbb{G}}(g(y_t + a_t - d_{t+1})) \right\} \ , \tag{13}$$

for $t = 0, \ldots, \tau - 1$, with $v_\tau^{\mathbb{G}} = 0$. We would expect that perfect information on demands would be quite valuable in this problem, so the lower bound with zero penalty primarily serves as a benchmark.

We also consider penalties generated by the myopic value function, i.e., we use $v = v^{\mathrm{m}}$. With this penalty, the inner problem can be solved through the recursion

$$\bar{v}_t^{\mathbb{G}}(y_t) = \min_{a_t \in A(y_t)} \left\{ c(y_t, a_t) + \delta \mathbb{E}[v^{\mathrm{m}}(g(y_t + a_t - d)) | \mathscr{D}_t] - v^{\mathrm{m}}(g(y_t + a_t - d_{t+1})) + \bar{v}_{t+1}^{\mathbb{G}}(g(y_t + a_t - d_{t+1})) \right\} , \tag{14}$$

for $t = 0, \ldots, \tau - 1$, now with $\bar{v}_\tau^{\mathbb{G}} = v^{\mathrm{m}}$ as the boundary condition.[3] Note that all demands $d_1, \ldots, d_\tau$ are known from $t = 0$ onward in (14). Thus, though expectations conditional on $\mathscr{D}_t$ are evaluated in each time period in (14), there is no need to carry $\mathscr{D}_t$ as an explicit state variable. Averaging over demand sequences, the value $\mathbb{E}[\bar{v}_0^{\mathbb{G}}(y_0)]$ provides a lower bound on $v^\star(y_0, \mathscr{D}_0)$.

In comparing (13) to (14), the role of this penalty becomes apparent: with demands fully known and zero penalty, the optimal orders in (13) will tend to match the next-period demands d_{t+1} so as to keep the next-period holding and backorder costs small. In (14), however, matching demands may be far from optimal. For example, if d_{t+1} is large (relative to its expectation conditional on $\mathscr{D}_t$) in a given period, the penalty $\delta \mathbb{E}[v^{\mathrm{m}}(g(y_t + a_t - d)) | \mathscr{D}_t] - v^{\mathrm{m}}(g(y_t + a_t - d_{t+1}))$ will be minimized at some value less than d_{t+1}, and vice versa if d_{t+1} is small. Overall the penalty leads to less extreme order quantities in (14) compared to the order quantities with zero penalty in (13).

With or without the penalty, the perfect information relaxation in this problem is much easier than the primal DP due to the fact that most of the complexity in the primal DP stems from the high-dimensional state space induced by the exogenous demand process. When demands are known, the state space collapses to the one-dimensional space of inventory levels.

3.3. Examples

We consider two examples, each with three discount factors of $\delta = 0.9$, $\delta = 0.95$, and $\delta = 0.99$, and cost parameters $c_o = 1$, $c_h = 0.2$, and $c_b = 1$. One example has Poisson demands and the other has geometric demands, both with means given by (10). The geometric distribution has much heavier tails than the Poisson distribution and leads to substantially more variation in demands. We take $k = 4$ in these examples, so the previous 4 realizations of demands determines the mean demand in every period. The specific numbers we use are $\mathbb{E}[d_{t+1} | \mathscr{D}_t] = 2 + 0.36 d_t + 0.27 d_{t-1} + 0.18 d_{t-2} + 0.09 d_{t-3}$; these numbers are chosen so that the sum of the coefficients multiplying previous demands equals 0.9, which implies the long-run mean of the expected demand process is $2/(1 - 0.9) = 20$. The initial state in all examples is $y_0 = 0$, and $\mathscr{D}_0 = (20, 20, 20, 20)$. We

[3] This formulation is equivalent to taking $\bar{v}_\tau^{\mathbb{G}} = 0$ and excluding the $-v^{\mathrm{m}}$ term at $\tau - 1$, since absorption occurs at time τ.

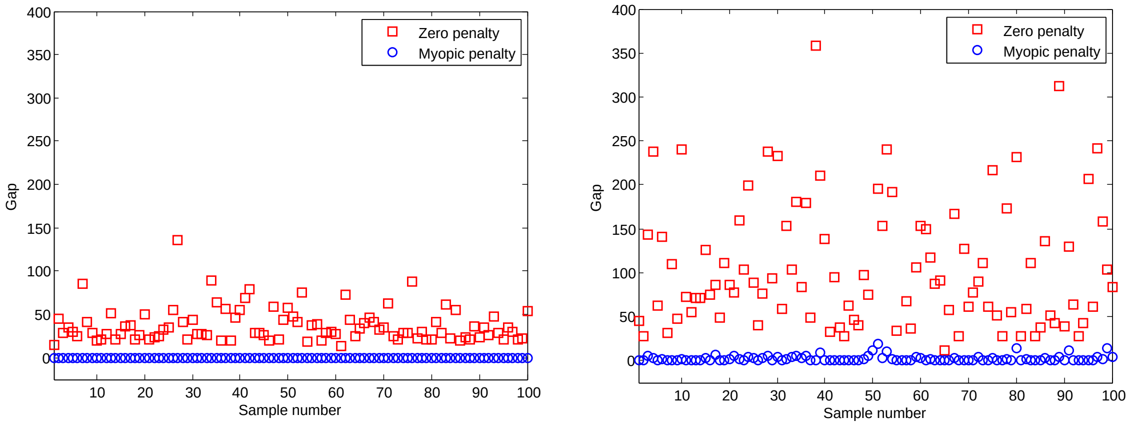
		Poisson Demands				Geometric Demands			
δ		Mean	MSE	Gap %	Time (s.)	Mean	MSE	Gap %	Time (s.)
0.9	Myopic policy cost	218.28	2.10		20.06	269.20	11.57		12.23
	Zero penalty gap	35.82	0.26	16.41	23.37	98.55	2.46	36.61	23.09
	Myopic penalty gap	0.00	0.00	0.00	60.39	2.45	0.25	0.91	59.31
0.95	Myopic policy cost	428.80	4.28		37.85	538.19	19.78		23.23
	Zero penalty gap	49.95	0.39	11.65	46.80	181.01	5.08	33.63	46.66
	Myopic penalty gap	0.00	0.00	0.00	120.01	8.95	0.94	1.66	120.46
0.99	Myopic policy cost	2150.90	31.19		186.67	2524.40	76.96		115.75
	Zero penalty gap	158.37	2.09	7.36	233.80	801.00	28.14	31.73	228.73
	Myopic penalty gap	0.00	0.00	0.00	605.49	53.85	2.75	2.13	591.82

Table 1: Results for inventory examples. Zero (myopic) penalty gap is the difference between the myopic policy cost and the information relaxation with zero (myopic) penalty. Gap % is percent relative to the myopic policy and MSE denotes mean standard error.

take $\underline{y} = \overline{y} = 250$, for 501 total inventory levels possible; the limits were rarely hit by the myopic policy in our examples. A full solution of the DP in these examples would be challenging to compute: even if we were to only consider demands up to 100 in each period, which would be somewhat crude in the geometric examples, a full solution of the DP would need to consider $\sim 10^8$ demand states alone.

For both distributions and for each of the three discount factors, we ran a simulation of 1000 samples, where each sample consists of a draw of the absorption time τ, geometric with parameter $1 - \delta$ as well as τ demand samples drawn from the corresponding demand process. For each sample, we evaluated (i) the cost associated with the myopic policy; (ii) the perfect information relaxation (zero penalty) cost, by solving (13); and (iii) the perfect information relaxation (myopic penalty) cost, by solving (14). We obtain an upper bound on $v^\star(y_0, \mathscr{D}_0)$ with the myopic policy, and lower bounds on $v^\star(y_0, \mathscr{D}_0)$ with the information relaxations. In evaluating the cost of the myopic policy, we add myopic penalty terms as control variates to the costs. Following the discussion in Section 2.3.1, by taking the difference between the myopic cost (including the penalty) and $\bar{v}_0^{\mathbb{G}}$, we get an estimate of the suboptimality gap of the myopic policy that is nonnegative in every sample path and, in expectation, is an upper bound on the suboptimality of the myopic policy.

The results are shown in Table 1. Regarding the gap estimates just discussed, the values of the myopic policy cost and the information relaxations are highly correlated (they are affected by common demand samples) and thus these gap differences have very low mean standard errors (MSEs), especially with the myopic penalty. In terms of bound quality, the perfect information relaxation with zero penalty provides loose bounds, relative to the relaxation with the myopic penalty (relative gaps around 7%-16% for Poisson demands, and around 30%-35% for geometric demands): perfect information on demands is quite valuable. With the myopic penalty, however, we obtain gaps of 0.00% in all Poisson examples - Table 1 reports all values



(a) Poisson demands (b) Geometric demands

Figure 1: Samples values of gaps for inventory control examples with $\delta = 0.9$; "Gap" is the difference between the myopic cost and the optimal value of the information relaxation, i.e., $v_0^{\mathbb{G}}$ in (13) or $\bar{v}_0^{\mathbb{G}}$ in (14).

to two decimal places - and gaps of 0.91%, 1.66%, and 2.13% in the geometric examples. The information relaxation bounds with the myopic penalty take somewhat longer to calculate than those with zero penalty due to the additional effort in calculating the conditional expectation terms in (14). Nonetheless, these lower bounds are still easy to calculate (about 1 minute, 2 minutes, and 10 minutes total for each discount factor, respectively), given the fact that we used far more samples than necessary to get good estimates of the gaps, and given the complexity of solving the full DP in these examples. Thus, with relatively modest computational effort, the information relaxation bounds show that on these examples, the myopic policy is, for all practical purposes, optimal with Poisson demands, and quite good with geometric demands.

Figure 1 shows the sample gap values for the myopic policy on a subset of 100 samples for the $\delta = 0.9$ case, using both zero penalty and the myopic penalty. As discussed, the gap estimates are nonnegative on every sample path, as they must be. The gap estimates with geometric demands are more variable than those with Poisson demands, as we might expect. These results underscore just how well the myopic penalty works on these examples, both in terms of providing tight gap estimates and in terms of low sample error.

4. Information Relaxations and Duality with Reformulations of the Primal DP

The primary motivation for the information relaxation approach is to obtain relaxed problems that are easier to solve than the primal DP. For example with perfect information, we need to repeatedly solve the deterministic inner problems (8), which in some problems - such as the inventory control examples just discussed - may be much easier to solve than the primal DP. In other problems, however, even with perfect information, the inner problems may be difficult to solve. For example in the multiclass queueing application we study in Section 5, even if arrivals and service outcomes are known in advance, the decisions about which customer class to serve at each point in time can still lead to many possible states that need to be considered over a long horizon. In this section, we discuss one approach to addressing this issue.

4.1. Changing the State Transition Function

We consider using different state transition functions $\tilde{s}_t(x, a, w)$ for the evolution of states; these may depend on time. We use the notation $\tilde{x}_t$ to indicate states generated by $\tilde{s}_t$ (and assume $\tilde{x}_0 = x_0$). As in the absorption time formulation, we assume there is an absorbing and costless state x^a, and we let τ denote the (generally random) absorption time to x^a when state transitions follow $\tilde{s}_t$. As with the standard absorption time formulation in Section 2.3, we assume that under $\tilde{s}_t$ that τ is (i) almost surely finite and (ii) does not depend on actions (but may depend on time). $\{\tilde{s}_t\}$ denotes the sequence of these new state transition functions.

To relate the costs associated with the problem where states transition according to $\tilde{s}_t$ to the costs of the primal DP, we need to correct for the change in state transition probabilities. Analgous to $p(x_{t+1}|x_t, a_t)$, we let $q_t(x_{t+1}|x_t, a_t) = \mathbb{P}(\{w_{t+1} \in \mathbb{W} : \tilde{s}_t(x_t, a_t, w_{t+1}) = x_{t+1}\}|\mathscr{F}_t)$ denote the probability of a time t state transition to $x_{t+1} \in \mathbb{X}$ under $\tilde{s}_t$ when the current state is x_t and action $a_t \in A(x_t)$ is selected. If, for all t, states $x \in \mathbb{X}$ and feasible actions $a \in A(x)$, $q_t(y|x, a) = 0$ implies $p(y|x, a) = 0$ for all $y \in \mathbb{X}$, we say $\{\tilde{s}_t\}$ *covers* s; this is our succinct way of stating that $p(\cdot|x, a)$ is absolutely continuous with respect to $q_t(\cdot|x, a)$ for all times and feasible state-action pairs.

If $\{\tilde{s}_t\}$ covers s, then we can equivalently estimate the expected cost with a policy $\alpha \in \mathscr{A}_{\mathrm{S}}$ with state transitions following $\tilde{s}_t$, provided we correct for the change in state transition probabilities. In particular, the probability of the state trajectory $\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_t$ is $\prod_{\tau=0}^{t-1} q_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))$ under $\tilde{s}_t$ and is $\prod_{\tau=0}^{t-1} p(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))$ under s. Thus, we can obtain an equivalent estimate of the cost with α under $\tilde{s}_t$ by multiplying costs in each period by the factors $\Phi_t(\alpha) := \prod_{i=0}^{t-1} \varphi_i(\tilde{x}_{i+1}|\tilde{x}_i, \alpha(\tilde{x}_i))$, where $\varphi_t(y|x, a) = \frac{p(y|x,a)}{q_t(y|x,a)}$, and $\Phi_0 := 1$. The basic idea of correcting for changes in measure is standard and is used widely in importance sampling in the simulation of Markov processes (e.g., Glasserman 2004, §4.6).

We refer to the DP that uses $\tilde{s}_t$ in place of s and the cost in each period is multiplied by the factor Φ_t as a *reformulation* of the primal DP. From the above discussion, if $\tilde{s}_t$ covers s, the reformulation is equivalent to the primal DP in that the expected cost of a given policy $\alpha \in \mathscr{A}_{\mathrm{S}}$ is the same in the reformulation as it is in the primal DP; this fact is formally established in Prop. A.2, which is used in the proof of Thm. 4.1.

4.2. Duality Results

In general, we may also consider reformulations that are not necessarily equivalent to the primal DP. In order to get lower bounds on $v^\star(x_0)$ using information relaxations in such cases, we need the generating functions $v \in \mathscr{V}$ in the penalties to satisfy an additional property. Specifically, we consider approximate value functions $v \in \mathscr{V}$ that satisfy $v(x) \le c(x, a) + \delta\mathbb{E}[v(f(x, a, w))]$ for all $x \in \mathbb{X}$, $a \in A(x)$. We call such a function $v \in \mathscr{V}$ a *subsolution*.

We now state the duality results, analogous to Lemma 2.1 (weak duality) and Prop. 2.1 (strong duality) but applied to reformulations of the primal DP.

Theorem 4.1 (Duality Results with Reformulations). *For any relaxation $\mathbb{G}$ of $\mathbb{F}$, the following hold:*

(i) Weak duality. If either (a) $\{\tilde{s}_t\}$ covers s or (b) v is a subsolution, then for any $\alpha \in \mathscr{A}_{\mathrm{S}}$ and $v \in \mathscr{V}$,

$$v_\alpha(x_0) \;\geq\; \inf_{\alpha_G \in \mathscr{A}_{\mathbb{G}}} \mathbb{E}\left[\, \textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha_G)\left(c(\tilde{x}_t, \alpha_{G,t}) + \mathbb{E}[v(s(\tilde{x}_t, \alpha_{G,t}, w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha_{G,t}) v(\tilde{x}_{t+1})\right)\right] \;. \quad (15)$$

(ii) Strong duality.

$$\sup_{v \in \mathscr{V}} \inf_{\alpha_G \in \mathscr{A}_{\mathbb{G}}} \mathbb{E}\left[\, \textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha_G)\left(c(\tilde{x}_t, \alpha_{G,t}) + \mathbb{E}[v(s(\tilde{x}_t, \alpha_{G,t}, w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha_{G,t}) v(\tilde{x}_{t+1})\right)\right] \;=\; v^\star(x_0) \;.$$

In addition, with $v = v^\star$ we attain the supremum almost surely.

Thm. 4.1 implies Prop. 2.2: by taking $\tilde{s}_t = s$ for all t, the factors φ_t equal 1 for all state transitions and actions, and we recover the duality results using the absorption time formulation.

To understand Thm. 4.1, it is again most instructive to think of the case with $\mathbb{G}$ as the perfect information relaxation. In this case, we simulate a sample path $\boldsymbol{w} = \{w_1, \ldots, w_\tau\}$ and then solve the inner problem

$$\min_{\boldsymbol{a} \in \mathbf{A}(\boldsymbol{w})} \textstyle\sum_{t=0}^{\tau} \Phi_t(\boldsymbol{a})\left(c(\tilde{x}_t, a_t) + \mathbb{E}[v(s(\tilde{x}_t, a_t, w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, a_t) v(\tilde{x}_{t+1})\right) \;. \quad (16)$$

(Note that Φ_t is almost surely finite, since we are taking state transitions to be those under $\tilde{s}_t$, and by definition there is zero probability of a state transition for which $q_t(\tilde{x}_{t+1}|\tilde{x}_t, a) = 0$ for any action.)

This inner problem in (16) can also be written as a deterministic dynamic program, but with state-and-action-dependent discounting due to the change of measure terms. In particular, since $\Phi_t(\boldsymbol{a})$ is nonnegative and only depends on actions up to period $t-1$, we can equivalently solve (16) using the recursion

$$v_t^{\mathbb{G}}(\tilde{x}_t) \;=\; \min_{a \in A(x_t)} \left\{ c(\tilde{x}_t, a) + \mathbb{E}[v(s(\tilde{x}_t, a, w))] + \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, a)\left(v_{t+1}^{\mathbb{G}}(\tilde{x}_{t+1}) - v(\tilde{x}_{t+1})\right)\right\}, \quad (17)$$

for $t = 0, \ldots, \tau - 1$, and the boundary condition $v_\tau^{\mathbb{G}}(\tilde{x}_\tau) = v_\tau^{\mathbb{G}}(x^a) = 0$, and states now evolve as $\tilde{x}_{t+1} = \tilde{s}_t(\tilde{x}_t, a_t, w_{t+1})$. The value $v_0^{\mathbb{G}}(x_0)$ then equals the optimal value of (16). The weak duality result in Thm. 4.1(i) then implies that by simulating sample paths $\{w_1, \ldots, w_\tau\}$, solving for $v_0^{\mathbb{G}}(x_0)$, and averaging, we obtain a lower bound on $v^\star(x_0)$, i.e., $\mathbb{E}\left[v_0^{\mathbb{G}}(x_0)\right] \leq v^\star(x_0)$, provided one of the conditions in Thm. 4.1(i) hold.

In comparing the recursion (17) to the perfect information recursion (8) using the absorption time formulation, the only difference is the presence of the change of measure factor φ_t. We can interpret this in terms of state-and-action-dependent discounting: if we can select an action that leads to a state transition that is much less likely to occur in the primal DP than to occur in the reformulation with $\tilde{s}_t$, then we will have substantial discounting (i.e., φ_t close to zero) to compensate for this difference in relative likelihoods.

Nonetheless, the strong duality result in Thm. 4.1(ii) shows that we can, in theory, obtain tight lower bounds using this approach. Using an inductive argument similar to that at the end of Section 2.3, we can show that when $v = v^\star$, the recursion (17) leads to $v_t^{\mathbb{G}}(\tilde{x}_t) = v^\star(\tilde{x}_t)$ for all time periods and in every possible state. Thus, using an ideal penalty, we again obtain a tight lower bound that equals $v^\star(x_0)$ almost surely.

This is true for any reformulation.

The potential advantage of this approach is tractability. Recall that in some problems, solving the recursions (8) may still require dealing with many states: though the problem is deterministic, many states may still be possible through the effect actions have on states. We could, however, take $\tilde{s}_t$ to be a function that does not depend on actions. In this case, notice that since $\tilde{x}_{t+1} = \tilde{s}_t(\tilde{x}_t, w_{t+1})$, the trajectory of states $\tilde{x}_t$ in (17) is completely independent of actions. Thus solving (17) only requires a minimization at a single state in each period and τ states in total: namely, the states $x_0, \tilde{x}_1, \ldots, \tilde{x}_{\tau-1}$. We call a formulation where $\tilde{s}_t$ is independent of actions an *uncontrolled formulation.* Rogers (2007) also considers uncontrolled formulations in the case of perfect information relaxations. Thm. 4.1 generalizes these ideas by allowing for arbitrary information relaxations and for $\tilde{s}_t$ to depend on actions.

Finally, an important aspect of the weak duality result in Thm. 4.1(i) are conditions (a) and (b). If we were to evaluate the performance of a given stationary policy $\alpha \in \mathscr{A}_{\mathrm{S}}$ using a reformulation $\tilde{s}_t$, it would be essential that, for any state for which we have positive probability of visiting using the primal formulation with s, we also have positive probability of visiting with $\tilde{s}_t$. This is the essence of condition (a) in Thm. 4.1(i). Thm. 4.1(i) asserts that even if we do not fulfill this condition, we nonetheless obtain lower bounds on $v^\star(x_0)$, provided penalties are generated from a subsolution v.

For example, we may use an uncontrolled formulation where state transitions are given by $\tilde{s}_t(\tilde{x}_t, a_t, w_{t+1}) = s(\tilde{x}_t, \alpha(\tilde{x}_t), w_{t+1})$ for a given heuristic policy $\alpha \in \mathscr{A}_{\mathrm{S}}$ whose suboptimality we wish to assess. Such an $\tilde{s}_t$ may violate condition (a), as there may be zero probability of visiting some states that could be visited by other feasible policies (in particular, the optimal policy). Nonetheless, provided we generate a penalty from a subsolution v, condition (b) ensures that we obtain a lower bound by applying an information relaxation to this uncontrolled formulation. Moreover, this state transition function is a particularly attractive choice from an implementation perspective: we can evaluate the costs of the heuristic policy α using simulation to get an upper bound on $v^\star(x_0)$. Using this uncontrolled formulation and a penalty from a subsolution v, we obtain a lower bound on $v^\star(x_0)$ by solving (17) on exactly the same state trajectories. We will pursue this approach in the multiclass queueing application in Section 5.

4.3. Quality of the Information Relaxation Bounds

A well-known fact (e.g., Puterman 1994, Prop. 6.3.2(a)) is that any subsolution provides a lower bound on the optimal value function $v^\star$ in every state. The information relaxation approach is guaranteed to (weakly) improve the lower bounds provided by any subsolution, as we now show.

Proposition 4.1 (Bound Guarantees). *If $v \in \mathscr{V}$ is a subsolution to the primal DP, then:*

(i) For any relaxation $\mathbb{G}$, the lower bound on the right-hand side of (15) is no smaller than $v(x_0)$.

(ii) If $\mathbb{G}$ is the perfect information relaxation, then the recursion (17) satisfies $v_t^{\mathbb{G}}(x_t) \geq v(x_t)$ almost surely for all possible states. Moreover, v is a subsolution to the finite horizon problem in (17).

Prop. 4.1(i) shows that if we use a penalty generated by v, then the resulting lower bound is at least as good (i.e., as large) as $v(x_0)$ itself. In many large-scale examples we may already consider the use of subsolutions to obtain lower bounds, and by using an information relaxation with a penalty from a subsolution, we can improve on these lower bounds. From Thm. 4.1(i)-(b), this holds for any reformulation.

For example, in the multiclass queueing application we study in Section 5, we obtain subsolutions by considering Lagrangian relaxations that relax the requirement that a server can serve at most one customer at a time. These Lagrangian relaxations are far easier to calculate than solving the primal DP, because they lead to problems that decouple over customer classes. We can then use these Lagrangian relaxation subsolutions as penalties in the information relaxation approach and are guaranteed to obtain lower bounds that are at least as good as the lower bound from the Lagrangian relaxation itself. Many other problems may also have a "weakly coupled" structure in which relaxations of certain coupling constraints lead to subsolutions that are easy to calculate and could be used in this approach. Alternatively, the approximate linear programming (ALP) approach (de Farias and Van Roy 2003) is a widely used approach in approximate DP that involves using linear programming to calculate subsolutions expressed as a linear combination of basis functions. In problems where ALP can be used, Prop. 4.1 tells us that we will improve the lower bounds from ALP using information relaxations.

Prop. 4.1(ii) adds that, in the case when $\mathbb{G}$ is the perfect information relaxation, the recursions discussed in (17) are almost surely larger than $v(x_t)$ in all possible states: in this case not only do we get a better lower bound than $v(x_0)$ on average, but also in every sample path, $v_0^{\mathbb{G}}(x_0)$ can never fall below $v(x_0)$. Finally, Prop. 4.1(ii) adds that in the perfect information case, v is also a subsolution to the finite horizon DP described in (17). This result implies that even if we only solve (17) approximately by focusing on a set of subsolutions including v, we will still obtain lower bounds no worse than $v(x_0)$. We will illustrate the potential usefulness of this result in the multiclass queueing application in Section 5.5.2.

4.3.1. Suboptimality Gap Estimates and Variance Considerations

Similar to the discussion in Section 2.3.1, we can use information relaxations with reformulations to obtain an estimate on the suboptimality of a given primal feasible policy through the use of a single simulation. For example, we can consider an uncontrolled formulation where state transitions follow a heuristic policy $\alpha \in \mathcal{A}_{\mathrm{S}}$, i.e., $\tilde{s}_t(\tilde{x}_t, a_t, w_{t+1}) = s(\tilde{x}_t, \alpha(\tilde{x}_t), w_{t+1})$, and where v is a subsolution. Again, we can add the penalty terms as control variates to the costs when using α, as the terms have zero mean for any $\alpha \in \mathcal{A}_{\mathbb{F}}$ (see Prop. A.1 in Appendix A.1). For a given sample path, we can then equivalently evaluate the cost of using α to be

$$\hat{v}_t(x_t) \quad = \quad c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w_{t+1})) + \hat{v}_{t+1}(s(\tilde{x}_t, \alpha(\tilde{x}_t), w_{t+1})) \ ,$$

for $t = 0, \ldots, \tau - 1$, with $\hat{v}_\tau = 0$. This is equivalent to (17) with actions a_t fixed at $\alpha(\tilde{x}_t)$ in each period: the change of measure terms φ_t all equal 1 at these actions, since the state transitions are those taken by the policy α. We then have $\mathbb{E}[\hat{v}_0(x_0)] = v_\alpha(x_0)$. Again, since the actions selected by α are feasible in (17), the value $\hat{v}_0(x_0) - v_0^{\mathbb{G}}(x_0)$ is almost surely nonnegative and its expectation is an upper bound on $v_\alpha(x_0) - v^\star(x_0)$.

As with any change of measure approach, there are concerns about potentially large increases in variance. With this uncontrolled formulation, however, variance issues are substantially mitigated: noting the above discussion and Prop. 4.1(ii), the inequalities $v(x_0) \leq v_0^{\mathbb{G}}(x_0) \leq \hat{v}_0(x_0)$ hold almost surely. Thus, the lower bound estimate can never fall below $v(x_0)$ and never above the (adjusted) estimate of the costs when using α. When v and α are nearly optimal, $\hat{v}_0(x_0)$ and $v(x_0)$ will be close, and the variance in estimating the perfect information relaxation lower bound will be small. We use this uncontrolled formulation approach in the multiclass queueing examples in Section 5 and in those examples, we find very small sample errors in the lower bound (and suboptimality gap) estimates.

4.4. The Distribution of Absorption Time

Thm. 4.1 does not place any restrictions on the distribution of τ, the time to absorption, that is used in the reformulations other than τ being almost surely finite and independent of actions. A natural candidate for τ is geometric with parameter $1-\delta$, which is what is used in the standard absorption time formulation discussed in Section 2.3, and what we used in the inventory control examples in Section 3.3. Another natural candidate is to choose τ to be deterministic and equal to some fixed horizon length T; formally, this corresponds to reformulated state transition function $\tilde{s}_t$ that transitions to x^a with probability 1 at $t = T$ and never earlier. We refer to this as a *truncated horizon formulation.* Although such an $\{\tilde{s}_t\}$ does not cover s (there is zero probability of a non-absorbing transition at time T), by Thm. 4.1(i), we nonetheless obtain a lower bound on $v^\star(x_0)$ if we apply an information relaxation to this formulation, provided we use a penalty generated from a subsolution (i.e., no cost corrections are required).

With a truncated horizon formulation, the perfect information relaxations inner problems are deterministic DPs with T periods. In the case when the non-absorbing state transition probabilities under $\tilde{s}_t$ match those under s, applying (17), the inner problems take the form

$$v_t^{\mathbb{G}}(\tilde{x}_t) \quad = \quad \min_{a \in A(\tilde{x}_t)} \Big\{ c(\tilde{x}_t, a) + \mathbb{E}\left[v(s(\tilde{x}_t, a, w))\right] - \delta v(\tilde{s}_t(\tilde{x}_t, a, w_{t+1})) + \delta v_{t+1}^{\mathbb{G}}(\tilde{s}_t(\tilde{x}_t, a, w_{t+1})) \Big\}, \qquad (18)$$

for $t = 0, \ldots, T-1$, with $v_T^{\mathbb{G}} = 0$, where we use the fact that $\varphi_t(\tilde{x}_{t+1} | \tilde{x}_t, a) = \delta$ for $t < T$, since there is probability δ of a non-absorbing transition under s and probability 1 of a non-absorbing transition under $\tilde{s}_t$. Recursion (18) is identical to (8) with $\tau = T$, with the exception that the discount factor has reemerged in the continuation value term.

Intuitively, we should expect the distribution of τ to have an impact on the variance of costs and, in particular, the computational effort required to get good enough estimates of the lower bounds. The distribution of τ in general also affects the value of the lower bound itself (i.e., the mean) not just the variability of the lower bound. To illustrate this, we revisit the inventory control example from Section 3.3 with $\delta = 0.9$ and geometrically distributed demands. Recall from Table 1 that using a perfect information relaxation with the myopic penalty (i.e., $v = v^{\mathrm{m}}$), we concluded that the myopic policy is at most about

	Mean	MSE	Mean Gap	Gap MSE	Gap %
Geometric absorption formulation					
Myopic policy cost	272.55	2.15	-	-	-
Information relaxation	269.93	2.13	2.62	0.06	0.96
Truncated horizon formulations					
$T = 10$: Information relaxation	179.96	0.67	92.59	NA	33.97
$T = 20$: Information relaxation	239.15	1.48	33.40	NA	12.26
$T = 40$: Information relaxation	265.23	2.35	7.32	NA	2.69

Table 2: Results for inventory example with $\delta = 0.9$ and geometric demands using different absorption time formulations. Gaps are differences between the myopic policy cost and the (perfect) information relaxations with $v = v^{\mathrm{m}}$. Gap % is relative to the myopic policy.

0.91% suboptimal. The results in Table 1 use the absorption time formulation (geometric with parameter $1 - \delta$) and lead to precise suboptimality gap estimates (mean gap of 2.45 with an MSE of 0.25, relative to an estimated myopic policy cost of 269.20).

In Table 2, we show similar results using truncated horizon formulations with $T = 10$, 20, and 40, along with geometric absorption. Although the 1,000 sample paths used for the results in Table 1 were sufficient for precise gap estimates, to ensure we could make statistically significant comparisons between the lower bounds themselves, the results in Table 2 are based on longer simulations. To make the comparisons evenhanded, we use the same computational budget for each lower bound calculation, based on a total of 250,000 total time periods (geometric uses stratified sampling on τ with 25,000 sample paths; truncated horizon with $T = 10$ uses 25,000 sample paths, $T = 20$ uses 12,500 sample paths, and $T = 40$ uses 6,250 sample paths) and each lower bound simulation takes about the same time (about 20 minutes on the desktop we used). The myopic policy cost is estimated using the geometric absorption sample paths.

The lower bounds using truncated horizons improve with longer horizons, as we would expect. The bound using $T = 10$ is poor, and the bound using $T = 40$ is competitive with the bound using geometric absorption (slightly worse, but within sampling error). The longer horizons also have higher MSEs, which might be expected as well, and the MSE of the $T = 40$ horizon is (slightly) higher than the MSE using geometric absorption. This is perhaps a bit surprising, since geometric absorption includes the additional variation due to the time horizon being random. However, Thm. 4.1(ii) tells us that the ideal penalty will lead to zero variance lower bounds, and we suspect the myopic penalty reduces much of this additional variance in these examples. Moreover, the fact that we use stratified sampling on τ in the geometric absorption case also helps to reduce this additional variability.

In general, with truncated horizon models, when v is a subsolution, the optimal value $v_t^{\mathbb{G}}(x_t)$ of the perfect information inner problem (18) is almost surely nondecreasing in T in every possible state: this follows from Prop. 4.1(ii), which implies that $\delta v_{t+1}^{\mathbb{G}}(\tilde{s}_t(\tilde{x}_t, a, w_{t+1})) - \delta v(\tilde{s}_t(\tilde{x}_t, a, w_{t+1})) \geq 0$ for all $t < T$, so an increase in T can only increase the optimal costs in the perfect information relaxation in every sample path. Of course, very long horizons may lead to more computational effort. In the examples we have considered, geometric absorption has performed similarly to truncated horizon formulations with sufficiently large T, and in general due to Thm. 4.1(ii), we would expect the mean and variance of the lower bounds to be fairly insensitive to these choices when paired with a good penalty.

Finally, As mentioned in Section 2.3.1, geometric absorption is also useful because it allows for a common sample comparison to the costs associated with feasible policies (here, the myopic policy) within a single simulation. This allows us to look at a full distribution of the gap values and, for example, calculate the MSE on the gap estimate. A truncated horizon formulation leads to bias in the estimates of a policy's costs and thus, absent any corrections to the costs, does not lead to meaningful gap estimates in each sample.

5. Application to Dynamic Service Allocation for a Multiclass Queue

We consider the problem of allocating service to different types, or "classes," of customers arriving at a queue. Many researchers have studied different variations of multiclass queueing models. A recurring theme in the extensive literature on such problems is that relatively easy-to-compute policies are optimal or often perform very close to optimal. Cox and Smith (1961) study an average cost model with linear delay costs and show optimality of a simple index policy that prioritizes customers myopically by their immediate expected cost reduction (the "$c\mu$ rule"). Harrison (1975) studies a discounted version of the problem with rewards and shows optimality of a static priority policy. van Mieghem (1995) studies a finite horizon model with convex delay costs and shows that a myopic policy that generalizes the $c\mu$ rule is optimal in the heavy traffic limit.

A number of researchers have studied variations of multiclass scheduling problems for which optimal policies are not characterized in any simple form, but nonetheless find that relatively easy-to-compute index policies perform quite well. Veatch and Wein (1996) consider scheduling of a machine to make different items to minimize discounted inventory costs; they investigate the performance of "Whittle index" (Whittle 1988) policies based on analyzing the problem as a restless bandit problem. Ansell et al. (2003) develop Whittle index policies for multiclass queueing models with convex costs in an infinite horizon setting, and find that these policies are very close to optimal in numerical examples. Niño-Mora (2006) develops Whittle index policies for multiclass queues with finite buffers. The calculation of these Whittle indices has connections to Lagrangian relaxations, and we will use Lagrangian relaxations in approximating the model we study.

5.1. The Model

The model we study closely follows the model in Ansell et al. (2003). The system operates in continuous time, and there are I customer classes, with independent Poisson arrival and service processes; λ_i and ν_i denote the mean arrival and service rates for class i customers. Customers arrive to the queue and await service, and we let $x_i \in \mathbb{Z}_+$ denote the number of class i customers in the system at a given time. Costs are discounted and accrue at rate $\sum_i c_i(x_i)$, where the cost functions c_i are assumed to be nonnegative, increasing, and convex in x_i. We will assume that there is a finite buffer limit B_i on the number of class i customers that are allowed in the system at any point in time.

At each point in time a single server must decide which customer class to serve. Preemption is allowed; this, together with the fact that costs are increasing in queue lengths, implies that we may restrict attention without loss of generality to service policies that never idle when the queue is nonempty. The goal is to minimize the expected value of total discounted costs from some given initial state x_0 (e.g., an empty queue).

Although this problem is naturally formulated in continuous time, it is straightforward to convert the problem to an equivalent discrete-time model through "uniformization" (see e.g., Puterman 1994, §11.5), with each point in time representing a potential arrival or a completed service. In this formulation, we normalize the total event rate so that $\sum_i (\lambda_i + \nu_i) = 1$. We let $x = (x_1, \ldots, x_I)$ denote the number of customers of each class; this is the state of the system. We denote the random next-period state given that we are currently serving customer class i and the current state is x by $f(x, i, w) := (f_1(x_1, i, w), \ldots, f_I(x_I, i, w))$, where w is a uniform $[0, 1]$ random variable. In each period, exactly one of the following events occurs:

(a) An arrival of any one customer class j occurs with probability λ_j. The next-period state satisfies $f_j(x, i, w) = \min(x_j + 1, B_j)$ and is unchanged for all other customer classes.

(b) The service of the class i customer being served (if any) is completed with probability ν_i. The next-period state satisfies $f_i(x, i, w) = x_i - 1$ and is unchanged for all other customer classes.

(c) No arrivals occur, and the service of the class i customer being served (if any) is not completed. This occurs with probability $1 - \nu_i - \sum_j \lambda_j$. In this case, $f(x, i, w) = x$.

We obtain these state transitions and probabilities with w as follows: since $\sum_i (\lambda_i + \nu_i) = 1$, we can partition the unit interval according to the λ_i and ν_i. The uniform $[0, 1]$ value of w for each period then represents a class i arrival if it falls in an interval corresponding to λ_i, and a class i service token if it falls in an interval corresponding to ν_i. Consistent with **(c)**, if w results in a service token for class $j \neq i$ while we are serving class i, then the state is unchanged, i.e. $f(x, i, w) = x$ in this case.

We can then write the optimal value function $v^\star$ of the system as

$$v^\star(x) \quad = \quad \min_{i \in I_+(x)} \left\{ \textstyle\sum_j c_j(x_j) + \delta \mathbb{E}[v^\star(f(x, i, w))] \right\}, \tag{19}$$

where $I_+(x)$ represents the set of customer classes currently present in the queue. If no customers are present in the queue, the server simply idles; by convention we take $I_+(x) = \{0\}$ in this case. The discount factor δ in (19) is a scaling of the continuous time discount factor; again, see, e.g. §11.5 of Puterman (1994).

In all our examples we will work with the absorption time formulation of the problem. As in the general setup in Section 2.3, we let x^a denote the absorbing state and s denote the state transition function including absorption, i.e., s transitions to x^a with probability $1 - \delta$ in each period and state and otherwise (i.e., with probability δ) transitions with the same transition probabilities as f. For all approximate value functions v we consider, we take $v(x^a) = 0$, which implies $\delta \mathbb{E}[v(f(x, i, w))] = \mathbb{E}[v(s(x, i, w))]$ for all states x and service decisions i.

5.2. Approximate Value Functions and Heuristic Policies

The number of states involved in solving (19) scales exponentially in the number of customer classes, which will be challenging when there are more than a few customer classes in the model. In such situations, we may instead consider heuristic policies that allocate service based on approximations to the optimal value function. Specifically, we will consider heuristic policies that are "greedy" with respect to an approximate value function, i.e., policies that, in each period, allocate service to a customer class $\alpha(x)$ satisfying

$$\alpha(x) \quad \in \quad \arg\min_{i \in I_+(x)} \left\{ \sum_j c_j(x_j) + \delta \mathbb{E}[v(f(x,i,w))] \right\}, \tag{20}$$

where v is an approximate value function. The right-hand side of (20) approximates (19), with v in place of $v^\star$. We will also use v to form a penalty in the information relaxations. In our examples, we will only use v's that are subsolutions to (19).

One simple approximation is a myopic approximation of $v^\star$. The approximate value function in this case is given by $v^{\mathrm{m}}(x) := \sum_i c_i(x_i)$ and, when in service, the myopic heuristic serves a customer class i that maximizes the quantity $\nu_i(c_i(x_i) - c_i(x_i - 1))$. This myopic approximation serves as a benchmark; we do not expect it will lead to particularly tight bounds. Since costs are nonnegative, v^{m} is a subsolution.

Another approximation we use is based on Lagrangian relaxations. Specifically, we consider a relaxation of the problem in which customer classes are aggregated into groups; at any point in time, the server can serve at most one customer in any given group, but can simultaneously serve customers across different groups and is charged a Lagrangian penalty $\ell \geq 0$ for serving customers from multiple groups simultaneously. To make this formal, let there be G groups of customer classes, with I_{g} representing the indices of customer classes in group g (formally, $\{I_1, \ldots, I_{\mathrm{G}}\}$ is a partition of $\{1, \ldots, I\}$). Using a Lagrange multiplier $\ell \geq 0$, we can show that this relaxed problem decouples across groups, with

$$v^\ell(x) \quad = \quad \frac{(\mathrm{G}-1)\ell}{1-\delta} + \sum_{\mathrm{g}} v^\ell_{\mathrm{g}}(x_{\mathrm{g}}), \tag{21}$$

where x_{g} represents the state vector (number of customers) for group g, and v^ℓ_{g} is the optimal value function for group g in the Lagrangian relaxation:

$$v^\ell_{\mathrm{g}}(x_{\mathrm{g}}) \quad = \quad \min_{a \in \{0\} \cup I_{\mathrm{g}}} \left\{ \sum_{i \in I_{\mathrm{g}}} c_i(x_i) + \delta \mathbb{E}[v^\ell_{\mathrm{g}}(f_{\mathrm{g}}(x_{\mathrm{g}}, a, w))] - \ell \mathbb{1}_{\{a=0\}} \right\}. \tag{22}$$

In (22), state transitions f_{g} are defined over each group in the analogous way as in the original model. This decoupling essentially follows from Hawkins (2003) or Adelman and Mersereau (2008) and relies on the fact that costs decouple across customer classes as well as the fact that state transitions for each class depend only on the state of that class and whether or not we are serving that class.[4]

[4] Both Hawkins (2003) and Adelman and Mersereau (2008) assume state transition probabilities factor into a product form that does not hold in this problem, due to the fact that at most one event (arrival or service) can happen per period. Nonetheless, the state transition probabilities for each class do not depend on the states or service decisions of other classes, and it can be

Solving for the value functions for a particular group involves dealing with a number of states that grows exponentially in the number of classes in the group, but we may still be able to solve (22) relatively easily if we only consider groups consisting of a small number of classes. For a given grouping, we can optimize over the Lagrange multiplier $\ell \geq 0$ (e.g. using bisection) to maximize $v^{\ell}(x_0)$. Moreover, for any $\ell \geq 0$, it is straightforward to argue that v^{ℓ} is a subsolution to (19); thus $v^{\ell}(x_0) \leq v^{\star}(x_0)$ for any $\ell \geq 0$.

The lower bounds provided by these grouped Lagrangian relaxations will get tighter (larger) as we move to coarser groupings. For instance, any grouping of customer classes into pairs can do no worse than the Lagrangian relaxation with individual groups, as the former imposes all the constraints of the latter, as well as some additional constraints (customers in the same group cannot be served simultaneously). In our examples we group together classes that appear to be most demanding of service, as we will explain shortly.

In our examples, we will consider four approximate value functions: the myopic approximation, as well as Lagrangian relaxations with groups of size one, two, and four. We will use each of these approximations as an approximate value function for a heuristic policy that selects actions as in (20), as well as in forming a penalty in the information relaxations.

5.3. Perfect Information Relaxations and Penalties

In describing how we will use information relaxations to obtain lower bounds, it is helpful to first describe how we generate sample paths. For each sample path, we first simulate an absorption time[5] time, τ, that is geometric with parameter $1-\delta$. Given τ, we then generate $\tau - 1$ IID uniform $[0,1]$ random variables $\{w_1, \ldots, w_{\tau-1}\}$, representing arrivals and service tokens as described in Section 5.1. These are the sample paths that we use in evaluating all lower and upper bounds.

We take $\mathbb{G}$ to be the perfect information relaxation unless stated otherwise (see Sec. 5.5.1 for an imperfect information relaxation). With this relaxation, both τ and $\{w_1, \ldots, w_{\tau-1}\}$ are fully revealed in each sample path before making any service decisions. We will use an approximate value function v to form a penalty and will use either $v = v^{\mathrm{m}}$ or $v = v^{\ell}$ for a given Lagrangian relaxation, as discussed in Section 5.2. In this case, the inner problems analogous to (8) here take the form

$$v_t^{\mathbb{G}}(x_t) \quad = \quad \min_{i \in I_+(x_t)} \Big\{ \textstyle\sum_j c_j(x_{t,j}) + \mathbb{E}[v(s(x_t,i,w))] - v(s(x_t,i,w_{t+1})) + v_{t+1}^{\mathbb{G}}(s(x_t,i,w_{t+1})) \Big\}, \tag{23}$$

for $t = 0, \ldots, \tau - 1$, with $x_\tau = x^a$ and $v_\tau^{\mathbb{G}}(x^a) = 0$. Recursion (23) is a deterministic, finite horizon DP with τ periods and state transitions encoded through $\{w_1, \ldots, w_{\tau-1}\}$. The state in this problem in each period is the vector x_t. Even though this is a deterministic problem, we may need to keep track of many possible values for x_t, especially when τ is large. This occurs because the service decision in each period affects x_t and we may need to consider many possible downstream states x_t due to this interaction.

One way around this difficulty is to consider an uncontrolled formulation, where state transitions are

shown this is sufficient to lead to the decomposition (21).

[5] Following the discussion in Section 4.4, we also tried truncated horizon formulations with a relatively long horizon T, but found this made little difference in these multiclass queueing examples.

not affected by service decisions. In our examples, we do this by taking state transitions to be fixed at the state transitions associated with the heuristic policy α, described in (20). Formally, we consider a new state transition function $\tilde{s}$, where $\tilde{s}(x_t, i, w_{t+1}) = s(x_t, \alpha(x_t), w_{t+1})$ for every state x_t, feasible action i, and outcome w_{t+1}. In this formulation of the problem, if the heuristic policy successfully completes a class j service in a given period in a given time period, then a class j customer departs the system, regardless of which customer i we actually choose to serve. Note that with $\tilde{s}$, we only visit the states visited by the heuristic policy, so $\tilde{s}$ certainly does not cover s; however, since we will only use subsolutions v in forming penalties, by virtue of Thm. 4.1(i)-(b), we nonetheless obtain lower bounds on $v^\star(x_0)$ when we apply an information relaxation to this uncontrolled formulation.

Although states are uncontrollable in this reformulation, service decisions do influence the change of measure factors $\varphi(x_{t+1}|x_t, j)$ (and therefore the incurred costs). These factors have the following form:

(i) Transitions due to absorption: $\varphi(x^a|x_t, j) = 1$ for all x_t, $j \in I^+(x_t)$.

(ii) Transitions due to arrivals: if an arrival of a class i customer occurs and $x_{t,i} < B_i$, then $\varphi(x_{t+1}|x_t, j) = 1$ for all $j \in I^+(x_t)$; this follows since arrivals are independent of service decisions.

(iii) Transitions due to services: if a class i customer is successfully served by α, then $\varphi(x_{t+1}|x_t, j) = \mathbb{1}_{\{j=i\}}$.

(iv) Unchanged state transitions: if $x_{t+1} = x_t$ when $\alpha(x_t) = i$, then $\varphi(x_{t+1}|x_t, j) = \frac{1-\Lambda-\nu_j}{1-\Lambda-\nu_i}$, where $\Lambda = \sum_{\{k \,:\, x_{t,k} < B_k\}} \lambda_k$.

For a given sample path, we let $(\tilde{x}_0, \ldots, \tilde{x}_\tau)$ denote the states following the heuristic policy, with $\tilde{x}_t = (\tilde{x}_{t,1}, \ldots, \tilde{x}_{t,I})$ for $t < \tau$, $\tilde{x}_0 = x_0$, and $\tilde{x}_\tau = x^a$. When evaluating the cost of the heuristic policy in simulation, these are exactly the states we will visit in the sample path. Applying the perfect information relaxation to this uncontrolled formulation, the inner problems (17) here take the form

$$\bar{v}_t^{\mathbb{G}}(\tilde{x}_t) \quad = \quad \min_{i \in I_+(\tilde{x}_t)} \Big\{ \textstyle\sum_j c_j(\tilde{x}_{t,j}) + \mathbb{E}[v(s(\tilde{x}_t, i, w))] + \varphi(\tilde{x}_{t+1}|\tilde{x}_t, i)\left(\bar{v}_{t+1}^{\mathbb{G}}(\tilde{x}_{t+1}) - v(\tilde{x}_{t+1})\right) \Big\}, \tag{24}$$

for $t = 0, \ldots, \tau - 1$, where $\bar{v}_\tau^{\mathbb{G}} = 0$ and $\tilde{x}_{t+1} = \tilde{s}(\tilde{x}_t, i, w_{t+1}) = s(\tilde{x}_t, \alpha(\tilde{x}_t), w_{t+1})$.

Solving the recursion (24) is quite easy: the choice of action i does not influence the evolution of states in (24), as these are already fixed at $(\tilde{x}_0, \ldots, \tilde{x}_\tau)$ according to the heuristic policy. Thus we need only deal with a *single state* in each period in each sample path. Moreover, given that we are selecting actions for the heuristic according to (20) with the approximate value function v, the expectations in the recursions in (24) will have already also been calculated at the states $(\tilde{x}_0, \ldots, \tilde{x}_\tau)$ for all feasible actions in the sample path. Thus, we need only perform τ minimizations - one per period - to solve (24), with all costs (including penalties) having been precalculated.

The change of measure factors can have an adverse impact on the quality of the information relaxation lower bounds. In particular, in a given sample path, if the heuristic successfully completes service of a class i customer at time t, then selecting any other customer class j in (24) at time t leads to $\varphi(\tilde{x}_{t+1}|\tilde{x}_t, j) = 0$, which would wipe away the remaining tail costs in (24) in that sample path. Nonetheless, we know from Proposition 4.1(i) that the optimal value of the inner problem must satisfy $\bar{v}_0^{\mathbb{G}}(x_0) \geq v(x_0)$ since our choice of v will always be a subsolution to (19). For example, if we take $v = v^\ell$, where v^ℓ is a Lagrangian relaxation as discussed in Section 5.2, this approach is guaranteed to improve upon the lower bound $v^\ell(x_0)$ from the Lagrangian relaxation.

5.4. Numerical Examples

We have applied these methods to some examples of this multiclass queueing model with $I = 16$ customer classes and $B_i = 9$ for each customer class; the full DPs (19) for these examples have 10^{16} states and would be very difficult to solve. Following Ansell et al. (2003), we use a quadratic cost function of the form $c_i(x_i) = c_{1i}x_i + c_{2i}x_i^2$, with c_{1i} and c_{2i} generated randomly and uniformly in $[1, 5]$ and $[0.1, 2.0]$, respectively (these are the ranges of the same parameters used in the examples in Ansell et al. 2003). Arrival rates and service rates are also generated randomly and scaled so that $\sum_i(\lambda_i + \nu_i) = 1$ and $\sum_i(\lambda_i/\nu_i) > 1$; this is an overloaded system and the problem is challenging due to the resulting heavy congestion (finite buffers ensure stability). We will use discount factors of 0.9, 0.99, and 0.999, and the initial state is an empty queue. The total cost in each example is scaled by the factor $1 - \delta$: this is a standard scaling that facilitates comparisons across the discount factors in that $1/(1 - \delta)$ is the expected number of time periods. All calculations have been done on a standard desktop PC using Matlab.

For all examples, we first calculate lower bounds and approximate value functions using Lagrangian relaxations (21) with groups of size 1, 2, and 4. The Lagrangian relaxations with groups of size 1 have 16 class-specific value functions, each with 10 states; the group of size 2 and size 4 have 8 and 4 group-specific value functions with 10^2 and 10^4 states each, respectively. Each of these approximations are calculated using value iteration and bisection on the Lagrange multiplier. To determine the groupings for each discount factor, we ran a short simulation of the Lagrangian heuristic using the Lagrangian relaxation with groups of size 1 and tracked how frequently each of the classes were served by the heuristic; we grouped together classes most frequently served in decreasing order. More sophisticated ways of grouping classes are no doubt possible and may perform even better.

For each discount factor, we generate $1,000$ sample paths according to the sampling procedure discussed at the start of Section 5.3. For each approximate value function v, in each sample path, we evaluate:

(i) *Heuristic policy.* We select actions as in (20) and calculate the cost incurred by this policy.

(ii) *Perfect information relaxation of uncontrolled formulation.* State transitions follow the heuristic policy as discussed in Section 5.3, and we solve the inner problem (24) on the states $(\tilde{x}_0, \ldots, \tilde{x}_\tau)$ visited by the heuristic, using a penalty formed by v.

	Approximate Value Function (v), Used in Heuristic Policy and in Penalty															
	Myopic				LR, Groups of Size 1				LR, Groups of Size 2				LR, Groups of Size 4			
	Mean	MSE	Gap %	Time (s.)	Mean	MSE	Gap %	Time (s.)	Mean	MSE	Gap %	Time (s.)	Mean	MSE	Gap %	Time (s.)
δ=0.9																
Cost of Heuristic Policy	14.05	1.10	-	1.6	13.20	0.05	-	1.6	13.23	0.07	-	2.1	13.21	0.05	-	1.6
Gap from Heuristic to v	14.05	1.10	100.0	-	1.30	0.05	9.84	1.5	1.22	0.07	9.21	0.5	1.00	0.05	7.58	18.6
Gap from Heuristic to Information Relaxation	6.12	0.90	43.6	0.5	**0.19**	**0.04**	**1.47**	**0.5**	0.32	0.06	2.44	1.1	0.25	0.04	1.86	0.7
δ=0.99																
Cost of Heuristic Policy	201.73	16.5	-	18.6	204.00	0.68	-	18.3	203.66	0.44	-	29.8	203.39	0.09	-	26.6
Gap from Heuristic to v	201.73	16.5	100.0	-	12.12	0.68	5.97	13.1	10.16	0.44	4.99	6.9	4.32	0.09	2.12	340.2
Gap from Heuristic to Information Relaxation	197.98	16.5	98.1	5.2	8.12	0.67	3.98	5.1	6.44	0.43	3.16	9.9	**1.24**	**0.06**	**0.61**	**6.5**
δ=0.999																
Cost of Heuristic Policy	1058.58	44.0	-	204.1	944.82	1.02	-	196.8	947.14	0.91	-	362.9	943.93	0.56	-	330.1
Gap from Heuristic to v	1058.58	44.0	100.0	-	25.98	1.02	2.75	113.8	23.47	0.91	2.48	60.1	13.36	0.56	1.42	3665.2
Gap from Heuristic to Information Relaxation	1058.10	44.0	99.9	51.8	24.25	1.02	2.57	50.5	21.79	0.91	2.30	98.5	**11.98**	**0.56**	**1.27**	**63.8**

Table 3: Multiclass queue example results. The perfect information relaxations use the uncontrolled formulation and the heuristic policy selects actions using v as an approximate value function in (20). Bold highlights the results for the best gap for each δ. LR denotes Lagrangian relaxation and MSE denotes mean standard error.

The average of the heuristic policy costs provides an upper bound on $v^\star(x_0)$, and, by Thm. 4.1(i), the average of the optimal values $\bar{v}^{\mathbb{G}}(x_0)$ provides a lower bound on $v^\star(x_0)$. Since our main goal is to assess the suboptimality of the heuristic policy, in the results below, we report the difference between the heuristic policy cost and $\bar{v}_0^{\mathbb{G}}(x_0)$. The average value of this gap is an upper bound on the suboptimality of the heuristic policy. Following the discussion in Section 4.3.1, we include the penalty terms as control variates in the heuristic policy costs, and the gaps must be nonnegative for every sample path, as the actions chosen by the heuristic policy are always feasible in (24).

Table 3 shows the results for these experiments, reporting the estimate of the heuristic policy cost, as well as the suboptimality gaps using both the subsolution directly as well as the perfect information relaxation. For each estimated value, we report the mean, mean standard error (MSE), and time in seconds required to calculate the value. For each of the two gaps in each case, we also report the % suboptimality relative to the estimate of the heuristic policy cost.

The gaps using the myopic approximate value function are, as expected, quite poor. The myopic lower bound itself is not helpful: the initial state is an empty queue, so this lower bound is just zero, and the suboptimality gap using this is 100%. The information relaxation bound improves on this lower bound, as it must, but the gaps are still quite large using the myopic approximation. The heuristic policy using this approximation does appear to perform somewhat well (and we may expect such a policy to be reasonably good), but it is difficult to be certain about this, given the relatively high MSEs. (Given that the myopic value function is a poor approximation of the optimal value function, it serves as a relatively bad control variate in estimating the cost of the heuristic policy).

The Lagrangian relaxations provide much better lower bounds. These lower bounds improve with larger group sizes and fare better as the discount factor increases. The relative gaps using these lower bounds range from about 9.84% ($\delta = 0.9$ with groups of size 1) to 1.42% ($\delta = 0.999$ with groups of size 4). The upper bounds from the heuristic policies do not change much with larger groupings and are estimated quite well in all cases: these Lagrangian relaxation value functions, regardless of group size, appear to provide good approximations for a heuristic policy and also perform well as control variates.

The information relaxation lower bounds improve upon the Lagrangian relaxation bounds, as they must by Proposition 4.1(i), with the amount of improvement decreasing as δ increases. The intuition for this is that, with longer time horizons, there will be more scenarios in which it is optimal in the inner problem (24) to serve a class i with an associated change of measure term $\varphi(\tilde{x}_{t+1}|\tilde{x}_t, i) = 0$; such scenarios may result in relatively modest improvements on the original Lagrangian relaxation. Nonetheless, the relative gaps using the information relaxation lower bounds can be considerably smaller than those from the Lagrangian relaxations: in the $\delta = 0.9$ case, for instance, the relative gaps are a factor of 4 to 6 smaller than the relative gaps from the Lagrangian relaxations.

In terms of computational effort, the approximate value functions are calculated and stored prior to the simulation (there is nothing to calculate for the myopic approximation). For the Lagrangian relaxations, the run times get longer as the discount factor gets larger due to the fact that a larger discount factor necessitates more iterations in the value iteration routine. The groups of size 4 take longest to compute: these relaxations only have 4 groups, but each group corresponds to a subproblem with 10,000 states. All of the Lagrangian relaxations take from about 1.5 seconds (groups of 1, $\delta = 0.9$) to a few minutes to compute, with the exception of groups of size 4 with a discount factor of 0.999 requiring about an hour; optimizing these Lagrangian relaxation calculations was not our focus and these times could probably be improved.

The run times in Table 3 for the other bounds represent the total time for all 1,000 sample paths in the simulations. The time horizons tend to increase with the discount factor, ranging from tens of periods for the $\delta = 0.9$ case to thousands of periods for the $\delta = 0.999$ case. The number of calculations required in evaluating the heuristic policy and the information relaxation scales linearly in the time horizon, and this is evident in the run times. These bounds can be calculated quickly: the information relaxation calculations take around one second total in the $\delta = 0.9$ cases and around a minute total in the $\delta = 0.999$ cases. These times are insensitive to the group sizes. (There are some differences in times across the columns for these calculations, which reflect differences in time required to iterate over the groups in the penalty calculations. These differences are specific to our implementation.) In terms of estimation error, 1,000 sample paths in these examples is far more than necessary to obtain relatively precise bound estimates, at least for the bounds using the Lagrangian relaxations - note the very low MSEs on the cost of the heuristic policy and all gaps in those columns. Thus we view these run times as quite conservative.

In summary, we improve upon the lower bounds from the Lagrangian relaxations, and, with modest computational effort (an extra few seconds to about a minute), we obtain precise gap estimates that indicate that the heuristic policy is no worse than 1.47%, 0.61%, and 1.27% suboptimal for each discount factor, respectively. This is assuring, given that a full solution of the DP (19) is impractical in these examples.

5.5. Variations

Although the lower bounds reported above are likely good enough for most practical purposes for these examples, we also considered some other approaches to obtaining lower bounds using information relaxations in these examples. These variations are instructive and may also be useful in other problems.

5.5.1. Partially Controlled Formulations

Perfect information bounds using uncontrolled formulations are easy to calculate in that actions do not affect state transitions. This simplicity, however, can come at a price in terms of the quality of the bounds, as there may be many inner problem actions with low values of $\varphi(\tilde{x}_{t+1}|\tilde{x}_t, i)$ in many sample paths. On the other hand, working with the original (i.e., controlled) formulation avoids the change of measure terms entirely and thus may lead to better lower bounds, but in these multiclass queueing examples the inner problems (23) are difficult to solve. We now discuss some *partially controlled formulations* that may be viewed as a compromise between these two extremes.

Specifically, we consider a reformulation in which the service decisions do affect states, but only for a subset $S \subseteq \{1, \ldots, I\}$ of classes. Customers in all other classes $\bar{S} := \{1, \ldots, I\} \backslash S$, on the other hand, cannot be successfully served. Formally, we take $\tilde{s}(x, i, w) = s(x, i, w)$ for all $i \in S$ and, for all $i \in \bar{S}$, we take $\tilde{s}(x, i, w) = s(x, i, w)$ if w corresponds to an arrival and $\tilde{s}(x, i, w) = x$ otherwise. When $S = \{1, \ldots, I\}$, we recover the original formulation of the problem, and when $S = \varnothing$, we recover an uncontrolled formulation in which the server always idles. Just like with the uncontrolled formulations discussed above, this $\tilde{s}$ does not cover s but we again obtain lower bounds on $v^\star(x_0)$ since we use subsolutions to construct penalties. As above, we take τ to be geometric with parameter $1 - \delta$.

With this formulation, customers from a class in $\bar{S}$ may arrive but can never be served, and when we apply a perfect information relaxation, the states of all customers in $\bar{S}$ in each period in each sample path are fixed. In contrast, we do need to account for the different states that are possible for the class S customers due to the fact that service decisions still affect those classes. The inner problems with this formulation are then deterministic DPs with at most $\prod_{i \in S}(B_i + 1)$ states in each period, with class $\bar{S}$ states fixed in each sample path. If $|S|$ is not too large, these deterministic DPs will be manageable.

Given that we will be working with all possible states for classes in S, we can take this one step further and consider an imperfect information relaxation in which all events (arrivals and service tokens) are known perfectly, except those for class S customers. In this imperfect information relaxation, in each period prior to absorption, w_t is known to be one of (i) an arrival for a customer in $\bar{S}$, (ii) a service token for a customer in $\bar{S}$ (which we cannot "use" under $\tilde{s}$), or (iii) a "class S event." If a class S event is known to occur, this event will be an arrival (resp., service token) for a customer $i \in S$ with probability λ_i / Γ_S (resp., μ_i / Γ_S), where $\Gamma_S := \sum_{j \in S}(\lambda_j + \mu_j)$. In this case the inner problems are now stochastic DPs with $\prod_{i \in S}(B_i + 1)$ states in each period, with these conditional probabilities at all periods corresponding to class S events, and deterministic transitions at all other periods (corresponding to class $\bar{S}$ events; recall that we cannot control the state of customers in $\bar{S}$). The factors $\varphi(\tilde{x}_{t+1} | \tilde{x}_t, i)$ equal 1 for all $i \in S$; for $i \in \bar{S}$, we have $\varphi(\tilde{x}_{t+1} | \tilde{x}_t, i) = (1 - \mu_i - \Lambda)/(1 - \Lambda)$ if $\tilde{x}_{t+1} = \tilde{x}_t$ and $\tilde{x}_{t,i} > 0$, and is 1 otherwise, where Λ is as in Section 5.3.

Table 4 shows the results for this approach on the same examples from Section 5.4. These results correspond to the same 1,000 sample paths for each discount factor, and we use the Lagrangian relaxations with groups of size 4 in the penalty. The set of classes S for each of the three discount factors is the first group of size 4 in the grouped Lagrangian relaxation. Thus, the inner problems using this partially controlled formulation have $10,000$ states in each period. Table 4 also lists the relative gaps for these lower bounds and also restates for comparison the relative gaps using information relaxations with the uncontrolled formulation and the Lagrangian relaxation with groups of size 4. These lower bounds must be better than the Lagrangian relaxation and, as we might expect, are also better than the lower bounds using the uncontrolled formulation, with a more marked improvement in the $\delta = 0.9$ and $\delta = 0.99$ cases. The runtimes (99 seconds, about 20 minutes, and about 3.75 hours, respectively) are substantially longer than with uncontrolled formulations, although these are not unreasonable times given the complexity of the full DP and given that 1,000 sample paths is evidently far more than necessary with such low MSEs. Overall, the lower bounds using these

	Information Relaxation with Partially Controlled Formulation				Using Previous Lower Bounds	
	Mean	MSE	Gap %	Time (s.)	Gap % (U)	Gap % (LR)
δ = 0.9	13.16	0.04	0.37	99	1.86	7.58
δ = 0.99	202.62	0.04	0.38	1,197	0.61	2.12
δ = 0.999	932.84	0.02	1.17	13,537	1.27	1.42

Table 4: Imperfect information relaxation results for multiclass queue examples with partially controlled formulation. Gaps are % relative to heuristic policy; Gap % (U) is the gap from the information relaxations with the uncontrolled formulation, and Gap % (LR) is the gap from the lower bounds from the Lagrangian relaxation with groups of size 4.

partially controlled formulations are very good (relative gaps of 0.37%, 0.38%, and 1.17%) and illustrate that using judiciously chosen "middle ground" reformulations with information relaxations may be effective.

5.5.2. Relaxations of the Information Relaxations

The motivation for using the uncontrolled formulation was to reduce the size of the state space that needed to be considered in the information relaxations; recall that we may need to consider many states in (23). Rather than introducing another formulation whose information relaxations we can solve easily, we could instead attempt to solve relaxations of (23) that are easier to solve. By obtaining a lower bound on $v_0^{\mathbb{G}}(x_0)$ in every sample path, we would then still obtain a lower bound on $v^\star(x_0)$.

For example, we could solve Lagrangian relaxations of (23) that relax the constraint that at most one customer class can be served in each time period. This is analogous to the Lagrangian relaxations v^ℓ in (5.2) with groups of size 1, except that these Lagrangian relaxations are for the deterministic, *finite horizon* DP in (23). This approach pairs nicely with penalties constructed from Lagrangian relaxations (21) with groups of size 1 in that with such penalties the Lagrangian relaxations of the inner problem also decouple by customer class. The Lagrange multipliers in each sample path do not depend on the queue state but may depend on time. Thus in each scenario we are optimizing over τ Lagrange multipliers $(\ell_0, \ldots, \ell_{\tau-1})$, where $\ell_t \geq 0$, to obtain the largest lower bound on $v_0^{\mathbb{G}}(x_0)$ in each sample path. This can be viewed as optimizing over a set of subsolutions to (23) in every sample path; from Proposition 4.1(ii), since v^ℓ itself is a subsolution to (23) and can be shown to be feasible here by taking $\ell_t = \ell$ for all t, the lower bound on $v_0^{\mathbb{G}}(x_0)$ we obtain with this approach can never be smaller than $v^\ell(x_0)$. Thus, we may still improve upon the lower bounds from the Lagrangian relaxations $v^\ell(x_0)$ using this additional relaxation. The full details of these inner problem Lagrangian relaxations are discussed in Appendix A.2.

We applied this approach on the examples above and obtained suboptimality gaps of 1.31%, 3.56%, and 2.49% for $\delta = 0.9$, $\delta = 0.99$, and $\delta = 0.999$, respectively, again with very low sample error on the gaps. We solved the Lagrangian relaxations of (23) as linear programs, which took much longer (about 27 seconds, 967 seconds, and 10 hours, respectively, for all 1,000 sample paths) than the uncontrolled formulation calculations. These times probably could be reduced substantially, e.g., by using a subgradient method in the optimization. These lower bounds do improve upon the Lagrangian relaxation bounds with groups of size 1 (relative gaps of 9.84%, 5.97%, and 2.75%, respectively), as they must.

Although this approach of combining relaxations was not better than the information relaxations using the reformulations we considered in these examples, this idea may be useful in other problems where information relaxations with uncontrolled formulations lead to weak bounds or are difficult to calculate, e.g. due to high-dimensional action spaces.

6. Conclusion

In this paper, we have shown how to use information relaxations to calculate performance bounds for infinite horizon MDPs with discounted costs. The general approach allows for reformulations of the state transition functions, which can help to simplify the information relaxation inner problems, both by yielding finite horizon inner problems and by reducing the number of states to consider in the inner problems. When the penalty is generated from a subsolution to the optimal value function, we can relax absolute continuity requirements that would normally be required in obtaining an equivalent formulation, and weak and strong duality still apply. Additionally, the method is guaranteed to improve the lower bounds from the subsolutions themselves. We have applied the method to large-scale examples in inventory control and multiclass queueing with encouraging results. The results in the multiclass queueing examples are, to our knowledge, the first proof of concept that information relaxations applied to an "uncontrolled formulation" (or a "partially controlled formulation") can be effective on a large-scale application.

Moving forward, there are a number of interesting research directions. First, there is considerable freedom in selecting reformulations of the state transition function, and it would be interesting to consider this approach in other applications. It may be possible in some problems to optimize over the reformulation parameters (e.g., reformulated state transition probabilities), perhaps jointly with the penalty, to yield the best lower bound. Developing the theory of information relaxations for more general infinite horizon problems, and showing how to successfully apply the method to such problems, would also be useful. In problems that have an absorbing state, such as stochastic shortest path problems, we could consider information relaxations applied to reformulations like those considered here, where absorption is treated as exogenous. Alternatively, Feinberg and Huang (2015) have recently shown that a broad class of infinite horizon problems, including total cost and average cost problems, can be equivalently expressed as discounted infinite horizon problems. This reduction uses and develops "similarity transformations" studied in Veinott (1969). Provided such transformations could be explicitly carried out on a given (large-scale) application, we could apply information relaxations as discussed here to the equivalent discounted formulations to obtain lower bounds on the optimal value in these problems.

Acknowledgements

We thank an anonymous associate editor, two anonymous referees, Jim Smith, Jeannette Song, and Peng Sun for many helpful comments.

References

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

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

Ansell, P.S., K.D. Glazebrook, J. Niño-Mora, M. O'Keeffe. 2003. Whittle's index policy for a multi-class queueing system with convex holding costs. *Math Meth Oper Res* 57 21-39.

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

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

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

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

Chen, N. and P. Glasserman. 2007. Additive and multiplicative duals for American option pricing. *Finance and Stochastics* 11 153-179.

Cox, D.R., and W.L. Smith. 1961. *Queues*. Methuen, London.

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

Desai, V.V., V.F. Farias, and C.C. Moallemi. 2011. Bounds for Markov decision processes. In *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, (F. L. Lewis, D. Liu, eds.), IEEE Press.

Desai, V.V., V.F. Farias, and C.C. Moallemi. 2012. Pathwise optimization for optimal stopping problems. *Management Science* 58(12) 2292-2308.

Devalkar, S., R. Anupindi, and A. Sinha. 2011. Integrated optimization of procurement, processing, and trade of commodities. *Operations Research* 59(6) 1369-1381.

Feinberg, E.A. 2011. Total expected discounted reward MDPs: Existence of optimal policies. *Wiley Encyclopedia of Operations Research and Management Science*, Wiley Online Library.

Feinberg, E.A., and J. Huang. 2015. On the reduction of total-cost and average-cost MDPs to discounted MDPs. Working paper, `http://arxiv.org/abs/1507.00664`.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York.

Graves, S.C. 1999. A single-item inventory model for a nonstationary demand process. *Manufacturing Service & Operations Management* 1(1) 50-61.

Harrison, J.M. 1975. Dynamic scheduling of a multiclass queue: discount optimality. *Operations Research* 23(2) 270-281.

Haugh, M.B., G. Iyengar, and C. Wang. 2014. Tax-aware dynamic asset allocation. Working paper, Columbia University.

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

Haugh, M. B. and C. Wang. 2014a. Dynamic portfolio execution and information relaxations. *SIAM J. Financial Math.* 5 316-359.

Haugh, M.B. and C. Wang. 2014b. Information relaxations and dynamic zero-sum games. Working paper, Columbia University.

Hawkins, J. 2003. *A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications.* Ph.D. thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Johnson, G. and H. Thompson. 1975. Optimality of myopic inventory policies for certain dependent demand processes. *Management Science* 21(11) 1303-1307.

Kim, M.J. and A.E.B. Lim. 2016. Robust multiarmed bandit problems. *Management Science* 62(1) 264-285.

Kogan, L. and I. Mitra. 2013. Accuracy verification for numerical solutions of equilibrium models. Working paper, Massachusetts Insitute of Technology.

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

Lu, X., J.-S. Song, and A. Regan. 2006. Inventory planning with forecast updates: approximate solutions and cost error bounds. *Operations Research* 54(6) 1079-1097.

Niño-Mora, J. 2006. Marginal productivity index policies for a scheduling a multiclass delay-/loss-sensitive queue. *Queueing Syst* 54 281-312.

Puterman, M.L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY.

Rogers, L.C.G. 2002. Monte Carlo valuation of American options. *Mathematical Finance* 12 271-286.

Rogers, L.C.G. 2007. Pathwise stochastic optimal control. *SIAM Journal on Control and Optimization* 46 1116-1132.

Rudin, W. 1987. *Real and Complex Analysis.* McGraw-Hill, Inc., New York, NY.

van Mieghem, J.A. 1995. Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *The Annals of Applied Probability* 5(3) 809-833.

Veatch, M.H., L.M. Wein. 1996. Scheduling a make-to-stock queue: Index policies and hedging points. *Operations Research* 44 634-547.

Veinott, A.F., Jr. 1965. Optimal policy for a multi-product, dynamic, nonstationary inventory system. *Management Science* 12(3) 206-222.

Veinott, A.F., Jr. 1969. Discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Stat.* 40 1635-1660.

Whittle, P. 1988. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability* 25 287-298.

Williams, D. 1991. *Probability with Martingales.* Cambridge University Press, Cambridge, UK.

Ye, F., Zhu, H., and E. Zhou. 2014. Weakly coupled dynamic program: Information and Lagrangian relaxations. Working paper, Georgia Institute of Technology.

A. Proofs and Detailed Derivations

A.1. Proofs

Proof of Proposition 2.1.

We show the result by considering a finite horizon problem with fixed horizon T, invoke the finite horizon results of BSS (2010) to this problem, and then let $T \to \infty$. Towards this end we let $v_t^T(x)$ denote the optimal value function for the corresponding primal DP with finite horizon T, initial state x and current time $t \le T$. In our definition of this problem, we define the time T terminal cost to be $v_T^T(x) = v^\star(x)$, where $v^\star$ is the optimal value function for the infinite horizon problem. This finite horizon problem then satisfies the recursion

$$v_t^T(x) \quad = \quad \min_{a \in A(x)} \left\{ c(x,a) + \delta \mathbb{E}[v_{t+1}^T(f(x,a,w))] \right\}$$

for $t = 0, \dots, T-1$, with the boundary condition $v_T^T = v^\star$. It is straightforward by induction to see that $v_t^T(x) = v^\star(x)$ for all $t \le T$ and any optimal stationary policy α^* to the infinite horizon problem is also optimal to this finite horizon problem.

Now consider any relaxation $\mathbb{G}$, and consider adding to the costs the penalty $\Pi_T := \sum_{t=0}^{T-1} \delta^t \pi_t(x_t, a_t, w_{t+1})$, where

$$\pi_t(x_t, a_t, w_{t+1}) \quad := \quad \delta \mathbb{E}[v^\star(f(x_t, a_t, w))] - \delta v^\star(f(x_t, a_t, w_{t+1})). \tag{25}$$

By (2), this penalty satisfies $\mathbb{E}[\Pi_T] = 0$ when evaluated over any primal feasible policy $\alpha_F \in \mathcal{A}_\mathbb{F}$. Moreover, Π_T is an ideal penalty for the finite horizon problem in that

$$\begin{aligned} v_0^T(x_0) \quad &= \quad \inf_{\alpha_G \in \mathcal{A}_\mathbb{G}} \mathbb{E}\Big[\textstyle\sum_{t=0}^{T-1} \delta^t (c(x_t, \alpha_{G,t}) + \pi_t(x_t, \alpha_{G,t}, w_{t+1})) + \delta^T v^\star(x_T) \Big] \\ &= \quad \mathbb{E}\Big[\textstyle\sum_{t=0}^{T-1} \delta^t (c(x_t, \alpha^*(x_t)) + \pi_t(x_t, \alpha^*(x_t), w_{t+1}) + \delta^T v^\star(x_T) \Big]. \end{aligned} \tag{26}$$

Eq. (26) states that the optimal primal policy α^* is also optimal with this ideal penalty; this follows from Thm. 2.3 of BSS (2010).

We now let $T \to \infty$ in (26). On the left-hand side we obtain $\lim_{T\to\infty} v_0^T(x_0) = \lim_{T\to\infty} v^\star(x_0) = v^\star(x_0)$, since $v_0^T(x_0) = v^\star(x_0)$ for any T in this construction. Since $\delta \in (0,1)$ and since c, $v^\star$, and π_t are uniformly bounded, we can apply the dominated convergence theorem on the right-hand side and obtain

$$v^\star(x_0) \quad = \quad \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \delta^t (c(x_t, \alpha^*(x_t)) + \pi_t(x_t, \alpha^*(x_t), w_{t+1})) \big]. \tag{27}$$

Suppose now that

$$v^\star(x_0) \quad > \quad \inf_{\alpha_G \in \mathcal{A}_\mathbb{G}} \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \delta^t (c(x_t, \alpha_{G,t}) + \pi_t(x_t, \alpha_{G,t}, w_{t+1})) \big]. \tag{28}$$

Since c, $v^\star$, and π_t are uniformly bounded and $\delta \in (0,1)$, (28) implies we can find a $T < \infty$ such that

$$v^\star(x_0) \quad > \quad \inf_{\alpha_G \in \mathcal{A}_\mathbb{G}} \mathbb{E}\Big[\textstyle\sum_{t=0}^{T-1} \delta^t (c(x_t, \alpha_{G,t}) + \pi_t(x_t, \alpha_{G,t}, w_{t+1})) + \delta^T v^\star(x_T) \Big]. \tag{29}$$

But (29) contradicts (26), and so (28) is false. This together with weak duality (Lemma 2.1) implies

$$v^\star(x_0) \quad = \quad \inf_{\alpha_G \in \mathcal{A}_\mathbb{G}} \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \delta^t (c(x_t, \alpha_{G,t}) + \pi_t(x_t, \alpha_{G,t}, w_{t+1})) \big] \tag{30}$$

as desired. Moreover by (27) we see that the infimum in (30) is achieved by the optimal primal policy α^*. □

Before proving Theorem 4.1 we need some additional results.

Proposition A.1. *For any $\alpha \in \mathscr{A}_{\mathbb{F}}$,*

$$v_\alpha(x_0) \quad = \quad \mathbb{E}\big[\textstyle\sum_{t=0}^{\tau} c(x_t, \alpha_t) + \mathbb{E}[v(s(x_t, \alpha_t, w))] - v(x_{t+1})\big], \tag{31}$$

where state transitions are given by $x_{t+1} = s(x_t, a_t, w_{t+1})$, i.e., the absorption time formulation described in Section 2.3.

Proof. First, note that by standard results (e.g., Puterman 1994, Prop. 5.3.1), for any $\alpha \in \mathscr{A}_{\mathbb{F}}$, we can equivalently express the expected cost with α as $v_\alpha(x_0) = \mathbb{E}\big[\sum_{t=0}^{\tau} c(x_t, \alpha_t)\big]$. Now consider adding the terms $\Pi_\tau := \sum_{t=0}^{\tau} \pi_t$, where $\pi_t = \mathbb{E}[v(s(x_t, a_t, w))] - v(x_{t+1})$. For any $\alpha \in \mathscr{A}_{\mathbb{F}}$, the sequence $\{\pi_t\}_{t\geq 0}$ are martingale differences under $\mathbb{F}$. Note that τ is an almost surely finite stopping time under $\mathbb{F}$, and since v is bounded, the martingale differences π_t are bounded. Thus by an application of the optional stopping theorem (e.g., §10.10 of Williams 1991), we have $\mathbb{E}[\Pi_\tau] = \mathbb{E}[\Pi_0] = 0$. □

We will also use the following lemma. Here, given a probability space (Ω, Σ, P), we say P is concentrated on $A \in \Sigma$ if, for every event $E \in \Sigma$ such that $A \cap E = \varnothing$, it holds that $P(E) = 0$.

Lemma A.1. *Consider a measure space (Ω, Σ) with two probability measures P and Q, and assume Q is concentrated on $A_Q \in \Sigma$. Let φ denote the unique Radon-Nikodym derivative of the absolutely continuous component of P with respect to Q. Let Y be a bounded random variable on this space such that $Y(\omega) \geq 0$ for all $\omega \notin A_Q$. Then $\mathbb{E}^P[Y] \geq \mathbb{E}^Q[\varphi Y]$.*

Proof. First, note that by the Lebesgue-Radon-Nikodym Theorem (see, e.g., Thm. 6.10 in Rudin 1987) we can uniquely decompose P as $P = P^a + P^o$, where P^a is absolutely continuous with respect to Q with unique density (or Radon-Nikodym derivative) φ, and P^o is concentrated on a set A_{P^o} such that $A_{P^o} \cap A_Q = \varnothing$.

We then have

$$\begin{aligned}
\mathbb{E}^P[Y] &= \int_{\omega \in \Omega} Y(\omega) dP(\omega) \\
&= \int_{\omega \in \Omega} Y(\omega) dP^o(\omega) + \int_{\omega \in \Omega} Y(\omega) dP^a(\omega) \\
&= \int_{\omega \in A_{P^o}} Y(\omega) dP^o(\omega) + \int_{\omega \in \Omega} Y(\omega) dP^a(\omega) \\
&\geq \int_{\omega \in \Omega} Y(\omega) dP^a(\omega) \\
&= \int_{\omega \in \Omega} \varphi(\omega) Y(\omega) dQ(\omega) \\
&= \mathbb{E}^Q[\varphi Y].
\end{aligned}$$

The second equality follows from the Radon-Nikodym theorem referenced above. The third equality follows from the fact that P^o is concentrated on A_{P^o} and the fact that Y is bounded. The inequality follows from the condition that $Y(\omega) \geq 0$ for all $\omega \notin A_Q$ and the fact that $A_{P^o} \cap A_Q = \varnothing$, and thus $\omega \in A_{P^o}$ implies $\omega \notin A_Q$. The second-to-last equality follows from absolute continuity of P^a with respect to Q and the Radon-Nikodym theorem, noting that Y is bounded and therefore integrable under P^a. □

Remark: When Ω is countable, Lemma A.1 reduces to the following. For a given probability measure Q, let Ω_Q be the set of outcomes for which $Q(\omega) > 0$. Then $P^a(\omega) = Q(\omega)$ if $\omega \in \Omega_Q$ and 0 for all $\omega \in \Omega \smallsetminus \Omega_Q$. Moreover, if $Y(\omega) \geq 0$ for all $\omega \in \Omega \smallsetminus \Omega_Q$, then

$$\mathbb{E}^P[Y] \;=\; \sum_{\omega \in \Omega \smallsetminus \Omega_Q} Y(\omega) P(\omega) + \sum_{\omega \in \Omega_Q} Y(\omega) P(\omega) \;\geq\; = \sum_{\omega \in \Omega_Q} Y(\omega) P(\omega) \;=\; \sum_{\omega \in \Omega_Q} Y(\omega) P^a(\omega) \;=\; \mathbb{E}^Q[\varphi Y].$$

Proposition A.2. *For any $\alpha \in \mathscr{A}_{\mathrm{S}}$:*

(i) If $\{\tilde{s}_t\}$ covers s, then

$$v_\alpha(x_0) \;=\; \mathbb{E}\big[\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha)\left(c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))v(\tilde{x}_{t+1})\right)\big] \;. \quad (32)$$

(ii) If v is a subsolution, then

$$v_\alpha(x_0) \;\geq\; \mathbb{E}\big[\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha)\left(c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))v(\tilde{x}_{t+1})\right)\big] \;. \quad (33)$$

Proof. For part (i) we note that

$$\begin{aligned}
v_\alpha(x_0) &\overset{(a)}{=} \mathbb{E}\left[\textstyle\sum_{t=0}^{\tau} c(x_t, \alpha(x_t)) + \mathbb{E}[v(s(x_t, \alpha(x_t), w))] - v(x_{t+1})\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\left(\textstyle\sum_{t=0}^{\tau} c(x_t, \alpha(x_t)) + \mathbb{E}[v(s(x_t, \alpha(x_t), w))] - v(x_{t+1})\right)\mathbb{1}\{\tau < \infty\}\right] \\
&\overset{(c)}{=} \sum_{\tau'=1}^{\infty} \mathbb{E}\Big[\Big(\textstyle\sum_{t=0}^{\tau'} c(x_t, \alpha(x_t)) + \mathbb{E}[v(s(x_t, \alpha(x_t), w))] - v(x_{t+1})\Big)\mathbb{1}\{\tau = \tau'\}\Big] \\
&\overset{(d)}{=} \sum_{\tau'=1}^{\infty} \mathbb{E}\Big[\Big(\textstyle\sum_{t=0}^{\tau'} \Phi_t(\alpha)(c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))v(\tilde{x}_{t+1})\Big)\mathbb{1}\{\tau = \tau'\}\Big] \\
&\overset{(e)}{=} \mathbb{E}\Big[\Big(\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha)(c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))v(\tilde{x}_{t+1})\Big)\mathbb{1}\{\tau < \infty\}\Big] \\
&\overset{(f)}{=} \mathbb{E}\Big[\Big(\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha)(c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))v(\tilde{x}_{t+1})\Big)\Big] \;.
\end{aligned}$$

Equality (a) follows by Proposition A.1. Equality (b) follows since c and v are uniformly bounded and τ has finite mean. Equality (c) follows from the law of total expectations. Equality (d) follows by application of the Radon-Nikodym theorem to each term in the sum in (d), which is justified since $\{\tilde{s}_t\}$ covers s. Equality (e) follows again by the law of total expectations, and equality (f) follows by our assumption that τ is almost surely finite when state transitions follow $\tilde{s}_t$.

For part (ii), recall that v being a subsolution means $v(x) \leq c(x,a) + \delta\mathbb{E}[v(f(x,a,w))] = c(x,a) + \mathbb{E}[v(s(x,a,w))]$ for all $a \in A(x)$ and $x \in \mathbb{X}$. All of the steps in the proof of part (i) hold, with the exception that (d) becomes an inequality. To see this, conditioned on $\tau = \tau'$, we have

$$\begin{aligned}
&\textstyle\sum_{t=0}^{\tau'} c(x_t, \alpha(x_t)) + \mathbb{E}[v(s(x_t, \alpha(x_t), w))] - v(x_{t+1}) \\
=\;& c(x_0, \alpha(x_0)) + \mathbb{E}[v(s(x_0, \alpha(x_0), w))] + \textstyle\sum_{t=1}^{\tau'} \left(c(x_t, \alpha(x_t)) + \mathbb{E}[v(s(x_t, \alpha(x_t), w))] - v(x_t)\right),
\end{aligned}$$

where we use the fact that $x_t = x^a$ for all $t \geq \tau$ and $v(x^a) = 0$. Note that the terms $c(x_t, \alpha(x_t)) + \mathbb{E}[v(s(x_t, \alpha(x_t), w))] - v(x_t)$ are nonnegative for all states due to the fact that v is a subsolution. Applying Lemma A.1 to each of these terms and noting that $\Phi_0 = 1$ and x_0 is known leads to (d) holding as an inequality. □

Remark: In the proof of Prop. A.2(ii), the subsolution property is stronger than is required to apply Lemma A.1: the result also holds if $c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - v(\tilde{x}_t) \geq 0$ for all states $\tilde{x}_t$ that occur with zero probability under the state transition process induced by $\tilde{s}_t$.

Proof of Theorem 4.1.

(i) With either condition (a) or (b) holding as given in the statement, we have

$$\begin{aligned} v_\alpha(x_0) &\geq \mathbb{E}[\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha)\,(c(\tilde{x}_t, \alpha(\tilde{x}_t)) + \mathbb{E}[v(s(\tilde{x}_t, \alpha(\tilde{x}_t), w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha(\tilde{x}_t))v(\tilde{x}_{t+1}))] \\ &\geq \inf_{\alpha_G \in \mathcal{A}_{\mathbb{G}}} \mathbb{E}\big[\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha_G)\,(c(\tilde{x}_t, \alpha_{G,t}) + \mathbb{E}[v(s(\tilde{x}_t, \alpha_{G,t}, w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha_{G,t})v(\tilde{x}_{t+1}))\big]. \end{aligned}$$

The first inequality follows by Prop. A.2, and the second inequality follows by the fact that $\alpha \in \mathcal{A}_{\mathrm{S}} \subseteq \mathcal{A}_{\mathbb{F}} \subseteq \mathcal{A}_{\mathbb{G}}$.

(ii) To show strong duality, we take $v = v^\star$. Clearly, $v^\star$ is a subsolution to the primal DP. Consider a fixed sample path $\{w_1, \ldots, w_\tau\}$ and any $\alpha_G \in \mathcal{A}_{\mathbb{G}}$. By definition, $\alpha_G \in \mathcal{A}$ (recall from Section 2.1 that $\mathcal{A}$ is the set of all policies, i.e., the set of all functions that map from $(w_1, \ldots, w_\tau)$ to feasible actions). When $\mathbb{G}$ is the perfect information relaxation, $\mathcal{A}_{\mathbb{G}} = \mathcal{A}$, and, by Prop. 4.1(ii), the optimal value of the inner problem with perfect information using $v = v^\star$ is no smaller than $v^\star(x_0)$, so

$$\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha_G)\,(c(\tilde{x}_t, \alpha_{G,t}) + \mathbb{E}[v(s(\tilde{x}_t, \alpha_{G,t}, w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha_{G,t})v(\tilde{x}_{t+1})) \quad \geq \quad v^\star(x_0)\ . \tag{34}$$

On the other hand, we claim that with $\alpha_G = \alpha^\star$, where $\alpha^\star$ is an optimal stationary policy to the primal DP, that

$$\textstyle\sum_{t=0}^{\tau} \Phi_t(\alpha_G)\,(c(\tilde{x}_t, \alpha_{G,t}) + \mathbb{E}[v(s(\tilde{x}_t, \alpha_{G,t}, w))] - \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, \alpha_{G,t})v(\tilde{x}_{t+1})) \quad = \quad v^\star(x_0) \tag{35}$$

holds almost surely. To see this, note that with $\alpha_G = \alpha^\star$ and $v = v^\star$, that $c(\tilde{x}_t, \alpha_{G,t}) + \mathbb{E}[v(s(\tilde{x}_t, \alpha_{G,t}, w))] = v(\tilde{x}_t)$ by (1). Using this fact, and noting that $v(x_\tau) = v(x^{a}) = 0$, all terms on the left hand side of (35) cancel except $v^\star(x_0)$. Since $\alpha^\star \in \mathcal{A}_{\mathrm{S}} \subseteq \mathcal{A}_{\mathbb{G}}$, $\alpha^\star$ is feasible in the $\mathbb{G}$-adapted problem, and since every $\mathbb{G}$-adapted policy also satisfies (34) almost surely, it follows that $\alpha^\star$ is an optimal $\mathbb{G}$-adapted policy with $v = v^\star$. The result then follows by (35). □

Proof of Proposition 4.1.

(i) This result is implied by part (ii) below. Since the optimal value of the perfect information relaxation recursion (17) satisfies $v_0^{\mathbb{G}}(x_0) \geq v(x_0)$ by part (ii), we have $\mathbb{E}[v_0^{\mathbb{G}}(x_0)] \geq v(x_0)$. Since $\mathcal{A}_{\mathbb{G}} \subseteq \mathcal{A}$, the right-hand side of (15) for any information relaxation $\mathbb{G}$ can never be lower than the value with perfect information, and the result follows. □

(ii) We fix a sample path $\{w_1, \ldots, w_\tau\}$ and prove the result by induction. At time τ, by definition, $x_\tau = x^{a}$, which implies that $v(x_\tau) = v(x^{a}) = 0$, so $v_\tau^{\mathbb{G}} = 0 = v(x_\tau)$. Now consider $t < \tau$, and assume that $v_{t+1}^{\mathbb{G}}(x_{t+1}) \geq v(x_{t+1})$ for all possible states x_{t+1} at time $t+1$. We have

$$\begin{aligned} v_t^{\mathbb{G}}(\tilde{x}_t) &= \min_{a \in A(x_t)} \{c(\tilde{x}_t, a) + \mathbb{E}[v(s(\tilde{x}_t, a, w))] + \varphi_t(\tilde{x}_{t+1}|\tilde{x}_t, a)\,(v_{t+1}^{\mathbb{G}}(\tilde{x}_{t+1}) - v(\tilde{x}_{t+1}))\} \\ &\geq \min_{a \in A(x_t)} \{c(\tilde{x}_t, a) + \mathbb{E}[v(s(\tilde{x}_t, a, w))]\} \\ &\geq v(\tilde{x}_t), \end{aligned}$$

where the equality follows from (17), the first inequality follows from the induction assumption and the fact that $\varphi_t \geq 0$, and the second inequality follows from the fact that v is a subsolution to the primal DP. Finally, setting $v_{t+1}^{\mathbb{G}} = v$ and $v_t^{\mathbb{G}} = v$, the same inequalities also show that v is a subsolution to (17).

□

A.2. Lagrangian Relaxations of Inner Problems for the Multiclass Queueing Application

We show how the inner problems with perfect information decouple with Lagrangian relaxations in the multiclass queueing application, as discussed in Section 5.5.2. We consider a fixed sample path $\{w_1, \dots, w_{\tau-1}\}$. In this case the inner problems corresponding to (8) have the form

$$v_t^{\mathbb{G}}(x_t) \quad = \quad \min_{i \in I^+(x_t)} \Big\{ \textstyle\sum_j c_j(x_{t,j}) + \mathbb{E}[v(s(x_t, i, w))] - v(s(x_t, i, w_{t+1})) + v_{t+1}^{\mathbb{G}}(s(x_t, i, w_{t+1})) \Big\}, \qquad (36)$$

for $t = 0, \dots, \tau - 1$ and where $v_\tau^{\mathbb{G}} = 0$. We will consider Lagrangian relaxations of (36) that relax the constraint that in each time period (and state) the server can serve at most one customer class. To this end, it will be convenient to express actions slightly differently than in (36); specifically, we let $a := (a_1, \dots, a_I)$ denote a binary vector indicating which of the I classes we choose to serve, and denote the action set by

$$A(x) \quad := \quad \Big\{ \{0,1\}^I \ : \ \textstyle\sum_i \mathbb{1}_{\{a_i = 0\}} \geq (I-1), \ a \neq 0 \text{ if } x_i > 0 \text{ for some } i = 1, \dots, I \Big\}.$$

We can then write (36) equivalently as

$$v_t^{\mathbb{G}}(x_t) \quad = \quad \min_{a \in A(x)} \big\{ \textstyle\sum_i c_i(x_{t,i}) + \mathbb{E}[v(s(x_t, a, w))] - v(s(x_t, a, w_{t+1})) + v_{t+1}^{\mathbb{G}}(s(x_t, a, w_{t+1})) \big\}, \qquad (37)$$

with the understanding that the transition function s is now defined on the actions a in the analogous way. We will consider a Lagrangian relaxation of (37) that relaxes the constraint set $A(x)$ to $\{0,1\}^I$ (i.e., the server can serve any class) in all states but adds a Lagrangian penalty $\ell_t((I-1) - \sum_i \mathbb{1}_{\{a_i=0\}})$ in each period for some Lagrange multipliers $\ell_t \geq 0$ (which depend on time but not states). Note that this Lagrangian penalty is less than or equal to zero for any $a \in A(x)$, so the optimal value of this relaxation will be no larger than $v_t^{\mathbb{G}}(x_t)$ in all states and times, for any $\boldsymbol{\ell} := (\ell_0, \dots, \ell_{\tau-1})$ with each $\ell_t \geq 0$. We let $v_t^{\boldsymbol{\ell}}$ denote the value function for this Lagrangian relaxation of (37); this relaxation satisfies

$$\begin{aligned} v_t^{\boldsymbol{\ell}}(x_t) \ &= \ (I-1)\ell_t \ + \\ &\min_{a \in \{0,1\}^I} \Big\{ \textstyle\sum_i (c_i(x_{t,i}) - \ell_t \mathbb{1}_{\{a_i=0\}}) + \mathbb{E}[v(s(x_t, a, w))] - v(s(x_t, a, w_{t+1})) + v_{t+1}^{\boldsymbol{\ell}}(s(x_t, a, w_{t+1})) \Big\}, \end{aligned} \qquad (38)$$

for $t = 0, \dots, \tau - 1$, with $v_\tau^{\boldsymbol{\ell}} = 0$. We argue that when v decouples by customer class, i.e., $v(x) = \sum_i v_i(x_i)$ (plus perhaps a constant, which we will omit to simplify notation), then (38) also decouples by customer class, i.e., $v_t^{\boldsymbol{\ell}}(x_t) = \theta_t + \sum_i v_{t,i}^{\boldsymbol{\ell}}(x_{t,i})$, where θ_t is a constant that does not depend on the state.

We argue this by induction. This is clearly true at $t = \tau$, since $v_\tau^{\boldsymbol{\ell}} = 0$. Now assume that the result holds for $v_{t+1}^{\boldsymbol{\ell}}$ for some $t + 1 \leq \tau$. Note that since each component s_i of $s(x, a, w)$ only depends on x through x_i and a through a_i, we have

$$v(s(x_t, a, w_{t+1})) \quad = \quad \textstyle\sum_i v_i(s_i(x_{t,i}, a_i, w_{t+1})),$$

and similarly for the $v_{t+1}^{\boldsymbol{\ell}}$ term in (38). Moreover, we have

$$\mathbb{E}[v(s(x_t, a, w))] \ = \ \mathbb{E}\big[\textstyle\sum_i v_i(s_i(x_{t,i}, a_i, w)) \big] \ = \ \textstyle\sum_i \mathbb{E}_i[v_i(s_i(x_{t,i}, a_i, w))],$$

where $\mathbb{E}_i$ denotes the expectation with respect to the state transition probabilities for class i, emphasizing the fact that these probabilities do not depend on the states or actions associated with other classes. Finally, the cost terms $c_i(x_{t,i}) - \ell_t \mathbb{1}_{\{a_i=0\}}$ in (38) also decouple by class, and there are no constraints on actions across classes. Altogether, this implies that $v_t^{\boldsymbol{\ell}}$ can be decomposed as stated; in particular,

$$\begin{aligned} v_{t,i}^{\boldsymbol{\ell}}(x_{t,i}) \quad &= \\ &\min_{a_i \in \{0,1\}} \big\{ c_i(x_{t,i}) - \ell_t \mathbb{1}_{\{a_i=0\}} + \mathbb{E}_i[v_i(s_i(x_{t,i}, a_i, w))] - v_i(s_i(x_{t,i}, a_i, w_{t+1})) + v_{t+1,i}^{\boldsymbol{\ell}}(s_i(x_{t,i}, a_i, w_{t+1})) \big\}. \end{aligned}$$

Thus the Lagrangian relaxation (38) decouples by customer class in each scenario. Each $v_{t,i}^{\boldsymbol{\ell}}$ has $B_i + 1$ states in each period, so solving the decoupling for each class involves dealing with $\tau \cdot (B_i + 1)$ total states when the scenario has τ periods.

We can then consider the problem $\max_{\boldsymbol{\ell} \geq \mathbf{0}} v_0^{\boldsymbol{\ell}}(x_0)$ in each scenario. There are different ways we could solve this problem; as stated in Section 5.5.2, we solve these inner problem Lagrangian relaxations as linear programs, with decision variables for the Lagrange multipliers $(\ell_0, \ldots, \ell_{\tau-1})$ and variables representing the value functions $v_{t,i}^{\boldsymbol{\ell}}(x_{t,i})$ for all class i states in each period. This results in a linear program with $\tau + \tau \sum_i (B_i + 1)$ variables. Moreover, if ℓ^* is the optimal Lagrange multiplier for v^{ℓ} in (21), then the choice $\ell_t = \ell^*$ in each period leads to $v_{t,i}^{\boldsymbol{\ell}} = v_i^{\ell^*}$ as defined in (21). This follows from Proposition 2.2(ii), noting that v^{ℓ^*} is the optimal value function for the Lagrangian relaxations with $\ell_t = \ell^*$. The Lagrangian relaxation (21) is therefore a feasible choice in every sample path and we can do no worse than $v^{\ell^*}(x_0)$ in every sample path with this approach.

Independently, Ye, Zhu, and Zhou (2014) show a similar decoupling for weakly coupled DPs and Lagrangian relaxations; like Hawkins (2003) and Adelman and Mersereau (2008) they assume a product form for state transition probabilities that does not hold in the multiclass queueing application we study in Section 5. Ye, Zhu, and Zhou (2014) study the use of subgradient methods to solve their decoupled inner problems and provide a "gap analysis" that describes a limit on how much slack can be introduced by using Lagrangian relaxations of the inner problems. In our multiclass queueing example with grouped Lagrangian relaxations, using a subgradient method with groups of size 4 would still require solving DPs with $10,000\tau$ states and the runtimes could still be substantial, particularly given that subgradient methods can be slow to converge. We have found it more fruitful in these multiclass queueing examples to pursue reformulations of the state transition function (e.g., uncontrolled formulations or partially controlled formulations) that lead to simpler inner problems. These reformulations lead to very good lower bounds that are significantly easier to calculate; moreover, the complexity of these calculations does not depend in a critical way on the choice of approximate value function, unlike, e.g., Lagrangian relaxations of the inner problems without any reformulations.