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: March 24, 2015  
First version: September 6, 2014

Abstract

We consider infinite horizon stochastic dynamic programs with discounted costs and study how to use information relaxations to calculate lower bounds on the performance of an optimal policy. We develop a general framework that allows for reformulations of the underlying state transition function. These reformulations can simplify the information relaxation calculations, both in leading to finite horizon subproblems and by reducing the number of states in these subproblems. We study as important special cases both “weak formulations” in which states are independent of actions and “strong formulations” which retain the original dependence on actions. Our reformulations incorporate penalties for information in a direct way via control variate terms. We show that the approach improves on the lower bounds from “Bellman feasible” approximate value functions when the control variates are built from these approximate value functions. We apply the approach to the problem of dynamic service allocation in a multiclass queue; such models are well-studied but challenging to solve. In our examples, we find the information relaxation lower bounds are relatively easy to calculate and are very close to the upper bounds obtained from simple heuristics. Finally, we discuss extensions of the approach to stochastic shortest path and average cost problems.

Keywords: infinite horizon dynamic programs, information relaxations, 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.

One way to obtain performance bounds is through the information relaxation approach developed in Brown, Smith, and Sun (2010) (hereafter, BSS). This approach involves working with policies that can take advantage of additional information that is not available to feasible policies. For 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. An important component of the approach in BSS is incorporating a penalty that charges for the use of the additional information. A dual problem is then to optimize penalties to obtain the best, i.e., largest, lower bound from the information relaxation. For finite horizon DPs, BSS show how to construct dual feasible penalties; weak duality then yields lower bounds on the optimal cost. BSS also show strong duality in that there exists an “ideal” penalty for which the resulting lower bound equals the optimal cost. Although this approach has been applied successfully in many applications, in some problems, solving the information relaxation inner problems may still require dealing with many states. This challenge can arise in problems where actions strongly influence states as well as in long-running systems such as those described by infinite horizon models, as many states may be possible at distant times.

In this paper, 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 by appropriate likelihood ratios (more formally, Radon-Nikodym derivatives). This idea is motivated in part by the work of Rogers (2007), who develops “weak formulations” in which state transitions do not depend on actions. This is in contrast to using state transition functions that depend on actions as in the original formulation of the primal DP, i.e., “strong formulations.” The framework we propose unifies these approaches and generalizes 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. In general, the key is selecting an information relaxation, a penalty, and a reformulation in order to obtain a relaxed problem that is easy enough to solve while still providing a good lower bound.

We study and develop these ideas for infinite horizon DPs with discounted costs. This includes finite horizon DPs as a special case and thus the basic information relaxation results (weak and strong duality) extend the corresponding results from BSS. Conceptually, the generalization of information relaxations to
infinite horizon problems is straightforward: for example, in the multiclass queueing application, when the server operates over an infinite horizon, a perfect information relaxation simply involves infinite sequences of arrivals and services. Practically, however, we need to deal with the fact that we can neither generate nor store infinitely long sequences. By using the reformulations, we can ensure the information relaxation inner problems have finite horizons (including, but not limited to, the standard (e.g., Puterman 1994) reformulation as a DP with no discounting and a geometric stopping time). The reformulations can also limit the number of states that are possible in the information relaxations. This can be important in infinite horizon DPs, as we may need reformulations with relatively long horizons to get good bounds when discounting is light.

We capture penalties for information implicitly by adding control variates to the primal DP. These control variates are zero mean for primal feasible policies but may have nonzero mean for the policies in the dual inner problems where the information has been relaxed. This approach simplifies the analysis of information relaxations; for example, there is no need to establish dual feasibility of these additional terms, since they are incorporated directly into the cost structure of the primal DP. This also allows for direct “gap” comparisons between the cost of a primal feasible policy and the cost of the information relaxation inner problem in each scenario. Control variates are standard and have been used in some applications involving information relaxations. Brown and Smith (2014) observed that similar gap comparisons can be made in some cases with their “gradient penalties,” but did not fully connect control variates and penalties. To our knowledge, this direct and natural connection between control variates and dual penalties had yet to be fully appreciated.

We contribute to the literature on information relaxations in several additional ways. Although the proofs of weak and strong duality use very similar arguments to those in BSS, we also identify conditions under which weak duality continues to hold even when the reformulated state transitions may not cover states visited by optimal policies. More specifically, weak duality continues to hold even when absolute continuity conditions are violated, provided we use “Bellman feasible” approximate value functions in constructing the control variates. We also provide an analysis of the quality of the resulting lower bounds: we show the information relaxation approach always improves on the lower bounds from Bellman feasible approximate value functions and we also show that this guaranteed bound improvement continues to hold if we can only solve certain relaxations of the corresponding inner problems. We discuss the impact that the inner problem horizon length has on the quality of the lower bounds. We discuss how to extend the approach to general stochastic shortest path (SSP) problems and also how this extension can be used for average cost problems.

Finally, as stated earlier, we apply the approach to the problem of service allocation in a multiclass queue. Multiclass queueing 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 between 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, but we show how to use information relaxations to obtain relatively easy-to-compute lower bounds that are nearly tight with upper bounds that result from relatively simple feasible heuristic policies.
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 that inspired the “weak 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. 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 to state transitions. We also use Lagrangian relaxations in our multiclass queueing examples and find state transition reformulations to be quite important there; more broadly, our bound improvement results apply to discounted infinite horizon DPs with any information relaxation, any Bellman feasible approximate value function, and any reformulation of the state transition function.

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 (2013) 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 infinite horizon primal DP, briefly review BSS, and develop the primal DP reformulations involving control variates and reformulated state transition functions. Section 3 develops the main theory related to information relaxations in our setting. Section 4 presents the multiclass queueing application and discusses how to apply information relaxations to this problem; numerical examples of the model and bounds are presented. Section 5 discusses directions for future research. Most proofs and some detailed derivations are presented in Appendix A. Appendix B provides some additional illustrative examples and offers a discussion on extensions of the approach to SSP and average cost problems.
2. Problem Formulation

In this section, we formulate the primal DP, review some of the results in BSS for finite horizon stochastic DPs, and develop the reformulations of the primal DP that will be useful for the information relaxation results we later develop.

2.1. The Primal Dynamic Program

We study an infinite horizon stochastic dynamic program with discounted costs. Time is discrete and indexed by \( t \), starting with \( t = 0 \). We will work with a Markov decision process formulation, with \( x_t \) and \( u_t \) denoting the state and action, respectively, at time \( t \). We let \( X \) and \( U \) denote the state and action spaces, respectively, and we assume each of them is a nonempty Borel space. For every \( x_t \in X \), \( u_t \) is restricted to be within the nonempty set \( U(x_t) \subseteq U \), and we assume the set of admissible state-action pairs \( \{(x, u) : x \in X, u \in U(x)\} \) is a Borel set in \( X \times U \). States evolve as

\[
x_{t+1} = s(x_t, u_t, w_{t+1}),
\]

where \( w_{t+1} \) is a real-valued random variable on a Borel space with probability measure \( P \), \( s \) is a Borel-measurable transition function, and \( x_0 \) is given. We assume (without loss of generality) the \( w_t \)'s are IID and we denote expectations over \( w_{t+1} \) (or, sometimes, just \( w \)) by \( E \). Costs depend on states and actions and are denoted by \( c(x_t, u_t) \), which we assume to be Borel-measurable on \( X \times U \) and uniformly bounded. The objective is to minimize the total expected discounted costs, where \( \alpha \in (0,1) \) is a known discount factor.

A policy \( \mu := \{\mu_t\} \) is a sequence of Borel-measurable functions, each mapping from \( \{w_t\}_{t \geq 1} \) to feasible actions in \( U \); we will focus on non-randomized policies. If each \( \mu_t \) can be written as a Borel-measurable function from \( X \) to \( U \), we say \( \mu \) is a Markov policy. Finally, a Markov policy is stationary if \( \mu_t \) does not depend on \( t \). We let \( \mathcal{U} \) denote the set of stationary policies. We denote the expected cost of a policy \( \mu \in \mathcal{U} \) starting in state \( x_0 \) by \( V_\mu(x_0) := E\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, \mu(x_t))\right] \), where the expectation is over \( \{w_t\}_{t \geq 1} \). Finally, \( V(x) \) denotes the optimal expected cost from state \( x \) among all policies.

The following results are standard (e.g., Bertsekas and Shreve 1996; we discuss details in Sec. A.1):

(a) For any \( \mu \in \mathcal{U} \), the associated value function \( V_\mu(x) \) is Borel-measurable on \( X \) and satisfies

\[
V_\mu(x) = c(x, \mu(x)) + \alpha E[V_\mu(s(x, \mu(x), w))].
\]

(b) The optimal value function \( V \) is finite and satisfies, for every \( x \in X \), the Bellman equation

\[
V(x) = \inf_{u \in U(x)} \{c(x, u) + \alpha E[V(s(x, u, w))]\}. \tag{1}
\]

We assume that the infimum on the right-hand side of (1) is achieved for every \( x \in X \); this implies (e.g., Bertsekas and Shreve 1996, Corollary 9.12.1) there exists a stationary policy that is optimal. Note that this
formulation includes finite horizon DPs simply by including time in the state variable; thus the results that follow apply to finite horizon problems in this setting as well.

An equivalent formulation of the primal DP that will be useful to us is one in which there is no discounting but there is an absorbing and costless state $x^a$ that is reached with probability $1 - \alpha$ from every state $x \in X$ and for every $u \in U(x)$, conditional on not absorbing, the distribution of state transitions is unchanged. In this formulation, we can equivalently express the expected cost of any $\mu \in U$ as

$$V_\mu(x_0) = \mathbb{E}[\sum_{t=0}^{T-1} c(\tilde{x}_t, \mu(\tilde{x}_t))],$$

where $\tau$ is the stopping (or absorption) time for reaching $x^a$ and is geometric with parameter $1 - \alpha$, and we use $\tilde{x}_t$ to denote states in this reformulation. We refer to this as the strong SSP formulation and the equivalence to the discounted formulation above is standard (e.g., Puterman 1994, Proposition 5.3.1). In Section 2.4, we consider more general reformulations that also involve changing the way actions influence states.

### 2.2. Information Relaxations and Review of BSS

The information relaxation approach in BSS, which we now briefly review, considers policies that use more information than is available to primal feasible policies. Specifically, we consider policies that can use advance information about realizations of the uncertain sequence $\{w_t\}_{t \geq 1}$; such policies are generally not feasible to the primal DP. By finding the best such policies, we can obtain lower bounds on the optimal value of the primal DP.

To make this idea concrete, we let $\mathcal{F}$ denote the underlying filtration associated with $\{w_t\}_{t \geq 1}$, i.e., $\mathcal{F} := \{\mathcal{F}_t\}_{t \geq 0}$, where $\mathcal{F}_t$ denotes the $\sigma$-algebra generated by $(w_1, \ldots, w_t)$ at time $t$. We will describe additional information structures via another filtration $\mathcal{G} := \{\mathcal{G}_t\}_t$. We say $\mathcal{G}$ is a relaxation of $\mathcal{F}$ if $\mathcal{F}_t \subseteq \mathcal{G}_t$ for all $t$, i.e., under $\mathcal{G}$, at least as much (and perhaps more) is known about $\{w_t\}_{t \geq 1}$ at all times $t$. We let $\mathbb{E}_{\mathcal{G}}$ denote expectations conditional on $\mathcal{G}_t$. We will refer to realizations of $\{w_t\}_{t \geq 1}$ as scenarios.

A policy is adapted to a filtration $\mathcal{F}$ (or $\mathcal{F}$-adapted) if the actions taken by the policy in every period $t$ are measurable with respect to $\mathcal{F}_t$. Note that stationary policies to the primal DP above only depend on $x_t$ and are therefore $\mathcal{F}$-adapted. We let $\mathcal{U}_F$ denote the set of feasible, $\mathcal{F}$-adapted policies. If $\mathcal{G}$ is a relaxation of $\mathcal{F}$, then $\mathcal{U}_F \subseteq \mathcal{U}_G$: a policy that is $\mathcal{F}$-adapted is also $\mathcal{G}$-adapted.

The primal DP formulation in BSS is somewhat different in that it is a finite horizon problem and there is no explicit definition of a state: costs are functions of uncertainties and past actions. To keep notation as consistent as possible, we will present results in BSS for a finite horizon counterpart of the primal DP above with horizon length $T$, cost functions $c_t$ and state transition functions $s_t$. In this case, the expected cost for any $\mu \in \mathcal{U}_F$ can be written as

$$V_0,\mu(x_0) = \mathbb{E}[\sum_{t=0}^{T-1} c_t(x_t, \mu_t(x_t))].$$
where the additional subscript emphasizes that this is a finite horizon problem. We let $V_t(x_t)$ denote the optimal cost at time $t$ from state $x_t$.

We let $\mathbf{u} := (u_0, \ldots, u_{T-1})$ and $\mathbf{w} := (w_1, \ldots, w_T)$. BSS then consider penalty functions $z(\mathbf{u}, \mathbf{w}) := \sum_{t=0}^{T-1} z_t(\mathbf{u}, \mathbf{w})$ that depend on actions and uncertainties. BSS call a penalty function $z$ \textit{dual feasible} if $\mathbb{E}[z(\mu)] \leq 0$ for all $\mu \in \mathcal{U}_F$.

The following result recaps some of the main results from BSS.

\textbf{Proposition 2.1. (Brown, Smith, and Sun 2010)} Let $\mathcal{G}$ be any information relaxation of $\mathbb{F}$.

(a) Weak duality: For any dual feasible penalty $z$,

$$V_0(x_0) \geq \inf_{\mu \in \mathcal{U}_G} \mathbb{E}\left[ \sum_{t=0}^{T-1} c_t(x_t, \mu_t(x_t)) + z_t(\mu) \right]. \quad (3)$$

(b) Dual feasible penalties: Let $g_1, \ldots, g_T$ be any sequence of measurable functions $g_t : X \rightarrow \mathbb{R}$. Then the penalty given by $z_t(\mathbf{u}, \mathbf{w}) = \mathbb{E}[g_{t+1}(s_{t+1}(x_t, u_t, w_{t+1}))] - \mathbb{E}_{\mathcal{G}}[g_{t+1}(s_{t+1}(x_t, u_t, w_{t+1}))]$ is dual feasible.

(c) Strong duality: When $g_t = V_t$, inequality (3) holds with equality.

Part (a) of Proposition 2.1 shows that we can get lower bounds with any information relaxation and any dual feasible penalty. For example, when $\mathcal{G}$ is the perfect information relaxation, we can simulate scenarios $(w_1, \ldots, w_T)$ and select actions “path-wise” that minimize the sum of costs plus penalties. The expected value of the optimal costs (plus penalties) then provides a lower bound on the optimal value $V_0(x_0)$. Part (b) of Proposition 2.1 is useful because it provides a method for constructing dual feasible penalties: we can obtain a dual feasible penalty by adding up differences in conditional expectations of a sequence of any “generating functions” $g_1, \ldots, g_T$. Finally, part (c) of Proposition 2.1 shows that this approach, in theory, provides tight bounds: by taking the generating functions to be the optimal cost functions, the optimal value from the relaxed problem equals the optimal cost of the primal DP. Of course, when applying the method, we would typically not know $V_t$, but we can take the generating functions to be approximate value functions and, by parts (a) and (b) of Proposition 2.1, we will nonetheless obtain a lower bound on the optimal cost.

In order to develop and apply results analogous to Proposition 2.1 to the infinite horizon primal DP from Section 2.1, two reformulations of the problem will be helpful. We develop these in the next two subsections.

\subsection*{2.3. Control Variates}

The use of control variates is a standard variance reduction technique in Monte Carlo simulation. Control variates will be useful to us for two reasons. First, we will use control variates for variance reduction when estimating the expected costs of policies. Second, and more importantly, these control variates will play the role of penalties in the information relaxations, as we will explain in Section 3.

Our use of control variates can be viewed as an application of the “approximating martingales” method of Henderson and Glynn (2002). Specifically, we will use approximate value functions $\bar{V}$ that we assume are
bounded and Borel-measurable, and we let \( v(x_t, u_t) := \alpha \mathbb{E}[\bar{V}(s(x_t, u_t, w))] - \bar{V}(x_t) \), where the expectation in the first term is taken over \( w \). We will add \( v(x_t, u_t) \) to the costs in each period. For any stationary policy \( \mu \) and any state \( x_t \in X \), an application of the law of iterated expectations shows that \( \mathbb{E}[\alpha \mathbb{E}[\bar{V}(s(x_t, \mu(x_t), w))] - \alpha \bar{V}(x_{t+1})] = 0 \). Thus in adding \( v \) to the costs in each period there will be a cancelation in expectation between adjacent terms and adding these terms does not change the expected costs of any primal feasible policy. The following simple result formalizes this. We let \( c(\mu) := \sum_{t=0}^{\infty} \alpha^t c(x_t, \mu(x_t)) \) and \( v(\mu) := \sum_{t=0}^{\infty} \alpha^t v(x_t, \mu(x_t)) \).

**Proposition 2.2.** For any \( \mu \in \mathcal{U} \),

\[
V_\mu(x_0) = \bar{V}(x_0) + \mathbb{E}[c(\mu) + v(\mu)].
\]

These control variates can be quite helpful in reducing variance when estimating \( V_\mu(x_0) \). In particular, when \( \bar{V} = V_\mu \), we have \( c(\mu) + v(\mu) = 0 \) almost surely, so (4) reduces to \( V_\mu(x_0) = \bar{V}(x_0) \) in every scenario. In this case, \( V_\mu(x_0) \) can be estimated with zero variance. Of course, in such cases there is no need for estimation in the first place, but in cases when \( V_\mu \) is unknown, we may nonetheless expect a reduction in sample variance if \( \bar{V} \) is a good approximation of \( V_\mu \).

We will always assume that any approximate value function \( \bar{V} \) we use satisfies, like costs, \( \bar{V}(x^*) = 0 \). This implies that the counterpart of (4) for the strong SSP formulation is

\[
V_\mu(x_0) = \bar{V}(x_0) + \mathbb{E}\left[ \sum_{t=0}^{T-1} c(\tilde{x}_t, \mu(\tilde{x}_t)) + v(\tilde{x}_t, \mu(\tilde{x}_t)) \right].
\]

Although we can use any approximate value function as a control variate, we will focus on \( \bar{V} \) satisfying

\[
\bar{V}(x) \leq c(x, u) + \alpha \mathbb{E}[\bar{V}(s(x, u, w))] \quad \text{for all } x \in X, \ u \in U(x).
\]

We will say an approximate value function \( \bar{V} \) satisfying (6) is **Bellman feasible**.

A well-known fact (e.g., Puterman 1994) is that Bellman feasible approximate value functions are a lower bound on the optimal value function in every state. We can obtain Bellman feasible approximate value functions in a variety of ways. For instance, the “approximate linear programming” approach (e.g., de Farias and Van Roy 2003) involves a linear programming formulation that imposes Bellman feasibility constraints on an approximate value function that is a weighted sum of basis functions. \(^1\) Wang, O’Donoghue, and Boyd (2014) show how to use semidefinite programming to obtain Bellman feasible approximate value functions for infinite horizon DPs. Finally, relaxations of the primal DP also may lead to Bellman feasible approximate value functions. For instance, in the multiclass queueing application in Section 4, we will work with Bellman feasible approximate value functions from Lagrangian relaxations that relax constraints on how many customers can be served simultaneously.

---

\(^1\) In some applications of approximate linear programming, ensuring Bellman feasibility may require dealing with a large number of constraints; in such cases, constraint sampling techniques are often used to approximate Bellman feasibility. de Farias and Van Roy (2003) use constraint sampling in their application to queueing networks.
2.4. Reformulating State Transitions

We will also work with reformulations of the primal DP that employ different state transition functions and thus different transition probabilities. As we will discuss, these reformulations can simplify solving the relaxed DPs in the information relaxations.

Specifically, in our reformulations, we will consider using different state transition functions \( \tilde{s}_t(x, u, w) \) (again, assumed Borel-measurable) for the evolution of states; these may depend on time. We will use the notation \( \tilde{x}_t \) to indicate states generated by \( \tilde{s}_t \). As in the strong SSP formulation discussed in Section 2.1, we assume there is an absorbing and costless state \( x^\tau \), and we let \( \tau \) denote the (potentially random) absorption time to \( x^\tau \) when state transitions follow \( \tilde{s}_t \). We assume that under \( \tilde{s}_t \), the absorption time \( \tau \) is almost surely finite. This places no practical limits on our reformulations: as we will see, one of the roles of these reformulations is to lead to finite horizon inner problems in the information relaxations.

We can interpret the strong SSP formulation as a change of measure in the primal DP, in which the probability of a transition to \( x^\tau \) becomes \( 1 - \alpha \), and transitions to non-absorbing states are scaled down by a factor of \( \alpha \). More general reformulations are possible, however. To formalize this, we let \( P_u(x, \mathcal{X}) \) denote the state transition stochastic kernel to any Borel set of states \( \mathcal{X} \) from state \( x \in X \) when action \( u \) is selected in the primal DP, i.e., \( P_u(x, \mathcal{X}) := \mathbb{P} \{ (w : s(x, u, w) \in \mathcal{X}) \} \). Analogously, let \( Q_{t,u}(x, \mathcal{X}) \) denote the state transition stochastic kernel when states evolve according to \( \tilde{s}_t \) at time \( t \), i.e., \( Q_{t,u}(x, \mathcal{X}) := \mathbb{P} \{ (w : \tilde{s}_t(x, u, w) \in \mathcal{X}) \} \). If \( P_u \) is absolutely continuous with respect to \( Q_{t,u} \), i.e., \( Q_{t,u}(x, \mathcal{X}) = 0 \) implies \( P_u(x, \mathcal{X}) = 0 \) for any Borel set \( \mathcal{X} \), an application of the Radon-Nikodym theorem implies that there exists a function \( \varphi_{t,u} \) such that for any real-valued, bounded, and Borel-measurable function \( f \),

\[
\mathbb{E}[f(s(x, u, w))] = \mathbb{E}[\varphi_{t,u}(x, \tilde{s}_t(x, u, w)) f(\tilde{s}_t(x, u, w))].
\]

where \( \varphi_{t,u}(x, y) \) is often written as \( dP_u(x, y)/dQ_{t,u}(x, y) \). Note that we have written (7) in such a way that expectations on both sides are over the same random variable \( w \), but the state transition probabilities differ; the factor \( \varphi_{t,u} \) adjusts for this. In models with a countable number of states, as in all our examples, \( \varphi_{t,u}(x, y) \) is simply the ratio of the transition probabilities from state \( x \) to state \( y \) under \( s \) and \( \tilde{s}_t \), respectively.

The following result relates this reformulation to our original formulation of the primal DP. In what follows, we use the notation \( \Phi_t(\mu) := \prod_{i=0}^{t-1} \varphi_{t,i}(\tilde{x}_i, \tilde{x}_{i+1}) \), with \( \Phi_0 := 1 \), where \( \mu \in \mathcal{U} \) is a given policy. Note we allow for the possibility that \( P_u \) may not be absolutely continuous with respect to \( Q_{t,u} \); in general, \( \varphi_{t,u} \) should be understood to mean the unique Radon-Nikodym derivative of the absolutely continuous component of \( P_u \) with respect to \( Q_{t,u} \) (see, e.g., Thm. 6.10 in Rudin 1987 and the appendix for more details).

**Proposition 2.3.** For any \( \mu \in \mathcal{U} \):

(i) If \( P_{\mu(x)} \) is absolutely continuous with respect to \( Q_{t,\mu(x)} \) for all \( x \in X \) and \( t \), then

\[
V_{\mu}(x_0) = \mathbb{E}[\tilde{c}_t(\mu)],
\]

where \( \tilde{c}_t(\mu) := \sum_{w \in \mathcal{W}} \tilde{s}_t(x, u, w) \tilde{c}_t(\mu) \).
where \( \tilde{c}_\tau(\mu) \) is defined as
\[
\tilde{c}_\tau(\mu) := \tilde{V}(x_0) + \sum_{t=0}^{\tau-1} \alpha^t \Phi_t(\mu) \left( c(\tilde{x}_t, \mu(\tilde{x}_t)) + \alpha \mathbb{E}[\tilde{V}(s(\tilde{x}_t, \mu(\tilde{x}_t), w))] - \tilde{V}(\tilde{x}_t) \right).
\]

(ii) If \( \tilde{V} \) is Bellman feasible, then
\[
V_\mu(x_0) \geq \mathbb{E}[\tilde{c}_\tau(\mu)].
\] (9)

Proposition 2.3(i) states that we can calculate the expected costs for any policy \( \mu \in \mathcal{U} \) with state transitions generated by a different state transition function \( \tilde{s}_t \) and associated transition kernel \( Q_{t,u(\cdot)} \) in each state \( x \). This builds on Proposition 2.2 in that we include the effect of control variates; note that due to the way we defined the control variates, the term involving the expectation of \( \tilde{V} \) uses the original state transition function \( s \). This reformulation also requires a scaling of costs by likelihood-ratio factors. Proposition 2.3(ii) states that we obtain a lower bound on \( V_\mu(x_0) \) for any state transition function \( \tilde{s}_t \) (i.e., absolute continuity may be violated), provided we use Bellman feasible approximate value functions in the control variates.

As stated, an important component of these reformulations is how we handle transitions to absorption, i.e., the distribution of \( \tau \). One possibility is to have \( \tau \) geometric with parameter \( 1 - \alpha \), as in the strong SSP formulation. Alternatively, we could simply take a “truncated horizon” model with some horizon length \( T \), in which \( \tilde{s}_t \) transitions to \( x^a \) with probability one for all actions at time \( T \), but never earlier. Although this violates the absolute continuity condition in Proposition 2.3(i) at time \( T \), by Proposition 2.3(ii), we would still obtain a lower bound on a given policy \( \mu \in \mathcal{U} \) if \( \tilde{V} \) is Bellman feasible. Such formulations may be useful not only for variance reduction, but may also improve the information relaxation lower bounds, as we explain in Section 3.4. Many other variations are possible. Although not required by Proposition 2.3, we will work exclusively with reformulations that have \( \tau \) independent of states and actions, but not necessarily time.

These reformulations can also lead to other simplifications in the information relaxation inner problems. For instance, a perfect information relaxation will involve revealing sequences of \( \{w_t\}_{t \geq 1} \) in advance and solving (deterministic) DPs along each sequence. When state transitions are generated by \( s \), solving these deterministic DPs may still require dealing with many states due to the effect of actions on states under \( s \). On the other hand, if \( \tilde{s}_t \) is constant across actions for all states, actions cannot influence states, and there is a unique set of states that will be encountered given a particular sequence \( \{w_t\}_{t \geq 1} \). We refer to the case when \( \tilde{s}_t \) is constant across actions for all states and times as a weak formulation (or “weak form”).

Conversely, we refer to the case when the transition probabilities conditional on not absorbing under \( \tilde{s}_t \) match those under \( s \) for all actions and times as a strong formulation (or “strong form”); strong formulations can differ in the way absorption is handled, but all strong formulations retain the action dependence on state transitions (conditional on not absorbing) from the primal DP formulation. Many variations between these weak formulation and strong formulation “extremes” are possible, and we consider one such example of a “quasi-strong formulation” in the multiclass queueing application in Section 4.5.

The terms “strong formulation” and “weak formulation” resemble terminology used in Rogers (2007), who focuses on finite horizon MDPs and the use of what we call weak formulations in that setting. The
reformulation in Proposition 2.3 is more general in that state transitions may be action dependent.

2.5. Summary of Reformulations

We quickly review the reformulations we will apply to the primal DP, an infinite horizon problem with discounted costs and state transition function \( s \).

(i) We add control variates to the cost in each period. These control variates have zero mean for primal feasible policies, are built from approximate value functions, and will serve as penalties in the information relaxations.

(ii) We include an absorbing state with zero costs and we allow for reformulated state transition functions \( \tilde{s}_t \) that may depend on time. We will use reformulations for which the absorption time \( \tau \) is independent of states and actions and is almost surely finite. The reformulation may also change the way actions influence state transitions.

From Proposition 2.3, for any \( \mu \in \mathcal{U} \), the expected costs remain unchanged with these reformulations (correcting for the change in probabilities as explained in the proposition), provided the absolute continuity condition holds. Without this absolute continuity condition, we obtain a lower bound on the expected costs of \( \mu \in \mathcal{U} \) as long as the control variates are built from Bellman feasible approximate value functions.

3. Information Relaxations and Duality

In this section, we present a development and analysis of information relaxations for the primal DP. Section 3.1 discusses the main duality results (weak and strong duality), and Section 3.2 provides an analysis of the quality of the resulting lower bounds. Section 3.3 then presents a simple example illustrating the approach. Section 3.4 discusses how the distribution of the absorption time can affect the lower bounds, and Section 3.5 deals with the case when additional relaxations are used in the information relaxation inner problems.

3.1. Main Duality Results

The next two results are analogs of the finite horizon results in BSS, with the differences here being that we consider reformulations to the state transition function, penalties for information are implicit in the control variates, and the time horizon may now be random. The arguments in the proofs of weak and strong duality are otherwise similar to those in BSS.

**Proposition 3.1. (Weak Duality)** Consider any \( \mu \in \mathcal{U} \) and assume at least one of the conditions of Proposition 2.3 holds. For any relaxation \( \mathcal{G} \) of \( \mathcal{F} \),

\[
V_\mu(x_0) \geq \inf_{\mu \in \mathcal{G}} \mathbb{E}[\tilde{c}_\tau(\mu)].
\]

(10)

In particular, the right-hand side of (10) is a lower bound on \( V(x_0) \) if there is an optimal policy \( \mu^* \) such \( P_{\mu^*} \) is absolutely continuous with respect to \( Q_{\mu^*} \), or \( \tilde{V} \) is Bellman feasible.
Weak duality implies we can obtain lower bounds on the optimal value $V(x_0)$ by solving the relaxed problem on the right-hand side of (10) for any relaxed filtration $\mathcal{G}$, provided either (a) the state transition probabilities under $\bar{s}_t$ satisfy the absolute continuity condition relative to any optimal policy or (b) the approximate value functions used for control variates are Bellman feasible. In our examples we will always work with Bellman feasible approximate value functions. In problems for which useful Bellman feasible approximate value functions are not readily obtained, the absolute continuity condition can be met in a variety of ways, thus still leading to lower bounds on $V(x_0)$. For instance, with the strong SSP formulation, this condition holds automatically. In weak formulations, $\bar{s}_t$ is constant across actions, and we could ensure absolute continuity with respect to an optimal policy by ensuring $\bar{s}_t$ can transition to all possible states for all feasible actions.

In the case of a perfect information relaxation, the right-hand side of (10) involves optimizing over policies that can make full use of the scenario $\{w_t\}_{t \geq 1}$ in advance. In particular, we can select optimal actions “path-wise” for each scenario. Since we are working with transition functions for which the absorption time is independent of states and actions, we can first sample an absorption time $\tau$, then sample a scenario $(w_1, \ldots, w_{\tau-1})$. In this case, the inequality in (10) reduces to

$$V(x_0) \geq \tilde{V}(x_0) + E \left[ \inf_{u \in U} \left\{ \sum_{t=0}^{\tau-1} \alpha \mathbb{I}(u) (c(x_t, u_t) + v(\tilde{x}_t, u_t)) \right\} \right],$$

(11)

where $u$ denotes $(u_0, \ldots, u_{\tau-1})$, $U$ denotes the set of feasible $u$ for a given scenario $(w_1, \ldots, w_{\tau-1})$, and $\tilde{x}_t$ depends on $u$ through $\tilde{s}_t$. Thus we can calculate a lower bound in this case by sampling scenarios $(w_1, \ldots, w_{\tau-1})$ and solving the deterministic “inner problem” inside the expectation on the right of (11) for each scenario; the average of the optimal values provides a lower bound (in expectation) on $V(x_0)$.

With an imperfect information relaxation, we can follow an analogous procedure with the difference that the resulting inner problems will be stochastic DPs with expectations conditional on $\mathcal{G}_t$ in each scenario. Note that with an imperfect information relaxation, the absorption time may or may not be revealed in advance. We say a relaxation $\mathcal{G}$ includes $\tau$ if, under $\mathcal{G}$, the absorption time $\tau$ is known with probability one at $t = 0$. Relaxations that include $\tau$ are useful because they lead to finite horizon inner problems.

The following result provides more structure on the relaxed problems and asserts strong duality.

**Proposition 3.2. (Inner Problem Recursions and Strong Duality)**

Let $\mathcal{G}$ be any relaxation of $\mathcal{F}$ that includes $\tau$. Then:

(i) The lower bound $\inf_{\mu \in \Delta L_t} \mathbb{E}[\tilde{\mathcal{C}}(\mu)]$ equals $\mathbb{E}[V^\tau_0(x_0)]$, where $V^\tau_0(x_0)$ is given by the $\tau$-period recursion:

$$V^\tau_t(x_t) = \inf_{u \in U(x_t)} \left\{ (c(x_t, u) + \alpha \mathbb{E}[\tilde{V}(s(x_{t+1}, u, w))] + \alpha \mathbb{E}_{\mathcal{G}_t} [\gamma(x_{t+1}) (V^\tau_{t+1}(x_{t+1}) - \tilde{V}(\tilde{x}_{t+1}))] \right\},$$

(12)

where $\tilde{x}_{t+1} := \tilde{s}_t(x_t, u, w)$ and $V^\tau_\tau := 0$.

(ii) If $\tilde{V} = V$, then $V^\tau_t(x_t) = V(x_t)$ for all $x_t$ and $t$, in every scenario.
Proposition 3.2(i) implies that we can calculate lower bounds on the primal DP by generating scenarios and solving the recursion (12) in each scenario; since $G$ includes $\tau$, each of these recursions is a finite horizon DP. It is important to remember that the inner problem value functions $V^G_t$ are random variables that depend on scenarios: the expected value of $V^G_0(x_0)$ provides a lower bound on the optimal value of the primal DP.

The weak duality result (Proposition 3.1) masks the effect of any penalty for using extra information. The recursion in Proposition 3.2 sheds light on the role of the approximate value function, originally added to the primal DP as a control variate, as a penalty in the information relaxation: absent $\bar{V}$, it would be optimal in (12) to greedily select actions that only minimize costs, conditional on $G$. The terms involving $\bar{V}$ “discourage” such choices, provided $\bar{V}$ is a relatively good approximate value function. For instance, with perfect information and the strong SSP formulation, these penalty terms have the form $\alpha E[\bar{V}(s(x_t, u, w))] - \bar{V}(x_{t+1})$, which may be relatively large if $u$ is chosen to solely drive the system towards low-cost states $x_{t+1}$ (note that for the strong SSP formulation, the discount factor in front of the $E_{x_t}$ term is canceled by $\varphi_{t,u}$ prior to absorption). Later results and examples will further underscore the importance of the control variates as a penalty in obtaining good lower bounds.

Though Proposition 3.2(i) provides a recursion whose solution leads to lower bounds, the result says nothing about how difficult it is to solve these recursions. The value functions $V^G_t$ are functions of both $G_t$ and $x_t$, and thus solving (12) may require dealing with policies that are not Markov with respect to the state space $X$: we must be able to deal with both state transitions as well as updates in the conditional expectations described by the relaxed filtration $G$. We could of course view these recursions as Markov, perhaps with respect to a larger state space: for instance, consider an information relaxation that reveals $\tau$ and $w_{t+1}$ at each time period $t < \tau$, i.e., $G_t = \sigma(F_t, w_{t+1})$. The resulting inner problems can be formulated as a Markovian DP with a state space that includes both $X$ as well as the realization of $w_{t+1}$ at each time period; we provide a simple example illustrating this in Appendix B.1. Such imperfect information relaxations would probably be unusual choices in applications but are possible.

Despite this apparent downside, note that Proposition 3.2(ii) also makes no statements about the number of states that need be considered in solving the recursions (12). In general, we need only consider states that can be encountered under $G$ with the transition functions $\hat{s}_t$, and this may be a relatively small number of states. For example, with perfect information and a weak formulation, we need only deal with a single state per period in solving the recursions (12) in each scenario: this follows from the fact that state transitions $\hat{s}_t$ do not depend on actions and the values of $w_t$ are known exactly. Thus in each scenario the states $(x_0, \ldots, x_{\tau-1})$ are fixed and solving (12) amounts to selecting an optimal action for each of the $\tau$ periods. More generally, the complexity of solving (12) will depend on the information relaxation, the chosen state transition function, and the approximate value function: the key is in balancing between choices that lead to inner problems that are easy enough to solve while leading to good lower bounds.

Proposition 3.2(ii) is a strong duality result: if the approximate value function in the penalty equals the optimal value function, then the lower bounds are tight in all states (including the initial state) and at
all times in the recursion (12), in every scenario. Thus, in theory, we can obtain tight lower bounds using this approach. This provides hope that in cases where we only have an approximation of the optimal value function we can still obtain good lower bounds. In addition, Proposition 3.2(ii) does not place restrictions on $\tau$, so good lower bounds may be possible even when the inner problems have very short horizons.

In problems with continuous action or state spaces, it may be easier to solve inner problems with perfect information as a single static optimization problem, rather than as a recursion. A convex structure for the inner problem (e.g., penalized costs are convex in actions) is necessary for this approach to be viable in general; see Brown and Smith (2014) for a detailed development of this approach for finite horizon problems.

3.2. Analysis of Information Relaxation Bounds

It is useful to develop some insights into the quality of the lower bounds we will obtain using information relaxations. The result below relates the lower bounds to bounds from the approximate value function $\bar{V}$ and also adds a comparison to feasible policies. We say $\bar{V}$ is $\epsilon$-Bellman feasible if (6) holds with $\bar{V} + \epsilon$ on the left side of the inequality; in this definition, $\epsilon$ represents the “slack” available in satisfying Bellman feasibility, with $\bar{V}$ being Bellman feasible if and only if $\bar{V}$ is $\epsilon$-Bellman feasible for some $\epsilon \geq 0$.

We also use the notation $\gamma := \alpha \inf_{u, x, y \neq x_a} \varphi_t(x, y)$. When $\gamma_t$ does not depend on time, we denote $\gamma_t$ simply by $\gamma$. We let $\delta := \inf_{u, x, y \neq x_a} dP_u(x, y) / dQ_u(x, y)$ in the case when $\tilde{s}_t$ does not depend on time, where $P_u$ and $Q_u$ denote the state transition kernels under $s$ and $\tilde{s}_t$, respectively.

We will also compare the lower bounds to the upper bounds associated with expected costs of feasible policies. To this end, we can write a recursion analogous to (12) which, rather than minimizing over actions, fixes actions according to a policy $\mu \in U$. We let $V^\mu_G$ denote the corresponding value function.

**Proposition 3.3. (Bound Guarantees)** Let $G$ be any relaxation of $F$ that includes $\tau$, and let $\bar{V}$ be $\epsilon$-Bellman feasible, with $\epsilon \geq 0$. Then, for all $t$ and $x_t$, in every scenario:

(i) The inner problem value functions $V^\mu_G$ satisfy

$$V^\mu_G(x_t) \geq \bar{V}(x_t) + \kappa_t \epsilon,$$

where $\kappa := \sum_{i=0}^{\tau-t} \gamma_t^i$ and $\kappa_{\tau-t} := 1$.

(ii) If $\tilde{s}_t$ does not depend on time, $\kappa_t = (1 - \gamma^{-1})/(1 - \gamma)$, and

$$\mathbb{E}[V^\mu_G(x_0)] \geq \bar{V}(x_0) + \frac{\epsilon}{1 - \alpha \delta}.$$

(iii) With the strong SSP formulation, (i) and (ii) hold for any $\epsilon$ with $\gamma = \delta = 1$, and (13) reduces to

$$V^\mu_G(x_t) \geq \bar{V}(x_t) + (\tau - t) \epsilon.$$

13
(iv) For any primal feasible policy $\mu$:

$$V^G_t(x_t) \leq V^\mu_t(x_t). \quad (16)$$

Moreover, $E[V^G_t(x_t)] \leq V^\mu_t(x_t)$, with equality holding if condition (i) of Proposition 2.3 holds for $\mu$.

The results in Proposition 3.3 hold in every scenario: these are “path-wise” results. Part (i) of Proposition 3.3 implies that the inner problem value functions will be no smaller than $\bar{V}$ in all states and in all scenarios, provided $\bar{V}$ is Bellman feasible. This result also provides a lower bound on how much larger the inner problem value functions will be compared to $\bar{V}$. Part (ii) of the result is the special case of part (i) when the transition functions $\tilde{s}_t$ do not depend on time; in this case, we obtain a guarantee on the lower bound from the information relaxation (i.e., the expected value of $V^\circ_0$) by the simple form in (14). Recall that we assume transitions to absorption do not depend on state or action; if transitions to absorption also do not depend on time, then $\tau$ has a geometric distribution, and note that (ii) holds for any geometric $\tau$ (i.e., with any parameter; parameters other than $1 - \alpha$ could be effective but would affect sample variance). Part (iii) specializes part (i) to the strong SSP formulation but also adds that (15) holds for any $\epsilon$, not just $\epsilon \geq 0$.

The lower bounds in parts (i)-(iii) may be coarse in that $\epsilon$ and $\gamma_t$ are minimum values across all states and actions; no doubt a more refined analysis is possible. In fact, in many problems, good choices of $\bar{V}$ may well have $\epsilon = 0$. Nonetheless, the results in Proposition 3.3 shed some light on when we may expect large or small improvements on $\bar{V}$. On the one hand, we may expect more substantial improvements on $\bar{V}$ if there are many states for which the “slack” in the corresponding Bellman feasibility inequality is large. On the other hand, when there are many states and actions such that $\varphi_{t,u}$ is small (and hence $\gamma_t$ is small), we may expect relatively modest improvements on $\bar{V}$. Regardless, we can never do worse than the lower bounds provided by any Bellman feasible approximate value function.

Part (iv) of Proposition 3.3 shows that the inner problem value functions are no larger than the corresponding value functions for any primal feasible policy. This result follows directly from the observation that primal feasible policies are feasible to the inner problems. This result is useful because it shows that we can make “path-wise” comparisons of costs for a primal feasible policy and the lower bounds from the information relaxation. The difference $V^G_{t,\mu}(x_t) - V^\circ_t(x_t)$ represents an estimate of the suboptimality gap of the policy $\mu$ in a given scenario. Eq. (16) ensures that these gap estimates are nonnegative in every scenario. Note that the estimates for the primal $V^G_{t,\mu}(x_t)$ also depend on $G$ as well, due to the expectations conditional on $G_t$ in (12); part (iv) adds that for primal feasible policies, the effect of $G$ “averages out” in expectation.

The pairing of the primal and dual calculations suggested by part (iv) may also be computationally attractive in that evaluating $V^G_{t,\mu}(x_t)$ requires relatively little additional work given the calculations for the relaxation value functions $V^\circ_t(x_t)$: the approximate value functions, change of measure terms $\varphi_{t,u}$, and conditional expectations will have already been evaluated at the relevant states. In some cases, actions under $\mu$ may have already been evaluated as well. For instance, we can use weak formulations with state
transitions taken according to a given feasible policy \( \mu \), i.e., \( \hat{s}_t(x_t, u_t, w) = s(x_t, \mu(x_t), w) \) for all \( u_t \in U(x_t) \).

We will take such an approach in the multiclass queueing application we study in Section 4.

Finally, Proposition 3.3 provides hard limits on the relaxed value functions in every state and scenario, which suggests that sample variances should not be unmanageable when estimating the lower bounds. With a Bellman feasible \( \bar{V}(x_n) \), the inner problem values cannot fall below \( \bar{V} \), nor can the inner problem values go above the primal upper bound values in part (iv). With the strong SSP formulation, or with weak formulations for which \( \hat{s}_t(x_t, u_t, w) = s(x_t, \mu(x_t), w) \), the factor \( \varphi_{t,\mu} \) equals 1 in all the recursions for evaluating the primal value functions \( V_{t,\mu} \). This ensures in these cases bounded variances for a given \( \tau \) and suggests the sample variance in the lower bounds should be low when \( \bar{V} \) is a good approximation to \( V \).

### 3.3. A Simple Example: Machine Repair

We illustrate the approach with a classic example in which a decision maker is trying to determine the optimal repair policy for a machine that deteriorates over time. The machine can be in any one of \( n \) states, denoted by \( 1, \ldots, n \). In each period the decision maker can either attempt to repair the machine or not. The cost of attempting to repair the machine is \( R \geq 0 \). In addition, if \( x \) is the state in a given period then a cost of \( c(x) \geq 0 \) is incurred for that period in addition to the possible repair cost. Costs are increasing in \( x \). We let \( P^r \) and \( P^f \) denote the transition probability matrices corresponding to the actions “repair” \( (r) \) and “do not repair” \( (\bar{r}) \), respectively.

The objective is to minimize the total expected discounted cost over an infinite horizon. If \( n \) is not too large, solving for the optimal value function can be done easily (e.g., value iteration or by solving a linear program with \( n \) variables). Nonetheless, the example is instructive. The optimal value function \( V \) satisfies

\[
V(x) = \min \{ R + c(x) + \alpha P_x^r V, \ c(x) + \alpha P_x^f V \},
\]

where \( P_x^r \) and \( P_x^f \) denote the \( x \)th rows of \( P^r \) and \( P^f \), respectively, and \( V \) is written as an \( n \times 1 \) vector.

In our examples, we use \( n = 10 \) states, with \( c(x) = x \), \( R = 10 \) and \( \alpha = 0.75 \). If a repair attempt is made, the next state will be state 1 with probability .9. The repair fails with probability .1, in which case the machine remains in its current state. Thus, \( P^r(1,1) = 1 \), and \( P^r(x,1) = .9 \) and \( P^r(x,x) = .1 \) for \( x > 1 \). If a repair attempt is not made, the machine may deteriorate to a worse (higher) state. We use \( P^f(x,x) = .8 \), \( P^f(x,x+1) = P^f(x,x+2) = .1 \) for \( x \leq n - 2 \), \( P^f(n-1,n-1) = .8 \), \( P^f(n-1,n) = .2 \) and \( P^f(n,n) = 1 \).

We calculate lower bounds using two information relaxations with the strong SSP formulation. For both information relaxations, we use control variates based on approximate value functions calculated from value iteration. Specifically, we take \( \bar{V} = B^k 0 \) for some number \( k \geq 0 \) of iterations of the Bellman operator, \( B \), starting with all zeros when \( k = 0 \). Since costs are nonnegative, the choice \( \bar{V} = 0 \) is Bellman feasible; from monotonicity and convergence of the Bellman operator (e.g., Bertsekas 2001), \( \bar{V} \) constructed in this way is Bellman feasible for any \( k \geq 0 \).

With a perfect information relaxation, the decision maker knows both the time horizon \( \tau \) as well as the
scenario \( (w_1, \ldots, w_{\tau-1}) \). Using the strong SSP formulation, an inner problem (12) is therefore obtained by first sampling \( \tau \) from a geometric distribution with parameter \( 1 - \alpha \) and then sampling \( w = (w_1, \ldots, w_{\tau-1}) \).

Here we take the \( w_i \)'s to be IID uniform \([0,1] \) random variables and use inverse transforms on the state transition matrices \( P^r \) and \( P^w \) to generate state transitions for each action in each time period. We use \( \Pi^r_t \) and \( \Pi^w_t \) to denote the associated (deterministic) state transition matrices in each scenario, and again use the subscript \( x \) to denote the \( x \)th row of these matrices.

We then solve the inner problem (12), which in this case reduces to solving the deterministic DP

\[
V^c_t(x) = \min\left\{ c(x) + R + \left( \alpha P^r_t - \Pi^r_t \right) \bar{V} + \Pi^r_t V^c_{t+1}, \ c(x) + \left( \alpha P^w_t - \Pi^w_t \right) \bar{V} + \Pi^w_t V^c_{t+1} \right\}
\]

for \( t = 0, \ldots, \tau - 1 \), with \( V^c_\tau = 0 \).

We also consider an imperfect information relaxation in which the decision maker knows \( \tau \) at all times but does not know any machine state transitions after time \( t \), i.e., \( \mathcal{G}_t = \sigma (F_t, \tau) \). We call this the absorption time relaxation. This is a tighter information relaxation than perfect information and therefore must lead to better lower bounds (BSS 2010, Proposition 2.3(i)). Under the strong SSP formulation, this relaxation reduces to only sampling \( \tau \) from a geometric distribution with parameter \( 1 - \alpha \) and then solving the \( \tau \)-period inner problem. The inner problems are now finite horizon stochastic DPs, and in this example have the form

\[
V^c_t(x) = \min\left\{ c(x) + R + \left( \alpha P^r_t \bar{V} + P^w_t V^c_{t+1}, \ c(x) + \left( \alpha P^w_t \bar{V} + P^r_t V^c_{t+1} \right) \right\}
\]

for \( t = 0, \ldots, \tau - 1 \), with \( V^c_\tau = 0 \).

Figure 1 shows the bounds for all initial states. The optimal value function is indicated in blue, and the black line indicates \( \bar{V} \) for some number \( k \) of iterations of value iteration, starting from all zeros. Since \( \bar{V} \) is Bellman feasible, we know from Proposition 3.3 that the lower bounds from both information relaxations will be at least as good as the lower bounds \( \bar{V} \); the green line shows the guarantee over \( \bar{V} \) from Proposition 3.3(iii). The solid and dashed red lines correspond to the perfect and absorption time information relaxations, respectively.

There are several noteworthy features in these results. First, even when the control variates are constructed from poor approximations to the optimal value function, the information relaxation bounds are relatively good. Second, the bound improvement guarantee from Proposition 3.3(iii) ranges from being quite slack in the \( k = 0 \) case to (relatively) tight in the \( k = 3 \) case. Third, even though the absorption time relaxation bounds are tighter than the perfect information bounds (as they must be) there is little difference in these bounds, particularly as \( k \) increases. In this example, it appears the benefit of additional information comes largely from knowing the absorption time: if the absorption time is relatively short and the penalty is somewhat weak, then regardless of the specific machine state transitions, it may not be optimal to repair.

Note also that in order to solve the inner problem recursions under either information relaxation, we generally need to calculate optimal actions in each time period for all states: even when the inner problem
is deterministic as in the perfect information case, state transitions are determined endogenously through actions, and there is no simple shortcut around keeping track of the relaxed value function in all states in each step of the recursions. This underscores a potential challenge in using information relaxations with strong formulations: sometimes (but not always) these inner problems may require dealing with many states. As discussed, however, with weak formulations and perfect information, we need only select actions on a single set of states in each scenario; the machine repair example with weak formulations is in Appendix B.2.

3.4. The Distribution of Absorption Time

Although it is natural to use state transitions with an absorption time that is geometric with parameter $1 - \alpha$, there is no requirement that this be the case. All else equal, changing the absorption time distribution is inconsequential in expectation when evaluating primal feasible policies: according to Proposition 2.3, we can use any distribution for absorption time and we will obtain the same expected costs for any primal feasible policy, provided the absolute continuity condition holds. (There may be other reasons, e.g., reducing sample variance, to change the absorption time distribution when evaluating primal policies.)

As suggested by the machine repair example, however, this observation does not carry over to the information relaxations: policies in the relaxation take advantage of additional information and, in particular, can take advantage of advance knowledge of $\tau$. For example, if $\tau$ is known to be short in some scenarios, it may be optimal to take actions that have low costs over a short horizon but would otherwise (i.e., without knowing $\tau$) risk exposure to costly states in later periods. This freedom in the relaxations may weaken the lower bounds, and changing the absorption time distribution to counteract this effect may be beneficial.

One way to do this involves working with longer horizons. For instance, rather than taking state transitions to have a geometrically distributed absorption time, we could instead use state transition functions $\tilde{s}_t$ that absorb with probability one at time $T \geq 1$ but never earlier. We call this a truncated horizon model.

Although such state transitions violate the absolute continuity condition in Proposition 2.3(i) (due to the
Figure 2: Perfect information bounds for machine repair strong formulations: SSP vs. truncated horizons.

fact that there is zero probability of a transition to a non-absorbing state at \( t = T \), by Proposition 2.3(ii) we nonetheless get lower bounds on the optimal value if control variates use Bellman feasible approximate value functions. The following result shows that, all else equal, longer horizons provide better lower bounds.

**Proposition 3.4.** Let \( \{\tilde{s}_1^t\} \) and \( \{\tilde{s}_2^t\} \) be truncated horizon models with horizons \( T_1 \) and \( T_2 \), where \( T_2 \geq T_1 \) and \( \tilde{s}_2^t = \tilde{s}_1^t \) for \( t = 0, \ldots, T_1 - 2 \). If \( \bar{V} \) is Bellman feasible, then for any relaxation \( \mathcal{G} \), in every scenario

\[
V^{\mathcal{G},1}_t(x_t) \geq V^{\mathcal{G},2}_t(x_t),
\]

for all \( t \) and \( x_t \), where \( V^{\mathcal{G},1}_t \) and \( V^{\mathcal{G},2}_t \) denote the value functions in (12) for each model.

Proposition (3.4) shows that an increase in the horizon length results in larger inner problem values in every state and scenario, which implies better lower bounds. Proposition 3.4 presents this idea for the case of truncated horizon models, but this can be generalized and many variations of this idea are possible (e.g., let the absorption time be geometric \( 1 - \alpha \) plus a fixed number of time periods \( T \)).

To illustrate this, consider again the machine repair example. In Figure 2, we show perfect information lower bounds based on a truncated horizon model with \( T = 3 \) and \( T = 25 \) periods. Relative to the discount factor of \( \alpha = 0.75 \), the horizon \( T = 3 \) is short and the horizon \( T = 25 \) is long. The control variates are zero in this case (i.e., the \( k = 0 \) case from Fig. 1), so there is no penalty for information. We see that that the lower bound from the \( T = 3 \) truncated model is not as good as the lower bound from the strong SSP formulation, which uses a geometric absorption time with parameter 0.25. When we increase the horizon to \( T = 25 \), however, the truncated model fares better than the strong form SSP bounds in all states.

As \( k \) increases, \( \bar{V} \) converges to the optimal value function, \( V \). If \( \bar{V} = V \), Proposition 3.2(ii) implies that, no matter what distribution of absorption we use, we would obtain tight lower bounds (e.g., even truncated models with very short horizons). When \( \bar{V} \) is Bellman feasible, we can never fall below \( \bar{V} \), and in general whether or not longer horizons will be worthwhile will depend on the quality of the approximation \( \bar{V} \).
3.5. Relaxations of the Inner Problems

As discussed, in some cases the inner problems may not be easy to solve. For instance, when using strong formulations and a perfect information relaxation, solving recursion (12) may require dealing with many states due to the dependence of state transitions on actions. In such cases we may instead attempt to find lower bounds on \( V_o(x_0) \) that are easier to calculate. Since the expected value of \( V_o(x_0) \) provides a lower bound on the primal DP, if we can calculate lower bounds on \( V_o(x_0) \) in each scenario, we can still obtain lower (albeit generally weaker) bounds on the primal DP.

There are different ways we could do this in a given problem and although the details of calculating lower bounds on \( V_o(x_0) \) will be problem specific, one general approach is to attempt to calculate approximate value functions \( \hat{V}_t(x_t) \) in each scenario such that \( \hat{V}_t(x_t) \leq V_t(x_t) \). We may then attempt to find the best (i.e., largest) such lower bound among a set \( \mathcal{V} \) of approximate value functions in each scenario. (Other approaches may involve finding a lower bound only on \( V_o(x_0) \), which would also be sufficient.)

When \( \bar{V} \) is Bellman feasible and \( \bar{V} \in \mathcal{V} \), this procedure can never yield a worse lower bound than \( \bar{V}(x_0) \) in each scenario. The following result formalizes this.

**Corollary 3.1.** If \( \bar{V} \) is Bellman feasible and \( \bar{V} \in \mathcal{V} \), then, in every scenario:

\[
\max_{\hat{V}_t \in \mathcal{V}} \{ \hat{V}_o(x_0) : \hat{V}_t \leq V_t \} \geq \bar{V}(x_0). \tag{18}
\]

This is a direct consequence of Proposition 3.3(i), which shows that if \( \bar{V} \) is Bellman feasible, then \( V_t(x_t) \leq \bar{V}(x_t) \) in every scenario. Thus, if \( \bar{V} \in \mathcal{V} \), choosing \( \hat{V}_t = \bar{V} \) is a feasible choice in the maximization in (18) and thus such relaxations of the inner problems will provide \( t = 0 \) values at least as large as \( \bar{V}(x_0) \) in every scenario.

We will follow this approach in the strong form inner problems for the multiclass queueing application we study in Section 4. In that problem, we will use Lagrangian relaxations of the original problem to generate a Bellman feasible penalty. These Lagrangian relaxations relax the constraint that a server can serve at most one customer at a time and lead to a decoupling of the value function across the customer classes. The associated strong form inner problems are often too difficult to solve exactly so instead we will solve for the analogous Lagrangian relaxations of these inner problems in each scenario; we then optimize over the associated Lagrange multipliers in the inner problem to obtain the largest lower bound in each scenario. By the above arguments, we can do no worse in each scenario than the lower bound provided by the original Lagrangian relaxation and, in our examples, in most scenarios we do strictly better than this lower bound.

In other problems, it may be effective to use an “approximate linear programming” approach that takes \( \mathcal{V} \) to be the set of weighted combinations of basis functions and optimizes over the weights while enforcing Bellman feasibility. Depending on the problem and the basis functions, ensuring Bellman feasibility may require dealing with far fewer states than a full solution of the primal DP. This could be done to generate both a \( \bar{V} \) as well as scenario-specific lower bounds on \( V_o(x_0) \) that are no smaller than \( \bar{V}(x_0) \).
4. 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.

4.1. 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; $\lambda_i$ and $\nu_i$ denote the mean arrival and service rates for class $i$ customers. Customers arrive to the queue and await service, and we let $x_i$ 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.

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., Bertsekas 2001), with points in time representing epochs at which arrivals or completed services may occur. 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 $s(x,i,w) := (s_1(x,i,w), \ldots, s_I(x,i,w))$, 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 $\lambda_j$. The next-period state satisfies

$$s_j(x,i,w) = \min(x_j + 1, B_j)$$

and is unchanged for all other customer classes.

(b) The service of the class $i$ customer being served (if any) is completed with probability $\nu_i$. The next-period state satisfies

$$s_i(x,i,w) = x_i - 1$$

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, $s(x,i,w) = x$.

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 $\lambda_i$ and $\nu_i$. The uniform $[0,1]$ value of $w$ for each period then represents a class $i$ arrival if it falls in an interval corresponding to $\lambda_i$, and a class $i$ service token if it falls in an interval corresponding to $\nu_i$. 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. $s(x,i,w) = x$ in this case.

We can then write the optimal value function $V$ of the system as

$$V(x) = \min_{i \in I_+(x)} \left\{ \sum_j c_j(x_j) + \alpha E[V(s(x,i,w))] \right\},$$

(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 $\alpha$ in (19) is a scaling of the continuous time discount factor; again, see, e.g. Sec. 5.1 of Bertsekas (2001).

4.2. Approximate Value Functions

The number of states involved in solving (19) scales exponentially in the number of customer classes; this 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 $\bar{V}$, i.e., policies that, in each period, allocate service to a customer class $i^*$ satisfying

$$i^* \in \arg \min_{i \in I_+(x)} \left\{ \sum_j c_j(x_j) + \alpha E[\bar{V}(s(x,i,w))] \right\},$$

(20)

where $\bar{V}$ is an approximate value function. The right-hand side of (20) approximates (19), with $\bar{V}$ in place of $V$. We can also use $\bar{V}$ to form a control variate in evaluating upper bounds from the heuristic, as well as lower bounds from an information relaxation. We will only consider $\bar{V}$ that are Bellman feasible.

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

Another approximation we use is based on Lagrangian relaxations. Specifically, we consider a relaxation of the problem in which the server can simultaneously serve multiple classes, but is charged a Lagrangian penalty for serving more than one class at any point in time. We let \( \ell \geq 0 \) represent the corresponding Lagrange multiplier; when \( \ell \) is independent of states, we can show that the relaxed problem decouples into a sum of value functions for each class; this essentially follows from Hawkins (2003) or Adelman and Mersereau (2008). This decoupling relies on costs decoupling by class and state transitions \( s_i \) depending only on \( x_i \) and whether or not we are serving class \( i \); the Lagrangian relaxations then have the form

\[
\hat{V}^\ell(x) = \frac{(I-1)\ell}{1-\alpha} + \sum_i \hat{V}^\ell_i(x_i),
\]

where \( \hat{V}^\ell_i(x_i) = \min_{a \in \{0,1\}} \{ c_i(x_i) + \alpha \mathbb{E}[\hat{V}^\ell_i(s_i(x_i, a, w))] - \ell \mathbbm{1}_{\{a=0\}} \} \). This expression for \( \hat{V}^\ell_i \) can be interpreted as the value function for a model with only class \( i \) in the system, with the inclusion of a cost reduction \( \ell \geq 0 \) that is received whenever the server idles. For a fixed Lagrange multiplier \( \ell \), solving for \( \hat{V}^\ell \) involves only dealing with \( B_i + 1 \) states. We may then optimize over \( \ell \geq 0 \) to maximize \( \hat{V}^\ell(x) \) in some particular state (or some weighted combination of states). This can be done relatively easily in a variety of ways, such as a bisection search on the Lagrange multiplier. For any \( \ell \geq 0 \), \( \hat{V}^\ell \) is Bellman feasible to the original model.

An extension of this idea is to aggregate customer classes into groups, and allow the server to simultaneously serve classes from multiple groups, while imposing the constraint that at most one class per group is served at any point in time. 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, \( \{I_1, \ldots, I_G\} \) is a partition of \( \{1, \ldots, I\} \)). We then consider a relaxed model that allows the server to simultaneously serve multiple groups, with the restriction that at most one class in each group can be served at any time, and includes a Lagrangian penalty for providing service to multiple groups simultaneously. Using a Lagrange multiplier \( \ell_G \geq 0 \), this relaxed problem decouples similarly as in (21), with

\[
\hat{V}^{\ell_G}(x) = \frac{(G-1)\ell_G}{1-\alpha} + \sum_g \hat{V}^{\ell_G}_g(x_g),
\]

where \( x_g \) represents the state vector (number of customers) for group \( g \), and

\[
\hat{V}^{\ell_G}_g(x_g) = \min_{a \in \{0,1\} \cup I_g} \left\{ \sum_{i \in I_g} c_i(x_i) + \alpha \mathbb{E}[\hat{V}^{\ell_G}_g(s_g(x_g, a, w))] - \ell_G \mathbbm{1}_{\{a=0\}} \right\}.
\]

In (23), state transitions \( s_g \) are defined over each group in the analogous way as in the original model.

---

\(^2\)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 probabilities of state transitions for each class do not depend on the states or service decisions of other classes, and it can be shown this is sufficient to lead to the decomposition (21). Similarly, the decoupling in the approximate linear program (ALP) bounds in Adelman and Mersereau (2008) goes through in this problem; these ALP bounds are the same as the Lagrangian relaxation bounds (21) in our single server examples, using Thm. 7 in Adelman and Mersereau (2008).
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 (23) relatively easily if we only group together a few classes. The Lagrangian relaxation described in (21) is the special case of each group consisting of an individual class. For any grouping, we can again optimize over the Lagrange multiplier \( \ell_G \geq 0 \), and for any \( \ell_G \geq 0 \), we obtain an approximate value function \( \bar{V}^{\ell_G} \) that is Bellman feasible.

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: a 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 control variates.

4.3. Information Relaxations and Inner Problems

In describing how we will solve the inner problems, it is helpful to first describe how we generate scenarios. For each scenario, we first simulate an absorption time that is geometric with rate \( 1 - \alpha \) (following the discussion in Section 3.4, we also tried truncated horizons with a relatively long horizon \( T \), but found this made little difference in these multiclass queueing examples; moreover, it is instructive to see how inner problem values vary with horizon length). Given an absorption time \( \tau \), we then generate \( \tau - 1 \) uniform \([0, 1]\) random variables \((w_1, \ldots, w_{\tau-1})\), representing arrivals and service tokens as described in Section 4.1. These lists of \( \tau - 1 \) class arrivals and service tokens are the scenarios that we use in evaluating all bounds.

We will work exclusively with perfect information relaxations. Solving the perfect relaxation inner problems exactly for strong formulations may be quite difficult here: discount factors close to one can lead to scenarios with many time periods and arrivals. Even though the inner problem is deterministic and may lead to fewer possible states to consider in a given scenario than contained in the full model (for instance, if there are few arrivals for some customer classes), the number of states we need to consider may nonetheless be quite large.

One way around this difficulty is to consider a weak formulation in which state transitions cannot be influenced by service decisions. In our examples, we will do this by taking state transitions in each scenario to be the fixed state transitions associated with the heuristic policy. In this formulation of the problem, if the heuristic successfully completes a class \( i \) service in a given period in a given scenario, then a class \( i \) customer departs the system, regardless of which customer we actually choose to serve. Note that this state transition function may not satisfy the absolute continuity condition associated with an optimal policy; however, we will be working exclusively with Bellman feasible \( \bar{V} \), which by Proposition 3.1 ensures the information relaxations...
using this weak formulation will lead to lower bounds on the optimal value function.

We let \((x_0, \ldots, x_{\tau-1})\) denote the states following the heuristic policy, with \(x_t = (x_{t,1}, \ldots, x_{t,i})\). Then these weak formulation inner problem values can be calculated by solving the recursion

\[
V^c_t(x_t) = \min_{i \in I_t(x_t)} \left\{ \sum_{j} c_j(x_{t,j}) + \alpha \mathbb{E} \left[ \tilde{V}(s(x_t,i,w)) \right] + \varphi_i(x_t, x_{t+1}) \left( V^c_{t+1}(x_{t+1}) - \tilde{V}(x_{t+1}) \right) \right\},
\]

for \(t = 0, \ldots, \tau - 1\), where \(V^c_0 = 0\). The inner problem value in the scenario is then given by \(V^c_0(x_0)\). The factor \(\varphi_i\) is equal to 1 at time \(\tau - 1\) for all actions. In earlier time periods, if the queue is empty at time \(t\), then \(\varphi_0(x_t, x_{t+1}) = 1\) (idling is the only feasible action for any feasible policy as well as in the recursion in (24)). Otherwise, if an arrival of a class \(i\) customer occurred at time period \(t\) and \(x_{t,i} < B_i\), then \(\varphi_j(x_t, x_{t+1}) = 1\) for all \(j\). If instead a service of class \(i\) was successfully completed under the heuristic policy from period \(t\) to \(t+1\), \(\varphi_j(x_t, x_{t+1}) = 1\) if \(j = i\) and 0 otherwise. Finally, if \(x_{t+1} = x_t\) under the heuristic policy when class \(i\) is served by the heuristic, then \(\varphi_j(x_t, x_{t+1}) = (1 - \Lambda - \nu_j)/(1 - \Lambda - \nu_i)\), where \(\Lambda = \sum_{k : x_{t,k} < B_k} \lambda_k\).

Solving the inner problems in (24) is quite easy: the choice of action \(i\) does not influence the evolution of states in (24) (these are already fixed according to the heuristic policy) and thus we need only deal with a single state in each period in each scenario. Moreover, given that we are selecting actions for the heuristic according to (20), the expectations in the recursions in (24) will have already been calculated at the relevant states for all feasible actions in the scenario as well.

The challenge with the lower bounds using this weak formulation is that there may be many actions in a given scenario in the inner problem with small (in particular, zero) values of \(\varphi_i\). In particular, if the heuristic successfully completes service of a class \(i\) customer in going from \(x_t\) to \(x_{t+1}\), then selecting any other customer class \(j\) in (24) leads to \(\varphi_j = 0\), which would wipe away the remaining tail costs in (24) in that scenario. Nonetheless, we know from Proposition 3.3(i) that the optimal value of the inner problem must satisfy \(V^c_0(x_0) \geq \bar{V}(x_0)\) if \(\bar{V}\) is Bellman feasible.

Another approach is to use the strong SSP formulation and instead calculate lower bounds on the inner problem value in every scenario. We do this by solving Lagrangian relaxations of the strong formulation inner problems: as with the Lagrangian relaxations (21), these relax the constraint that at most one customer class can be served in each time period, but unlike (21), these Lagrangian relaxations are for a deterministic, finite horizon dynamic program. This approach works well with penalties constructed from Lagrangian relaxations (21) of the original problem in that with such penalties the Lagrangian relaxations of the inner problem also decouple by customer class. The Lagrange multipliers in each scenario do not depend on the queue state but may depend on time. Thus in each scenario we are optimizing over \(\tau - 1\) Lagrange multipliers to obtain the largest lower bound. From Corollary 3.1, the optimal value of these inner problem Lagrangian relaxations can be no smaller than the lower bound from (21), despite the additional relaxation. The full details of these inner problem Lagrangian relaxations are discussed in Appendix A.3.
4.4. 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
\[
c_i(x) = c_{1i} x_i + c_{2i} x_i^2,
\]
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.

For all examples, we first calculate lower bounds and approximate value functions using Lagrangian relaxations (21) as well as grouped Lagrangian relaxations (22), with groups of size 2 and groups of size 4. The Lagrangian relaxations (21) represent groups of size 1 and have 16 class-specific value functions, each with 10 states; the groupings 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 (21) 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 scenarios according to the sampling procedure discussed at the start of Section 4.3. We adjust costs using control variates based on each of the four approximate value functions (myopic and the three Lagrangian relaxations). For each approximate value function, in each scenario, we evaluate:

(i) **Heuristic policy.** We select actions as in (20); the (adjusted) expected cost is an upper bound on (19).

(ii) **Perfect information relaxation of weak formulation.** State transitions follow the heuristic policy.

(iii) **Perfect information relaxation of strong formulation with inner problem Lagrangian relaxation.** The inner problem Lagrangian relaxations are solved as linear programs. We only calculate these lower bounds with the myopic approximation and the Lagrangian relaxations with groups of size 1: calculating these bounds using larger groupings is possible but would be more time consuming since we would need to group together classes in the inner problem Lagrangian relaxations (see Appendix A.3).

Note that in evaluating the heuristic policy, it is immaterial whether we take state transitions to be strong form or weak form: for the heuristic, the state transitions are the same in either case, and the factors \( \varphi_{t,i} = 1 \) in all states and times for the heuristic using either form. Thus in each scenario the upper bound values from the heuristic (which we denote by \( \mu \)) are simply given by costs adjusted with control variates, i.e.,
\[
V^\mu_{\theta_i}(x_0) = V(x_0) + \sum_{t=0}^{t-1} \left( \sum_i c_i(x_{t,i}) + \alpha \mathbb{E}[\hat{V}(s(x_t, \mu(x_t), w)) - \hat{V}(x_t)] \right).
\]
### 4.4.1. Runtimes

Table 1 shows the time to calculate the bounds. All calculations are done on a desktop computer (a Dell PC with a 3.07GHz quad-core processor and 12.0GB of RAM) using Matlab 7.12.0 (R2011a), with the Mosek 7.0 Optimization Toolbox used for solving linear programs in the Lagrangian relaxations of the strong formulation inner problems. 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 runtimes 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 one second 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 runtimes could probably be improved.

The runtimes in Table 1 for the other three bounds are the total runtimes for all 1,000 scenarios in the simulations. The time horizons tend to increase with the discount factor, ranging from tens of periods for the $\alpha = 0.9$ case to thousands of periods for the $\alpha = 0.999$ case. The number of calculations required in evaluating the heuristic policy and weak formulation perfect information relaxation scales linearly in the time horizon, and this is evident in the runtimes. These bounds are calculated relatively quickly: the weak formulation calculations take around one second total in the $\alpha = 0.9$ case and around a minute total in the $\alpha = 0.999$ case. The linear programs for the strong formulation Lagrangian relaxations require much more time to calculate than the weak formulation inner problems and these runtimes scale more than linearly with the time horizon; the times for the strong formulation calculations for all 1,000 scenarios range from about 30 seconds when $\alpha = 0.9$ to about 10 hours when $\alpha = 0.999$.

The runtimes for the weak formulation information relaxations and heuristic policy are relatively insensi-

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Myopic</th>
<th>Lagrangian Relaxations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Groups of Size 1</td>
<td>Groups of Size 2</td>
</tr>
<tr>
<td>$\alpha = 0.9$</td>
<td>Approximate Value Function</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>WF with Perfect Information</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>SF+LR with Perfect Information</td>
<td>27.2</td>
</tr>
<tr>
<td></td>
<td>Heuristic Policy</td>
<td>1.6</td>
</tr>
<tr>
<td>$\alpha = 0.99$</td>
<td>Approximate Value Function</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WF with Perfect Information</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>SF+LR with Perfect Information</td>
<td>869.0</td>
</tr>
<tr>
<td></td>
<td>Heuristic Policy</td>
<td>18.6</td>
</tr>
<tr>
<td>$\alpha = 0.999$</td>
<td>Approximate Value Function</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WF with Perfect Information</td>
<td>51.8</td>
</tr>
<tr>
<td></td>
<td>SF+LR with Perfect Information</td>
<td>26,498.0</td>
</tr>
<tr>
<td></td>
<td>Heuristic Policy</td>
<td>204.1</td>
</tr>
</tbody>
</table>

Table 1: Runtimes (seconds) for multiclass queue examples.
### Table 2: Bounds for multiclass queue examples.

There are some differences in runtimes across the columns for these calculations, which reflect differences in time required to iterate over the groups in the penalty and control variate calculations. These differences are specific to our implementation.

Finally, as will be evident in the mean standard errors in the next section, 1,000 scenarios in these examples is far more than necessary to obtain relatively precise bound estimates, at least for the bounds using the Lagrangian relaxations as control variates. Thus we view these runtimes as quite conservative.

#### 4.4.2. Results

Table 2 shows the results for all bounds. The bound values in these results are all scaled by the factor $1 - \alpha$. This is a standard scaling that facilitates comparisons across the discount factors in that $1/(1 - \alpha)$ is the expected number of time periods. The estimated bounds are shown in Table 2, with mean standard errors in parentheses. For each discount factor, the top row indicates the value of each of the approximate value function to the group sizes.
functions in the initial state; these approximate value functions are Bellman feasible and are therefore lower bounds. For the lower bounds from the information relaxations, we also report the gap to the upper bound for the heuristic policy, the mean standard error for this gap, and the relative gap in percent (by relative gap, we mean the gap divided by the heuristic policy upper bound). For each discount factor and approximate value function, we highlight in bold the smallest relative gap, and underline the best relative gap overall for each discount factor.

The bounds using the myopic approximate value function \(\bar{V}(x) = \sum_i c_i(x_i)\) 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. The information relaxation bounds (weak and strong formulation) improve on this lower bound, as they 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 mean standard errors. Given that the myopic value function is a relatively bad approximation of the optimal value function, it serves as a relatively bad control variate as well; this in turn leads to relatively large mean standard errors for the simulated bounds. The myopic results thus underscore the importance of having relatively good approximations to the value function in both the lower bound and upper bound calculations.

The Lagrangian relaxations provide much better lower bounds. These lower bounds get increasingly better with larger group sizes and fare better as the discount factor increases. The relative gaps for these lower bounds range from about 10\% (\(\alpha = 0.9\) with groups of size 1) to 1.42\% (\(\alpha = 0.999\) 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 weak formulation lower bounds improve upon the Lagrangian relaxation bounds, as they must by Proposition 3.3(i), with the amount of improvement decreasing as \(\alpha\) 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 to serve a class \(i\) with an associated value of \(\varphi_i\) of zero; such scenarios may result in relatively modest improvements on the original Lagrangian relaxation. Nonetheless, the relative gaps for the weak formulation bounds can be considerably smaller than those from the Lagrangian relaxations: in the \(\alpha = 0.9\) case, for instance, the relative gaps are a factor of 4 to 6 smaller than the relative gaps from the Lagrangian relaxations.

The strong formulation lower bounds (with Lagrangian relaxation) appear to be somewhat better than the corresponding weak formulation lower bounds, and, like the weak formulation lower bounds, are guaranteed to outperform the Lagrangian relaxation lower bounds. Also like the weak formulation lower bounds, the improvement over the Lagrangian relaxation lower bounds tends to diminish as the discount factor increases. This should not be surprising: larger \(\alpha\) means longer time horizons, which leads to many more possible downstream states in the inner problems, and thus the Lagrangian relaxation of the inner problem becomes an increasingly weaker relaxation. We suspect that if possible, exact solutions of the strong formulation
inner problems would lead to very small gaps (at the very least, this would not be any worse than the strong formulation results in Table 2).

Finally, Figure 3 shows sample gap values (relative to the heuristic value; in these plots, 0.1 means 10%, etc.) for some of the weak and strong formulation lower bounds (specifically, the weak and strong formulation bounds with Lagrangian relaxation with groups of size 1, as well as the weak formulation bounds with the Lagrangian relaxation with groups of size 4). These sample gap values are nonnegative in every scenario, as they must be by Proposition 3.3(iv), and it is apparent from Figure 3 that longer time horizons tend to lead to larger gap values. The intuition for this is that longer time horizons generally lead to larger costs for the heuristic policy, but, as explained above, may have much less effect on the information relaxation bounds.

Common scenarios are used for all the bound calculations for each discount factor, so we can also make comparisons between the sample gap values for the different bounds in Figure 3. With the same control variate (Lagrangian relaxations with groups of size 1), the strong formulation bounds generally perform a bit better (smaller gaps) than the weak formulation bounds, although there is not much visible difference in the $\alpha = 0.999$ case. Between the two weak formulation bounds, those with the groups of size 4 Lagrangian relaxation tend to perform better, as expected (the exception is the $\alpha = 0.9$ case, for which these two lower bounds are about the same). Also noteworthy are the gap values in the case of the weak formulation bounds with groups of size 4 and $\alpha = 0.99$: even for long time horizons in the plot, the gap values are close to zero. This suggests the approximate value function is very close to the optimal value function in this case; recall from Proposition 3.2(ii) that with the optimal value function, the gap values would be zero in all scenarios.

4.5. Quasi-Strong Formulations and an Imperfect Information Relaxation

Perfect information bounds using weak 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 $\phi_i$ in many scenarios. Strong formulations avoid this and thus may lead to better lower bounds, but in these multiclass queueing examples the strong formulation inner problems are very difficult to solve (although we did solve their Lagrangian relaxations). We now discuss reformulations that may be viewed as a compromise between strong and weak formulations. Although the best information relaxation bounds in Table 2 are already probably good enough, these quasi-strong
formulations provide even better bounds and illustrate the flexibility of the approach.

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 \( \bar{S} := \{1, \ldots, I\} \setminus S \), 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 a strong formulation, and when \( S = \emptyset \), we recover a weak formulation in which the server always idles.

This reformulation, just like the weak formulation above, does not satisfy the absolute continuity condition in Proposition 2.3(i) but we again obtain lower bounds on \( V(x_o) \) by using Bellman feasible approximate value functions within the control variates. As above, we take \( \tau \) to be geometric with parameter \( 1 - \alpha \).

With this reformulation, customers from a class in \( \bar{S} \) may arrive but can never be served, and when used with a perfect information relaxation, the states of all customers in \( \bar{S} \) in each period of each scenario are fixed; although we could make states on \( \bar{S} \) fixed with other reformulations, this reformulation also leads to no zero values of \( \varphi_i \), as we discuss shortly. In contrast, we do need to account for the different states that are possible for the class \( S \) customers due to service decisions for those classes. The inner problems (12) are then deterministic DPs with \( \prod_{i \in S} (B_i + 1) \) states in each period, with class \( \bar{S} \) states fixed in each scenario. 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, \( w_t \) 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 \( \hat{s} \)), 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 \( \lambda_i/\Gamma_S \) (resp., \( \mu_i/\Gamma_S \)), where \( \Gamma_S := \sum_{j \in S} (\lambda_j + \mu_j) \).

The inner problems (12) in this case are now stochastic DPs with \( \prod_{i \in S} (B_i + 1) \) states in each period, using 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 \( \varphi_i(x_t, \tilde{x}_{t+1}) \) (\( = \varphi_i(x_t, \hat{s}(x_t, i, w_t)) \)) now satisfy \( \varphi_i = 1 \) for all \( i \in S \); for \( i \in \bar{S} \), we have \( \varphi_i(x_t, \tilde{x}_{t+1}) = (1 - \mu_i - \Lambda)/(1 - \Lambda) \) if \( \tilde{x}_{t+1} = x_t \) and \( x_{t,i} > 0 \), and \( \varphi_i = 1 \) otherwise, where \( \Lambda \) is as in Section 4.3.

Table 3 shows the results for this approach on the same examples from Section 4.4. These results correspond to the same 1,000 scenarios for each discount factor, and we use the Lagrangian relaxations with groups of size 4 for control variates. The set of classes \( S \) for each of the three discount factors is the first group of size 4 in the grouped Lagrangian relaxation (recall that the Lagrangian relaxation groups were set by ranking classes according to how often they were served by the heuristic (20) using the Lagrangian relaxation with individual classes as groups). Thus, the inner problems using this quasi-strong formulation (QSF) have \( 10,000 \) states in each period. Table 3 also lists the relative gaps for these lower bounds and also restates for comparison the relative gaps from the weak formulation and the Lagrangian relaxation with groups of size 4. These lower bounds must be better than the Lagrangian relaxation and, as we would expect.
Table 3: Imperfect information relaxation bounds for multiclass queue examples using quasi-strong formulation.

(although this is likely not a general result - see Appendix B.3), are also better than the lower bounds from the weak formulation, with a more marked improvement in the $\alpha = 0.9$ and $\alpha = 0.99$ cases. The runtimes (99 seconds, about 20 minutes, and about 3.75 hours, respectively) are substantially longer than for the weak formulations, although these are not unreasonable runtimes given the complexity of the full DPs and given that 1,000 scenarios is evidently far more than necessary with such low MSEs. Overall, these quasi-strong formulations illustrate that judiciously chosen “middle ground” reformulations are likely to provide better lower bounds than the lower bounds based on weak formulations.

4.6. Summary

In summary, we are able to obtain relatively small optimality gaps in these examples. For each of the three discount factors, the best gaps from the perfect information relaxation lower bounds in Table 2 are 1.31%, 0.61%, and 1.27%. By comparison, by only using Lagrangian relaxation lower bounds, the best gaps are 7.58%, 2.12%, and 1.42%. Moreover, the information relaxation bounds that are best in each case do not require much additional time to evaluate (26.8 seconds, 6.5 seconds, and 63.8 seconds, respectively, for all 1,000 scenarios). We then discussed how to use quasi-strong formulations and imperfect information relaxations to improve these lower bounds (gaps of 0.37%, 0.38%, and 1.17%), albeit at the expense of longer runtimes. There are of course many possible variations of this model (e.g., multiple servers, network models, stochastic arrival and service rates), and it would be interesting to see how the approach fares on some of these variations.

5. Conclusion

In this paper, we have shown how to use information relaxations to calculate performance bounds for infinite horizon stochastic DPs with discounted costs. The 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. Moreover, the method is guaranteed to provide improved lower bounds when the control variates are built from Bellman feasible approximate value functions. We have applied the method to a challenging multiclass queueing application with encouraging results; to our knowledge, this is the first proof of concept that the “weak formulations” in particular can in fact be effective on a large-scale application.

Moving forward, there are a number of interesting research directions. First, the framework provides considerable freedom in reformulating the state transition function, and we have only scratched the surface here. On a related note, it may be possible in some problems to optimize over the reformulation parameters
(e.g., reformulated state transition probabilities) to yield the best lower bound. It may even be possible to jointly optimize over these parameters and parameterized control variates. Finally, average cost problems remain a challenge: we know in principle how to apply the method to average cost problems (see App. B.4), but there may be other ways to approach these problems, and a successful implementation of information relaxations on a challenging average cost application has yet to be done.

Acknowledgements
We thank an anonymous associate editor, two anonymous referees, Jim Smith, and Peng Sun for many helpful comments on earlier versions of the paper.

References


A. Proofs and Detailed Derivations

A.1. Discussion of Standard Primal DP Results

The setup of the primal DP in Sec. 2.1 is a special case of the “infinite horizon Borel model (D)” (discounting and bounded costs) in chapter 9 of Bertsekas and Shreve (1996). Specifically, this model requires the following (we augment with notation from Bertsekas and Shreve 1996 as necessary):

1. $X$ and $U$ are nonempty Borel spaces.
2. The set \{(x, u) : x \in X, u \in U(x)\} is analytic (Definition 7.16 in Bertsekas and Shreve 1996) in $X \times U$.
3. The disturbance space $(W)$ is a nonempty Borel space, and the disturbance kernel $(p(\text{d}w|x,u))$ is a Borel-measurable stochastic kernel on $W$ given $X \times U$.
4. $s$ is Borel-measurable from $X \times U \times W$ to $X$.
5. $c$ is a lower semianalytic function (Definition 7.21 in Bertsekas and Shreve 1996) and uniformly bounded on $X \times U$.
6. $\alpha \in (0, 1)$.

For the primal DP in Section 2.1, we are assuming conditions (i), (iv), and (vi) directly. Condition (ii) follows from the fact that we assume the set \{(x, u) : x \in X, u \in U(x)\} is a Borel set, which implies it is also an analytic set (Proposition 7.36 in Bertsekas and Shreve 1996). Condition (iii) follows since we take disturbances $w_t$ to be real-valued random variables on a Borel space; note that Bertsekas and Shreve (1996) allow for state and action dependence in the transition kernel, whereas we incorporate all such dependence entirely within the state transition function $s$. The condition (v) that $c$ be lower semianalytic is implied by our assumption that $c$ is Borel-measurable and Lemma 7.30 in Bertsekas and Shreve (1996).

Statement (a) in Section 2.1 then follows from the discussion on p. 215 of Bertsekas and Shreve (1996) (for Borel-measurability of $V_\alpha$) and Proposition 9.9 in Bertsekas and Shreve (1996). Statement (b) is Proposition 9.8 in Bertsekas and Shreve (1996). Finally, Corollary 9.12.1 in Bertsekas and Shreve (1996) asserts that there exists an optimal policy that is stationary if and only if the infimum in the right-hand side of (1) is attained for every state.

It should be noted that the definition of policies in Bertsekas and Shreve (1996) is somewhat broader in that their policies are only assumed to be “universally measurable” (Definition 7.20 in Bertsekas and Shreve 1996), which is weaker than Borel-measurability. It may be the case that an optimal stationary policy is Borel-measurable; see, for instance, Section 9.5 of Bertsekas and Shreve (1996). Finally, all examples above involve finite state and action spaces and thus fit into the framework assumed in Section 2.1.

A.2. Proofs

Proof of Proposition 2.2. For any $\mu \in \mathcal{U}$,

$$
\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \left(c(x_t, \mu(x_t)) + v(x_t, \mu(x_t))\right)\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \left(c(x_t, \mu(x_t)) + \alpha \mathbb{E}[V(s(x_t, u_t, w)_t)] - \bar{V}(x_t)\right)\right] \\
= V_\mu(x_0) + \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \left(\alpha \mathbb{E}[V(s(x_t, u_t, w))] - \alpha \bar{V}(x_{t+1})\right)\right] - \bar{V}(x_0) \\
= V_\mu(x_0) + \sum_{t=0}^{\infty} \alpha^t \mathbb{E}\left[\alpha \mathbb{E}[V(s(x_t, u_t, w))] - \alpha \bar{V}(x_{t+1})\right] - \bar{V}(x_0) \\
= V_\mu(x_0) - \bar{V}(x_0),
$$

34
where the interchange of summation and expectation follows from the bounded convergence theorem and we use the fact that, for \( \mu \in \mathcal{U} \), \( \mathbb{E}[\alpha \mathbb{E}[\bar{V}(s(x_t, \mu(x_t), w))] - \alpha \bar{V}(x_{t+1})] = 0 \). Adding \( \bar{V}(x_o) \), we obtain (4). \( \square \)

**Proof of Proposition 2.3.** For Proposition 2.3(i), the first key step in the argument involves rewriting \( V_\mu(x_0) \) in an equivalent way in which costs are not discounted but there is a random absorption time \( \tau \) that is independent of actions and geometric with parameter \( 1 - \alpha \) (and thus almost surely finite). This follows from Proposition 5.3.1 in Puterman (1994). Note that although the treatment in Chapter 5 of Puterman (1994) assumes a countable state space, Proposition 5.3.1 in Puterman (1994) holds more generally and the argument only requires bounded costs and a discount factor strictly less than one. The remaining steps of the proof essentially adapt the arguments in pp. 257-258 of Glasserman (2004) to the setting here.

Specifically, we have

\[
V_\mu(x_0) - \bar{V}(x_0) \quad \overset{(a)}{=} \quad \mathbb{E}[c(\mu) + v(\mu)] \\
\overset{(b)}{=} \quad \mathbb{E} \left[ \sum_{t=0}^{\tau-1} c(x_t, \mu(x_t)) + v(x_t, \mu(x_t)) \right] \\
\overset{(c)}{=} \quad \mathbb{E} \left[ \left( \sum_{t=0}^{\tau-1} c(x_t, \mu(x_t)) + v(x_t, \mu(x_t)) \right) 1 \{ \tau < \infty \} \right] \\
\overset{(d)}{=} \quad \sum_{\tau'=1}^{\infty} \mathbb{E} \left[ \left( \sum_{t=0}^{\tau'-1} c(x_t, \mu(x_t)) + v(x_t, \mu(x_t)) \right) 1 \{ \tau = \tau' \} \right] \\
\overset{(e)}{=} \quad \sum_{\tau'=1}^{\infty} \mathbb{E} \left[ \left( \sum_{t=0}^{\tau'-1} \alpha^t \Phi_t(\mu)(c(x_t, \mu(x_t)) + v(x_t, \mu(x_t))) \right) 1 \{ \tau = \tau' \} \right] \\
\overset{(f)}{=} \quad \mathbb{E} \left[ \left( \sum_{t=0}^{\tau-1} \alpha^t \Phi_t(\mu)(c(x_t, \mu(x_t)) + v(x_t, \mu(x_t))) \right) 1 \{ \tau < \infty \} \right] \\
\overset{(g)}{=} \quad \mathbb{E}[\tilde{c}_\tau(\mu)] - \bar{V}(x_0).
\]

Equality (a) follows by Proposition 2.2. Equality (b) follows by Proposition 5.3.1 in Puterman (1994), where \( \tau \) in the expression on the right is geometric with parameter \( 1 - \alpha \); note that in the expression on the right of (b), we are taking the state transitions to be absorption with probability \( 1 - \alpha \) from any other state and for any feasible action, but otherwise the state transition stochastic kernel follows \( P_\mu \). Equality (c) follows since costs (and control variates) are bounded and \( \tau \) has finite mean. Equality (d) follows from the law of total expectations. Equality (e) follows by application of the Radon-Nikodym theorem to each term in the sum in (d), which is justified since \( P_\mu \) is absolutely continuous with respect to \( Q_{1,\mu} \). Note that the factor of \( \alpha^t \) arises in the change of measure terms because \( \Phi_t \) is defined relative to the original state transition kernel \( P_\mu \), which transitions to non-absorbing states with probability one, whereas the reformulation on the right side of (d) only has probability \( \alpha \) of a transition to non-absorbing states. Equality (f) follows again by the law of total expectations, and equality (g) follows by our assumption that \( \tau \) is almost surely finite when state transitions follow \( s_t \) and the definition of \( \tilde{c}_\tau(\mu) \).

For Proposition 2.3(ii), we first observe that if \( \bar{V} \) is Bellman feasible, then \( c(x, u) + v(x, u) \geq 0 \) for all \( x \in X \), \( u \in U(x) \); this follows from the definition of \( v(x, u) \) and the fact that Bellman feasibility of \( \bar{V} \) requires \( \bar{V}(x) \leq c(x, u) + \alpha \mathbb{E}[\bar{V}(s(x, u, w))] \) for all \( x \in X \), \( u \in U(x) \).

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

**Lemma A.1.** Consider a measure space \( (\Omega, \Sigma) \) with two probability measures \( P \) and \( Q \), and assume \( Q \) is concentrated on \( A_Q \in \Sigma \). Let \( \varphi \) 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 "Theorem of Lebesgue-Radon-Nikodym" (see, e.g., Thm. 6.10 in Rudin 1987) we can uniquely decompose \( P \) as \( P = P^a + P^c \), where \( P^a \) is absolutely continuous with respect to \( Q \) with unique density (or Radon-Nikodym derivative) \( \varphi \), and \( P^c \) is concentrated on a set \( A' \) such that \( A_P \cap A_Q = \emptyset \).
We then have
\[
\mathbb{E}^\pi [Y] = \int_{\omega \in \Omega} Y(\omega) dP(\omega) = \int_{\omega \in \Omega} Y(\omega) dP^\pi(\omega) + \int_{\omega \in \Omega} Y(\omega) dP^\nu(\omega) \\
\geq \int_{\omega \in \Omega} Y(\omega) dP^\nu(\omega) = \int_{\omega \in \Omega} \varphi(\omega) Y(\omega) dQ(\omega) = \mathbb{E}^Q [\varphi Y].
\]

The second equality follows from the Lebesgue decomposition theorem referenced above. The third equality follows from the fact that $P^\pi$ is concentrated on $A_{P^\nu}$ 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 $A_{P^\nu} \cap A_Q = \emptyset$, and thus $\omega \notin A_Q$. The second-to-last equality follows from absolute continuity of $P^\nu$ with respect to $Q$ and the Radon-Nikodym theorem, noting that $Y$ is bounded and therefore integrable under $P^\nu$.

To complete the proof of Proposition 2.3(ii), the steps above in the proof of Proposition 2.3(i) all hold, with the exception that equality ($e$) becomes an inequality: this follows from the fact (as noted above) that Bellman feasibility of $\hat{\overline{V}}$ implies $c(x, u) + v(x, u) \geq 0$ in all feasible state-action pairs and by application of Lemma A.1 to each term in the sum on the left side of ($e$), which completes the proof.

It is perhaps worth noting that nonnegativity of the terms $c(x, u) + v(x, u)$ in all feasible state-action pairs is more than is needed to apply Lemma A.1: the result holds more generally, provided $c(\hat{x}_t, \mu(\hat{x}_t)) + v(\hat{x}_t, \mu(\hat{x}_t)) \geq 0$ for all states $\hat{x}_t$ that occur with zero probability under the state process induced by $Q_{t, \omega}$.

**Proof of Proposition 3.1.** Note that any policy $\mu \in \mathcal{U}$ is stationary and therefore $\mathcal{F}$-adapted, i.e., $\mu \in \mathcal{U}_\mathcal{F}$. The inequality then follows by Proposition 2.3 and the fact that the right-hand side takes the infimum over $\mathcal{U}_\mathcal{G}$, which includes $\mathcal{U}_\mathcal{F}$ and therefore includes $\mu$. The final statement follows by the same argument noting that for any policy $\mu^* \in \mathcal{U}$ that is optimal, $V_{\mu^*}(x_0) = V(x_0)$ (a formal derivation of this fact follows from Propositions 9.8 and 9.9 in Bertsekas and Shreve 1996).

**Proof of Proposition 3.2.** For part (i), if $\mathcal{G}$ includes $\tau$, then in the information relaxation, we need only choose actions for time periods $0$ through $\tau - 1$, and thus the problem is a finite horizon problem. The goal is to select a $G$-adapted policy so as to minimize the expected costs of the overall cost function
\[
\hat{c}_\tau(\mathbf{u}_\tau) = \hat{V}(x_0) + \sum_{t=0}^{\tau-1} \alpha^t \Phi_t(\mathbf{u}_t)(c(\hat{x}_t, u_t) + \alpha \mathbb{E}(V(s(\hat{x}_t, u_t, w)))) - \hat{V}(\hat{x}_t),
\]
where $\mathbf{u}_t := (u_0, \ldots, u_{t-1})$. We can solve this problem over $\mathcal{G}$-adapted policies recursively\(^3\) as
\[
\hat{V}_\tau(x_0, \mathbf{u}_\tau) = \inf_{u_t \in \mathcal{U}(x_t)} \left\{ \alpha^t \Phi_t(\mathbf{u}_t)(c(x_t, u_t) + \alpha \mathbb{E}(V(s(x_t, u_t, w)))) - \hat{V}(x_t) + \mathbb{E}_{\mathcal{G}_t}[\hat{V}_{\tau+1}(\hat{x}_{t+1}, \mathbf{u}_{t+1})] \right\},
\]
with $\hat{V}_\tau(x_0, \mathbf{u}_\tau) := 0$. Notice that $\hat{V}_\tau$ depends on past actions $(u_0, \ldots, u_{t-1})$, as these affect the value of the density terms $\Phi_t(\mathbf{u}_t)$. Solving this recursion provides the optimal value of the $G$-adapted policy conditional on $\mathcal{G}_0$, solving this recursion conditional on $\mathcal{G}_0$ (which includes $\tau$ by assumption) and averaging gives the

\(^3\)Note that (25) involves the sum of costs over all time periods, and thus we use $\hat{x}_t$ to emphasize that state transitions evolve according to $\hat{s}_t$, whereas the recursion (26) is evaluated at a given initial state; for consistency with (12) we write this initial state as $x_t$. It should be understood, however, that in both (25) and (26), states evolve according to $\hat{s}_t$. Finally, recall that the terms $\mathbb{E}(V(\cdot))$ from the control variates are by definition over the original state transition function $s$.

36
optimal value on the right-hand side of (10), i.e.,
\[
\inf_{\mu \in \mathcal{U}_t} \mathbb{E}[\tilde{c}_t(\mu_t)] = \mathbb{E}[\tilde{V}_0^\circ(x_0)] + \tilde{V}(x_0),
\]
where we account for the additional term \( \tilde{V}(x_0) \) in the definition of \( \tilde{c}_t \).

We now show that solving this recursion over \( \tilde{V}_t^\circ \) is equivalent to solving the recursion (12). To see this, we claim the relationship
\[
\tilde{V}_t^\circ(x_t, u_t) = \alpha^t \Phi_t(u_t) \left( V_t^\circ(x_t) - \tilde{V}(x_t) \right)
\]
holds for all \( x_t, u_t, \mathcal{G}_t \), and \( t = 0, \ldots, \tau \), where \( V_t^\circ \) is as defined in (12). To see this, we argue by induction. The base case is trivial as \( \tilde{V}_0^\circ = 0 \), \( V_0^\circ = 0 \), and \( \tilde{V}(x) = \tilde{V}(x) = 0 \) by assumption. We now assume (28) holds at \( t + 1 \). We then have
\[
\begin{align*}
\hat{V}_t^\circ(x_t, u_t) &= \inf_{u_t \in U(x_t)} \{ \alpha^t \Phi_t(u_t) \{ c(x_t, u_t) + \alpha \mathbb{E}[\tilde{V}(s(x_t, u_t, w))] - \tilde{V}(x_t) \} + \mathbb{E}_{\mathcal{G}_t}[\hat{V}_{t+1}^\circ(\bar{x}_{t+1}, u_{t+1})] \} \\
&= \inf_{u_t \in U(x_t)} \{ \alpha^t \Phi_t(u_t) \{ c(x_t, u_t) + \alpha \mathbb{E}[\tilde{V}(s(x_t, u_t, w))] - \tilde{V}(x_t) \} + \mathbb{E}_{\mathcal{G}_t}[\alpha^{t+1} \Phi_{t+1}(u_{t+1})(V_{t+1}^\circ(\bar{x}_{t+1}) - \tilde{V}(\bar{x}_{t+1}))] \} \\
&= \alpha^t \Phi_t(u_t) \{ c(x_t, u_t) + \alpha \mathbb{E}[\tilde{V}(s(x_t, u_t, w))] - \tilde{V}(x_t) \} + \alpha \mathbb{E}_{\mathcal{G}_t}[\varphi_{t,u}(x_t, \bar{x}_{t+1})(V_{t+1}^\circ(\bar{x}_{t+1}) - \tilde{V}(\bar{x}_{t+1}))] \\
&= \alpha^t \Phi_t(u_t) \left( \hat{V}_t^\circ(x_t) - \tilde{V}(x_t) \right),
\end{align*}
\]
which verifies the induction assumption (28). The first equality is the definition of \( \hat{V}_t^\circ \) from above, the second equality uses the fact that \( \Phi_{t+1}(u_{t+1}) = \Phi_t(u_t) \cdot \varphi_{t,u}, \) and the fourth equality uses the definition of \( V_t^\circ \) from (12). Thus, by solving for \( \hat{V}_0^\circ(x_0) \) in the recursion in (12), we obtain \( \hat{V}_0^\circ(x_0) = V_0^\circ(x_0) - \tilde{V}(x_0) \), which, through (27), yields the right-hand side of the weak duality inequality (10).

Part (ii) of the result follows by using part (i) and a similar induction. If \( \hat{V} = V \), then \( \hat{V}(x_t) = \inf_{u_t \in U(x_t)} \{ c(x_t, u_t) + \alpha \mathbb{E}[\hat{V}(s(x_t, u, w))] \} \) in all states. Clearly then, \( V_{t-1}^\circ = \hat{V} = V \) in all states. Assuming that \( V_{t-1}^\circ = \hat{V} = \tilde{V} = \tilde{V} = V \) as well, and the result follows.

\section*{Proof of Proposition 3.3} We show part (i) by induction on \( t \) with a fixed \( \tau \). For \( t = \tau - 1 \), the right-hand side of (13) reduces to \( V(x_{t-1}) + \epsilon \). Using (12), we have
\[
\begin{align*}
V_{t-1}^\circ(x_{t-1}) &= \inf_{u_{t-1} \in U(x_{t-1})} \{ c(x_{t-1}, v_{x_{t-1}}) + \alpha \mathbb{E}[\hat{V}(s(x_{t-1}, u, w))] \} \geq V(x_{t-1}) + \epsilon,
\end{align*}
\]
where we use \( \epsilon \)-Bellman feasibility of \( \hat{V} \). This shows the base case. Now assume the result is true for \( t+1 < \tau \), i.e., assume \( V_{t+1}^\circ(x_{t+1}) \geq \tilde{V}(x_{t+1}) + \kappa_{t+1} \epsilon \). Then:
\[
\begin{align*}
V_t^\circ(x_t) &= \inf_{u_t \in U(x_t)} \{ c(x_t, u) + \alpha \mathbb{E}[\hat{V}(s(x_t, u, w))] \} + \alpha \mathbb{E}_{\mathcal{G}_t}[\varphi_{t,u}(x_t, \bar{x}_{t+1})(V_{t+1}^\circ(\bar{x}_{t+1}) - \tilde{V}(\bar{x}_{t+1}))] \\
&\geq \inf_{u_t \in U(x_t)} \{ c(x_t, u) + \alpha \mathbb{E}[\hat{V}(s(x_t, u, w))] \} + \alpha \mathbb{E}_{\mathcal{G}_t}[\varphi_{t,u}(x_t, \bar{x}_{t+1})] \cdot \epsilon \\
&\geq \inf_{u_t \in U(x_t)} \{ c(x_t, u) + \alpha \mathbb{E}[\hat{V}(s(x_t, u, w))] \} + \alpha \mathbb{E}_{\mathcal{G}_t}[\varphi_{t,u}(x_t, \bar{x}_{t+1})] \cdot \epsilon \\
&\geq \hat{V}(x_t) + \epsilon + \alpha \kappa_{t+1} \gamma_t \epsilon \\
&= \hat{V}(x_t) + \kappa_t \epsilon.
\end{align*}
\]
The first equality uses (12). The first inequality uses the induction assumption and nonnegativity of \( \varphi_{t,u} \) and \( \epsilon \). The second inequality uses nonnegativity of \( \kappa_{t+1} \) and \( \epsilon \), the fact that we are minimizing, and the definition of \( \gamma_t \). The final inequality uses \( \epsilon \)-Bellman feasibility of \( \hat{V} \) and the equality follows from the definition of \( \kappa_t \). This completes the proof of part (i).

For part (ii), \( \gamma_t \) does not depend on time, and the expression for \( \kappa_t \) follows directly from evaluating the
where the first inequality uses the fact that there is zero probability of a transition to \( x^n \) under \( s \), the first inequality uses Lemma A.1 (with the random variable \( Y = 1 \)), and the second inequality uses the definition of \( \gamma; \gamma < 1/\beta \) then follows from \( \alpha < 1 \). We then have

\[
\begin{align*}
\mathbb{E} \left[ \frac{1 - \gamma^r - t}{1 - \gamma} \right] &= \frac{1}{1 - \gamma} \mathbb{E}[1 - \gamma^r] \\
&= \frac{1}{1 - \gamma} \left( 1 - \frac{(1 - \beta)\gamma}{1 - \beta\gamma} \right) \\
&= \frac{1}{1 - \beta\gamma} \\
&= \frac{1}{1 - \alpha\delta}.
\end{align*}
\]

The first equality follows from the fact that the distribution of \( \tau \) is memoryless. The second equality follows by evaluating the moment generating function for \( \tau \), which converges since \( \gamma < 1/\beta \). The next equality follows from basic algebra. The final equality follows by noting that

\[
\gamma = \alpha \inf_{\{u, x \neq x^n, y \neq x^n\}} \phi_{t,u}(x,y) = (\alpha/\beta) \inf_{\{u, x \neq x^n, y \neq x^n\}} dP_u(x,y) / dQ_u(x,y) = (\alpha/\beta)\delta.
\]

The result then follows from (13).

Part (iii) is straightforward simply by observing that in this case \( \phi_{t,u}(x,y) = 1/\alpha \) for all times, actions, and all non-absorbing state pairs; this implies \( \gamma = 1 \) and \( \delta = 1 \). The proof of part (i) goes through for any \( \epsilon \) in this case, as there is no need to bound \( \phi_{t,u} \) from below in the induction step.

For part (iv), (16) follows from the fact that \( \mu \in \mathcal{U} \) is feasible to (12) in all states and scenarios. That \( \mathbb{E}[V_{t,\mu}(x_1)] \leq V_{\mu}(x_1) \) follows from iterated expectations and Proposition 2.3, noting that \( \tilde{V} \) is Bellman feasible if \( \epsilon \geq 0 \).

**Proof of Proposition 3.4.** By construction \( \tau \) is included in both models, since a transition to \( x^n \) occurs with probability one at \( T_1 \) in one model and \( T_2 \) in the other model (and never earlier). Thus under \( F \), as well as under the relaxation \( G \), \( \tau \) is known from time zero in either case. Now fix a scenario \( (w_1, \ldots, w_{T_1}) \) (note that in the \( T_1 \) model, absorption occurs at time \( T_1 \) so we only need \( (w_1, \ldots, w_{T_1}) \) in evaluating the inner problem in that case). We will show (17) by induction.

Time \( T_1 - 1 \) is the base case. Consider any state \( x_{T_1-1} \) at \( T_1 - 1 \). Then by (12)

\[
\begin{align*}
V_{T_1-1}^{\leq}(x_{T_1-1}) &\geq \inf_{u \in U(x_{T_1-1})} \left\{ c(x_{T_1-1}, u) + \alpha \mathbb{E}[\tilde{V}(s(x_{T_1-1}, u, w))] + \alpha \mathbb{E}_{\tilde{X}_{T_1}}[\phi_{T_1-1,u}(x_{T_1-1}, \tilde{x}_{T_1})(V_{T_1-1}^{\leq}(\tilde{x}_{T_1}) - \tilde{V}(\tilde{x}_{T_1}))] \right\} \\
&= V_{T_1-1}^{\leq}(x_{T_1-1}).
\end{align*}
\]

The first equality is the definition of \( V_{T_1-1}^{\leq} \). The inequality follows by Bellman feasibility of \( \tilde{V} \), Proposition
3.3(i), and nonnegativity of $\varphi_{t,u}$. The second equality follows from the fact that, under $\hat{s}_1$, absorption occurs with probability one at $T_1$, and the definition of $V_{39}^{(2)}$.

Now assume the result holds for time $t + 1 \leq T_1 - 1$, i.e., $V_{t+1}^{(2)} \geq V_{t+1}^{(1)}$ in all states. Then

$$V_{t}^{(2)}(x_t) = \inf_{u \in U(x_t)} \left\{ \sigma(x_t, u) + \alpha E[\hat{V}(s(x_t, u, w))] + \alpha E_{x_t}[\varphi_{x_{t}}(x_t, \hat{x}_{t+1})(V_{t+1}(\hat{x}_{t+1}) - \hat{V}(\hat{x}_{t+1}))] \right\},$$

$$\geq \inf_{u \in U(x_t)} \left\{ \sigma(x_t, u) + \alpha E[\hat{V}(s(x_t, u, w))] + \alpha E_{x_t}[\varphi_{x_{t}}(x_t, \hat{x}_{t+1})(V_{t+1}(\hat{x}_{t+1}) - \hat{V}(\hat{x}_{t+1}))] \right\},$$

$$= V_{t}^{(1)}(x_t).$$

The first and final equalities use the definitions of $V^{(2)}$ and $V^{(1)}$. The inequality follows by the induction assumption and the fact that the state transitions are identical for time periods up to $T_1 - 1$. This completes the proof of (17).

We note that Proposition 3.3 provides more than $V_{t+1}^{(2)} \geq \hat{V}$ if $\hat{V}$ is $\epsilon$-Bellman feasible with $\epsilon > 0$, so we could strengthen (17) in that case, although additional notation would be required.

### A.3. Lagrangian Relaxations of Strong Form Inner Problems for the Multiclass Queueing Application

We show how the strong (SSP) form inner problems with perfect information decouple with Lagrangian relaxations of strong form inner problems for the multiclass queueing application, as discussed in Section 4.3. We consider a fixed scenario $(w_1, \ldots, w_{\tau-1})$. The inner problems given in (12) in this case have the form

$$V_t(x_t) = \min_{a \in A(x)} \left\{ \sum_i c_i(x_t, i) + \alpha E[\hat{V}(s(x_t, a, w))] - \hat{V}(s(x_t, a, w_{t+1})) + V_{t+1}(s(x_t, a, w_{t+1})) \right\},$$

for $t = 0, \ldots, \tau - 1$, where $V_x = 0$ and we have dropped the superscript $G$ in (12) for ease of notation. We will consider Lagrangian relaxations of (29) 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 (29); specifically, we let $a := (a_1, \ldots, a_I)$ denote a binary vector indicating which of the $I$ classes we choose to serve, and denote the action set by

$$A(x) := \{0, 1\}^I : \sum_i I_{(a_i = 0)} \geq (I - 1), \ a \neq 0 \text{ if } x_i > 0 \text{ for some } i = 1, \ldots, I \}.$$  

We can then write (29) equivalently as

$$V_t(x_t) = \min_{a \in A(x)} \left\{ \sum_i c_i(x_t, i) + \alpha E[\hat{V}(s(x_t, a, w))] - \hat{V}(s(x_t, a, w_{t+1})) + V_{t+1}(s(x_t, a, w_{t+1})) \right\},$$

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 (30) 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 $\ell_i((I - 1) - \sum_i I_{(a_i = 0)})$ in each period for some Lagrange multipliers $\ell_i \geq 0$ (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 $V_t(x_t)$ in all states and times, for any $\ell := (\ell_0, \ldots, \ell_{\tau-1})$ with each $\ell_i \geq 0$. We let $V_t^\ell$ denote the value function for this Lagrangian relaxation of (30); this relaxation satisfies

$$V_t^\ell(x_t) = (I - 1)\ell_t + \min_{a \in \{0,1\}^I} \left\{ \sum_i (c_i(x_t, i) - \ell_i I_{(a_i = 0)}) + \alpha E[\hat{V}(s(x_t, a, w))] - \hat{V}(s(x_t, a, w_{t+1})) + V_{t+1}(s(x_t, a, w_{t+1})) \right\},$$

for $t = 0, \ldots, \tau - 1$, with $V^\ell_\tau = 0$. We argue that when $\hat{V}$ decouples by customer class, i.e., $\hat{V}(x) = \sum_i \hat{V}(x_i)$ (plus perhaps a constant, which we will omit to simplify notation), then (31) also decouples by customer class, i.e., $V_t^\ell(x_t) = \theta_t + \sum_i V_t^{\ell_i}(x_i)$, where $\theta_t$ is a constant that does not depend on the state.

We argue this by induction. This is clearly true at $t = \tau$, since $V^\ell_\tau = 0$. Now assume that the result holds for $V_{t+1}^\ell$ 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 $a_t$, we have

$$
\bar{V}(s(x_t, a, w_{t+1})) = \sum_i \bar{V}_i(s_i(x_{t,i}, a_i, w_{t+1})),
$$

and similarly for the $V_{t,i}^\ell$ term in (31). Moreover, we have

$$
\mathbb{E}[\bar{V}(s(x_t, a, w))] = \mathbb{E}\left[\sum_i \bar{V}_i(s_i(x_{t,i}, a_i, w))\right] = \sum_i \mathbb{E}_i[V_i(s_i(x_{t,i}, a_i, w))],
$$

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 $c_i(x_{t,i}) - \ell_t I_{(a_i = 0)}$ in (31) also decouple by class, and there are no constraints on actions across classes. Altogether, this implies that $V_{t,i}^\ell$ can be decomposed as stated; in particular,

$$
V_{t,i}^\ell(x_{t,i}) = \min_{a_i \in \{0, 1\}} \left\{ c_i(x_{t,i}) - \ell_t I_{(a_i = 0)} + \alpha \mathbb{E}_i[\bar{V}_i(s_i(x_{t,i}, a_i, w))] - \bar{V}_i(s_i(x_{t,i}, a_i, w_{t+1})) + V_{t+1,i}^\ell(s_i(x_{t,i}, a_i, w_{t+1})) \right\}.
$$

Thus the Lagrangian relaxation (31) decouples by customer class in each scenario. Each $V_{t,i}^\ell$ has $B_t + 1$ states in each period, so solving the decoupling for each class involves dealing with $\tau \cdot (B_t + 1)$ total states when the scenario has $\tau$ periods.

We can then consider the problem $\max_{x_0} V_{t,i}^\ell(x_0)$ in each scenario. There are different ways we could solve this problem; as stated in Section 4.3, 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 $V_{t,i}^\ell(x_{t,i})$ for all class $i$ states in each period. This results in a linear program with $\tau \cdot \sum_i (B_t + 1)$ variables. Moreover, if $\ell^*$ is the optimal Lagrange multiplier for $\bar{V}^\ell$ in (21), then the choice $\ell_t = \ell^*$ in each period leads to $V_{t,i}^\ell = V_{t,i}^{\ell^*}$: this follows from Proposition 3.2(ii), noting that $\bar{V}^{\ell^*}$ is the optimal value function for the Lagrangian relaxations with $\ell_t = \ell^*$. The Lagrangian relaxation (21) is a therefore a feasible choice in every scenario and by Cor. 3.1, we can do no worse than $\bar{V}^{\ell^*}(x_0)$ in every scenario with this approach.

Finally, when using grouped Lagrangian relaxations as in (22), it is not hard to see that this same decoupling goes through with the difference that (31) decouples into group-specific value functions in each period. Thus, we could also use this approach with approximate value functions based on grouped Lagrangian relaxations, although if solved as linear programs, there could be many more variables. For example, with groups of size 4 and each with 10 states, there would be $\tau \cdot 10^4 = 10,000\tau$ variables for the value functions associated with each group in each scenario.

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 4. 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 this multiclass queueing application to pursue reformulations of the state transition function (e.g., weak formulations or quasi-strong formulations as discussed in Section 4.5) that lead to simpler inner problems, rather than trying to work with relaxations of the strong formulation 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 strong formulation inner problems.
B. Further Examples and Extensions

B.1. Inner Problems May Require Larger State Spaces

Here we provide an illustrative example that shows that for some information relaxations, it may be necessary to expand the state space to solve the inner problem recursions (12). This example introduces a new imperfect information relaxation, and the calculations with this relaxation can be instructive in terms of the mechanics of the inner problems with imperfect information. For simplicity, we present the example in a finite horizon setting and with zero control variates.

The example is the following: you can throw a fair \( n \)-sided die up to a maximum of \( T \) times. After each throw, you must choose to stop or continue. If you choose to stop, then the game terminates and you obtain a prize equal in dollars to the value you just threw. For example, if you throw a 4 and then choose to stop, then you obtain 4 dollars and the game is over; otherwise, you continue and can roll again (up to \( T \) times). We want to compute the optimal expected value of this game.

We denote by \( x_t \) the value of the \( t \)th throw, which is both the uncertainty in the problem as well as the state variable. This problem is a simple dynamic programming problem with an optimal policy that has a threshold structure. When \( n = 6 \) and \( T = 3 \), for example, it is easy to show the optimal \( t = 0 \) value (i.e., the value before the first throw) is 4.667.

Suppose now that we wish to calculate an information relaxation (upper) bound for this problem using the information relaxation \( G \), where \( \mathcal{G}_t := \sigma(\mathcal{F}_t, x_{t+1}) \) for \( t < T \) and \( \mathcal{G}_T := \mathcal{F}_T \), i.e., the relaxed filtration allows us to see the outcome of the die one period ahead at all times \( t < T \). In solving the inner problem recursion under this \( G \), we need to account for both the value \( x_t \) (known under \( \mathcal{F} \)) as well as the value \( x_{t+1} \). This can be handled by increasing the state space to be all values of \((x_t, x_{t+1})\) for each period \( t < T \). We then can solve a stochastic DP on this increased state space that takes an expectation over \( x_{t+2} \) at all periods \( t \leq T - 2 \). Note that because only \( x_1 \) is revealed under \( G \) at \( t = 0 \), we would average over \( x_1 \) in calculating the optimal \( t = 0 \) value under \( G \); the remaining uncertainty over \((x_2, \ldots, x_T)\) would be “averaged out” in the expectations conditional on \( G_0 \); there is no Monte Carlo simulation here. (We omit the detailed calculations but leave it as an instructive exercise to verify that under this information relaxation with \( n = 6 \) and \( T = 3 \), the optimal value at \( t = 0 \) is 4.870).

As stated, this example is illustrative: the primary motivation for information relaxations is to lead to problems that are much easier to solve than the primal DP. Thus although information relaxations can in theory lead to problems that are harder than the primal DP itself, it would be unusual to consider such relaxations in applications.

B.2. Machine Repair Example with Weak Form State Transitions

We demonstrate the machine repair example of Section 3.3 using weak formulation state transitions. In these examples, we work directly with the state transition probability matrix (rather than the state transition function), which we denote by \( Q \). Since this is a weak formulation, state transition probabilities do not depend on actions (i.e., whether a repair attempt is made or not). We assume under \( Q \) that absorption occurs with probability \( 1 - \alpha \) from every non-absorbing state.

The weak formulation inner problems (12) with perfect information then are

\[
V^*_i(x_t) = \min \left\{ c(x_t) + \alpha P^F_{x_t} \bar{V} + \frac{P^F(x_t, x_{t+1})}{Q(x_t, x_{t+1})} \left( V^*_i(x_{t+1}) - \bar{V}(x_{t+1}) \right), \right. \\
R + c(x_t) + \alpha P^F_{x_t} \bar{V} + \frac{P^F(x_t, x_{t+1})}{Q(x_t, x_{t+1})} \left( V^*_i(x_{t+1}) - \bar{V}(x_{t+1}) \right) \}
\]

for \( 0 \leq t \leq \tau - 1 \) with \( V^*_i \) = 0.

Under the absorption time relaxation, the lower bounds under the strong SSP formulation and these weak formulations coincide. This follows from the fact that only \( \tau \) is known under the absorption time relaxation and we are using the same distribution of \( \tau \) in these weak formulation examples.

Figure B.1 shows the lower bounds. The optimal value function is indicated in blue, and the black line indicates \( \bar{V} \) for some number, \( k \), of value iterations beginning with the zero vector. Since (as discussed in
Section 3.3) $\bar{V}$ is Bellman feasible, the results of Proposition 3.3 apply. In particular, the lower bounds from each information relaxation must be strictly better than $\bar{V}$; this guaranteed improvement over $\bar{V}$ is indicated with the green line. The red lines represent the lower bounds using the perfect and absorption time information relaxations.

The plotted lower bounds are in fact an average of ten separate weak formulation examples, each resulting from a randomly generated state transition matrix $Q$. For each weak formulation example, we used stratified sampling on $\tau$ in generating scenarios to help reduce sample variance; for each example, $Q$ is generated uniformly and normalized so that rows sum to one (since every element of each $Q$ was strictly greater than zero, the absolute continuity condition in Proposition 2.3(i) holds relative to any feasible policy; thus even without Bellman feasible control variates, we would be ensured lower bounds on the optimal value here).

A notable feature of Figure B.1 is that the perfect information bound only provides a modest improvement over the improvement guarantee provided by Proposition 3.3, somewhat in contrast to the strong formulation case in Figure 1. This is perhaps not surprising given that $\gamma = 0$ for each of the transition matrices $Q$ in each weak formulation example (recall that $\gamma$ was defined in Section 3.2 and denotes $\alpha$ times the minimum value across all actions and states of change of measure factors, here given by $P^u(x_t, x_{t+1})/Q(x_t, x_{t+1})$, where $u$ is $r$ or $\bar{r}$).

It is interesting to observe that, unlike upper bounds from any primal feasible policy, the perfect information lower bounds depend on $Q$. In Figure B.2 we show the perfect information lower bounds for each of the weak formulation examples, each corresponding to a different $Q$. The blue line again indicates the optimal value function. More variation in the perfect information lower bounds are apparent in the $k = 0$ case (zero control variates) than in the $k = 3$ case. We should expect an interaction between the approximate value function forming the control variates and the weak formulation state transition probabilities. In particular, when we have relatively poor approximations to the optimal value function, differences in the weak formulation state transition probabilities can have a more pronounced effect on the quality of the lower bounds than when the approximations are relatively good. The example in Section B.3 demonstrates explicitly how the reformulated state transition probabilities in weak formulations can affect the lower bounds.

### B.3. Bounds from Weak Formulations Can Be Better than Bounds from Strong Formulations

Given the results of the machine repair example and Proposition 3.3, it is be natural to conjecture that, for the same distribution of absorption time, lower bounds from strong formulations are always at least as good as those from weak formulations (for a given information relaxation and control variates). We now provide a simple counterexample to this conjecture. We consider a simple problem for which there are two periods, $t = 0$ and $t = 1$, two possible states, “good” (g) and “bad” (b), and two possible actions, 1 and 2. Costs depend only on the current state and the current action in each period. In particular, costs are: $c(g, 1) = 0$, $c(g, 2) = 0.2$, $c(b, 1) = 1$ and $c(b, 2) = 1$. The initial state is $g$ and if the time $t = 0$ action is 1, then the time
$t = 1$ state will be $g$ with probability $p_1 = 0.6$ and $b$ otherwise; if the time $t = 0$ action is 2, then the time $t = 1$ state will be $g$ with probability $p_2 = 0.7$ and $b$ otherwise. In time $t = 1$, we can select either action 1 or 2 one final time and the problem then ends. The objective is to minimize the expected costs. The optimal cost $c^*$ is

$$c^* = \min_{i=1,2} \left\{ c(g, i) + p_i \min_{j=1,2} c(g, j) + (1 - p_i) \min_{j=1,2} c(b, j) \right\}$$

$$= \min \{ 0 + 0.6 \cdot 0 + 0.4 \cdot 1, 0.2 + 0.7 \cdot 0 + 0.3 \cdot 1 \}$$

$$= 0.4,$$

an optimal policy is to always select action 1 (it is irrelevant which action is taken at time $t = 1$ if the state is $b$). In each of the lower bound calculations below we use a perfect information relaxation and zero penalty (control variate).

**Bound from Strong Formulation**

Using the strong formulation, with a perfect information relaxation, we learn the outcome of the state transition from $t = 0$ to $t = 1$ under each action. With probability $p_1 = 0.6$ we discover that both actions lead to $g$. With probability $p_2 - p_1 = 0.1$ we learn that action 1 leads to $b$ but action 2 leads to $g$. With probability $1 - p_2 = 0.3$ we learn that both actions lead to $b$. In each of these three scenarios, we can choose the best action. The resulting lower bound is thus

$$\mathcal{L}_{\text{Strong}} = p_1 \min_{i=1,2} \left\{ 2c(g, i) \right\} + (p_2 - p_1) \min_{j=1,2} \left\{ c(g, 1) + c(b, 1), c(g, 2) + \min_{j=1,2} c(b, j) \right\}$$

$$+ (1 - p_2) \min_{j=1,2} \left\{ c(g, 1) + \min_{j=1,2} c(b, j), c(g, 2) + \min_{j=1,2} c(b, j) \right\}$$

$$= 0.6(2 \cdot 0) + 0.1(0.2) + 0.3(1)$$

$$= 0.32.$$
bound in this case as

\[
\sum_{\text{Weak}} = \min_{i=1,2} \left\{ c(g,i) + \frac{p_i}{q} \min_{j=1,2} c(g,j) \right\} + (1-q) \min_{i=1,2} \left\{ c(g,i) + \frac{(1-p_i)}{(1-q)} \min_{j=1,2} c(b,j) \right\}
\]

\[
= 0 + \min \{0.2(1-q) + 0.3, 0.4\}
\]

\[
= \begin{cases} 
  0.4 & \text{if } q \leq 0.5 \\
  0.2(1-q) + 0.3 & \text{otherwise.}
\end{cases}
\]

We thus have \(\sum_{\text{Weak}} = c^*\) for all \(q \leq 0.5\) and \(\sum_{\text{Weak}} > \sum_{\text{Strong}}\) for all \(q < 0.9\). It is interesting to note that even with zero penalty for information, the weak formulation is capable (with a well-chosen \(q\), i.e. \(q \leq 0.5\)) of providing a tight lower bound. This observation helps highlight the fact that the performance of the lower bounds from weak formulations may depend jointly on the control variates (i.e., penalty) as well as the reformulation.

### B.4. Stochastic Shortest Path Problems and Average Cost Problems

We now briefly discuss how the approach developed for discounted infinite horizon problems can be extended to general stochastic shortest path (SSP) problems and average cost (per stage) problems. We will begin with SSP problems and then use these results to discuss one potential approach to handling average cost problems.

#### B.4.1. General Stochastic Shortest Path Problems

The general SSP problem has an absorbing state, again denoted by \(x^a\), which is entered at some absorption time \(\tau\). No costs are incurred after absorption. In contrast to the strong SSP formulation of the discounted problem in Section 2.1, however, the distribution of \(\tau\) now depends on actions in general rather than being geometric and independent of actions. Moreover it is possible that \(\tau = \infty\) under some policies. Under suitable assumptions, however, an optimal policy is guaranteed to exist under which \(\tau\) will be finite almost surely. One such set of general assumptions is described by Bertsekas (2001) and we will adopt these assumptions here. The treatment of SSP problems in Bertsekas (2001) assumes finite state spaces, and we will make that assumption in what follows in this section and in Section B.4.2.

**Definition B.1.** A stationary policy \(\mu\) is said to be proper if, when using this policy, there is positive probability that the absorbing state \(x^a\) will be reached after at most \(|X|\) periods, regardless of the initial state. Otherwise the stationary policy is said to be improper.

It follows from Definition B.1 that a proper policy will reach \(x^a\) with probability 1. We also note that whether or not a policy is proper will depend on the state transition probabilities that we use (i.e., the definition could be in reference to a reformulation of the problem as in Section 2.4). We are assuming the definition holds for the original state transition probabilities of the SSP problem under the state transition function \(s\); we denote these transition probabilities by \(P_u(x,y)\). We will adopt the following two assumptions throughout this subsection. The first assumption is that there exists at least one proper policy. The second assumption is that for every improper policy, \(\mu\), the corresponding expected cost \(V_\mu(x)\) is infinite for at least one state, \(x\). An optimal stationary proper policy is known to exist under these two assumptions (see Bertsekas 2001) and the optimal value function for the SSP problem may be obtained from the usual value iteration algorithm beginning with any initial vector. It follows that we can restrict our optimization to the space of proper policies.

To use information relaxations in these general SSP problems, we can follow the development of Section 3. We can obtain analogous results to Proposition 2.3 (but without the discount factor \(\alpha\)) as long as \(\mu\) is a proper policy. Then weak and strong duality, i.e. Props. 3.1 and 3.2, also hold with the additional condition that we are restricted to the space of proper policies. This additional restriction does not present a problem since we know there exists an optimal proper policy as long as the two assumptions stated above hold.

How might we put these observations into practice? One strong formulation that may work in some problems is given by using a reformulation with transition matrix \(Q_u\) under which \(\tau\) is geometrically distributed with parameter \(\alpha\). Conditional on not hitting \(x^a\) in a given period, however, we take \(Q_u\) and \(P_u\) to
be identical. Together these assumptions imply

$$Q_u(x, \bar{x}) = \begin{cases} \alpha \frac{P_u(x, \bar{x})}{1 - P_u(x, x^a)}, & \bar{x} \neq x^a \\ 1 - \alpha, & \bar{x} = x^a \end{cases}$$

(1)

for $x \neq x^a$. One potential problem with this approach is that there may exist a $(u, x)$ pair such that $P_u(x, x^a) = 1$, in which case $Q_u(x, x^a)$ cannot be defined according to (1). In this case, taking action $u$ in state $x$ will drive the system to the absorbing state with probability 1 and so we cannot generate $\tau$ exogenously, i.e. in an action-independent manner. This problem is easy to resolve, however, with one simple solution requiring the introduction of a new fictitious and costless state, $x^a$, that can only self-transition or transition to $x^a$. Costs would be unchanged and transition probabilities are also identical except for $(u, x)$ pairs where $P_u(x, x^a) = 1$. For such $(u, x)$ pairs the probability $P_u(x, x^a)$ would be divided between $P_u(x, x^a)$ and $P_u(x, x^a)$, resulting in an equivalent problem where $P_u(x, x^a) \neq 1$ for any $u$ and any $x \neq x^a$. This approach could also be easily extended by making the absorption probability depend on time (i.e., replace $1 - \alpha$ above with $1 - \alpha_1$).

An inner problem is then obtained by first simulating $\tau$ from the geometric distribution with parameter $1 - \alpha$ and then simulating $w = (w_1, \ldots, w_{\tau-1})$. Although this could work, a geometric absorption time is not necessarily a natural choice for a general SSP problem, however, and a difficulty we have encountered with this approach is that sample variances may be very large, possibly unbounded. Alternative and perhaps better approaches include: (i) using truncated horizons as outlined in Section 3.4 and (ii) using weak formulations where state transitions are taken to be $\tilde{s}(x_1, u_1, w) = s(x_1, \mu(x_1), w)$ for some proper policy $\mu$. These approaches would in general violate absolute continuity relative to optimal policies and would rely on an analogous result to part (ii) of Proposition 2.3 in that they would therefore require the control variates to be constructed from Bellman feasible approximate value functions. Given our experience with these methods for discounted problems, we expect these approaches could yield very good lower bounds for general SSP problems.

B.4.2. Average Cost Problems

Average cost problems are a well-studied class of problems where the goal is to minimize

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} c(x_k, \mu(x_k))$$

over all stationary policies $\mu$. In this section, we will pursue a connection between SSP’s and average cost problems, and briefly outline how we might use information relaxations and this connection to obtain lower bounds for average cost problems. Our analysis here builds on chapter 4 of Bertsekas (2001).

We say a policy $\mu$ is unichain if the associated Markov chain has a single recurrent class and a possibly empty set of transient states. A well known and intuitive result states that the average cost of a unichain policy is independent of the initial state. It therefore follows that if every optimal policy within the class of stationary policies is unichain then the optimal average cost is independent of the initial state. In addition, Bertsekas (2001) then also shows\(^4\) that the average cost problem can be associated with a particular SSP problem. This SSP problem requires the identification of a special state, which we now characterize.

Let $\lambda_\mu$ denote the average cost associated with any stationary policy $\mu$ and let $\lambda^* := \min_\mu \lambda_\mu$ denote the optimal average cost. Then $M := \{\mu \mid \lambda_\mu = \lambda^*\}$ is the collection of optimal policies. If there is a state, $x^*$, that is simultaneously recurrent for all $\mu \in M$, then we can take $x^*$ to be our special state. Otherwise we select a $\bar{\mu} \in M$ that is minimal\(^5\) in the sense that there is no $\mu \in M$ whose recurrent class is a strict subset of the recurrent class of $\bar{\mu}$. We can then take the special state $x^*$ to be any state that is recurrent under $\bar{\mu}$. Although this could be hard in general, in many applications of interest it should be straightforward to identify such an $x^*$; for example, in the multiclass queueing model in Section 4, given that we can restrict without loss to policies that never idle, an empty queue would be such a state.

We now form an SSP problem, which we call the $\lambda$-SSP, by first creating a new state, $x^*$, which plays the role of the absorption state and essentially represents a copy of $x^*$, with corresponding state transition

\(^4\)See his discussion of the $\lambda$-SSP in Section 4.3 and also Exercise 4.13.

\(^5\)It is sufficient that $\bar{\mu}$ has a minimal number of recurrent states over all $\mu \in M$. 

45
probabilities constructed in a straightforward manner (see Bertsekas 2001 for details). The costs are now
\(c_\lambda(x, u) := c(x, u) - \lambda\) for some constant \(\lambda\) with \(c(x, u)\) being the cost function for the original average cost problem. We let \(h_\lambda(x)\) denote the optimal cost vector for this \(\lambda\)-SSP problem. The goal is to find \(\lambda^*\) such that \(h_\lambda(x^*) = 0\). Such a \(\lambda^*\) will be the optimal average cost for the original problem.

For any fixed \(\lambda\), we can find upper and lower bounds, \(\bar{h}_\lambda(x^*)\) and \(\underline{h}_\lambda(x^*)\), respectively, for \(h_\lambda(x^*)\) by simulating a feasible policy for the upper bound and using information relaxations and control variates as in Appendix B.4.1 for the lower bound. These bounds satisfy
\[
\underline{h}_\lambda(x^*) \leq h_\lambda(x^*) \leq \bar{h}_\lambda(x^*).
\] (2)

If we assume we can evaluate the bounds in (2) exactly (i.e., without sample error from Monte Carlo simulation), then we can use a simple iterative scheme to bound \(\lambda^*\). Towards this end, we first define \(f(\lambda) := h_\lambda(x^*)\). It is not hard to show that \(f\) is monotonic decreasing in \(\lambda\). Suppose now that we have found a \(\lambda^*\) such that \(\underline{h}_\lambda(x^*) = 0\). By (2) it then follows that \(h_\lambda(x^*) \geq 0\). But since \(f(\lambda)\) is monotonic decreasing in \(\lambda\) it follows that \(\lambda^* \leq \lambda^*\), i.e., \(\lambda^*\) is a lower bound for the optimal average cost problem. We can find such a \(\lambda^*\) by repeatedly applying the information relaxation approach to the SSP for different values of \(\lambda\). By using the monotonicity of \(f(\lambda)\) we could use a bisection search on \(\lambda\), initialized to the interval \([\min_{x, u} c(x, u), \max_{x, u} c(x, u)]\), which must contain \(\lambda^*\). We also note that an upper bound could be obtained in a similar manner by seeking a value of \(\lambda\) for which \(\bar{h}_\lambda(x^*) = 0\).

There are additional issues that would need to be addressed with this approach. The \(\lambda\)-SSP’s will in general have some negative costs and, unlike for discounted problems, we cannot convert a general SSP problem to an equivalent SSP with nonnegative costs. Thus, to ensure we are obtaining a lower bound on the optimal cost in general, we either need to ensure the absolute continuity condition relative to an optimal policy is satisfied, or use (nonzero) Bellman feasible approximate value functions as control variates. As \(\lambda\) is updated, however, the costs in the \(\lambda\)-SSP would also be updated and so we would also need to update the approximate value functions to maintain Bellman feasibility. One way to do this is as follows. Suppose we know \(\lambda_1 < \lambda^* < \lambda_2\) and that we have Bellman feasible \(\bar{V}_i\) for the \(\lambda_i\)-SSP where \(i = 1, 2\). Then for any value of \(\lambda := a\lambda_1 + (1 - a)\lambda_2\) with \(a \in [0, 1]\) it is not hard to show that \(a\bar{V}_1 + (1 - a)\bar{V}_2\) will be Bellman feasible for the \(\lambda\)-SSP. We can therefore obtain the necessary control variates for the \(\lambda\)-SSP using an appropriate convex combination of \(\bar{V}_1\) and \(\bar{V}_2\). Another issue is the assumption that the bounds in (2) can be computed exactly; in general this will not be possible and we would need to deal with estimation error in the search over \(\lambda\).

In conclusion, the foregoing suggests a viable approach for using information relaxations to obtain lower bounds for average cost problems. Nonetheless, there are challenges and the computational burden could be substantial in some problems. Alternatively, the theory of Blackwell optimal policies provides a well known link between average cost and discounted cost problems and it may be possible to use this connection for average cost problems. For example, we could apply information relaxations to a discounted version of the problem with \(\alpha\) very close to 1 (and then multiply the resulting bound by \(1 - \alpha\)). The issue with this route is that, although it would provide a lower bound on the discounted problem, there is no guarantee in general that this value (scaled by \(1 - \alpha\)) would be a lower bound on the average cost problem, even if \(\alpha\) is very close to 1. In particular, the lower bound from the discounted problem would depend on the initial state. There may be ways to correct for this in a given problem, but it is not immediately clear how to do so in general.