Index Policies and Performance Bounds for Dynamic Selection Problems

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

James E. Smith
Tuck School of Business
Dartmouth College
jim.smith@dartmouth.edu

This version: November 26, 2018
First version: September 18, 2017

Abstract

We consider *dynamic selection problems*, where a decision maker repeatedly selects a set of items from a larger collection of available items. A classic example is the dynamic assortment problem with demand learning, where a retailer chooses items to offer for sale subject to a display space constraint. The retailer may adjust the assortment over time in response to the observed demand. These dynamic selection problems are naturally formulated as stochastic dynamic programs (DPs) but are difficult to solve because the optimal selection decisions depend on the states of all items. In this paper, we study heuristic policies for dynamic selection problems and provide upper bounds on the performance of an optimal policy that can be used to assess the performance of a heuristic policy. The policies and bounds that we consider are based on a Lagrangian relaxation of the DP that relaxes the constraint limiting the number of items that may be selected. We characterize the performance of the Lagrangian index policy and bound and show that, under mild conditions, these policies and bounds are asymptotically optimal for problems with many items; mixed policies and tiebreaking play an essential role in the analysis of these index policies and can have a surprising impact on performance. We demonstrate these policies and bounds in two large scale examples: a dynamic assortment problem with demand learning and an applicant screening problem.

Keywords: Dynamic programming, restless bandits, Lagrangian relaxations, Gittins index, Whittle index.

We would like to thank conference and seminar participants at INFORMS (2016, 2017, and 2018), University of Maryland, Dartmouth, University of Southern California, Northwestern, the Advances in Decision Analysis Conference 2017 (Austin), UCLA, Princeton, UT Austin, Duke, and University of Illinois-Chicago for helpful comments and questions. We are also grateful for the helpful comments and suggestions from an anonymous Associate Editor and three anonymous referees.

1. Introduction

In this paper, we consider *dynamic selection problems*, where a decision maker repeatedly selects a set of items from a larger collection of available items. A classic example is the dynamic assortment problem with demand learning, where a decision maker (DM) – prototypically a retailer – chooses products to offer for sale, selecting from many possible products, but is limited by display space. In this problem, product demand rates are uncertain, and the retailer may want to update the assortment over the course of the selling season in response to demands observed in previous periods. Similar problems arise in internet advertising (which ads should be displayed on a news site?), in yield trials for experimental crop varieties (which experimental varieties should be planted in a trial?), and in hiring or admissions decisions (which applicants should be interviewed, hired or admitted?).

These dynamic selection problems are naturally formulated as stochastic dynamic programs (DPs) but are difficult to solve to optimality. Even when the reward processes are independent across items, the competition for limited resources (e.g., display space) links the selection decisions: the selection decision for one item will depend on the states of the other available items. In this paper, we study heuristic policies for dynamic selection problems and provide upper bounds on the performance of an optimal policy. We focus on problems with a finite horizon, but also consider an extension to an infinite horizon setting with discounting.

Our methods and analysis are based on a Lagrangian relaxation of the DP that relaxes the constraint limiting the number of items that can be selected. This Lagrangian relaxation decomposes into item-specific DPs that are not difficult to solve and the value of the Lagrangian provides an upper bound on the value of an optimal policy. We can solve the Lagrangian dual problem (a convex optimization problem) to find Lagrange multipliers that give the best possible Lagrangian bound. This optimal Lagrangian can also be used to generate a heuristic policy that performs well and, if we mix policies and break ties appropriately, is asymptotically optimal: under mild conditions, as we increase the number of items available and the number that can be selected, the relative performance of the heuristic approaches the Lagrangian upper bound.

We illustrate these results with two example problems. The first is based on the dynamic assortment model with demand learning from Caro and Gallien (2007). The second is an applicant screening problem where a DM must decide which applicants (e.g., for a college or job) should be screened (e.g., reviewed or interviewed) and which applicants should be admitted or hired.

1.1. Literature Review

Our paper builds on and contributes to two related streams of literature. First, the dynamic selection problem can be viewed as a special case of a weakly coupled DP. For example, Hawkins (2003), Adelman and Mersereau (2008) and Bertsimas and Mišić (2016) study DPs that are linked through global resource constraints. The dynamic selection problem can be viewed as a weakly coupled DP where the linking constraint is a cardinality constraint that limits the number of items that can be selected in a period. Hawkins (2003), Adelman and Mersereau (2008) and Bertsimas and Mišić (2016) all consider Lagrangian

relaxations of weakly coupled DPs, similar to the Lagrangian relaxation in §3 below. Lagrangian relaxations of DPs have been used in a number of applications including network revenue management (e.g., Topaloglu 2009) and marketing (e.g., Bertsimas and Mersereau 2007 as well as Caro and Gallien 2007).

The dynamic selection problem can also be viewed as a finite-horizon, non-stationary version of the restless bandit problem introduced in Whittle (1988). The restless bandit problem is an extension of the classical multiarmed bandit problem where (i) the DM may select multiple items in any given period and (ii) items may change states when not selected. Whittle (1988) introduced an index policy where items are prioritized for selection according to an index that is essentially equivalent to the Gittins index. Whittle (1988) motivates this policy through a Lagrangian analysis, viewing the index as a breakeven Lagrange multiplier (see §4.2) and conjectured that in the infinite-horizon average reward setting these policies are asymptotically optimal for problems with many items. Weber and Weiss (1990) showed that this conjecture is true under certain conditions but need not be true in general. Caro and Gallien (2007) studied Whittle indices in the dynamic assortment problem. Bertsimas and Niño-Mora (2000) study restless bandit problems with discounted rewards over an infinite horizon and develop performance bounds based on a hierarchy of linear programming (LP) relaxations; they show that the first-order LP relaxation corresponds to the Lagrangian relaxation studied by Whittle (1988) and they use this relaxation to generate an index policy. Hodge and Glazebrook (2015) develop and analyze an index policy for an extension of the restless bandit model where each item can be "activated" at different levels. For a comprehensive discussion of the restless bandit problem, see Gittins et al. (2011).

1.2. Contributions and Outline

Our main contributions are (i) a detailed analysis of the Lagrangian relaxation of the dynamic selection problem and, building on this, (ii) the development of an *optimal Lagrangian index policy* that performs well in examples and is proven to be asymptotically optimal. Specifically, we consider limits where we increase both the number of items available (S) and the number of items that may be selected (N) with a growth condition (for example, N is a fixed fraction of S). We show that the performance gap (the difference between the Lagrangian bound and the performance of the heuristic policy) grows with the same rate as $\sqrt{N}$ for the optimal Lagrangian index, whereas the gaps for Whittle index policy (Whittle 1988) grows linearly with N. Mixed policies and tiebreaking play a surprising and important role in the analysis and in the numerical results. For example, a Lagrangian index policy that breaks ties randomly may also exhibit linear growth in the performance gap.[1]

[1] During the review process for this paper, we became aware of a working paper, Hu and Frazier (2017), that studies the use of Lagrangian relaxations for finite-horizon restless bandit problems. The model studied in Hu and Frazier (2017) is a special case of a dynamic selection problem where all items are *a priori* identical and state transition probabilities and resource constraints are constant over time. Hu and Frazier (2017) consider an index policy based on varying the Lagrange multiplier for the current time period, keeping all future Lagrange multipliers fixed. This policy appears to be equivalent to our optimal Lagrangian index policy where policies are mixed according to Markov policies (see §4.4). Hu and Frazier (2017) provide a proof of asymptotic optimality of this policy based on the convergence of occupation measures of the index policy to that of the Lagrangian relaxation. Our proof of asymptotic optimality is based on explicit bounds on the suboptimality of the Lagrangian index policy. These bounds provide a rate of convergence for the Lagrangian index policies and provide additional insight into the nature of this convergence that is helpful, for example, when considering the infinite horizon extension of §7.2.

We begin in §2 by defining the dynamic selection problem and introducing the dynamic assortment and applicant screening problems. In §3, we describe the Lagrangian relaxation and discuss its theoretical properties; we describe a cutting-plane method for efficiently solving the Lagrangian dual optimization problem in Appendix A. In §4, we define a number of heuristic policies including the Whittle index policy and the optimal Lagrangian index policy. In §5, we characterize the performance of the optimal Lagrangian index policy and present results on the asymptotic optimality of this policy. In §6, we simulate the heuristic policies of §4 in the context of the two example problems and evaluate their performance. In §7, we discuss the applicability of these methods in problems with long time horizons, considering Whittle (1988)'s conjecture on the asymptotic optimality of the Whittle index policy and the counterexample of Weber and Weiss (1990). We also present an extension of the asymptotic optimality of the Lagrangian index policy to an infinite horizon setting with discounting. In Electronic Companion (EC) §E, we describe information relaxation performance bounds (see, e.g., Brown et al. 2010) based on the Lagrangian relaxation and show how they improve on the standard Lagrangian bounds; these bounds are illustrated in the numerical examples of §6. Most proofs and some other detailed discussions are also provided in the EC.

2. The Dynamic Selection Problem

We first describe the general dynamic selection problem and then discuss the dynamic assortment and applicant screening problems as examples of this general framework.

2.1. General Model

We consider a finite horizon with periods $t = 1, \ldots, T$. In period t, the DM can select a maximum of N_t items out of S available. The DM's state of information about item s is summarized by a state variable x_s. To avoid measurability and other technical issues, we will assume that the state variables x_s can take on a finite number of values. We define a binary decision variable u_s where 1 (0) indicates item s is (is not) selected. In each period, item s generates a reward $r_{t,s}(x_s, u_s)$ that depends on the state x_s, the selection decision u_s, and the period t. Between periods, the state variables x_s transition to a random new state $\widetilde{\chi}_{t,s}(x_s, u_s)$ with transitions depending on the current state, the selection decision, and period. We let $\boldsymbol{x} = (x_1, , \ldots, x_S)$ denote a vector of item states, $\boldsymbol{u} = (u_1, \ldots, u_S)$ a vector of selection decisions, and $\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}, \boldsymbol{u}) = (\widetilde{\chi}_{t,1}(x_1, u_1),$ $\ldots, \widetilde{\chi}_{t,S}(x_S, u_S))$ the corresponding random vector of next-period item states.

The DM selects items with the goal of maximizing the expected total reward earned over the given horizon. Though a policy for making these selections can depend on the whole history of states and actions and could be randomized, standard DP arguments (e.g., Puterman 1994) imply there is an optimal policy that is deterministic and Markovian i.e., of the form $\pi = (\pi_1, \ldots, \pi_T)$, where $\pi_t(\boldsymbol{x})$ specifies a vector of selection decisions $\boldsymbol{u}$ given state vector $\boldsymbol{x}$, where $\boldsymbol{u}$ must be in

$$\mathcal{U}_t \equiv \left\{ \boldsymbol{u} \in \{0,1\}^S \ : \ \sum_{s=1}^{S} u_s \leq N_t \right\}. \tag{1}$$

Taking the terminal value $V^*_{T+1}(\boldsymbol{x}) = 0$, we can write the optimal value function for earlier periods as

$$V^*_t(\boldsymbol{x}) = \max_{\boldsymbol{u} \in \mathcal{U}_t} \left\{ r_t(\boldsymbol{x}, \boldsymbol{u}) + \mathbb{E}\big[\, V^*_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}, \boldsymbol{u}))\,\big] \right\}, \tag{2}$$

where the total reward for a given period is the sum of item-specific rewards $r_t(\boldsymbol{x}, \boldsymbol{u}) = \sum_{s=1}^{S} r_{t,s}(x_s, u_s)$. We will also consider variations of the problem where the DM must select exactly N_t items in period t; i.e., where the inequality constraint in (1) is replaced by an equality constraint.

For an arbitrary policy π, we can write the corresponding value function $V^\pi_t(\boldsymbol{x})$ recursively as

$$V^\pi_t(\boldsymbol{x}) = r_t(\boldsymbol{x}, \pi_t(\boldsymbol{x})) + \mathbb{E}\big[\, V^\pi_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}, \pi_t(\boldsymbol{x})))\,\big], \tag{3}$$

where the terminal case is $V^\pi_{T+1}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$. A policy π is *optimal for initial state* $\boldsymbol{x}$ if it always satisfies the linking constraint (1) and $V^\pi_1(\boldsymbol{x}) = V^*_1(\boldsymbol{x})$.

As mentioned in the introduction, the dynamic selection problem can be viewed as a nonstationary, finite-horizon version of the restless bandit problem of Whittle (1988). Whittle mentions a number of potential applications of restless bandits including clinical trials, aircraft surveillance, and worker scheduling. Bertsimas and Niño-Mora (2000) mentions applications of restless bandits in controlling drug markets and in controlling a make-to-stock production facility. We will illustrate our general framework by considering two specific applications that we describe next.

2.2. Dynamic Assortment Problem with Demand Learning

Following Caro and Gallien (2007, CG for the remainder of this section), in the dynamic assortment problem with demand learning, we consider a retailer who repeatedly chooses products (items) to display (select) from a set of S products available, subject to a shelf space constraint that requires the number of products displayed in a period to be less than or equal to N_t. The demand rate for products is unknown and the DM updates beliefs about these rates over time using Bayes' rule. The retailer's goal is to maximize the expected total profit earned. As in CG (2007), we assume the demand for product s follows a Poisson distribution with an unknown product-specific rate γ_s. The demand rates are assumed to be independent across products and have a gamma prior with shape parameter m_s and inverse scale parameter α_s ($m_s, \alpha_s > 0$), which implies the mean and variance of γ_s are m_s/α_s and m_s/α_s^2. The state variable x_s for product s is the vector (m_s, α_s) of parameters for its demand rate distribution. If a product is displayed, its reward for that period is assumed to be proportional to the mean demand m_s/α_s; if a product is not displayed, its reward is zero.

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

In our numerical examples, we will consider parameters similar to those in CG (2007). We consider horizons $T = 8$, 20 and 40. We assume that all products are *a priori* identical with gamma distribution parameters $(m_s, \alpha_s) = (1.0, 0.1)$ (so the mean and standard deviation for the demand rate are both 10) and rewards are equal to the mean demand m_s/α_s (i.e., the profit margin is \$1 per unit).[2] We will vary the number of products available S and assume that the DM can display 25% of the products available in each period, i.e., $N_t = 0.25S$.

CG (2007) considered several extensions of this basic model that also fit within the framework of dynamic selection problems. One such extension introduced a lag of l periods between the time a display decision is made and when the products are available for sale. In this extension, the item-specific state variable x_s is augmented to keep track of the display decisions in the previous l periods. CG (2007) also considered an extension with switching costs, which requires keeping track of whether a product is currently displayed.

Of course, there are many variations on the assortment problem (see Kök et al. 2008 for a review) that do not fit within the framework of dynamic selection problems. Although CG (2007) modeled aggregate demand for a retailer over the course of a fixed time period (say, a week), recent work on dynamic assortment problems have modeled the arrivals of individual customers, e.g., to a web page. For example, Rusmevichientong et al. (2010) consider a dynamic assortment model with capacity constraints (like (1)) but where demands are modeled using a multinomial logit choice model with unknown customer preferences. Bernstein et al. (2015) consider a dynamic assortment problem with demand modeled using a multinomial logit choice model where products have limited inventories. The multinomial choice model used in these two papers captures substitution effects and the rewards cannot be decomposed into the sum of item-specific rewards as required in the dynamic selection model.

2.3. Applicant Screening Problem

In this example, we consider a set of S applicants seeking admission at a competitive college or applying for a prestigious job. The DM's goal is to identify and admit (or hire) the best possible set of applicants. Each applicant has an unknown quality level $q_s \in [0, 1]$, with uncertainty given by a beta distribution with parameters $x_s = (\alpha_s, \beta_s)$ where $\alpha_s, \beta_s > 0$; the mean quality is then equal to $\alpha_s/(\alpha_s + \beta_s)$.

In the first $T - 1$ periods, the DM can screen up to N_t applicants. Screening an applicant yields a signal about the applicant's quality; the signals are drawn from a binomial distribution with n trials and probability of success q_s on each trial. The number of trials n in the binomial distribution can be interpreted as a measure of the informativeness of the signals. For example, a binomial signal with $n = 5$ provides as much information as 5 signals from a Bernoulli signal (a binomial with $n = 1$). After screening an applicant and observing a signal d_s, the applicant's state is updated using Bayes' rule to $(\alpha_s + d_s, \beta_s + n - d_s)$. In the Bernoulli case, we can think of the signal as being a "thumbs up" or "thumbs down" indicating whether the screener thought the applicant should be admitted (or hired) or not. An applicant's state does not change

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

when not screened. The rewards are assumed to be zero during the screening periods. In the final period, the DM can admit up to N_T applicants. The reward for admitted applicants is their mean quality $(\alpha_s/(\alpha_s+\beta_s))$ and the reward for those not admitted is zero.

In our numerical examples, we will focus on examples with $T = 5$ and *a priori* identical applicants with $(\alpha_s, \beta_s) = (1, 1)$. We will vary the number of applicants S and assume 25% of the applicants can be admitted and 25% can be screened in each of the four screening periods (i.e., $N_t = 0.25S$). We will also vary the informativeness of the signals, taking $n = 1$ or 5 in the binomial distribution for the signal process. We will also consider an example case with Bernoulli signals ($n = 1$) and a longer time horizon ($T = 51$) where a smaller fraction of applicants can be screened in each period ($N_t = 0.02S$) and just 2% can be admitted. In all of these examples, the DM needs to strike a balance between a desire to screen each applicant at least once (which is feasible) and the desire to identify and admit the best applicants, a process which typically requires multiple screenings. With the chosen parameters, the DM can screen applicants more than once only if some other applicants are not screened at all.

3. Lagrangian Relaxations

The DP (2) is difficult to solve because the constraint (1) limiting the number of items selected links decisions across items: the selection decision for one item depends on the states of the other items. In this section, we consider Lagrangian relaxations of this problem where we relax this linking constraint and decompose the value functions into computationally manageable subproblems. This Lagrangian relaxation can then be used to generate a heuristic selection policy (as described in §4) as well as an upper bound on the performance of an optimal policy. Propositions 1-3 are fairly standard in the literature on Lagrangian relaxations of DPs (e.g., Hawkins 2003, Caro and Gallien 2007, and Adelman and Mersereau 2008). Proposition 4 provides a detailed analysis of the gradient structure of the Lagrangian that is important in later analysis.

3.1. The Lagrangian DP

Though one could in principle consider Lagrange multipliers that are state dependent, to decompose the DP we focus on Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T) \geq \mathbf{0}$ that depend on time but not states. As we will see in Proposition 4 below, the assumption that the Lagrange multipliers are constant across states means that an optimal set of Lagrange multipliers requires the linking constraint (1) to hold "on average" (or in expectation) rather than in each state. Taking $L^{\boldsymbol{\lambda}}_{T+1}(\boldsymbol{x}) = 0$, the Lagrangian (dual) DP has period-t value function that is given recursively as

$$L^{\boldsymbol{\lambda}}_t(\boldsymbol{x}) = \max_{\boldsymbol{u}\in\{0,1\}^S} \left\{ r_t(\boldsymbol{x},\boldsymbol{u}) + \mathbb{E}\big[\, L^{\boldsymbol{\lambda}}_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x},\boldsymbol{u}))\,\big] + \lambda_t\left(N_t - \sum_{s=1}^{S} u_s\right)\right\}. \tag{4}$$

Compared to the DP (2), we have made two changes. First, we have incorporated the linking constraint into the objective by adding $\lambda_t(N_t - \sum_{s=1}^{S} u_s)$; with $\lambda_t \geq 0$, this term is nonnegative for all policies satisfying the linking constraint. Second, we have relaxed the linking constraint, allowing the DM to select as many

items as desired (we require $\boldsymbol{u} \in \{0,1\}^S$ rather than $\boldsymbol{u} \in \mathcal{U}_t$). Both of these changes can only increase the optimal value so the Lagrangian value function provides an upper bound on the true value function.

Proposition 1 (Weak duality)**.** *For all $\boldsymbol{x}$, t, and $\boldsymbol{\lambda} \geq \mathbf{0}$, $V_t^*(\boldsymbol{x}) \leq L_t^{\boldsymbol{\lambda}}(\boldsymbol{x})$.*

Thus, for any $\boldsymbol{\lambda} \geq 0$, $L_t^{\boldsymbol{\lambda}}(\boldsymbol{x})$ can be used as a performance bound to assess the quality of a feasible policy.

The advantage of the Lagrangian relaxation is that, for any fixed $\boldsymbol{\lambda}$, we can decompose the Lagrangian dual function into a sum of item-specific problems that can be solved independently.

Proposition 2 (Decomposition)**.** *For all $\boldsymbol{x}$, t, and $\boldsymbol{\lambda} \geq \mathbf{0}$,*

$$L_t^{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{\tau=t}^{T} \lambda_\tau N_\tau + \sum_{s=1}^{S} V_{t,s}^{\boldsymbol{\lambda}}(x_s) \tag{5}$$

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

$$V_{t,s}^{\boldsymbol{\lambda}}(x_s) = \max\left\{ r_{t,s}(x_s,1) - \lambda_t + \mathbb{E}\big[\, V_{t+1,s}^{\boldsymbol{\lambda}}(\widetilde{\chi}_{t,s}(x_s,1))\,\big] \;,\; r_{t,s}(x_s,0) + \mathbb{E}\big[\, V_{t+1,s}^{\boldsymbol{\lambda}}(\widetilde{\chi}_{t,s}(x_s,0))\,\big] \right\}. \tag{6}$$

The first term in the maximization of (6) is the value if the item is selected and the second term is the value if the item is not selected. Intuitively, the period-t Lagrange multiplier λ_t can be interpreted as a charge for using the constrained resource in period t. We will let ψ denote an optimal deterministic (Markovian) policy for the Lagrangian relaxation (4) and ψ_s denote an optimal deterministic policy for the item-specific problem (6); we reserve π for policies that respect the linking constraints (1).

3.2. The Lagrangian Dual Problem

As discussed after Proposition 1, the Lagrangian DP can be used as an upper bound to assess the performance of heuristic policies. Although any $\boldsymbol{\lambda}$ provides a bound, we want to choose $\boldsymbol{\lambda}$ to provide the best possible bound. We can write this Lagrangian dual problem as

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \; L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) \;. \tag{7}$$

To say more about this Lagrangian dual problem (7), we will consider a fixed initial state $\boldsymbol{x}$ and focus on properties of $L_1^{\boldsymbol{\lambda}}(\boldsymbol{x})$ and $V_{1,s}^{\boldsymbol{\lambda}}(x_s)$ with varying $\boldsymbol{\lambda}$. Accordingly, for the remainder of this section, we will let $V_s(\boldsymbol{\lambda}) = V_{1,s}^{\boldsymbol{\lambda}}(x_s)$ and $L(\boldsymbol{\lambda}) = L_1^{\boldsymbol{\lambda}}(\boldsymbol{x})$.

First, we note that the item-specific value functions are convex functions of the Lagrange multipliers so the Lagrangian dual problem is a convex optimization problem.

Proposition 3 (Convexity)**.** *For all $\boldsymbol{x}$, t, and $\boldsymbol{\lambda} \geq \mathbf{0}$, $L(\boldsymbol{\lambda})$ and $V_s(\boldsymbol{\lambda})$ are piecewise linear and convex in $\boldsymbol{\lambda}$.*

Proof. See EC §B.1. □

In (6) we see that the Lagrange multipliers λ_t appear as costs paid whenever an item is selected; thus the gradients of $V_s(\boldsymbol{\lambda})$ and $L(\boldsymbol{\lambda})$ will be related to the probability of selecting items under an optimal policy for the item-specific DPs (6) for the given $\boldsymbol{\lambda}$. These selection probabilities are not difficult to compute when solving the DP. Since a convex function is differentiable almost everywhere, for "most" $\boldsymbol{\lambda}$ these gradients will be unique. However, as piecewise linear functions, there will be places where $V_s(\boldsymbol{\lambda})$ and $L(\boldsymbol{\lambda})$ are not differentiable and the optimal solution for the Lagrangian dual (7) will typically be at such a "kink." These kinks correspond to values of $\boldsymbol{\lambda}$ where there are multiple optimal solutions for the item-specific DPs. The following proposition describes the sets of subgradients for the Lagrangian and their relationships to optimal policies for the item-specific DPs.

Proposition 4 (Subgradients)**.** *Let $p_{t,s}(\psi_s)$ be the probability of selecting item s in period t when following a policy ψ_s for the item-specific DP* (6) *and let $\Psi_s^*(\boldsymbol{\lambda})$ be the set of deterministic policies that are optimal for the item-specific DP* (6) *in the initial state with Lagrange multipliers $\boldsymbol{\lambda}$.*

(i) Subgradients for item-specific problems: *For any $\psi_s \in \Psi_s^*(\boldsymbol{\lambda})$,*

$$\nabla_s(\psi_s) = -(p_{1,s}(\psi_s), \ldots, p_{T,s}(\psi_s)) \tag{8}$$

is a subgradient of V_s at $\boldsymbol{\lambda}$; that is,

$$V_s(\boldsymbol{\lambda}') \geq V_s(\boldsymbol{\lambda}) + \nabla_s(\psi_s)^\intercal(\boldsymbol{\lambda}' - \boldsymbol{\lambda}) \quad \textit{for all } \boldsymbol{\lambda}'. \tag{9}$$

The subdifferential (the set of all subgradients) of V_s at $\boldsymbol{\lambda}$ is

$$\partial V_s(\boldsymbol{\lambda}) = \mathbf{conv}\{\nabla_s(\psi_s) : \psi_s \in \Psi_s^*(\boldsymbol{\lambda})\} \tag{10}$$

where $\mathbf{conv}A$ *denotes the convex hull of the set A.*

(ii) Subgradients for the Lagrangian. *The subdifferential of L at $\boldsymbol{\lambda}$ is*

$$\partial L(\boldsymbol{\lambda}) = \boldsymbol{N} + \sum_{s=1}^{S} \partial V_s(\boldsymbol{\lambda}) = \boldsymbol{N} + \mathbf{conv}\left\{\sum_{s=1}^{S} \nabla_s(\psi_s) : \psi_s \in \Psi_s^*(\boldsymbol{\lambda}) \quad \forall s\right\} \tag{11}$$

where the sums are setwise (i.e., Minkowski) sums and $\boldsymbol{N} = (N_1, \ldots, N_T)$.

(iii) Optimality conditions. *$\boldsymbol{\lambda}^*$ is an optimal solution for the Lagrangian dual problem* (7) *if and only if, for each s, there is a set of policies $\{\psi_{s,i}\}_{i=1}^{n_s}$ with $\psi_{s,i} \in \Psi_s^*(\boldsymbol{\lambda}^*)$ ($n_s \leq T+1$) and mixing weights $\{\gamma_{s,i}\}_{i=1}^{n_s}$ (with $\gamma_{s,i} > 0$ and $\sum_{i=1}^{n_s} \gamma_{s,i} = 1$) such that*

$$\sum_{s=1}^{S} \sum_{i=1}^{n_s} \gamma_{s,i}\ p_{t,s}(\psi_{s,i}) \ = \ N_t \quad \textit{for all } t \textit{ such that } \lambda_t^* > 0 \quad \textit{and}$$

$$\sum_{s=1}^{S}\sum_{i=1}^{n_s}\gamma_{s,i}\ p_{t,s}(\psi_{s,i})\ \leq\ N_t\ \ \textit{for all } t \textit{ such that } \lambda_t^*=0\ .$$

Proof. See EC §B.1. □

We can interpret the result of Proposition 4(iii) as saying that the optimal policies for the Lagrangian DP must satisfy the linking constraints (1) "on average" for a *mixed policy* $\widetilde{\psi} = (\widetilde{\psi}_1, \ldots, \widetilde{\psi}_S)$ where the item-specific mixed policies $\widetilde{\psi}_s$ are independently generated as a mixture of deterministic policies $\psi_{s,i}$ with probabilities given by the mixing weights $\gamma_{s,i}$. Here, when we say the linking constraints must hold on average (or in expectation), this average includes the uncertainty in the state evolution when following a given item-specific policy $\psi_{s,i}$ (this determines $p_{t,s}(\psi_{s,i})$) and the probability $\gamma_{s,i}$ of following policy $\psi_{s,i}$.[3]

Although the result of Proposition 4(iii) suggests a mixture of policies where the DM randomly selects a deterministic policy $\psi_{s,i}$ for each item in advance (i.e., before period 1) and follows that policy throughout, we could use the policies and mixing weights of the proposition to construct item-specific Markov random policies that randomly decide whether to select an item in each period, with state-dependent selection probabilities; see EC §B.2. In both representations, we randomize independently across items.

In the special case where all items are *a priori* identical (i.e., identical item-specific DPs (6) with the same initial state), the Lagrangian computations simplify because we no longer need to consider distinct item-specific value functions. In this case, we can drop the subscript s indicating a specific item and the optimality condition of Proposition 4(iii) reduces to: $\boldsymbol{\lambda}^*$ is an optimal solution for the Lagrangian dual problem (7) if and only if there is a set of policies $\{\psi_i\}_{i=1}^n$ with $\psi_i \in \Psi^*(\boldsymbol{\lambda}^*)$ $(n \leq T+1)$ and mixing weights $\{\gamma_i\}_{i=1}^n$ such that

$$\begin{aligned} S\sum_{i=1}^{n}\gamma_i\ p_t(\psi_i)\ &=\ N_t\ \ \text{for all } t \text{ such that } \lambda_t^*>0\ \ \text{and} \\ S\sum_{i=1}^{n}\gamma_i\ p_t(\psi_i)\ &\leq\ N_t\ \ \text{for all } t \text{ such that } \lambda_t^*=0\ . \end{aligned} \tag{12}$$

Here we can interpret the mixing weights γ_i as the probability of assigning an item to policy ψ_i or we can view it as the fraction of the population of items that are assigned to this policy. Alternatively, as discussed above, we can assign all items a Markov random policy that selects according to state-contingent selection probabilities. If some, but not all, items are identical, we get partial simplifications of this form.

Given the piecewise linear, convex nature of the Lagrangian and the fact that subgradients are readily available, it is natural to use cutting-plane methods (see, e.g., Bertsekas et al. 2003) to solve the Lagrangian dual problem (7). Alternatively, one could use subgradient methods (as in, e.g., Topaloglu 2009 and Brown and Smith 2014), a Nelder-Mead method (as in Caro and Gallien 2007), or an LP formulation (as in Hawkins

[3] If the DM must select exactly N_t items in each period (rather than less than or equal to N_t items), we drop the nonnegativity constraint on $\boldsymbol{\lambda}$ in the dual problem (7) and the optimality conditions require the linking constraint to hold with equality in expectation for all t, regardless of the sign of the optimal λ_t^*.

2003, Adelman and Mersereau 2008, and Bertsimas and Mišić 2016); we discuss the LP formulation in more detail in EC §B.3. In Appendix A, we describe a cutting-plane method that exploits the structure of the subgradients described in Proposition 4 and exploits the separability (over items and time) in the Lagrangian dual problem. Unlike the subgradient or Nelder-Mead methods, the cutting-plane method is guaranteed to terminate in a finite number of iterations with a provably optimal solution. The cutting-plane method also provides the item-specific value functions (6) as well as the set of optimal policies and mixing weights of Proposition 4(iii). The LP formulation provides an exact solution to the Lagrangian dual and may be more efficient in problems with long time horizons and small state spaces (such as the example of Weber and Weiss 1990 in §7.1), but in our dynamic assortment and applicant screening examples, the LP formulation was typically much less efficient than the cutting plane method. For instance in the dynamic assortment problem with horizon $T = 20$, solving the Lagrangian dual as an LP took about 16 hours using a commercial LP solver (MOSEK) and exploiting the simplifications due to having identical items. In contrast, the cutting-plane method took less than 2 minutes with this example.

3.3. Applicant Screening Example

To illustrate the Lagrangian DP and the role of mixed policies, we consider the applicant screening problem described in §2.3 in the case with horizon T and Bernoulli signals. Here the DM can screen 25% of the applicants in each of the first four periods and can admit 25% in the final period. As discussed in §2.3, with these assumptions the DM must choose between screening each applicant once or screening some applicants more than once with the hope of identifying better applicants to admit. Using the cutting-plane method to solve the Lagrangian dual problem (7), we find an optimal solution with $\boldsymbol{\lambda}^* = (0.0333, 0.0333, 0.0333, 0.0333, 0.60)$ and optimal policies ψ_i with selection probabilities $p_t(\psi_i)$ and mixing weights γ_i shown in Table 1. This mixture of policies selects 25% of the applicants in each period in expectation, as required by the optimality condition (12).

		Selection probabilities by period ($p_t(\psi_i)$)					Mixing
	Policy (ψ_i)	1	2	3	4	5	weight (γ_i)
a:	Never screen	0	0	0	0	0	0.300
b:	Screen once	0	0	1	0	0.5	0.025
c:	Screen once	0	0	0	1	0.5	0.075
d:	May screen twice	1	0	0	0.5	0.333	0.250
e:	May screen twice	0	1	0.5	0	0.333	0.250
f:	May screen twice	0	0	1	0.5	0.333	0.100
	Weighted average:	0.25	0.25	0.25	0.25	0.25	1.000

Table 1: Selection probabilities for policies involved in an optimal mixture for the applicant screening example

Figure 1 depicts the mean field evolution of the screening process with the optimal mixture of policies shown in Table 1.[4] The blue rectangles represent possible applicant states in each period and the flows

[4] In the mean field limit, the system state evolves deterministically with the fractions of items making a given state transition matching the transition probability under the selected control policy. As the number of items S increases, the fraction of items in a given state will converge to the mean field limit; see, e.g., Le Boudec et al. (2007).

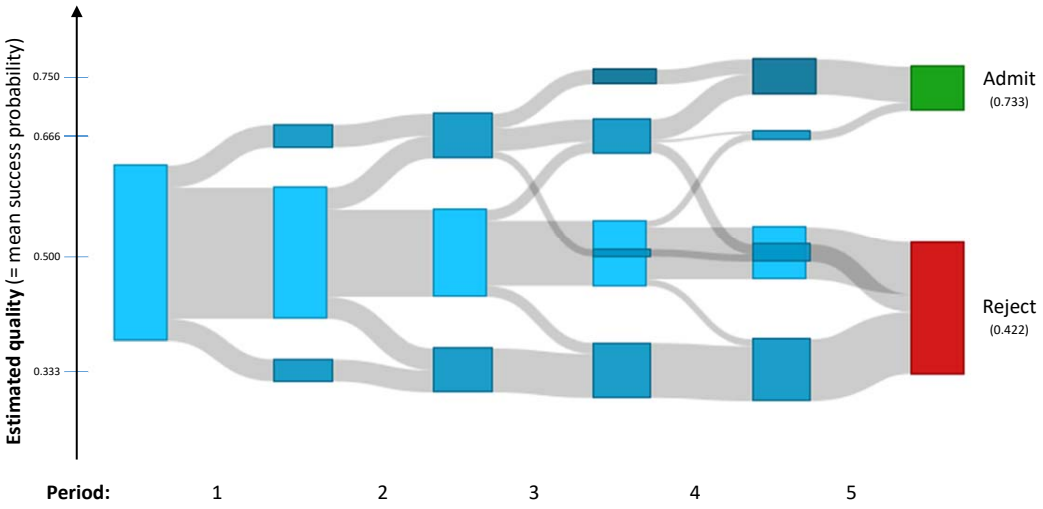


Figure 1: Optimal flows for the Lagrangian relaxation of the applicant screening example

represent state transitions; the widths of the flows represent the number of applicants making the given transition. The midpoint of each rectangle on the vertical dimension represents the expected quality for applicants in that state. Initially all applicants are unscreened and have a beta (1,1) prior, which implies an expected quality of 0.5. In the first period, the 25% of the applicants following policy (d) are screened; in expectation, half of them receive positive signals and half receive negative signals. The screened applicants then move to higher or lower states for the next period, with expected qualities equal to 0.666 and 0.333, respectively. In the second period, the 25% of applicants following policy (e) are screened and are similarly split into higher and lower states. In the third period, there is a mix of applicants being screened for the first time (from policies (b) and (f)) and a second time (from policy (e)). The last screening period ($t = 4$) also includes a mix of applicants being screened for the first and second time.

In the final period, those applicants who have received two positive signals in two screenings and those who have received one positive signal in one screening are admitted. All others are rejected. On average, 20% of those admitted have one positive signal in one screening (with expected quality 0.666) and 80% have two positive signals in two screenings (with expected quality 0.75): the Lagrangian value is $0.20 \times 0.666 + 0.8 \times 0.75 = 0.7333$ per admitted applicant. Rejected applicants have an average expected quality of 0.4222.

Though the optimal Lagrange multipliers $\boldsymbol{\lambda}^*$ in this example are unique, the optimal policies and mixing weights in Table 1 are not unique. Other optimal mixtures may, for example, involve policies that schedule follow-up screenings differently or screen some of those who will be screened once in the first or second period. Some alternative optimal solutions may induce the same flows shown in Figure 1, but others may induce different flows. However, in all optimal mixtures, the policies involved must be optimal for the item-specific DP (6) and the set of policies must be coordinated to ensure that, on average, 25% of the applicants are selected (i.e., screened or admitted) in each period, as required by the optimality condition (12).

4. Heuristic Policies

The optimal policies for the Lagrangian DP cannot be implemented because they regularly violate the linking constraint (1). For instance in the applicant screening problem, the optimal policy for Lagrangian selects N_t applicants on average, but if more applicants receive positive signals than expected, the Lagrangian policy will screen or admit more applicants than is allowed. In this section, we consider heuristic policies that respect the linking constraint in every scenario and hence can be implemented. We analyze the performance of the optimal Lagrangian index policy (introduced in §4.4) in §5 and evaluate the performance of these heuristics for the dynamic assortment and applicant screening problems in a simulation study in §6.

4.1. Index Policies

The heuristics we consider can all be viewed as index policies. In an index policy, we calculate a *priority index* $i_{t,s}(x_s)$ that indicates the relative attractiveness of selecting item s in period t when the item is in state x_s. Given priority indices for all items, the policies proceed as follows: (a) if there are more than N_t items with nonnegative indices, select the N_t items with the largest indices; (b) otherwise, select all items with nonnegative indices.[5] The linking constraints will thus be satisfied and these index policies will be feasible for the dynamic selection problem (2). We will generally break ties among items with the same priority index randomly, with the exception of the optimal Lagrangian index policy described in §4.4.

The indices we consider all approximate the value added by selecting item s in period t when the item is in state x_s,

$$i_{t,s}(x_s) = (r_{t,s}(x_s,1) + \mathbb{E}[\, W_{t+1,s}(\widetilde{\chi}_{t,s}(x_s,1))\,]) - (r_{t,s}(x_s,0) + \mathbb{E}[\, W_{t+1,s}(\widetilde{\chi}_{t,s}(x_s,0))\,])\ , \tag{13}$$

using some item-specific approximation $W_{t+1,s}$ of the next-period value function. We generate different heuristic policies by considering different approximate value functions. For example, the *Lagrangian index policy for* $\boldsymbol{\lambda}$ takes the approximate value function $W_{t+1,s}(x_s)$ to be the item-specific value function $V^{\boldsymbol{\lambda}}_{t+1,s}(x_s)$ given by (6). The *myopic policy* simply takes $W_{t+1,s}(x_s) = 0$.

Though we describe these heuristics as index policies, we can also view these heuristics as being "greedy" with respect to an approximate value function $W_t(\boldsymbol{x}) = \sum_{s=1}^{S} W_{t,s}(x_s)$. That is, in each period, the DM solves an optimization problem that respects the linking constraint and uses this function to approximate the continuation value:

$$\max_{\boldsymbol{u}\in\mathcal{U}_t} \left\{ r_t(\boldsymbol{x},\boldsymbol{u}) + \mathbb{E}[\, W_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x},\boldsymbol{u}))\,] \right\}. \tag{14}$$

Ranking items by priority index and selecting N_t items with the largest (nonnegative) indices solves the optimization problem (14) exactly. In the case of the Lagrangian index policy, the approximate value function $W_{t+1}(\boldsymbol{x})$ differs from the Lagrangian value function $L^{\boldsymbol{\lambda}}_{t+1}(\boldsymbol{x})$ by a constant. Thus a Lagrangian index policy can be viewed as using the Lagrangian as an approximate value function (as in Hawkins 2003

[5] If the DM must select exactly N_t items, we select the N_t items with the largest indices.

and Adelman and Mersereau 2008).

4.2. Whittle Index Policy

The Whittle index policy (Whittle 1988) can be seen as a variation of the Lagrangian index policy where the Lagrange multipliers are assumed to be constant over time (i.e., $\lambda_t = w$ for all t or $\boldsymbol{\lambda} = w\mathbf{1}$ where $\mathbf{1}$ is a T-vector of ones) and w is a breakeven Lagrange multiplier for the given period and state. Specifically, the Whittle index $i_{t,s}(x_s)$ is the w that makes the DM indifferent between selecting and not selecting an item,

$$r_{t,s}(x_s,1) - w + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s,1))\,\big] \;=\; r_{t,s}(x_s,0) + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s,0))\,\big]$$

or, equivalently, in the form of (13),

$$w = \big(r_{t,s}(x_s,1) + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s,1))\,\big]\big) - \big(r_{t,s}(x_s,0) + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s,0))\,\big]\big)\;. \tag{15}$$

The intuition behind this follows that of the Gittins index for the standard multiarmed bandit problem: the breakeven Lagrange multiplier represents the most the DM would be willing to pay for use of the constrained resource and the policy prioritizes by this willingness to pay.

It is important to note that these Whittle indices may not be well defined. For example, Whittle (1988) describes an example where some items are not "indexable" because there are multiple w satisfying (15). Even when well defined, these Whittle indices can be very difficult to compute exactly: to find the breakeven w for a state x_s in period t, we must repeatedly solve the item-specific DPs (6) with $\boldsymbol{\lambda} = w\mathbf{1}$ with varying w to identify the breakeven w. If we want to calculate indices for all periods and states, we can streamline this process by using a parametric approach (see EC §C.1 for details), but this still essentially requires solving item-specific DPs once for each period and state. As mentioned in §1.1, Whittle (1988) conjectured that the Whittle index policy is asymptotically optimal for restless bandit problems when the items are all indexable; this conjecture was shown to be false by Weber and Weiss (1990). We will discuss Whittle's conjecture and Weber and Weiss's counterexample in more detail in §7.1.

Caro and Gallien (2007) showed that Whittle indices are well defined in the dynamic assortment problem (i.e., the model is indexable) and noted that computing the indices is a "complicated task." Rather than using these hard-to-compute Whittle indices, Caro and Gallien (2007) proposed an approximate index that is based on approximating the expected continuation values in (15) with a one-step lookahead value function and a normal distribution. In our numerical examples for the dynamic assortment problem in §6, we will focus on exact Whittle indices but will briefly describe some results for Caro and Gallien's approximation.

In the applicant screening problem, the Whittle indices are also well defined but, perhaps surprisingly, are not helpful in determining applicants to screen. In period T, the Whittle index for any applicant is the applicant's mean quality (i.e., the expected reward for admitting the applicant). In all earlier (i.e., screening) periods, however, the Whittle index for every applicant equals zero, regardless of the state x_s

of the applicant. Intuitively, $w = 0$ is the Whittle index for screening periods because with $w = 0$ (a) all applicants would be admitted in the final period and (b) given this, it does not matter whether an applicant is screened or not in any period because the information provided by screening does not affect the admission decision or value obtained; thus (15) is satisfied with $w = 0$. (See Proposition 7 in EC §C.2 for a formal statement and proof of this claim.)

Although this failure of the Whittle index policy initially surprised us, it perhaps should not have been surprising: the setting here – with a finite horizon and time-varying rewards – is quite far removed from the classical multiarmed bandit where these index policies are optimal and also quite different from the infinite-horizon stationary restless bandits that Whittle (1988) considered.

4.3. Modified Whittle Index Policy

Given a model with time-varying rewards, constraints, and/or state transitions, it seems natural to consider Lagrange multipliers that are varying over time rather than constant over time, as assumed in the Whittle index. Accordingly, we define a *modified Whittle index* of this sort. The indices are calculated recursively. To find the index $m_{t,s}(x_s)$ for period t and state x_s, we set all future Lagrange multipliers λ_τ (for $\tau > t$) to be equal to the previously calculated period-τ indices, i.e., $\boldsymbol{m} = (m_{t+1,s}(x_s), \ldots, m_{T,s}(x_s))$ for this same state x_s. We then take

$$m_{t,s}(x_s) = \big(r_{t,s}(x_s, 1) + \mathbb{E}\big[\, V^{\boldsymbol{m}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big]\big) - \big(r_{t,s}(x_s, 0) + \mathbb{E}\big[\, V^{\boldsymbol{m}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 0))\,\big]\big)\ . \tag{16}$$

The vector $(m_{1,s}(x_s), \ldots, m_{T,s}(x_s))$ of modified Whittle indices for a given state x_s can thus be calculated using a recursive procedure that is similar to solving one item-specific DP (6).

These modified Whittle indices are thus much easier to calculate than the standard Whittle index. The modified Whittle indices require effort on the order of solving one item-specific DP per state, whereas the standard Whittle indices require solving one DP per state, per period. Moreover, indexability is not an issue with the modified Whittle indices because the period-t index is uniquely defined by (16).[6]

In our dynamic assortment examples, the modified Whittle index policies appear to outperform the Whittle index policies in problems with short time horizons; the two policies tend to perform similarly with longer time horizons. In the applicant screening examples, with our specific numerical assumptions, the modified Whittle index policy prioritizes screening unscreened applicants, so it recommends screening every applicant once; this is true for both Bernoulli ($n = 1$) and binomial ($n = 5$) signal processes. However, with other prior distributions or constraints, the modified Whittle index policy may give higher priority to applicants who have been previously screened than those who have not yet been screened.

[6] Note that the definition of the modified Whittle indices implicitly assumes that the state space for the items is constant or growing over time: when calculating the index $m_{t,s}(x_s)$, we reference indices $(m_{t+1,s}(x_s), \ldots, m_{T,s}(x_s))$ for future periods for this same state x_s. This assumption is true in all of the examples that we consider.

4.4. The Optimal Lagrangian Index Policy

Although we can define a Lagrangian index policy for any $\boldsymbol{\lambda}$, intuitively, we might expect Lagrange multipliers $\boldsymbol{\lambda}$ that lead to better performance bounds would lead to better approximate value functions and tend to generate better heuristics. We will show that the Lagrange multipliers $\boldsymbol{\lambda}^*$ that solve the Lagrangian dual problem (7), do in fact generate an index policy that is asymptotically optimal (in a sense to be made precise in §5), but we need to take care when breaking ties if there are items with equal priority indices. Recall that, in the Lagrangian relaxation, optimal policies are typically mixed policies where the mixing coordinates actions across items to ensure that N_t items are selected on average in each period (assuming $\lambda_t^* > 0$; see Proposition 4(iii)). Our proposed tiebreaking scheme for the Lagrangian index policy mimics this mixing to coordinate actions in the heuristic.

To illustrate the importance of tiebreaking, consider implementing the Lagrangian index policy for $\boldsymbol{\lambda}^*$ in the applicant screening example discussed in §3.3. In the first period, all applicants are in the same state and have the same priority index. In this first period, it does not matter which applicants are screened so long as N_t are selected. In later screening periods, some applicants will have been screened before and the priority indices are equal for (i) those applicants who have been screened once and had a positive signal and (ii) those who have not been screened before. In both states, the priority indices are equal to the Lagrange multiplier ($\lambda_t = 0.0333$) because screening and not screening are both optimal actions in these states in the Lagrangian DP. Here, tiebreaking is important. If we consistently break ties in favor of screening unscreened applicants, all applicants will be screened once and in the final period the DM will choose applicants to admit from the many applicants with a single positive signal. Consistently breaking ties in favor of rescreening applicants with a positive signal is also not ideal.

In this applicant screening example, the ties are a result of there being multiple optimal policies for the Lagrangian DP. As discussed in §3.2, the optimal Lagrange multipliers $\boldsymbol{\lambda}^*$ will typically lead to multiple optimal policies for the relaxed DP (4). Whenever there are multiple optimal policies, there must be *indifference states* – like those in the applicant screening example – where selecting and not selecting are both optimal and the selection index is equal to that period's Lagrange multiplier λ_t. If there are two such indifference states in the same period, then items in these two states will be tied. It is difficult to predict how many indifference states there will be, how these indifference states will be allocated over time, and how likely ties will be. In the applicant screening example with $T = 5$ and Bernoulli signals, ties are common and, as we will see in our numerical experiments, tiebreaking is important; in the Bernoulli case with $T = 51$, tiebreaking is even more important. In the applicant screening example with $T = 5$ and binomial signals (with $n = 5$), applicants wind up being more spread out over the state space and ties occur but less frequently than with Bernoulli signals ($n = 1$); tiebreaking plays a role but is less important than in the Bernoulli case. In the dynamic assortment examples, there are many indifference states but they tend to be spread out over time and tiebreaking makes little or no difference.

Given an index policy π defined by priority indices $i_{t,s}(x_s)$, we can define a new index policy π' that *uses*

a policy $\psi = (\psi_1, \ldots, \psi_S)$ *as a tiebreaker* by defining a new index

$$i'_{t,s}(x_s) = i_{t,s}(x_s) - \epsilon \cdot (1 - \psi_{t,s}(x_s)) \ , \tag{17}$$

for a small $\epsilon > 0$. Here ϵ is chosen to be small enough (e.g., smaller than the smallest difference between unique values of the original indices $i_{t,s}(x_s)$) so the tiebreaker does not change the rankings of items that do not have the same index value. With this modified index, ties will be broken to match the choice with policy ψ_s: items not selected by ψ_s in a given period/state are penalized slightly, so they will "lose" on this tiebreaker. Also note that items with an original priority index $i_{t,s}(x_s)$ equal to zero will not be selected with this new index policy if ψ_s does not select the item. We break any remaining ties randomly.

We define an *optimal Lagrangian index policy* $\widetilde{\pi}$ as a Lagrangian index policy for $\boldsymbol{\lambda}^*$ that uses an optimal mixed policy $\widetilde{\psi}$ for the Lagrangian dual problem (7) as a tiebreaker. Note that with the optimal Lagrangian index policy, the only states where tiebreaking is relevant are the indifference states where the selection indices $i_{t,s}(x_s)$ are equal to the λ^*_t. If $i_{t,s}(x_s) > (<)\ \lambda^*_t$, then all optimal policies for the Lagrangian relaxation (6) will select (not select) the item and all tied items will have the same index value $i'_{t,s}(x_s)$, after taking into account the tiebreaker as in (17).

We can generate a mixed policy $\widetilde{\psi}$ for tiebreaking using any of the three methods discussed after Proposition 4:

- Simple random mixing: independently randomly assign each item s a policy ψ_s according to the mixing weights of Proposition 4(iii) in each scenario.
- Markov random mixing: $\psi_{t,s}(x_s)$ in (17) is randomly selected from $\{0, 1\}$ with state-dependent probabilities given in EC §B.2.
- Proportional assignment: if some or all of the items are identical, we can sometimes construct a non-random tiebreaking policy ψ where items are assigned different policies with proportions reflecting the desired mixing weights.

In our numerical examples, we will generate tiebreaking policies ψ_s using proportional assignments, using simple random mixing to allocate non-integer remainders when necessary. For instance in the applicant screening problem with the optimal policy mixture in Table 1, if S=1000, we assign (300, 25, 75, 250, 250, 200) applicants to the 6 policies listed in Table 1. If S=100, the desired proportions are not integers, so we randomize, assigning (30, 3, 7, 25, 25, 20) or (30, 2, 8, 25, 25, 20) items to these 6 policies, each 50% of the time. In §6, we use proportional assignments because it reduces the uncertainty in the model and seems to lead to slightly better performance (see §6.4).

5. Analysis of the Optimal Lagrangian Index Policy

In this section, we characterize the performance of the optimal Lagrangian index policy and study asymptotic properties as we grow the size of the problem. The main result is the following proposition that relates the performance of the optimal Lagrangian index policy to the Lagrangian bound. Here we let $\bar{r}$ and $\underline{r}$ denote

upper and lower bounds on the rewards (across all items, states, periods, and actions) and let $N = \max_t\{N_t\}$.

Proposition 5. *Let $\boldsymbol{\lambda}^*$ denote an optimal solution for the Lagrangian dual problem (7) with initial state $\boldsymbol{x}$. Let $\widetilde{\psi}$ denote an optimal mixed policy for this Lagrangian and $\widetilde{\pi}$ an optimal Lagrangian index policy that uses $\widetilde{\psi}$ as a tiebreaker. Then*

$$L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}) - V_1^{\widetilde{\pi}}(\boldsymbol{x}) \;\le\; \underbrace{(\bar{r}-\underline{r})\sum_{t=1}^{T}\beta_t\sqrt{\bar{N}_t(1-\bar{N}_t/S)}}_{\equiv\,\Delta^{\widetilde{\psi}}(\boldsymbol{x})} \;\le\; (\bar{r}-\underline{r})\sum_{t=1}^{T}\beta_t\sqrt{N}, \tag{18}$$

where $\bar{N}_t$ is the expected number of items selected by $\widetilde{\psi}$ in period t ($\bar{N}_t = N_t$ if $\lambda_t^* > 0$ and $\bar{N}_t \le N_t$ if $\lambda_t^* = 0$), and the β_t are nonnegative constants that depend only on t and T.

Proof. See EC §D.1. □

The proof of Proposition 5 considers the states $\widetilde{\boldsymbol{x}}_t$ visited using the policy $\widetilde{\psi}$ that is optimal for the Lagrangian relaxation and characterizes the differences in rewards generated by $\widetilde{\psi}$ and those generated by the corresponding optimal Lagrangian index policy $\widetilde{\pi}$. The key observations in the proof are:

- The selection decisions made by the heuristic $\widetilde{\pi}$ are based on priority indices that are *aligned* with the decisions made by $\widetilde{\psi}$. Let n_t denote the number of items selected by the relaxed policy $\widetilde{\psi}$ in period t in a particular state; this may be larger or smaller than N_t. From (6) and (13) and taking into account the tiebreaking rule (17), we see that items with priority indices $i'_{t,s}(x_s) \ge (<)\ \lambda_t$ will (will not) be selected by $\widetilde{\psi}$. If $n_t < N_t$ items are selected by $\widetilde{\psi}$, then these n_t items will be among the N_t items with the largest selection indices and will also be selected by $\widetilde{\pi}$. If $n_t \ge N_t$ items are selected by $\widetilde{\psi}$, then $\widetilde{\pi}$ will select a subset of size N_t of those selected by $\widetilde{\psi}$. In both cases, the number of items with different decisions is bounded by $\left|n_t - N_t\right|$. Note that the tiebreaker is essential in ensuring alignment when there are ties in the original indices.
- Let $\tilde{n}_t$ represent the random number of items selected when using the relaxed policy $\widetilde{\psi}$. With an optimal policy $\widetilde{\psi}$ for the Lagrangian and $\lambda_t^* > 0$, the difference $\tilde{n}_t - N_t$ has zero mean (by Proposition 4(iii)) and the expectation of $\left|\tilde{n}_t - N_t\right|$ is bounded by a standard deviation term of the form $\sqrt{N_t(1-N_t/S)}$. The assumptions that the state transitions and the mixing of policies are independent across items ensure that the standard deviations grow with $\sqrt{N}$.
- The β_t terms in (18) reflect the maximum possible number of changes in item states caused by the selection decision of $\widetilde{\pi}$ deviating from $\widetilde{\psi}$ for a single item in period t. These β_t terms grow with the number of periods remaining as a change in decision in period t can have downstream implications on decisions and state transitions for other items in later periods. Specifically, with an index policy, changing the state (hence the index) of one item may affect the selection decisions for two items, as the changed item may become one of the top N_t items and be selected, thereby forcing another item out of the top N_t (or vice versa). In the worst case, this doubling of changed states can cascade through

all remaining periods and thus

$$\beta_t = 1 + 2 + 2^2 + \ldots + 2^{T-t} = 2^{T-t+1} - 1 \ . \tag{19}$$

This implies that the $\sum_{t=1}^{T} \beta_t$ term in (18) is equal to $2(2^T - 1) - T$.

- Finally, the $(\bar{r} - \underline{r})$ terms provide an upper bound on the possible loss in value caused by the state of a single item under $\widetilde{\pi}$ deviating from the state under $\widetilde{\psi}$ in single period t. This upper bound reflects the possibility that the DM may earn the minimum reward $\underline{r}$ rather than the maximum reward $\bar{r}$ as a result of the change in state.

This bound may seem quite conservative, but we will see that in the applicant screening examples, the gap $L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}) - V_1^{\widetilde{\pi}}(\boldsymbol{x})$ appears to grow with $\sqrt{N}$. Moreover, we have developed simple analytic examples where the gap between the Lagrangian and optimal Lagrangian policies asymptotically grows with $\sqrt{N}$; see EC §D.2. Thus $\sqrt{N}$ is the best possible growth rate for these performance gaps for general dynamic selection problems.[7]

We can use Proposition 5 to relate the performance of the optimal Lagrangian value function, the rewards generated by the corresponding optimal Lagrangian index policy, and the optimal value function $V_1^*(\boldsymbol{x})$.

Theorem 1 (Performance guarantees)**.** *In the setting of Proposition 5,*

$$V_1^*(\boldsymbol{x}) - \Delta^{\widetilde{\psi}}(\boldsymbol{x}) \ \leq \ L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}) - \Delta^{\widetilde{\psi}}(\boldsymbol{x}) \ \leq \ V_1^{\widetilde{\pi}}(\boldsymbol{x}) \ \leq \ V_1^*(\boldsymbol{x}) \ \leq \ L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}) \ . \tag{20}$$

Proof. The second inequality was established in Proposition 5. Proposition 1 implies the first and last inequalities. The remaining inequality (the third one) follows from the fact that $\widetilde{\pi}$ is feasible for the DP (2), i.e., it satisfies the linking constraint (1). □

Since $\Delta^{\widetilde{\psi}}(\boldsymbol{x})$ is bounded by a term that grows with $\sqrt{N}$, Proposition 5 and Theorem 1 provide insight into the asymptotic performance of the optimal Lagrangian index policy and bound for large problems. In our numerical experiments in §6, we consider problems where the items are all identical and we increase S and N_t in proportion. The next result establishes asymptotic optimality for large problems in a more general setting. Specifically, we consider a sequence of dynamic selection problems where we expand the set of items available (indexing these sets by their cardinality S) and simultaneously increase the number of items $N_t(S)$ that may be selected in period t, while holding the time horizon T constant.

Corollary 1 (Asymptotic optimality)**.** *Consider a growing sequence of dynamic selection problems (indexed by S) and let $V_t^*(\boldsymbol{x}; S)$, $L_t^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S)$ and $V_t^{\widetilde{\pi}}(\boldsymbol{x}; S)$ denote the corresponding optimal value functions, values for the optimal Lagrangian, and value for the corresponding optimal Lagrangian index policy $\widetilde{\pi}$. If the $V_1^*(\boldsymbol{x}; S)$ are positive and satisfy*

$$\lim_{S \to \infty} \frac{V_1^*(\boldsymbol{x}; S)}{\sqrt{N(S)}} = \infty, \tag{21}$$

[7] The result of Proposition 5 also applies in the case where the DM must select exactly N_t items, but we take $\bar{N}_t = N_t$ regardless of the sign of λ_t^*. Theorem 2 and Corollary 1 below follow with no additional changes.

then

$$\lim_{S\to\infty} \frac{L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};S) - V_1^{\widetilde{\pi}}(\boldsymbol{x};S)}{V_1^*(\boldsymbol{x};S)} = 0 \ . \tag{22}$$

Since $V_1^{\widetilde{\pi}}(\boldsymbol{x}) \le V_1^*(\boldsymbol{x}) \le L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x})$, (22) *implies*

$$\lim_{S\to\infty} \frac{V_1^*(\boldsymbol{x};S) - V_1^{\widetilde{\pi}}(\boldsymbol{x};S)}{V_1^*(\boldsymbol{x};S)} = 0 \quad \textit{and} \quad \lim_{S\to\infty} \frac{L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};S) - V_1^*(\boldsymbol{x};S)}{V_1^*(\boldsymbol{x};S)} = 0 \ .$$

Proof. See EC §D.1. □

This corollary implies that, when the growth condition (21) is satisfied, the gaps between $V_1^*(\boldsymbol{x};S)$, $L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};S)$ and $V_1^{\widetilde{\pi}}(\boldsymbol{x};S)$, when normalized by $V_1^*(\boldsymbol{x};S)$, all converge to zero. Therefore, we can view both the optimal Lagrangian index policy and the Lagrangian bound as being asymptotically optimal in this sense. The growth condition (21) is mild. For example, if the expected reward associated with selecting an item is bounded away from zero and $\lim_{S\to\infty} N_t(S) = \infty$, then (21) will be satisfied. We could normalize the ratios in Corollary 1 by the Lagrangian $L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};S)$ rather than $V_1^*(\boldsymbol{x};S)$ (because $V_1^*(\boldsymbol{x};S) \le L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};S)$) and find these ratios also converge to zero. Finally, if we are adding identical items and increasing S and N_t in proportion (as we will in §6.2), the Lagrangian increases in proportion to S and N_t and we can normalize by S or N_t and again find the ratios converge to zero.

6. Numerical Examples

In this section, we evaluate the performance of the heuristic policies considered in §4 in the context of the dynamic assortment and applicant screening problems. Specifically we consider: (i) the myopic policy, (ii) the Whittle index policy, (iii) the modified Whittle index policy, (iv) the Lagrangian index policy for an optimal solution $\boldsymbol{\lambda}^*$ to the Lagrangian dual (7) which randomly breaks ties among items with the same priority index, and (v) an optimal Lagrangian index policy, which breaks ties as discussed in §4.4. As discussed in §2.2-2.3, we consider three versions of the dynamic assortment problem (with horizon T equal to 8, 20 and 40) and three versions of the applicant screening problem (with $T = 5$ and binomial signal with $n = 1$ and 5 as well as a case with $T = 51$ and $n = 1$). We will vary the number of items considered (S) in all cases.

6.1. Run Times

To implement the Whittle, modified Whittle and Lagrangian index policies, we must first calculate their respective indices; Table 2 reports the times required to calculate these indices for all states for each example. All calculations were performed using Matlab on a personal computer.[8] In these examples, the items are identical so we need only calculate indices for a single item, regardless of the number of items S considered.

In these index calculations, the run times are dominated by the time required to solve the item-specific DPs (6). The time to required to solve these DPs is primarily determined by the number of states that

[8] Detailed specifications for the computer: 64-bit Intel Xeon E5-2697 v4 (2.30 GHz) CPU; 64.0 GB of RAM; running Windows 10 Enterprise, Matlab R2016b. We used MOSEK (Version 7.1.0.60) within Matlab to solve the LP (33) in the cutting-plane method when calculating Lagrangian indices.

	Dynamic Assortment			Applicant Screening		
	$T = 8$	$T = 20$	$T = 40$	$n = 1$ $T = 5$	$n = 5$ $T = 5$	$n = 1$ $T = 51$
Run times (seconds)						
Whittle	24.0	7,039	904,989	0.0073	0.0171	85.7
Modified Whittle	8.8	982	47,387	0.0024	0.0100	0.71
Lagrangian	0.9	126	2,716	0.0157	0.0179	3.79
States in item-specific DP	12,636	199,710	1,599,820	35	115	23,426
Cutting plane iterations	70	530	826	14	16	540

Table 2: Run times, problem sizes, and related statistics for index calculations

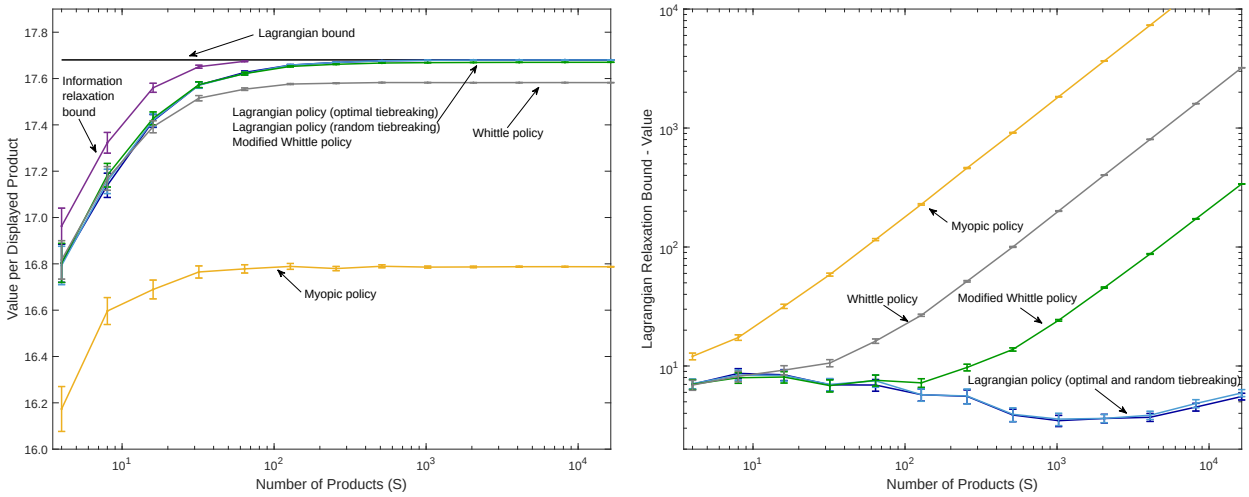
must be considered (also shown in Table 2). In problems with a fixed state space (such as Weber and Weiss 1990's example discussed in §7.1), the time required to solve the item-specific DPs will grow linearly with T. In the dynamic assortment and the applicant screening problem, the possible state space in period t grows quadratically in t (e.g., in the dynamic assortment problem, the number of possible α_s values grows linearly, as does the number of possible m_s values), so the computational effort in the item-specific DPs scales with T^3. The time required to compute the Whittle indices grows with T^6 (one must solve an item-specific DP with $\sim T^3$ states for each of $\sim T^3$ states). The cutting plane method used in the Lagrangian index calculation requires repeatedly solving these DPs, once in each iteration of the algorithm: the number of iterations required to find an optimal solution is hard to predict but typically increases with the the horizon T, which corresponds to the dimension of the Lagrange multiplier vector $\boldsymbol{\lambda}$ that is being optimized.

In the dynamic assortment examples, we find that with $T = 20$, the Whittle indices require about 2 hours to compute, the modified Whittle indices require about 16 minutes, and the Lagrangian indices require about two minutes. The differences are more pronounced in the $T = 40$ case: the Whittle indices require 10.5 days to compute whereas the Lagrangian indices require about 45 minutes. In the applicant screening examples, the item-specific DPs are much simpler and the calculations take much less time.

6.2. Simulation Results

Figures 2-5 describe the performance of the heuristic policies with the number of items S (products or applicants) equal to 4, 8, 16, ..., 16,384 ($= 2^{14}$). In all cases, we scale N_t (the number of products displayed or applicants screened/admitted) with S, taking N_t to be a fixed proportion of S. Note the horizontal axes in the figures showing S are plotted on a log scale. The heuristics are evaluated using simulation, with a sample of 1000 trials. The samples are common across heuristics: for any given S, the products have the same randomly generated demands (and applicants have the same signals) for all policies. The expected total rewards $V_1^{\pi}(\boldsymbol{x})$ for the policies are estimated from these simulations and adjusted using a control variate based on the Lagrangian; see (56) in EC §E. The error bars in the figures represent 95% confidence intervals for these estimated values. The Lagrangian bounds $L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x})$ are calculated exactly.

The (a) panels of Figures 2-5 show the relative performance of the heuristics, normalizing the total reward by dividing by the total number of products displayed in the assortment examples and by the number of applicants admitted in the screening examples. The Lagrangian bound scales linearly with S and, hence, is



(a) Relative performance

(b) Performance gaps

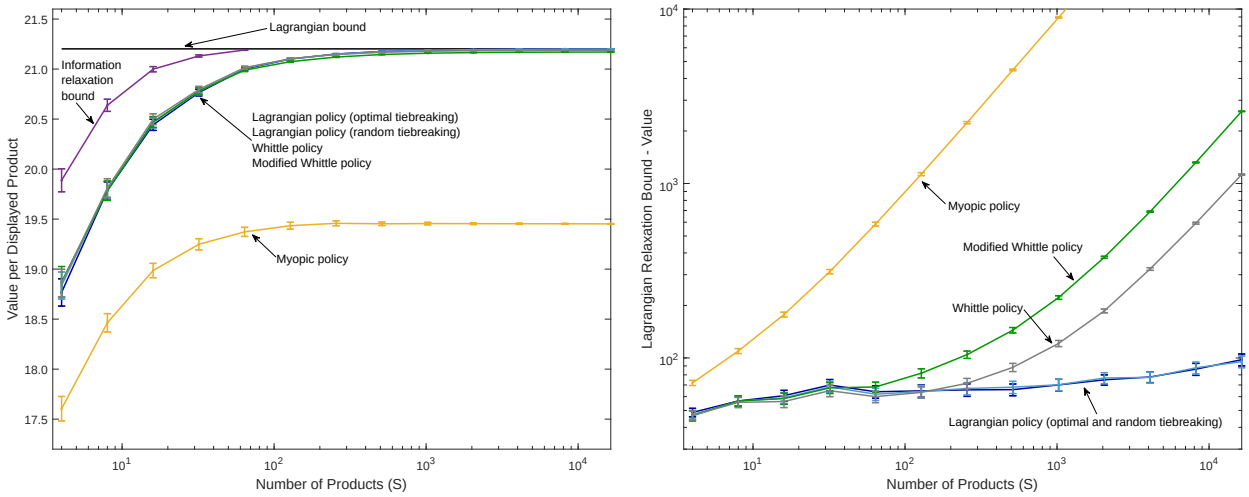
Figure 2: Results for the dynamic assortment examples with horizon T=8

constant when normalized. The (b) panels of these figures show estimates of the performance gap for the index policies, $L_1^{\lambda^*}(\boldsymbol{x}) - V_1^{\pi}(\boldsymbol{x})$, where the estimates of these gaps are plotted on a log scale.

Dynamic Assortment Examples. In the dynamic assortment examples with $T = 8$, in Figure 2(a) we see that the myopic policy is the worst of the heuristics considered. Intuitively, the myopic policy fails to explore enough to find the best products to display. The other heuristics – the two versions of the Whittle index policies and the two versions of the Lagrangian index policy – all perform similarly for small S. For large S, the Whittle index policies are significantly below the Lagrangian bound whereas the two Lagrangian bounds and the modified Whittle index appear to approach the Lagrangian bound. If we look at the performance gaps in Figure 2(b) in absolute terms rather than relative terms, we see that the gaps for both Whittle index policies grow linearly in S (linear growth corresponds to a slope of one in the log-log plot). In contrast, the performance gaps for the Lagrangian index policies grow sublinearly. This implies that in Figure 2(a), the modified Whittle index policy approaches an asymptote below the Lagrangian bound, whereas the two Lagrangian index policies truly approach the Lagrangian bound. In this example, there is no difference between the two Lagrangian index policies because there are no scenarios where products in different states have the same priority indices, so the tiebreaking rules do not matter.

Note that the optimal Lagrangian index policies perform *very well* for large S. For example with S=16,384, the total reward for the optimal Lagrangian policy is approximately \$579,348 (with a mean standard error of \$0.18) and the Lagrangian bound is \$579,354; this implies the optimal Lagrangian index policy is within \$6 of the optimal value!

Figures 3(a) and (b) are like Figures 2(a) and (b), but consider horizon $T = 20$ rather than $T = 8$. The results are similar, but the Whittle index policy fares somewhat better: the Whittle index policy outperforms the modified Whittle index policy for large S, but again both exhibit linear growth in the performance gap.



(a) Relative performance

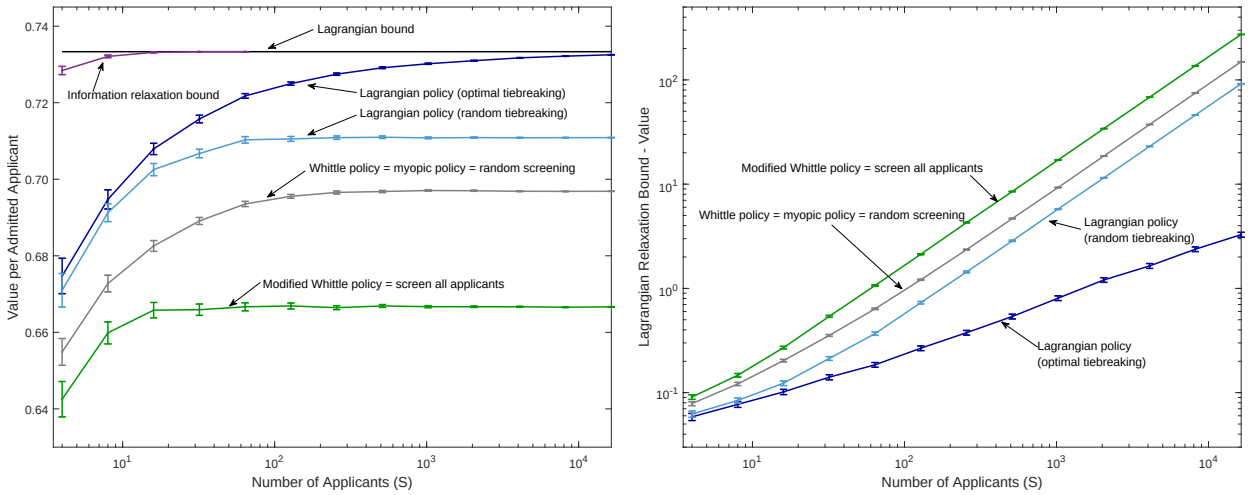
(b) Performance gaps

Figure 3: Results for the dynamic assortment examples with horizon T=20

The performance gaps for the Lagrangian index policies again grow sublinearly. With $S = 16,384$, the total reward for the optimal Lagrangian index policy is approximately \$1,736,761 (with a mean standard error of \$4) and the Lagrangian bound is \$1,736,858, so the optimal Lagrangian index policy is within \$97 of the optimal value. The results for the case with $T = 40$ are similar and are provided in EC §F.

Finally, although we do not show these results in Figures 2 and 3, we also simulated the one-period look-ahead/normal approximation of the Whittle index developed by Caro and Gallien (2007) (see §4.2) on these assortment planning examples. The performance was similar to that of the Whittle index: on the assortment planning examples with $T = 8$, we found that Caro and Gallien's approximate Whittle index policy performs approximately 0.2% worse on average than the Whittle index policy, ranging from 0.17% to 0.21% for the different values of S. For the assortment planning examples with $T = 20$ and $T = 40$, we found little difference in performance for the exact and approximate Whittle indices.

Applicant Screening Examples. The performance of the heuristics is more varied in the applicant screening problem. We first consider the case with $T = 5$ and Bernoulli signals ($n = 1$). In Figure 4(a), we see that all of the heuristic policies other than the optimal Lagrangian index policy approach an asymptote below the Lagrangian bound. As discussed in §4.2, the modified Whittle index policy here reduces to screening every applicant once, which typically leaves the DM choosing applicants to admit from those who receive a positive signal when screened; for large S, this has an expected value of 0.666 per applicant admitted. (With small S, there is some chance that fewer than 25% of the applicants will receive a positive signal so the expected value is less than 0.666 per applicant admitted.) As discussed in §4.2, the Whittle indices during the screening stages are all zero, so the Whittle index policy reduces to randomly selecting applicants to screen. Since the rewards are zero during the screening periods, the myopic policy also reduces to random screening. This random screening policy outperforms "screen all applicants" (as suggested by



(a) Relative performance

(b) Performance gaps

Figure 4: Results for the applicant screening examples with $T = 5$ and Bernoulli signals (n=1)

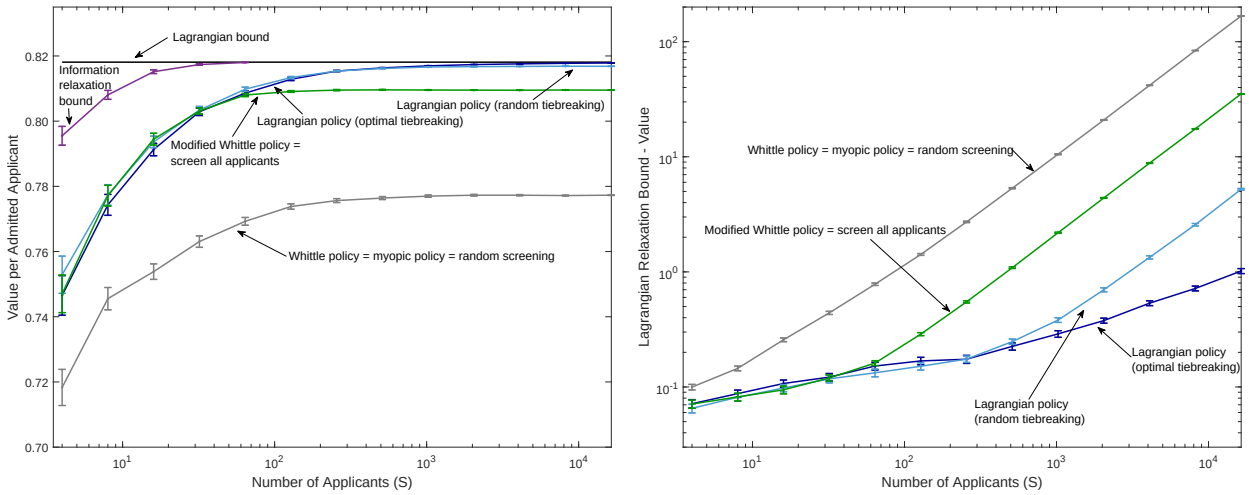
the modified Whittle index policy) because it generates some applicants with two or more positive signals who will be preferred to those with a single positive signal. The difference between the Lagrangian index policies with optimal and random tiebreaking highlights the importance of tiebreaking, as discussed in §4.4. In Figure 4(b), we see that the performance gaps grow linearly in S for all of the heuristics other than the optimal Lagrangian index policy, as we would expect given the results in Figure 4(a). The performance gap for the optimal Lagrangian index policy appears to grow with $\sqrt{S}$ (the line has slope 0.5 in the log-log plot) which is consistent with our theoretical analysis in §5.

Figures 5(a) and (b) show the same results for the case with $T = 5$ and binomial signals where $n = 5$. Here the results are similar but the policy that screens all applicants (as suggested by the modified Whittle index policy) outperforms random screening (as suggested by the standard Whittle index policy). With $n = 5$, the signals are much more informative and screening all applicants gives the DM more information about the applicants than in the Bernoulli case. For large S, "screen all applicants" is still worse than the Lagrangian index policies. The difference between the two tiebreaking methods in the Lagrangian index policy is also smaller here, as ties are less common with the more informative signals. However the performance gap for the random tiebreaking Lagrangian index policy still grows linearly for large S.

The results for the case with $T = 51$ and Bernoulli signals are similar to those with $T = 5$ and Bernoulli signals and are provided in Figure 8 of EC §F. In this case, proper tiebreaking makes a big difference.

6.3. Information Relaxation Bounds

In these numerical examples, the gaps between the optimal Lagrangian index policy and Lagrangian bound are very small (in relative terms) for large S, but are more substantial for small S. One might wonder whether these gaps are due to the policies being suboptimal or due to slack in the Lagrangian bound. In EC §E, we consider the use of information relaxations (e.g., Brown et al. 2010) with dynamic selection



(a) Relative performance

(b) Performance gaps

Figure 5: Results for the applicant screening examples with $T = 5$ and binomial signals (n=5)

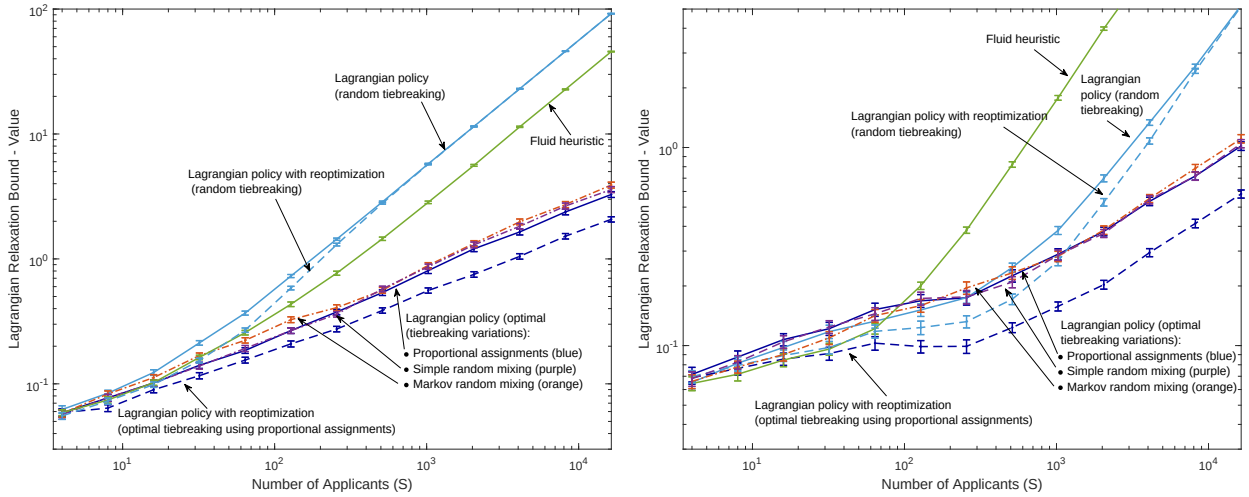
problems. These information relaxation bounds (i) relax the nonanticipativity constraints in the DP that require the DM to make decisions based only information known at the time the decision is made and (ii) impose a penalty that "punishes" the DM for violating these constraints. In the assortment planning example, we consider an information relaxation where demands for all products are known in advance. In the applicant screening example, we consider an information relaxation where all signals are known in advance. In both cases, we consider penalties based on the Lagrangian approximation of the value function. We show that these information relaxation bounds are guaranteed to (weakly) improve on the Lagrangian bounds. Lagrangian relaxations and the cutting plane method of Appendix §A play important roles in the analysis and computation.

In our numerical examples, these information relaxation bounds are shown in the (a) panels of Figures 2-5. In these results, we see that the information relaxation bounds improve on the Lagrangian dual, particularly when S is small. The improvement is greatest in the dynamic assortment example with $T = 8$ and $S = 4$. In this case, the Lagrangian bound ensures that the Lagrangian index policy is within (approximately) \$0.88 per product displayed of the value given by an optimal solution. The information relaxation bound tells us that the Lagrangian index policy is in fact within \$0.16 per product displayed of an optimal solution. These results are discussed in more detail in EC §E.

6.4. Variations on the Heuristics

Figure 6 shows results for several variations on the heuristics discussed above, focusing on the applicant screening problem. The format of the figure is the same as the (b) panels of Figures 2-5.

First we consider the optimal Lagrangian index policy *with reoptimization*. That is, in each simulated scenario, in each period, we solve the Lagrangian dual problem (7) with the current state for all items, breaking ties as in the optimal Lagrangian index policy. As one might expect, this policy with reoptimiza-



(a) With Bernoulli signals (n=1)

(b) With binomial signals (n=5)

Figure 6: Results for applicant screening examples with variations on the heuristics

tion appears to outperform the optimal Lagrangian policy without reoptimization, but they both appear to exhibit $\sqrt{S}$ growth in the performance gap. These applicant screening examples are small enough to allow reoptimization (the run times range from 9 to 46 seconds for the results reported in the figure), but reoptimization would be very time consuming in the dynamic assortment examples. With reoptimization, we have to solve the Lagrangian dual problem in every period in every simulated scenario and these problems become more complex as the items that are initially identical will transition to different states over time and no longer be identical. The figures also show results for a policy that reoptimizes the Lagrangian, but breaks ties randomly rather than using an optimal tiebreaking method: for large S the performance of this policy matches the performance of the Lagrangian policy without reoptimization using random tiebreaking and the errors grow linearly in S. Thus reoptimization is not a substitute for proper tiebreaking.

We also show results for the three different methods described in §4.4 for generating a mixed policy for tiebreaking with the optimal Lagrangian index policy, without reoptimization. As discussed in §4.4, proportional assignment seems to outperform simple random mixing and Markov random mixing, though the differences are small.

Finally, we show results for a "fluid heuristic" similar to that described in Bertsimas and Mišić (2016). This fluid heuristic is based on reoptimization of the Lagrangian dual problem (7), solving the dual LP formulation (39) (given in EC §B.3) in each period. The heuristic then selects items to maximize the total "flow" for the system for a given period and state, where these flows are given by the solution to the dual LP; see EC §B.3 for a more detailed description. The intuition behind this heuristic is that these flows are positive for items that would be selected in the Lagrangian relaxation and maximizing the flow would, in some sense, lead the heuristic to mimic the actions selected by the Lagrangian relaxation. In the example results in the figure, we see that the fluid heuristic is competitive with the other heuristics for small S, but

the performance gap grows linearly with S like the other policies that do not use an optimal tiebreaking method, rather than growing with $\sqrt{S}$ like the Lagrangian policies with optimal tiebreaking.

7. Problems with Long Time Horizons

In this section, we first consider Whittle (1988)'s conjecture on the asymptotic optimality of the Whittle index policy and Weber and Weiss (1990)'s counterexample. We use this example to motivate the extension of the results of §5 to the infinite horizon case with discounting, which we consider in §7.2.

7.1. Whittle's Conjecture and Weber and Weiss's Counterxample

It is interesting to compare the asymptotic optimality result of Corollary 1 to that conjectured in Whittle (1988). Whittle focused on an infinite-horizon average-reward formulation where the DM had to select exactly N items in each period and he considered a single Lagrange multiplier. The solution to the Lagrangian dual problem in this average reward setting yields a Lagrangian relaxed policy that selects N items per period, in expectation for the long-run average (see Whittle's Proposition 1). In his asymptotic analysis, Whittle considered a growing sequence of problems where items may be of different types but the proportion of items of each type is held constant as the total number of items S increases; the number of items selected N is assumed to be a constant fraction α of S.

Whittle conjectured that, if the items are indexable, then

$$\lim_{S\to\infty} \frac{L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S) - V_1^{\widetilde{\pi}}(\boldsymbol{x}; S)}{S} = 0 \ , \tag{23}$$

where $\widetilde{\pi}$ is the Whittle index policy, rather than the Lagrangian index policy. Adapting Whittle (1988, p. 293) to our notation and terminology, the intuition behind his conjecture was as follows.

> The Whittle index policy selects exactly the $N = \alpha S$ items of largest index. Under the assumption of indexability, the optimal policy $\widetilde{\psi}$ for the Lagrangian relaxation selects the $\tilde{n}$ items of largest index, where $\tilde{n}$ deviates from N only by a term of probable order $\sqrt{N}$ or, equivalently, $\tilde{n}/N$ deviates from α only by a term of probable order $1/\sqrt{N}$.

Whittle's intuition is closely related to the intuition behind Proposition 5, as discussed following that result: the key condition that ensures asymptotic convergence is that the heuristic policy and the optimal policy $\widetilde{\psi}$ for the Lagrangian relaxation are aligned so the two policies typically make the same selection decisions, with the number of different decisions growing at a rate less than $\sqrt{N}$. Whittle's intuition is consistent with the logic of Proposition 5 but, in the finite-horizon setting that we consider, the Whittle index policy need not be aligned with $\widetilde{\psi}$, whereas the optimal Lagrangian index policy is, by construction, aligned with $\widetilde{\psi}$. Weber and Weiss (1990) showed that optimal policies asymptotically converge to the Lagrangian bound in the average reward setting (in the relative sense of (23)) but provided an example that showed that the Whittle index policy need not be asymptotically optimal.

In EC §G, we consider a finite-horizon adaptation of the example from Weber and Weiss (1990) with $T = 20,000$. The key takeaway from this example is that, even in problems with constant rewards and transition matrices and long horizons, we may need time-varying Lagrange multipliers to optimally control selection decisions over time. Here again mixed policies and careful tiebreaking play an important role. The initial distribution of items across states affects the optimal Lagrange multipliers and a full set of Lagrange multipliers is required to align the optimal Lagrangian index policy with the optimal policy for the Lagrangian relaxation in every period. The Whittle indices depend on the state of a given item but, by construction, are independent of the states of all other items and of the distribution of items and the policy need not be aligned with that for the Lagrangian relaxation.

7.2. Asymptotic Optimality for Infinite-Horizon Dynamic Selection Problems

We now consider the extension of the results of §5 to an infinite-horizon setting with discounting, assuming a discount factor δ. We assume that the rewards for all items are bounded above and below by $\bar{r}$ and $\underline{r}$ and the number of items that may be selected N_t is bounded above by N.

There are two key challenges that must be addressed in the infinite-horizon setting. The first challenge, suggested by Weber and Weiss (1990)'s example above, is that to achieve asymptotic optimality, we may need to consider an infinite sequence of Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$. This leads to a Lagrangian dual problem (7) that is practically difficult (or impossible) to solve to optimality. The second challenge is that the β_t terms (19) appearing in the performance bound of Proposition 5 grow rapidly with the horizon T, reflecting the possible cascading of changes in selection decisions through subsequent periods. Incorporating discounting in the finite-horizon model (with horizon T), the result of Proposition 5 holds as stated, but with

$$\beta_t(T) = \frac{\delta^{t-1}}{2\delta - 1}\left((2\delta)^{T-t+1} - 1\right) . \tag{24}$$

(See EC §H for a more detailed derivation.) If $\delta > 1/2$, these β_t terms will grow without bound as T grows and the performance bound becomes increasingly slack. In our discussion, we will focus on this problematic case where $\delta \in (1/2, 1)$. (We present results for $\delta \in (0, 1/2]$ in EC §H.)

We will address these challenges by considering a series of finite-horizon approximations with horizon T and taking the limit as the horizon T and problem size S increase simultaneously. Let $L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; T)$ denote the optimal Lagrangian with finite horizon T, defined as in (4) (but with discounting) where $\boldsymbol{\lambda}^*$ solves the corresponding Lagrangian dual problem (7). Let $V_1^{\tilde{\pi}}(\boldsymbol{x}; T)$ denote the present value generated by the corresponding optimal Lagrangian index policy over the same finite horizon. Further, let $L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; \infty)$ denote the optimal infinite-horizon Lagrangian with the optimal infinite sequence of Lagrange multipliers and let $V_1^*(\boldsymbol{x})$ denote the optimal value function. Then, for any horizon T, we have

$$\underbrace{V_1^{\tilde{\pi}}(\boldsymbol{x}; T) + \frac{\delta^T}{1-\delta}\underline{r}S}_{\equiv \bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; T)} \;\leq\; V_1^*(\boldsymbol{x}) \;\leq\; L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; \infty) \;\leq\; \underbrace{L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; T) + \frac{\delta^T}{1-\delta}\bar{r}S}_{\equiv \bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; T)} . \tag{25}$$

Here the term on the left, $\bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T)$, represents a lower bound on the discounted rewards associated with following the optimal Lagrangian index policy based on horizon T for T periods and then following any policy thereafter (which generates rewards of at least $\underline{r}S$ in each period). Such a policy is feasible for the infinite-horizon problem, hence the first inequality above. The second inequality follows from Lagrangian duality, as in Proposition 1. The final term $\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T)$ represents the finite-horizon Lagrangian value for T followed by an upper bound on the rewards for all subsequent periods. The final inequality follows from the facts that the Lagrange multipliers $(\lambda_1^*,\ldots,\lambda_T^*)$ that are optimal for the finite-horizon dual problem are a feasible starting sequence $(\lambda_1^*,\ldots,\lambda_T^*,\ldots)$ for the infinite-horizon dual problem but are not necessarily optimal and that $\bar{r}$ is an upper bound on the item rewards.

As in §5, we will show that, in relative terms, $\bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T)$ approaches $\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T)$ as we increase S and T; since the optimal value function $V_1^*(\boldsymbol{x})$ is bracketed by these terms in (25), this will imply the desired asymptotic optimality result. In our analysis, we will consider sums of cash flows in the difference of $\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T)-\bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T)$ over a horizon $T' \leq T$ and obtain a bound of the form

$$\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T) - \bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T) \;\leq\; (\bar{r}-\underline{r})\left(\sum_{t=1}^{T'}\beta_t(T')\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right). \tag{26}$$

This follows from the argument underlying Proposition 5. We then choose T' to provide a good bound in (26). Intuitively, we want to choose the horizon T' to balance two objectives: we want short horizons to keep the finite-horizon performance gap ($\sum_{t=1}^{T'}\beta_t(T')\sqrt{N}$) small, but want longer horizons to reduce the effect of considering a finite rather than an infinite horizon (represented by $\delta^{T'}S/(1-\delta)$). By choosing the horizon T' to (approximately) minimize the bound of (26), we have the following infinite-horizon analog of Proposition 5.

Proposition 6. *Let $\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T)$ and $\bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T)$ be defined as in* (25) *and let $\lfloor z \rfloor$ denote the largest integer less than or equal to z. For any $T \geq \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$,*

$$\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T) - \bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T) \;\leq\; \gamma(\bar{r}-\underline{r})S\left(\frac{\sqrt{N}}{S}\right)^{\log_2\frac{1}{\delta}} \tag{27}$$

where γ is a positive constant that depends only on δ.

Although we would intuitively expect larger T to result in better heuristics and bounds, the bound of (27) does not improve if we increase T beyond $T \geq \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$. Like Proposition 5, this bound assumes the maximum possible loss in rewards when the Lagrangian relaxation and Lagrangian index polices are in different states and assumes the maximum possible cascading of differences in states through horizon T'. The bound also makes no assumptions about the performance of the heuristic or Lagrangian after period T', again assuming the maximum possible difference in rewards.

Proposition 6 leads to the following asymptotic optimality result that is analogous to Corollary 1.

Corollary 2 (Infinite-horizon asymptotic optimality)**.** *Consider a growing sequence of infinite-horizon dynamic selection problems (indexed by S) and let $T(S) \geq \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$. Let $\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S) = \bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; T(S))$ and $\bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; S) = \bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; T(S))$, as defined in* (25)*. If the optimal value functions $V_1^*(\boldsymbol{x}; S)$ are positive and satisfy*

$$V_1^*(\boldsymbol{x}; S) \geq \kappa S, \tag{28}$$

for some constant $\kappa > 0$, then

$$\lim_{S \to \infty} \frac{\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S) - \bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; S)}{V_1^*(\boldsymbol{x}; S)} = 0 \ . \tag{29}$$

Since the optimal value function $V_1^*(\boldsymbol{x}; S)$ lies between $\bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; S)$ and $\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S)$, this result implies asymptotic optimality of the sequence of finite-horizon Lagrangian index policies when normalized by the optimal value; as discussed following Corollary 1, we could also normalize in other ways. The growth condition on the optimal value function (28) is stronger than that in Corollary 1, as we require $V_1^*(\boldsymbol{x}; S)$ to scale in proportion with S (versus simply faster than $\sqrt{N(S)}$). For example, this stronger condition would hold if $N(S)$ scales in fixed proportion with S (i.e., $N(S) = \alpha S$ for some $\alpha \in (0, 1)$), the reward for not selecting is nonnegative, and the expected reward associated with selecting an item is bounded away from zero.

Though the asymptotic result of Corollary 2 suggests that the optimal Lagrangian index policies will perform well in problems with many items, provided we take the horizon T in the Lagrangian model to be sufficiently large. However, the guaranteed convergence rate is much slower in the infinite-horizon setting than the finite-horizon setting. For example if $N(S) = \alpha S$, in the infinite-horizon setting

$$\lim_{S \to \infty} \frac{\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S) - \bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; S)}{S} \ ,$$

converges to zero at rate $(\sqrt{1/S})^{\log_2(1/\delta)}$ (if we increase T with S accordingly), which is much slower than the $\sqrt{1/S}$ rate that we found in the finite-horizon setting. In particular, $\log_2(1/\delta)$ approaches 0 as δ approaches 1, implying slow convergence for large discount factors.

The slow convergence in the infinite-horizon result is primarily caused by the exponential growth in the β_t terms with the horizon T, reflecting the maximum possible cascading of differences in states visited by the Lagrangian relaxation and Lagrangian index polices. If the problem has a structure where the item states are (in some sense) recurrent, these differences may not cascade in this way and may no longer have such an exponential effect; perhaps then we would again obtain $\sqrt{1/S}$ convergence for large problems. We leave this as a topic for future research.

8. Conclusions

The numerical and theoretical results of this paper suggest that the optimal Lagrangian index policies are the most appropriate heuristic policies for use in dynamic selection problems, particularly for problems with many items. The optimal Lagrangian index policies are both easier to compute and perform better than the

popular Whittle index policies. The logic of the Lagrangian index policy is intuitive. First, find a set of prices for the constrained resources (Lagrange multipliers $\boldsymbol{\lambda}^*$) that lead to the required usage of the resource "on average." For large problems, the deviations from these averages will tend to be small in relative terms and policies that are based on these prices will tend to perform well. There are however some important subtleties that must be addressed, both in theory and in implementation. Notably, optimal prices often induce ties where the DM will be indifferent to selecting or not selecting some items and optimal performance requires careful coordination of the selection decisions across items when breaking ties.

A natural next step in this line of research would be to consider weakly coupled DPs with more general decision variables and resource constraints. For example, one might consider problems where items have multiple possible actions (rather than just select or not) with multidimensional budget constraints. The analysis of the Lagrangian in §3 would seem to generalize directly to this more complex setting, but it is not immediately clear how to generalize the Lagrangian index policies or the performance analysis of §5.

A. Cutting-Plane Method for Solving the Lagrangian Dual Problem

In the cutting-plane method, we proceed iteratively through a series of trial points $\boldsymbol{\lambda}_k$, calculating the item-specific value functions $V_s(\boldsymbol{\lambda}_k)$ and a subgradient $\nabla_{s,k} \in \partial V_s(\boldsymbol{\lambda}_k)$ at these points; as discussed in Proposition 4, these subgradients correspond to selection probabilities for an optimal policy for the given $\boldsymbol{\lambda}_k$. By (9), we know $V_s(\boldsymbol{\lambda}) \geq V_s(\boldsymbol{\lambda}_k) + \nabla_{s,k}^\intercal(\boldsymbol{\lambda} - \boldsymbol{\lambda}_k)$ for each k, i.e., the subgradients provide a linear approximation of $V_s(\boldsymbol{\lambda})$ from below. We then approximate the Lagrangian $L(\boldsymbol{\lambda}) = \boldsymbol{N}^\intercal\boldsymbol{\lambda} + \sum_{s=1}^S V_s(\boldsymbol{\lambda})$ as

$$\boldsymbol{N}^\intercal\boldsymbol{\lambda} + \sum_{s=1}^S V_s(\boldsymbol{\lambda}_{i_s}) + \nabla_{s,i_s}^\intercal(\boldsymbol{\lambda} - \boldsymbol{\lambda}_{i_s}) \tag{30}$$

where we use the value and subgradient from iteration i_s, $i_s \in \{1, \ldots, k\}$, to approximate $V_s(\boldsymbol{\lambda})$. Taking the upper envelope of these linear approximations, we have the cutting-plane model

$$\ell_k(\boldsymbol{\lambda}) \equiv \max_{i_1,\ldots,i_S \in \{1,\ldots,k\}} \left\{ \boldsymbol{N}^\intercal\boldsymbol{\lambda} + \sum_{s=1}^S \left(V_s(\boldsymbol{\lambda}_{i_s}) + \nabla_{s,i_s}^\intercal(\boldsymbol{\lambda} - \boldsymbol{\lambda}_{i_s}) \right) \right\}. \tag{31}$$

Since the $V_s(\boldsymbol{\lambda})$ are approximated from below, we know that $\ell_k(\boldsymbol{\lambda}) \leq L(\boldsymbol{\lambda})$, for all $\boldsymbol{\lambda}$.[9]

The cutting-plane method proceeds by taking the next trial point $\boldsymbol{\lambda}_{k+1}$ to be the point that minimizes the cutting-plane model $\ell_k(\boldsymbol{\lambda})$, i.e.,

$$\boldsymbol{\lambda}_{k+1} = \operatorname*{arg\,min}_{\boldsymbol{\lambda} \geq \mathbf{0}} \ \ell_k(\boldsymbol{\lambda}) \ . \tag{32}$$

We then calculate the item-specific value functions $V_s(\boldsymbol{\lambda}_{k+1})$ and subgradients $\nabla_{s,k+1} \in \partial V_s(\boldsymbol{\lambda}_{k+1})$ for this new point, as well as the Lagrangian $L(\boldsymbol{\lambda}_{k+1}) = \boldsymbol{N}^\intercal\boldsymbol{\lambda}_{k+1} + \sum_{s=1}^S V_s(\boldsymbol{\lambda}_{k+1})$. The process continues until $\ell_k(\boldsymbol{\lambda}_{k+1}) = L(\boldsymbol{\lambda}_{k+1})$. In this terminal case, since $\boldsymbol{\lambda}_{k+1}$ minimizes $\ell_k(\boldsymbol{\lambda})$ and $\ell_k(\boldsymbol{\lambda}) \leq L(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \geq 0$, we know that $\boldsymbol{\lambda}_{k+1}$ is an optimal solution for (7). If $\ell_k(\boldsymbol{\lambda}_{k+1}) < L(\boldsymbol{\lambda}_{k+1})$, we add the newly calculated values $V_s(\boldsymbol{\lambda}_{k+1})$ and gradients $\nabla_{s,k+1}$ to form a new cutting-plane model $\ell_{k+1}(\boldsymbol{\lambda})$. Note that in this case, we will

[9] The standard cutting-plane method takes the maximum in (31) using values and subgradients of the objective function, here $L(\boldsymbol{\lambda})$, at each stage. Effectively this requires using the values and gradients from the same iteration i_s for all items in (31) rather than allowing the use of results from different iterations for different items. The flexibility to choose different approximations for each item improves the bound given by the cutting-plane model (31) and thereby accelerates convergence of the algorithm. This is particularly important when reoptimizing (as in §6.4) or calculating information relaxation bounds (EC §E) where the items will necessarily be distinct.

have a *new* cutting plane for L (corresponding to a new optimal policy for at least one item) since the new subgradient will support L at $\boldsymbol{\lambda}_{k+1}$ whereas $\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \ell_k(\boldsymbol{\lambda}) = \ell_k(\boldsymbol{\lambda}_{k+1}) < L(\boldsymbol{\lambda}_{k+1})$. Since $L(\boldsymbol{\lambda})$ is piecewise linear with a finite number of pieces, the cutting-plane method will converge to the optimal solution in a finite number of iterations.

The cutting-plane optimization problem (32) can be formulated as a linear program (LP) as

$$\begin{aligned} \min_{\boldsymbol{\lambda}, v_s} \quad & \boldsymbol{N}^\intercal \boldsymbol{\lambda} + \sum_{s=1}^{S} v_s \\ \text{s.t.} \quad & v_s \;\geq\; V_s(\boldsymbol{\lambda}_i) + \nabla_{s,i}^\intercal (\boldsymbol{\lambda} - \boldsymbol{\lambda}_i) \qquad \forall i \in \{1, \ldots, k\}, \forall s \in \{1, \ldots, S\} \;, \\ & \boldsymbol{\lambda} \geq 0 \;. \end{aligned} \tag{33}$$

As we proceed iteratively in the cutting-plane method, we add additional constraints for the new values $V_s(\boldsymbol{\lambda}_{k+1})$ and subgradients $\nabla_{s,k+1}$ at the new trial value $\boldsymbol{\lambda}_{k+1}$. We solve (33) using the dual simplex method, using the optimal dual basis from one iteration as an initial dual basis for the next iteration.

We can write the dual of the LP (33) as

$$\begin{aligned} \max_{\gamma_{s,i}} \quad & \sum_{s=1}^{S} \sum_{i=1}^{k} (V_s(\boldsymbol{\lambda}_i) - \nabla_{s,i}^\intercal \boldsymbol{\lambda}_i) \gamma_{s,i} \\ \text{s.t.} \quad & -\sum_{s=1}^{S} \sum_{i=1}^{k} \gamma_{s,i} \nabla_{s,i} \leq \; \boldsymbol{N} \\ & \sum_{i=1}^{k} \gamma_{s,i} = \; 1 \qquad\qquad \forall s \in \{1, \ldots, S\} \;, \\ & \gamma_{s,i} \geq 0 \qquad\qquad \forall i \in \{1, \ldots, k\}, \forall s \in \{1, \ldots, S\} \;. \end{aligned} \tag{34}$$

In the final step of the cutting-plane method where $\ell_k(\boldsymbol{\lambda}_{k+1}) = L(\boldsymbol{\lambda}^*)$, the optimal dual variables $\gamma_{s,i}$ will correspond to mixing weights satisfying the conditions of Proposition 4(c). Counting constraints, we see that in a basic solution for (34) at most $S + T$ of these mixing weights $\gamma_{s,i}$ will be positive and these will correspond to the item-specific policies $\psi_{s,i}$ that are optimal given $\boldsymbol{\lambda}_i$ and also optimal given $\boldsymbol{\lambda}^*$.

If some or all items are identical, the cutting-plane method can be simplified as the DP and its gradients need only be evaluated once for the identical items; the LPs (33) and (34) similarly simplify. If we let S' denote the number of distinct items, the simplified version of the LP (33) will have $S' + T$ decision variables and $k \times S'$ constraints. The basic solutions for the simplified version of the dual LP (34) will have at most $S' + T$ positive mixing weights, corresponding to item-specific policies $\psi_{s,i}$ that are optimal given $\boldsymbol{\lambda}^*$. In our numerical examples, we have found that optimal solutions typically have exactly $S' + T$ positive mixing weights when the linking constraints (1) are binding.

In our numerical examples, the computational bottleneck when solving the Lagrangian dual problems using the cutting-plane method is calculating the item-specific value functions (6) and their subgradients. The LPs (33) are typically easy to solve, even if the item-specific DPs have large state spaces.

References

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

Bernstein, F., Kök, A. G. and Xie, L. (2015), 'Dynamic assortment customization with limited inventories', *Manufacturing & Service Operations Management* **17**(4), 538–553.

Bertsekas, D. P., Nedić, A. and Ozdaglar, A. E. (2003), *Convex Analysis and Optimization*, Athena Scientific.

Bertsimas, D. and Mersereau, A. J. (2007), 'A learning approach for interactive marketing to a customer segment', *Operations Research* **55**(6), 1120–1135.

Bertsimas, D. and Mišić, V. V. (2016), 'Decomposable markov decision processes: A fluid optimization approach', *Operations Research* **64**(6), 1537–1555.

Bertsimas, D. and Niño-Mora, J. (2000), 'Restless bandits, linear programming relaxations, and a primal-dual index heuristic', *Operations Research* **48**(1), 80–90.

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

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

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

Gittins, J., Glazebrook, K. and Weber, R. (2011), *Multi-armed bandit allocation indices*, John Wiley & Sons.

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

Hodge, D. J. and Glazebrook, K. D. (2015), 'On the asymptotic optimality of greedy index heuristics for multi-action restless bandits', *Advances in Applied Probability* **47**(3), 652–667.

Hu, W. and Frazier, P. (2017), 'An asymptotically optimal index policy for finite-horizon restless bandits'. Working paper, https://arxiv.org/abs/1707.00205.

Kök, A. G., Fisher, M. L. and Vaidyanathan, R. (2008), Assortment planning: Review of literature and industry practice, *in* 'Retail supply chain management', Springer, pp. 99–153.

Le Boudec, J.-Y., McDonald, D. and Mundinger, J. (2007), A generic mean field convergence result for systems of interacting objects, *in* 'Quantitative Evaluation of Systems, 2007. QEST 2007. Fourth International Conference on the', IEEE, pp. 3–18.

Puterman, M. L. (1994), *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons.

Rusmevichientong, P., Shen, Z.-J. M. and Shmoys, D. B. (2010), 'Dynamic assortment optimization with a multinomial logit choice model and capacity constraint', *Operations Research* **58**(6), 1666–1680.

Topaloglu, H. (2009), 'Using lagrangian relaxation to compute capacity-dependent bid prices in network revenue management', *Operations Research* **57**(3), 637–649.

Weber, R. R. and Weiss, G. (1990), 'On an index policy for restless bandits', *Journal of Applied Probability* **27**(3), 637–648.

Whittle, P. (1988), 'Restless bandits: Activity allocation in a changing world', *Journal of Applied Probability* **25**, 287–298.

Electronic Companion

B. Selected Proofs and Details for §3

B.1. Proofs for §3: Lagrangian Relaxations

Proof of Proposition 3. We can write the item-specific DP (6) as a maximization over item-specific policies ψ_s:

$$V_s(\boldsymbol{\lambda}) = \max_{\psi_s} \sum_{t=1}^{T} \mathbb{E}[\, r_{t,s}(\widetilde{x}_{t,s}(x_{1,s};\psi_s), \psi_{t,s}(\widetilde{x}_{t,s}(x_{1,s};\psi_s)) - \lambda_t \ \psi_{t,s}(\widetilde{x}_{t,s}(x_{1,s};\psi_s))\,] \tag{35}$$

where $\widetilde{x}_{t,s}(x_{1,s};\psi_s)$ is the random state for item s in period t when starting in state $x_{1,s}$ and following policy ψ_s. For a fixed policy ψ_s, the objective in (35) is linear in $\boldsymbol{\lambda}$. The pointwise maximum over these linear functions yields a piecewise linear and convex function. The Lagrangian $L(\boldsymbol{\lambda})$, as a finite sum of piecewise linear convex functions $V_s(\boldsymbol{\lambda})$ (plus additional linear terms), is also piecewise linear and convex. □

Proof of Proposition 4. (i): Consider the representation of the item-specific DP given in equation (35) in the proof of Proposition 3. There, for a fixed policy ψ_s, the objective in (35) is linear in $\boldsymbol{\lambda}$ and the t^{th} element of the gradient $\nabla_s(\psi_s)$ with policy ψ_s is $-\mathbb{E}[\,\psi_{t,s}(\widetilde{x}_{t,s}(x_{1,s};\psi_s))\,]$, which is $-p_{t,s}(\psi_s)$. The subdifferential result (10) then follows from Danskin's Theorem (see, e.g., Bertsekas et al. 2003 Proposition 4.5.1, p. 245). This subdifferential result implies $\nabla_s(\psi_s)$ is a subgradient of V_s at $\boldsymbol{\lambda}$ for any $\psi_s \in \Psi_s^*(\boldsymbol{\lambda})$.

(ii) The first equality follows from the fact the subdifferential of a sum of convex functions is the sum of the subdifferentials for the component functions (see, e.g., Bertsekas et al. 2003, Proposition 4.2.4, p. 232). The second equality follows from (i) and the fact that the Minkowski sum of the convex hulls of a collection of sets is equal to the convex hull of the sum of the sets.

(iii) A necessary and sufficient condition for $\boldsymbol{\lambda}^*$ to be optimal for the Lagrangian dual problem (7) is

$$0 \ \in \ \partial L(\boldsymbol{\lambda}^*) + \mathcal{N}_{\{\boldsymbol{\lambda} \geq \mathbf{0}\}}(\boldsymbol{\lambda}^*)$$

where $\mathcal{N}_{\{\boldsymbol{\lambda} \geq \mathbf{0}\}}(\boldsymbol{\lambda}^*)$ is the normal cone of $\{\boldsymbol{\lambda} \geq \mathbf{0}\}$ at $\boldsymbol{\lambda}^*$ (see, e.g., Bertsekas et al. 2003 Proposition 4.7.2, p. 257). The result then follows from (11) and the form of this normal cone: the normal cone terms are zero when $\lambda_t > 0$ and negative when $\lambda_t = 0$. The specific mixture representation here reflects the first representation of $\partial L(\boldsymbol{\lambda})$ in (11); we could obtain a different form of mixture using the second representation in (11). The limit on the number of points involved in the mixtures ($n_s \leq T+1$) follows from Caratheodory's theorem. □

B.2. Constructing a Markov Random Policy

Here we describe how to use the simple mixed policy representation of Proposition 4(iii) to construct a corresponding Markov random policy that makes selection decisions with state-contingent selection probabilities. First, let $\rho_{t,s}(x_s, \psi_s)$ denote the probability of item s occupying state x_s at time t when following a deterministic policy ψ_s; these probabilities are straightforward to compute. The probability of selecting item s in state x_s at time t with policy ψ_s is then $\rho_{t,s}(x_s, \psi_s)\psi_{t,s}(x_s)$ and the probability of not selecting is $\rho_{t,s}(x_s, \psi_s)(1 - \psi_{t,s}(x_s))$.

Let $\widetilde{\psi}$ denote a simple mixed policy representation of Proposition 4(iii) where $\gamma_{s,i}$ is the mixing weight associated with a deterministic policy $\psi_{s,i}$. Let $\nu_{t,s}(x_s, u_s; \widetilde{\psi})$ denote the probability of item s being in state x_s and choosing action u_s with the simple mixed policy $\widetilde{\psi}$. This is given by:

$$\nu_{t,s}(x_s, 1; \widetilde{\psi}) = \sum_{i=1}^{n_s} \gamma_{s,i} \ \rho_{t,s}(x_s, \psi_{s,i}) \ \psi_{t,s,i}(x_s)$$

$$\nu_{t,s}(x_s, 0; \widetilde{\psi}) = \sum_{i=1}^{n_s} \gamma_{s,i} \ \rho_{t,s}(x_s, \psi_{s,i}) \ (1 - \psi_{t,s,i}(x_s))$$

Thus the probability of being in state x_s with this mixed policy is $\nu_{t,s}(x_s, 0; \widetilde{\psi}) + \nu_{t,s}(x_s, 1; \widetilde{\psi})$. If $\widetilde{\psi}$ is an optimal mixed policy for the Lagrangian dual problem, $\nu_{t,s}(x_s, u_s; \widetilde{\psi})$ is an optimal solution for the LP (39).

For a Markov random policy that corresponds to the mixed distribution $\widetilde{\psi}$, we can take the probability of selecting an item s in state x_s in period t to be:

$$\frac{\nu_{t,s}(x_s, 1; \widetilde{\psi})}{\nu_{t,s}(x_s, 0; \widetilde{\psi}) + \nu_{t,s}(x_s, 1; \widetilde{\psi})} \tag{36}$$

By construction, this will generate the same state-action probabilities as $\widetilde{\psi}$, will select the same number of items on average in each period as $\widetilde{\psi}$, and will have the same expected total reward as $\widetilde{\psi}$. Note that these selection probabilities will be undefined when the probability of being in state x_s in period t (in the denominator of (36)) is zero. These undefined selection probabilities are irrelevant for evaluating policies for the Lagrangian relaxation, but may be relevant when we use the policy for the Lagrangian relaxation as a tiebreaker for the optimal Lagrangian index policy (as discussed in §4.4P. In our numerical examples, we take these undefined probabilities to be 0.5.

B.3. Linear Programming Formulation of the Lagrangian Dual Problem

We can also formulate the Lagrangian dual problem (7) as an LP; Hawkins (2003), Adelman and Mersereau (2008), and Bertsimas and Mišić (2016) considered similar LP formulations. First, following the standard LP formulation of a DP, we can write the item-specific DP (6) for item s with Lagrange multipliers $\boldsymbol{\lambda}$ as

$$\begin{aligned} \min_{V^{\boldsymbol{\lambda}}_{t,s}(x_s)} \quad & V^{\boldsymbol{\lambda}}_{s,1}(x^0_s) \\ \text{s.t.} \quad & V^{\boldsymbol{\lambda}}_{t,s}(x_s) \geq \; r_{t,s}(x_s, u_s) - \lambda_t u_s + \sum_{\widetilde{\chi}_{t,s}} p_t(\widetilde{\chi}_{t,s} \,|\, x_s, u_s)\; V^{\boldsymbol{\lambda}}_{t+1,s}(\widetilde{\chi}_{t,s}) \quad \forall\; t, x_s, u_s \;, \end{aligned} \tag{37}$$

where x^0_s is the initial state of item s and $p_t(\widetilde{\chi}_{t,s} \,|\, x_s, u_s)$ is the conditional probability of state $\widetilde{\chi}_{t,s}$ occurring when starting in state x_s and taking action u_s (with $u_s \in \{0, 1\}$). The decision variables in this LP are the values $V^{\boldsymbol{\lambda}}_{t,s}(x_s)$ for each period t and state x_s and the constraints represent the Bellman equations (6). (We assume $V^{\boldsymbol{\lambda}}_{T+1,s}(x_s) = 0$.) The value function constraints will be binding for optimal actions in states that are visited when following the optimal policy, but need not be binding for any action in states that are not visited by the optimal policy.

Building on this LP representation of the item-specific DPs, we can write the Lagrangian dual problem as an LP by combining these item-specific DPs and including the Lagrange multipliers $\boldsymbol{\lambda}$ as decision variables:

$$\begin{aligned} \min_{\boldsymbol{\lambda},\; V^{\boldsymbol{\lambda}}_{t,s}(x_s)} \quad & \sum_{t=1}^{T} \lambda_t N_t + \sum_{s=1}^{S} V^{\boldsymbol{\lambda}}_{1,s}(x^0_s) \\ \text{s.t.} \quad & V^{\boldsymbol{\lambda}}_{t,s}(x_s) \geq \; r_{t,s}(x_s, u_s) - \lambda_t u_s + \sum_{\widetilde{\chi}_{t,s}} p_t(\widetilde{\chi}_{t,s} \,|\, x_s, u_s)\; V^{\boldsymbol{\lambda}}_{t+1,s}(\widetilde{\chi}_{t,s}) \qquad \forall\; s, t, x_s, u_s \;, \\ & \lambda_t \geq \; 0 \qquad\qquad \forall\; t \;. \end{aligned} \tag{38}$$

If we let $|X_s|$ be the size of the state space for item s, this LP has $T \times \left(1 + \sum_{s=1}^{S} |X_s|\right)$ decision variables and $2 \times T \times \sum_{s=1}^{S} |X_s|$ constraints. (If some or all of the items are identical, this LP can be simplified.) Though this LP formulation delivers optimal values for $\boldsymbol{\lambda}$ and the initial values $V^{\boldsymbol{\lambda}}_{1,s}(x^0_s)$ for the item-specific DPs, it does not provide a full optimal value function for all periods and states because values for states that are not visited under the optimal policy do not affect the objective function. The Lagrangian index policy defined in §4 requires a full value function. To calculate these value functions using this LP formulation, we need to fix $\boldsymbol{\lambda}$ at the optimal value from (38) and solve LPs like (37) with an objective function that includes positive weights on the values $V^{\boldsymbol{\lambda}}_{t,s}(x_s)$ for all items, states, and periods.

Taking $\nu_{t,s}(x_s, u_s)$ to be the dual variables for the constraints in (38), we can write the dual of (38) as:

$$\begin{aligned}
\max_{\nu_{t,s}(x_s,u_s)} \quad & \sum_t \sum_s \sum_{x_s} \sum_{u_s} r_{t,s}(x_s, u_s)\nu_{t,s}(x_s, u_s) \\
\text{s.t.} \quad & \sum_{u_s} \nu_{1,s}(x_s^0, u_s) = 1 && \forall\ s\ , \\
& \sum_{u_s} \nu_{t,s}(\widetilde{\chi}_{t,s}, u_s) = \sum_{x_s} \sum_{u_s} p_t(\widetilde{\chi}_{t,s} \mid x_s, u_s)\ \nu_{t-1,s}(x_s, u_s) && \forall\ s, t > 1, \widetilde{\chi}_{t,s}\ , \\
& \sum_s \sum_{x_s} \nu_{t,s}(x_s, 1) \leq N_t && \forall\ t\ , \\
& \nu_{t,s}(x_s, u_s) \geq 0 && \forall\ s, t, x_s, u_s\ .
\end{aligned} \tag{39}$$

The dual variables here have a natural interpretation as flows: $\nu_{t,s}(x_s, u_s)$ can be interpreted as the probability of being in state x_s at time t and choosing action u_s. The objective in (39) is the expected total reward. The first two constraints are flow conservation conditions: the total flow in the initial state x_s^0 for each item $(\sum_{u_s} \nu_{1,s}(x_s^0, u_s))$ is equal to 1 and the total flow into a later state $\widetilde{\chi}_{t,s}$ must have come from a transition from some previous state. The third constraint requires the linking constraint to hold "on average" and complementary slackness ensures that this linking constraint holds with equality in period t whenever $\lambda_t > 0$. This average linking constraint is thus equivalent to the necessary and sufficient conditions for optimality in the Lagrangian dual given in Proposition 4(iii). Complementary slackness also implies that if the optimal flow $\nu_{t,s}(x_s, u_s)$ is positive, the corresponding value function inequality in (38) holds with equality: that is, the action u_s is optimal in state x_s in period t. The optimal flows $\nu_{t,s}(x_s, u_s)$ given by the LP (39) can also be calculated from the policies $\psi_{s,i}$ and mixing weights $\gamma_{s,i}$ given by the cutting-plane method of Appendix A; see Appendix B.2.

The Fluid Heuristic: Given this LP formulation, we can now describe the fluid heuristic that was discussed in §6.4. Bertsimas and Mišić (2016) considered problems were the state dynamics are independent across items, but the actions need not decompose across items. In dynamic selection problems these actions would be vectors of decision variables $\boldsymbol{u} = (u_1, \ldots, u_S)$ satisfying the linking constraint (1), i.e., $\boldsymbol{u} \in \mathcal{U}_t$. This is not a practical way to formulate large dynamic selection problems as there are $\binom{S}{N} + \binom{S}{N-1} + \cdots + \binom{S}{0}$ different actions to be considered.

In our numerical examples of §6.4, we consider a decomposed version of the fluid heuristic where we solve the Lagrangian dual problem (39) in each period and select items to maximize the total flow,

$$\boldsymbol{u} \in \arg\max_{\boldsymbol{u} \in \mathcal{U}_t} \sum_s \nu_{t,s}(x_s, u_s),$$

where the $\nu_{t,s}(x_s, u_s)$ are the optimal flows for the given period and state given by the solution to (39). Any ties are broken randomly. As noted after (39), complementary slackness implies that if the optimal flow $\nu_{t,s}(x_s, u_s)$ in the LP is positive, the action u_s is optimal in state x_s in period t. The heuristic chooses items to maximize this flow.

An issue with this heuristic is that in the applicant screening examples is that in the first period, the flow is maximized by not screening any applicants: because just 25% of the applicants can be screened and all applicants are in the same initial state, the optimal flows in this first period are $\nu_{1,s}(x_s, 1) = 0.25$ (select) and $\nu_{1,s}(x_s, 0) = 0.75$ (don't select) for all applicants s in the initial (unscreened) state x_s. Similar problems arise in other periods. We address this issue by requiring the choice of exactly N_t applicants in each period, rather than less than or equal to N_t applicants.

C. Notes on Whittle Indices

C.1. Calculating Whittle Indices

Our procedure for calculating Whittle indices assumes the model is "indexable" – that is, the set of periods and states (t, x_s) where no selection is optimal is monotonically increasing from the empty set to all periods and states as w increases from $-\infty$ to $+\infty$. Given this, if we want to calculate Whittle indices for all periods and states for item s, we can proceed as follows:

(i) Start with a small w such that it is optimal to select in all periods and all states. Set $\psi_{t,s}(x_s; w) = 1$ for all t, and x_s, indicating that it is optimal to select in all time periods and states at the initial w.

(ii) For all t and x_s, calculate $V^{w\mathbf{1}}_{t,s}(x_s)$ (by solving the DP (6)) and $\eta^w_{t,s}(x_s) = \partial V^{w\mathbf{1}}_{t,s}(x_s)/\partial w$. These partial derivatives can be evaluated using backward recursion given the policy ψ_s, starting with $\eta^w_{T,s}(x_s) = -1$ for all x_s such that $\psi_{T,s}(x_s; w) = 1$ and $\eta^w_{T,s}(x_s) = 0$ otherwise. In addition, for all t and x_s such that $\psi_{t,s}(x_s; w) = 1$, calculate

$$\Delta^w_{t,s}(x_s) = \big(r_{t,s}(x_s, 1) + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big]\big) - \big(r_{t,s}(x_s, 0) + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 0))\,\big]\big)$$
$$\sigma^w_{t,s}(x_s) = \mathbb{E}\big[\, \eta^w_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big] - \mathbb{E}\big[\, \eta^w_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 0))\,\big]\ .$$

Here $\Delta^w_{t,s}(x_s)$ is the difference on the right side of (15) and $\sigma^w_{t,s}(x_s)$ is the partial derivative of $\Delta^w_{t,s}(x_s)$ with respect to w.

(iii) We next find a new value of w that sets $\Delta^w_{t,s}(x_s) = 0$ for a new period and state. Calculate

$$\delta^\star = \min_{t,x_s}\left\{ \frac{\Delta^w_{t,s}(x_s) - w}{1 - \sigma^w_{t,s}(x_s)} : \psi_{t,s}(x_s; w) = 1 \right\}\ . \tag{40}$$

For all periods t and states x_s achieving this minimum, the Whittle index $w^*_{t,s}(x_s)$ is $w + \delta^\star$. (We explain this calculation after the description of the algorithm.)

(iv) Set w to $w + \delta^\star$ and $\psi_{t,s}(x_s; w) = 0$ for all periods t and states x_s achieving the minimum in (iii).

(v) If there are no states for which selection is optimal, we are done. Otherwise, go to (ii).

The breakpoint calculation in (40) can be understood as follows: for any states and periods satisfying $\psi_{t,s}(x_s; w) = 1$, selection is strictly optimal at the current w, and hence $\Delta^w_{t,s}(x_s) > w$ in such states. Since $\sigma^w_{t,s}(x_s)$ represents the partial derivative of $\Delta^w_{t,s}(x_s)$ with respect to w, we seek a value δ such that $w + \delta$ is a new Whittle index, i.e., δ satisfies

$$\Delta^w_{t,s}(x_s) + \sigma^w_{t,s}(x_s) \cdot \delta = w + \delta\ .$$

The ratio in (40) represents the largest increase to w such that the policy ψ_s remains optimal. For times and states attaining this value in (40), we are indifferent between selecting and not selecting the item at $w + \delta^\star$.

The efficiency of this procedure is improved by noting some properties of the value functions and derivatives when updating in step (ii), i.e., as w is replaced with $w' = w + \delta^\star$. First, we need only update $\eta^{w'}_{t,s}(x_s)$ and $\sigma^{w'}_{t,s}(x_s)$ in time periods up to $t^\star$, where $t^\star$ is the earliest time period attaining the minimum in (iv). The partial derivatives for later periods are unchanged because no decisions change after period $t^\star$. Second, we can update the differences as $\Delta^{w'}_{t,s}(x_s) = \Delta^w_{t,s}(x_s) + \sigma^w_{t,s}(x_s) \cdot \delta^\star$. This follows from the fact that the policy ψ_s is optimal from w to $w + \delta^\star$ and thus the value functions are linear functions of w in this range.

Even with these improvements to efficiency, the procedure can be time consuming when there are many states, because we have to repeatedly update the system of partial derivatives in step (ii), potentially once for each period and state in the problem.

C.2. Whittle Indices for the Applicant Screening Example

Here we show that in the applicant screening example, the Whittle indices have a particularly simple form. We let $\mu(x_s)$ denote an applicant's mean quality in state x_s, which we assume to be positive. For example, with a beta prior $\mu(x_s) = \alpha_s/(\alpha_s + \beta_s)$. The item-specific DP (6) with $\boldsymbol{\lambda} = w\mathbf{1}$ is given recursively as

$V^{w\mathbf{1}}_{T,s}(x_s) = \max\{\mu(x_s) - w, 0\}$ and, for $t < T$,

$$V^{w\mathbf{1}}_{t,s}(x_s) = \max\{-w + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big],\ V^{w\mathbf{1}}_{t+1,s}(x_s)\}\ . \tag{41}$$

A Whittle index for state x_s in period t is a w that equates the screen and do not screen options in this DP:

$$\begin{aligned} -w + \mu(x_s) &= 0 && \text{for } t = T, \text{ and} \\ -w + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big] &= V^{w\mathbf{1}}_{t+1,s}(x_s) && \text{for } t < T. \end{aligned} \tag{42}$$

We show the following.

Proposition 7. *In the applicant screening example, for all s, t, and x_s, the Whittle index is unique.*
(i) *In the final period ($t = T$), the Whittle index is $\mu(x_s)$.*
(ii) *In screening periods ($t < T$), the Whittle index is zero.*

In the proof, we will use the facts that $\mu(x_s) > 0$ in all states x_s and that $\mathbb{E}[\,\mu(\widetilde{\chi}_{t,s}(x_s, 1))\,] = \mu(x_s)$, i.e., the expected posterior quality after screening is equal to the prior expected quality.

Proof. (i) For $t = T$, the result follows directly from the definition of the Whittle index.

(ii) We first show that $w = 0$ is a Whittle index for $t < T$. In this case, $V^{w\mathbf{1}}_{T,s}(x_s) = \mu(x_s)$, since $\mu(x_s) > 0$. By induction and using the fact that the posterior mean is equal to the prior mean, for $t < T$, we have $V^{w\mathbf{1}}_{t,s}(x_s) = \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big] = \mathbb{E}[\,\mu(\widetilde{\chi}_{t,s}(x_s, 1))\,] = \mu(x_s)$. Thus (42) holds for $w = 0$.

We next rule out $w < 0$ and $w > 0$ as possible Whittle indices. Suppose $w < 0$. In this case, we claim that is strictly optimal to screen and collect the "reward" $-w$ in every period and $V^{w\mathbf{1}}_{t,s}(x_s) = \mu(x_s) - (T-t+1)w$. Given this as an induction hypothesis for period $t+1$, in period t screening yields

$$-w + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big] = -w + \mathbb{E}[\,\mu(\widetilde{\chi}_{t,s}(x_s, 1)) + (T-t)w\,] = \mu(x_s) - (T-t+1)w$$

where the first inequality follows from the induction hypothesis and the second from the fact that the expected posterior mean is equal to the prior mean. This is clearly true in the final period as all applicants would be admitted. From the induction hypothesis, not screening in period t yields

$$V^{w\mathbf{1}}_{t+1,s}(x_s) = \mu(x_s) - (T-t)w$$

which, since $w < 0$ is strictly less than screening. Thus screening strictly dominates not screening in every period and $w < 0$ cannot be a Whittle index.

Now suppose $w > 0$. In the final period, $V^{w\mathbf{1}}_{T,s}(x_s) = \max\{\mu(x_s) - w, 0\}$. We claim that not screening strictly dominates screening in all screening periods; if this is true, then $V^{w\mathbf{1}}_{t,s}(x_s) = \max\{\mu(x_s) - w, 0\}$ for $t \le T$. For the induction hypothesis, assume this is true for period $t+1$. Then for period t, not screening yields

$$V^{w\mathbf{1}}_{t+1,s}(x_s) = \max\{\mu(x_s) - w, 0\}$$

and screening yields:

$$\begin{aligned} -w + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big] &= -w + \mathbb{E}[\,\max\{\mu(\widetilde{\chi}_{t,s}(x_s, 1)) - w, 0\}\,] \\ &< -w + \mathbb{E}[\,\mu(\widetilde{\chi}_{t,s}(x_s, 1))\,] \\ &= -w + \mu(x_s) \\ &\le \max\{\mu(x_s) - w, 0\} \end{aligned}$$

The first equality follows from the induction hypothesis. The inequality follows from observing that, since $w > 0$, we have $\max\{x - w, 0\} < x$ for all $x > 0$; this implies the strict inequality above, since $\mu(\widetilde{\chi}_{t,s}(x_s, 1)) > 0$ for all $\widetilde{\chi}_{t,s}(x_s, 1)$. The next equality follows from the fact that the posterior mean is equal to the prior mean. The final inequality is straightforward. Notice this last term is equal to the value of not screening. Thus, if $w > 0$, not screening strictly dominates screening and $w > 0$ cannot be a Whittle index. □

D. Proofs for §5: Analysis of the Optimal Lagrangian Index Policy

D.1. Proof of Proposition 5

The proof of Proposition 5 relies on three key steps which we state in Lemmas 1, 3, and 4 below. Lemma 2 supports Lemma 3. In this discussion, we let $n(\boldsymbol{u}_t) = \sum_{s=1}^{S} u_{t,s}$ denote the number of items selected with action vector $\boldsymbol{u}_t$.

Lemma 1. *For any $\boldsymbol{\lambda} \geq \mathbf{0}$ and initial state $\boldsymbol{x}$, let ψ be an optimal policy for the Lagrangian (5), and let $\widetilde{\boldsymbol{x}}_t$ denote the state transition process generated by ψ. Then, for any policy π,*

$$L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) - V_1^{\pi}(\boldsymbol{x}) = \sum_{t=1}^{T} \mathbb{E}\big[\, d_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t), \pi_t(\widetilde{\boldsymbol{x}}_t))\,\big] \tag{43}$$

where

$$\begin{aligned} d_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}, \boldsymbol{u}_t^{\pi}) = \lambda_t\big(N_t - n(\boldsymbol{u}_t^{\psi})\big) + r_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}) - r_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\pi}) \\ + \mathbb{E}\Big[\, V_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}))\Big] - \mathbb{E}\big[\, V_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\pi}))\,\big] \ . \end{aligned} \tag{44}$$

Here the d_t terms are the differences in total rewards with actions $\boldsymbol{u}_t^{\psi}$ and $\boldsymbol{u}_t^{\pi}$ in period t, reflecting the differences in immediate rewards as well the differences in expected continuation values under π. The difference in total values, $L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) - V_1^{\pi}(\boldsymbol{x})$, is the expected total of these period-specific differences.

Proof. Since ψ is an optimal policy for the Lagrangian $L_t^{\boldsymbol{\lambda}}$ starting in state $\boldsymbol{x}$, we have

$$L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{t=1}^{T} \mathbb{E}\big[\, \lambda_t\big(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))\big) + r_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t))\,\big] \ . \tag{45}$$

We also have

$$\begin{aligned} V_1^{\pi}(\boldsymbol{x}) &= V_1^{\pi}(\boldsymbol{x}) + \sum_{t=2}^{T} \mathbb{E}[\, V_t^{\pi}(\widetilde{\boldsymbol{x}}_t)\,] - \sum_{t=2}^{T} \mathbb{E}[\, V_t^{\pi}(\widetilde{\boldsymbol{x}}_t)\,] \\ &= \sum_{t=1}^{T} \mathbb{E}[\, V_t^{\pi}(\widetilde{\boldsymbol{x}}_t)\,] - \sum_{t=1}^{T} \mathbb{E}\big[\, V_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t)))\,\big] \\ &= \sum_{t=1}^{T} \mathbb{E}\big[\, r_t(\widetilde{\boldsymbol{x}}_t, \pi_t(\widetilde{\boldsymbol{x}}_t)) + V_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \pi_t(\widetilde{\boldsymbol{x}}_t))) - V_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t)))\,\big] \ . \end{aligned}$$

The second equality uses the fact that $V_{T+1}^{\pi} = 0$ and the definition of $\widetilde{\boldsymbol{x}}_t$ as the state process under policy ψ, so $\widetilde{\boldsymbol{x}}_{t+1} = \widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t)))$. The last line uses the definition of the heuristic value function V_t^{π} given in (3) and the law of iterated expectations. The result of the lemma then follows by taking the difference $L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) - V_1^{\pi}(\boldsymbol{x})$ using these expressions. □

The next lemma provides a bound on the differences in heuristic values $V_t^{\pi}(\boldsymbol{x})$ as a function of the number of states x_s that differ. This bound is valid for any index policy, i.e., any policy that ranks items based on item-specific indices and selects up to N_t of these items.

Lemma 2. *Let π be an index policy and suppose states $\boldsymbol{x}'$ and $\boldsymbol{x}''$ differ for m or fewer items. Then, for any t, there exists a nonnegative constant k_t (that depends only on t and T) such that:*

$$\big|\, V_t^{\pi}(\boldsymbol{x}') - V_t^{\pi}(\boldsymbol{x}'')\,\big| \leq k_t \cdot (\bar{r} - \underline{r})\ m \ .$$

Proof. We prove this result using an induction argument on t. For the terminal case with $t = T+1$, we have $V_{T+1}^{\pi}(\boldsymbol{x}') - V_{T+1}^{\pi}(\boldsymbol{x}'') = 0$ since $V_{T+1}^{\pi}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$. Thus we can take $k_{T+1} = 0$.

We then assume the result is true for $t+1$ and show that it holds for period t. We have:

$$\begin{aligned}\big|V_t^\pi(\boldsymbol{x}') - V_t^\pi(\boldsymbol{x}'')\big| &= \big|r_t(\boldsymbol{x}', \pi(\boldsymbol{x}')) - r_t(\boldsymbol{x}'', \pi(\boldsymbol{x}'')) + \mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}', \pi(\boldsymbol{x}')))\,\big] - \mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}'', \pi(\boldsymbol{x}')))\,\big]\big| \\ &\le \big|r_t(\boldsymbol{x}', \pi(\boldsymbol{x}')) - r_t(\boldsymbol{x}'', \pi(\boldsymbol{x}''))\big| + \big|\,\mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}', \pi(\boldsymbol{x}')))\,\big] - \mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}'', \pi(\boldsymbol{x}')))\,\big]\big| \\ &\le 2(\bar{r} - \underline{r})m \;+\; 2k_{t+1}(\bar{r} - \underline{r})m \end{aligned} \tag{46}$$

The first inequality above follows from the triangle inequality. The second inequality follows from the following observations. First note that if states $\boldsymbol{x}'$ and $\boldsymbol{x}''$ differ for m items, then with an index policy π, the actions for at most $2m$ items will differ. (In the worst case, the differences lead all m items to change from not selected to selected (or vice versa) and m other items make the reverse change.) Thus the item-specific rewards differ for at most $2m$ items and

$$\big|r_t(\boldsymbol{x}', \pi(\boldsymbol{x}')) - r_t(\boldsymbol{x}'', \pi(\boldsymbol{x}''))\big| \le 2(\bar{r} - \underline{r})m \;.$$

With differences for at most $2m$ item decisions and state transitions that are independent across items, the random continuation states $\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}', \pi(\boldsymbol{x}'))$ and $\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}'', \pi(\boldsymbol{x}''))$ will differ for at most $2m$ items. (Here we are assuming that items in the same state in $\boldsymbol{x}'$ and $\boldsymbol{x}''$ make the *same* stochastic transitions.) Then, using the induction hypothesis, we have

$$\big|\,\mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}', \pi(\boldsymbol{x}')))\,\big] - \mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}'', \pi(\boldsymbol{x}')))\,\big]\big| \le 2k_{t+1}(\bar{r} - \underline{r})m \;,$$

completing the proof of the inequality (46). Then taking $k_t = 2(1 + k_{t+1}) = 2^{T-t+2} - 2$, we obtain the result of the lemma. □

We next use the previous lemma to establish an upper bound on the differences in Lemma 1 in the case where the policy π is a Lagrangian index policy with a tiebreaker that is an optimal policy ψ for the Lagrangian for any $\boldsymbol{\lambda}$. The key observation in the proof is to note that though ψ and π may select different numbers of items in a given state, the choices will differ for at most $\big|n(\psi_t(\boldsymbol{x}_t)) - N_t\big|$ items.

Lemma 3. *For any $\boldsymbol{\lambda} \ge \mathbf{0}$ and initial state $\boldsymbol{x}$, let ψ be an optimal policy for the Lagrangian* (5)*, and let π be the Lagrangian index policy for $\boldsymbol{\lambda}$ with ψ as a tiebreaker. For each t, there exists a nonnegative constant c_t (depending only on t and T), such that for all $\widetilde{\boldsymbol{x}}_t$ that may be realized when following policy ψ,*

$$d_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t), \pi_t(\widetilde{\boldsymbol{x}}_t)) \le \lambda_t\big(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))\big) + c_t(\bar{r} - \underline{r})\big|n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\big| \;. \tag{47}$$

If $\lambda_t = 0$, we have a tighter bound:

$$d_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t), \pi_t(\widetilde{\boldsymbol{x}}_t)) \le c_t(\bar{r} - \underline{r}) \max\{n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t, 0\} \;.$$

Proof. Fix period t and state $\widetilde{\boldsymbol{x}}_t$. First note that since the policy ψ is optimal for the Lagrangian, it will select all items that have priority indices $i_{t,s}(x_{t,s})$ such that $i_{t,s}(x_{t,s}) > \lambda_t$ and perhaps some items such that $i_{t,s}(x_{t,s}) = \lambda_t$. (It is important that $\widetilde{\boldsymbol{x}}_t$ be a state that may be visited under the policy ψ. An optimal policy ψ need not satisfy this condition in states that are not visited when using ψ.)

We consider two cases. Case (i): Suppose the Lagrangian policy ψ selects $n(\psi_t(\widetilde{\boldsymbol{x}}_t)) < N_t$ items. Those items selected by ψ with $i_{t,s}(x_{t,s}) > \lambda_t$ will be included in the top N_t items as ranked by the priority index and will thus also be selected by the heuristic π. The tiebreaking rules ensure that any other items selected by ψ with $i_{t,s}(x_{t,s}) = \lambda_t$ will also be selected by π. π may also select up to $N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))$ additional items with nonnegative priority indices that were not selected by ψ. (We note for future reference that if $\lambda_t = 0$, then in this case ψ and π will select exactly the same items.)

Case (ii): If the Lagrangian policy ψ selects $n(\psi_t(\widetilde{\boldsymbol{x}}_t)) \ge N_t$ items, these items selected by ψ will all have nonnegative priority indices and the heuristic π will select N_t of these items: the tiebreaking rules ensure that the N_t selected by π will be a subset of those selected by ψ. Thus, in both cases (i) and (ii), ψ and π will select no more than $\big|n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\big|$ different items in period t and state $\widetilde{\boldsymbol{x}}_t$.

The desired result (47) can now be established as follows:

$$\begin{aligned}
d_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t), \pi_t(\widetilde{\boldsymbol{x}}_t)) &= \underbrace{\lambda_t\big(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))\big)}_{(a)} + \underbrace{r_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t)) - r_t(\widetilde{\boldsymbol{x}}_t, \pi_t(\widetilde{\boldsymbol{x}}_t))}_{(b)} \\
&\qquad + \underbrace{\mathbb{E}\big[\, V^{\pi}_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t)))\,\big] - \mathbb{E}\big[\, V^{\pi}_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \pi_t(\widetilde{\boldsymbol{x}}_t)))\,\big]}_{(c)} \\
&\le \underbrace{\lambda_t(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t)))}_{(a)} + \underbrace{(\bar{r} - \underline{r})\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|}_{(b')} \\
&\qquad + \underbrace{2(\bar{r} - \underline{r})k_{t+1}\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|}_{(c')} \\
&= \lambda_t\big(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))\big) + (\bar{r} - \underline{r})(1 + 2k_{t+1})\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|
\end{aligned}$$

The inequality above follows term by term, using the terms identified above.

- The (a) term is unchanged.
- $(b) \le (b')$: Because ψ and π will select no more than $\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$ different items, we have

$$r_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t)) - r_t(\widetilde{\boldsymbol{x}}_t, \pi_t(\widetilde{\boldsymbol{x}}_t)) \;\le\; (\bar{r} - \underline{r})\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big| \;.$$

- $(c) \le (c')$: Because ψ and π will select no more than $\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$ different items and state transitions are independent across items, the random continuation states $\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}', \pi(\widetilde{\boldsymbol{x}}'))$ and $\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}'', \pi(\widetilde{\boldsymbol{x}}''))$ will differ for at most $\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$ items. Lemma 2 then implies

$$\mathbb{E}\big[\, V^{\pi}_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t)))\,\big] - \mathbb{E}\big[\, V^{\pi}_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}_t, \pi_t(\widetilde{\boldsymbol{x}}_t)))\,\big] \;\le\; (\bar{r} - \underline{r})k_{t+1}\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$$

  where k_t is as defined in Lemma 2.

The desired result then follows by taking $c_t = (1 + k_{t+1})$.

In the case where $\lambda_t = 0$, as discussed above in Case (i), ψ and π will select the same items, so combining Cases (i) and (ii), ψ and π will select no more than $\max\{n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t, 0\}$ different items. The proof then proceeds as before. □

The final lemma provides a bound on the expectations of the $\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$ terms appearing in Lemma 3 by calculating the variance of these quantities.

Lemma 4. *Let $\boldsymbol{\lambda}^*$ denote an optimal solution for the Lagrangian dual problem* (7) *with initial state $\boldsymbol{x}$ and let $\widetilde{\psi}$ denote an optimal mixed policy. Let $\tilde{n}_t(\widetilde{\psi}) = n(\widetilde{\psi}_t(\widetilde{\boldsymbol{x}}_t(\boldsymbol{x}, \widetilde{\psi})))$.*

(i) *If $\lambda_t > 0$, then*

$$\mathbb{E}\big[\, \big|\, \tilde{n}_t(\widetilde{\psi}) - N_t\,\big|\,\big] \le \sqrt{N_t(1 - N_t/S)} \;. \tag{48}$$

(ii) *If $\lambda_t = 0$, then*

$$\mathbb{E}[\max\{\tilde{n}_t(\widetilde{\psi}) - N_t, 0\}\,] \le \sqrt{\bar{N}_t(1 - \bar{N}_t/S)} \;, \tag{49}$$

where $\bar{N}_t = \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] \le N_t$.

Proof. We first characterize the variance of $\tilde{n}_t(\widetilde{\psi})$. Since the state transitions are independent across items and the policy mixing is also independent across items,$\tilde{n}_t(\widetilde{\psi})$ is the sum of S independent Bernoulli trials with probabilities of success $p_{t,s} = \mathbb{E}[\,p_{t,s}(\widetilde{\psi}_s)\,]$ where, as in Proposition 4, $p_{t,s}(\psi_s)$ is the probability of selecting item s in period t when following a policy ψ_s. We then have $\mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] = \sum_{s=1}^{S} p_{t,s}$ and

$$\mathbb{V}\mathrm{ar}[\,\tilde{n}_t(\widetilde{\psi})\,] = \sum_{s=1}^{S} p_{t,s}(1 - p_{t,s})$$

$$
\begin{aligned}
&= \sum_{s=1}^{S} p_{t,s} - \sum_{s=1}^{S} p_{t,s}^2 \\
&= \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] - \sum_{s=1}^{S} p_{t,s}^2 \\
&\le \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] - \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,]^2/S \\
&= \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,](1 - \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,]/S)
\end{aligned}
$$

The inequality follows from choosing $p_{t,s}$ to minimize $\sum_{s=1}^{S} p_{t,s}^2$ subject to the constraint that $\sum_{s=1}^{S} p_{t,s} = \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,]$. The minimum is obtained when $p_{t,s} = \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,]/S$ for all s.

We then apply this inequality for the two different cases for λ_t. Case (i): If $\lambda_t > 0$, by Proposition 4(iii), we know that $\mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] = N_t$. Then we have

$$
\begin{aligned}
\mathbb{E}\big[\,\big|\,\tilde{n}_t(\widetilde{\psi}) - N_t\,\big|\,\big]^2 &\le \mathbb{V}\mathrm{ar}[\,\tilde{n}_t(\widetilde{\psi}) - N_t\,] \\
&= \mathbb{V}\mathrm{ar}[\,\tilde{n}_t(\widetilde{\psi})\,] \\
&\le \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,](1 - \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,]/S) \\
&= N_t(1 - N_t/S)
\end{aligned}
$$

The first inequality follows from Jensen's inequality and the fact that $\mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] = N_t$.

Case (ii): If $\lambda_t = 0$, by Proposition 4(iii), we know that $\bar{N}_t \equiv \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] \le N_t$. Then, following the same logic as in the $\lambda_t > 0$ case after two preliminary steps:

$$
\begin{aligned}
\mathbb{E}[\,\max\{\tilde{n}_t(\widetilde{\psi}) - N_t, 0\}\,]^2 &\le \mathbb{E}\big[\,\max\big\{\tilde{n}_t(\widetilde{\psi}) - \bar{N}_t, 0\big\}\,\big]^2 \\
&\le \mathbb{E}\big[\,\big|\,\tilde{n}_t(\widetilde{\psi}) - \bar{N}_t\,\big|\,\big]^2 \\
&\le \mathbb{V}\mathrm{ar}\big[\,\tilde{n}_t(\widetilde{\psi}) - \bar{N}_t\,\big] \\
&= \mathbb{V}\mathrm{ar}[\,\tilde{n}_t(\widetilde{\psi})\,] \\
&\le \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,](1 - \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,]/S) \\
&= \bar{N}_t(1 - \bar{N}_t/S)
\end{aligned}
$$

□

Finally, we can assemble these results and prove Proposition 5.

Proof of Proposition 5. Using the notation of Lemmas 1, 3, and 4 and applying these results in that order, we have:

$$
\begin{aligned}
L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}) - V_1^{\widetilde{\pi}}(\boldsymbol{x}) &= \sum_{t=1}^{T} \mathbb{E}[\,d_t(\widetilde{\boldsymbol{x}}_t, \widetilde{\psi}, \widetilde{\pi})\,] \\
&\le \sum_{t=1}^{T} \left\{ \begin{array}{ll} \lambda_t^* \mathbb{E}[\,N_t - n(\widetilde{\psi}(\widetilde{\boldsymbol{x}}_t))\,] + c_t(\bar{r} - \underline{r})\mathbb{E}\big[\,\big|\,n(\widetilde{\psi}(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|\,\big] & \text{if } \lambda_t^* > 0 \\ c_t(\bar{r} - \underline{r})\mathbb{E}[\,\max\{n(\psi_t(\boldsymbol{x}_t)) - N_t, 0\}\,] & \text{if } \lambda_t^* = 0 \end{array} \right\} \\
&\le \sum_{t=1}^{T} c_t(\bar{r} - \underline{r})\sqrt{\bar{N}_t(1 - \bar{N}_t/S)}
\end{aligned}
$$

where $\bar{N}_t = N_t$ if $\lambda_t^* > 0$ and $\bar{N}_t = \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] \le N_t$ if $\lambda_t^* = 0$. In the final step above, we also use the fact that $\mathbb{E}[\,N_t - n(\widetilde{\psi}(\widetilde{\boldsymbol{x}}_t))\,] = 0$ when $\lambda_t^* > 0$; see Proposition 4(iii). When considering expectations involving the mixed policies, we assume that the realizations of $\widetilde{\psi}$ and $\widetilde{\pi}$ are coordinated so the realized π is the Lagrangian index policy with the realized ψ as tiebreaker: this is necessary when applying Lemma 3 in the second line above. Taking $\beta_t = c_t = 2^{T-t+1} - 1$, we obtain the result of the proposition.

The final inequality in (18) then follows from the fact that $\sqrt{\bar{N}_t(1 - \bar{N}_t/S)} \le \sqrt{\bar{N}_t} \le \sqrt{N}$. □

Proof of Corollary 1. Theorem 1 implies

$$\begin{aligned} 0 \;&\leq\; \frac{L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};S) - V_1^{\widetilde{\pi}}(\boldsymbol{x};S)}{V_1^*(\boldsymbol{x};S)} \\ &\leq\; (\bar{r}-\underline{r})\frac{\sum_{t=1}^{T}\beta_t\sqrt{\bar{N}_t(S)(1-\bar{N}_t(S)/S)}}{V_1^*(\boldsymbol{x};S)} \\ &\leq\; (\bar{r}-\underline{r})\sum_{t=1}^{T}\beta_t\frac{\sqrt{N(S)}}{V_1^*(\boldsymbol{x};S)}. \end{aligned}$$

The growth assumption implies $\lim_{S\to\infty}\sqrt{N(S)}/V_1^*(\boldsymbol{x};S)=0$, which gives the desired result (22). □

D.2. Example Showing the Lagrangian Performance Gap of $\sqrt{N}$ is Tight

We consider an example with $T=2$ and assume the number of items S is divisible by 4. The DM can select $N_1=N_2=N=S/2$ items in each period. There are three types of items:

(i) $S/2$ items are *a priori* identical and yield rewards $r_{t,s}(x_s^0,1)=1$ in their initial state x_s^0. If selected in period one, in period two these items transition to state $\bar{x}$ with probability 1/2 and to state $\underline{x}$ with probability 1/2, with $r_{2,s}(\bar{x},1)=2$ and $r_{2,s}(\underline{x},1)=0$. If not selected, these items do not change state. Let $\mathcal{S}_1$ denote this set of items.

(ii) $S/4$ items are identical and yield deterministic rewards $r_{t,s}(x_s^0,1)=1/2$ if selected in either period, and never transition from their initial state x_s^0, whether selected or not. Let $\mathcal{S}_2$ denote this set of items.

(iii) The remaining $S/4$ items are identical and yield deterministic rewards $r_{t,s}(x_s^0,1)=1/4$ if selected in either period, and never transition from their initial state x_s^0, whether selected or not. Let $\mathcal{S}_3$ denote this set of items.

All items yield zero reward when not selected.

Solution of the Lagrangian Dual. First, we claim that the Lagrange multipliers $\boldsymbol{\lambda}^*=(\lambda_1^*,\lambda_2^*)=(1/2,1/4)$ are optimal for the Lagrangian dual (7) for this example. To see this, note that with this choice of $\boldsymbol{\lambda}^*$, we have the following optimal Lagrangian value functions and policies:

(i) For $s\in\mathcal{S}_1$: In period two, $V_{2,s}^{\boldsymbol{\lambda}^*}(\bar{x})=7/4$, $V_{2,s}^{\boldsymbol{\lambda}^*}(\underline{x})=0$, $\mathbb{E}\big[V_{2,s}^{\boldsymbol{\lambda}^*}(\widetilde{\chi}_{1,s}(x_s^0,1))\big]=7/8$, and it is strictly optimal to select in state $\bar{x}$ and not select in state $\underline{x}$. In period 1, it is strictly optimal to select: the value of selecting is $r_{1,s}(x_s^0,1)-\lambda_1^*+\mathbb{E}\big[V_{2,s}^{\boldsymbol{\lambda}^*}(\widetilde{\chi}_{1,s}(x_s^0,1))\big]=11/8$ and the value of not selecting is $0+V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)=1-\lambda_2^*=3/4$. Thus, for $s\in\mathcal{S}_1$, there is a single optimal policy ψ_s for $s\in\mathcal{S}_1$.

(ii) For $s\in\mathcal{S}_2$: In period two, $V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)=1/4$ and it is strictly optimal to select. In period one, selecting or not selecting are both optimal: the value for selecting is $r_{1,s}(x_s^0,1)-\lambda_1^*+V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)=1/4$ and the value for not selecting is $V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)=1/4$. For all $s\in\mathcal{S}_2$, we take ψ_s to be the optimal policy that does not select these items in period one.

(iii) For $s\in\mathcal{S}_3$: In period two, $V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)=0$ and selecting and not selecting are both optimal. In period one, not selecting is strictly optimal. For all $s\in\mathcal{S}_3$, we take ψ_s to be the optimal policy that does not select these items in period two.

With these optimal policies, we select exactly $N=S/2$ items (all items in $\mathcal{S}_1$) in period one. In period two, we select those items in $\mathcal{S}_1$ that transition to $\bar{x}$ (expected number equal to $S/4$) and select all $S/4$ items in $\mathcal{S}_2$, for a total of $S/2$ items in expectation. By Proposition 4(iii), this implies that $\boldsymbol{\lambda}^*=(1/2,1/4)$ is optimal.

Total Reward with the Optimal Policy for the Lagrangian Relaxation. In the Lagrangian relaxation, it is optimal to select all items in $\mathcal{S}_1$ in the first period. We let Y denote the random variable corresponding to the number of items in $\mathcal{S}_1$ that transition to $\bar{x}$ in period two. The distribution of Y is binomial with $S/2$ trials and probability 1/2.

The first period rewards are simply $S/2$, as exactly $N=S/2$ items with reward 1 are selected. In the second period, all Y items in $\mathcal{S}_1$ are selected and yield reward 2, and all $S/4$ items in $\mathcal{S}_2$, each yielding reward 1/2, are selected. The Lagrangian penalty in period two is $\lambda_2^*(S/2-Y-S/4)=S/16-Y/4$. Putting this together, the total reward in the Lagrangian relaxation given Y is $(7/4)Y+(11/16)S$.

Total Reward with the Optimal Lagrangian Index Policy. In the first period, the priority index values are:

$$s \in \mathcal{S}_1 : i_{1,s}(x_s^0) = (r_{1,s}(x_s^0, 1) + \mathbb{E}\Big[\, V_{2,s}^{\boldsymbol{\lambda}^*}(\widetilde{\chi}_{1,s}(x_s^0, 1))\,\Big]) - (r_{1,s}(x_s^0, 0) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) = (1 + 7/8) - (0 + 3/4) = 9/8,$$
$$s \in \mathcal{S}_2 : i_{1,s}(x_s^0) = (r_{1,s}(x_s^0, 1) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) - (r_{1,s}(x_s^0, 0) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) = (1/2 + 1/4) - (0 + 1/4) = 1/2,$$
$$s \in \mathcal{S}_3 : i_{1,s}(x_s^0) = (r_{1,s}(x_s^0, 1) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) - (r_{1,s}(x_s^0, 0) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) = (1/4 + 0) - (0 + 0) = 1/4,$$

and thus all items in $\mathcal{S}_1$ are selected in the first period by the optimal Lagrangian index policy.

In the second period, the selection indices in the optimal Lagrangian index policy equal the item's rewards in their current state. Thus, in period two, the optimal Lagrangian index policy selects all Y items in $\mathcal{S}_1$ that yield reward 2, possibly in addition to some other items, which differ in two cases:

(a) If $Y < S/4$, then all $S/4$ items in $\mathcal{S}_2$ are also selected, each yielding reward $1/2$, as well as $S/2 - (Y + S/4) = S/4 - Y$ items in $\mathcal{S}_3$ are selected, each yielding reward $1/4$. The total reward (including period one) in this case is $(7/4)Y + (11/16)S$, equal to the Lagrangian relaxation value.

(b) If $Y \geq S/4$, then $S/2 - Y \leq S/4$ items from $\mathcal{S}_2$ are also selected, yielding a total reward (including period one) of $(3/2)Y + (3/4)S$.

Difference in Total Rewards. It follows that the difference between the Lagrangian relaxation value $L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x})$ and optimal Lagrangian index policy $V_1^{\widetilde{\pi}}(\boldsymbol{x})$ is

$$\begin{aligned}
L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}) - V_1^{\widetilde{\pi}}(\boldsymbol{x}) &= \mathbb{E}\left[\, \mathbb{1}\{Y \geq S/4\}\left(\frac{7}{4}Y + \frac{11}{16}S - \frac{3}{2}Y - \frac{3}{4}S\right)\,\right] \\
&= \mathbb{E}\left[\, \mathbb{1}\{Y \geq S/4\}\left(\frac{Y}{4} - \frac{S}{16}\right)\,\right] \\
&= \frac{1}{4}\mathbb{E}\left[\, \mathbb{1}\{Y \geq S/4\}\left(Y - \frac{S}{4}\right)\,\right] \\
&= \frac{1}{4}\mathbb{E}\left[\, \left(Y - \frac{S}{4}\right)^+\,\right].
\end{aligned}$$

Y follows a binomial distribution with $S/2$ trials and probability $1/2$ so, as $S \to \infty$, $Y - S/4$ approaches a normal distribution with mean zero and variance $S/8$. Then in the limit as $S \to \infty$, $\big|Y - S/4\big|$ follows a half-normal distribution generated by a normal random variable with variance $S/8$; thus, as $S \to \infty$,

$$L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S) - V_1^{\widetilde{\pi}}(\boldsymbol{x}; S) = \frac{1}{4}\mathbb{E}\big[\,(Y - S/4)^+\,\big] = \frac{1}{8}\mathbb{E}\big[\,\big|Y - S/4\big|\,\big] = \frac{\sqrt{2S}}{8\sqrt{8\pi}} = \frac{\sqrt{N}}{8\sqrt{2\pi}}\,.$$

E. Information Relaxation Bounds

As discussed briefly in §6.3, in the numerical examples of §6 the gaps between the optimal Lagrangian index policy and Lagrangian bound were very small (in relative terms) for large S, but were more substantial for small S. One might wonder whether these gaps are due to the policies being suboptimal or due to slack in the Lagrangian bound. In this section, we develop information relaxation bounds to provide tighter bounds. Here we follow the general approach developed in Brown, Smith and Sun (2010, BSS hereafter) but the application to dynamic selection problems poses some problem-specific challenges which we address here.

BSS (2010) generalized earlier applications of information relaxations for valuing American options (see, e.g., Haugh and Kogan 2004 and Rogers 2002). Our application to dynamic selection problems can be viewed as a new application in a growing list of applications of information relaxation methods. In addition to the many applications to valuing options and other derivative securities, recent applications of information relaxations include managing natural gas storage (Lai et al. 2010 and Lai et al. 2011), dynamic portfolio optimization with transaction costs or taxes (Brown and Smith 2011 and Haugh et al. 2016), and inventory and pricing models with lead time and backorders (Brown and Smith 2014 and Bernstein et al. 2015). Our application of information relaxations to the dynamic selection problem combines information relaxations and Lagrangian relaxations. Information relaxations and Lagrangian relaxations were similarly combined in a network revenue management problem in Brown and Smith (2014), in a multiclass queueing problem in Brown and Haugh (2017), and in Ye et al. (2018).

In this section, we first briefly and informally review the theory of information relaxation bounds as developed in BSS (2010), discuss the application to our examples, and the discuss numerical results for the examples considered in §6.

E.1. Information Relaxation Bounds

The key idea of information relaxation bounds is to consider models that relax the *nonanticipativity* constraints that require the DM to make decisions based only on information that is available at the time the decision is made. For instance in the dynamic assortment problem, in the real model, the DM observes demands for products that are displayed, when they are displayed, and uses this information to guide future display decisions. We will consider a relaxed model where the DM knows the demands for all products in all periods in advance, before making any display decisions.

The basic results on information relaxations are easiest to state if we take a high-level view of policies. If we let $\Pi_{\mathbb{F}}$ denote the set of policies that respect the nonanticipativity constraints (as well as the linking constraints) in the original problem, we can write the DP (2) as

$$V_1^*(\boldsymbol{x}) = \max_{\pi \in \Pi_{\mathbb{F}}} \mathbb{E}[\, r(\pi)\,]$$

where $r(\pi)$ denotes the random total reward under policy π, i.e., $r(\pi) = \sum_t r_t(\widetilde{\boldsymbol{x}}_t(\pi), \pi_t(\widetilde{\boldsymbol{x}}_t(\pi)))$ where $\widetilde{\boldsymbol{x}}_t(\pi)$ represents the random state-evolution process when starting in state $\boldsymbol{x}$ and following policy π and $\pi_t(\boldsymbol{x})$ is the period-t vector of selection decisions in state $\boldsymbol{x}$ when using policy π.

If we let $\Pi_{\mathbb{G}}$ denote a larger set of policies ($\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$) that can use additional information,[10] we can solve a relaxed version of the DP to obtain an upper bound on the primal DP:

$$V_1^*(\boldsymbol{x}) = \max_{\pi \in \Pi_{\mathbb{F}}} \mathbb{E}[\, r(\pi)\,] \le \max_{\pi \in \Pi_{\mathbb{G}}} \mathbb{E}[\, r(\pi)\,]\ . \tag{50}$$

Unfortunately, the bounds given by (50) will be weak if the extra information provided in the relaxation is valuable. To counter this, we incorporate a penalty that "punishes" the DM for using information that would not actually be available when making decisions. The penalty $z(\pi)$ is a policy-dependent random variable, like the rewards, i.e., $z(\pi) = \sum_t z_t(\widetilde{\boldsymbol{x}}_t(\pi), \pi_t(\widetilde{\boldsymbol{x}}_t(\pi)))$ for some set of period-t penalty terms $z_t(x_t, u_t)$. A penalty $z(\pi)$ is *dual feasible* if $\mathbb{E}[\, z(\pi)\,] \le 0$ for all $\pi \in \Pi_{\mathbb{F}}$; that is, if the expected penalty is nonpositive for all nonanticipative policies.

[10] To formalize the definitions of these sets of policies, a policy can be defined as a mapping from the underlying outcome space to selection decisions $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ for each product and each period (with $\boldsymbol{u}_t \in \mathcal{U}_t$). Policies in the DP (2) that make selections as a function of the current state of the system can be viewed as imposing measurability restrictions on this more general set of policies. The relaxed model imposes a weaker set of measurability restrictions. See BSS (2010) for more discussion.

The following weak duality result from BSS (2010) is the key tool for generating performance bounds using information relaxations.

Proposition 8 (Weak duality)**.** *Suppose $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. If policy π is nonanticipative (i.e., $\pi \in \Pi_{\mathbb{F}}$) and penalty z is dual feasible then*

$$\mathbb{E}[\, r(\pi)\,] \;\leq\; \max_{\pi' \in \Pi_{\mathbb{G}}} \mathbb{E}[\, r(\pi') - z(\pi')\,] \;. \tag{51}$$

Proof. We have:

$$\mathbb{E}[\, r(\pi)\,] \;\leq\; \mathbb{E}[\, r(\pi) - z(\pi)\,] \;\leq\; \max_{\pi' \in \Pi_{\mathbb{G}}} \mathbb{E}[\, r(\pi') - z(\pi'))\,] \;.$$

Given $\pi \in \Pi_{\mathbb{F}}$, the first inequality follows from the definition of dual feasibility ($\mathbb{E}[\, z(\pi)\,] \leq 0$) and the second inequality follows from the fact that $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. □

BSS (2010) provide a strong duality result that shows that there is a penalty such that the value for the relaxed model is exactly equal to the optimal value for the original, but these penalties require knowledge of the optimal value function (more on this in the next subsection).

We also note that if we can restrict attention to a subset of the available policies $\Pi_{\mathbb{F}}$ in the original problem without loss of optimality, we can impose these same restrictions on the policies $\Pi_{\mathbb{G}}$ for the relaxed model. For example, if all items are initially identical in the dynamic assortment or applicant screening examples, we can restrict the policies to a set of policies that select the first (in label index order) N_t items in the initial period (i.e., $s \leq N_t$), without loss of optimality. More generally, we can restrict the DM to policies to selecting items with $s \leq \sum_{\tau=1}^{t} N_\tau$ in period t. In our numerical examples, we will impose these restrictions on selections in the relaxed model. Enforcing these constraints can improve the information relaxation bound (i.e., lead to a lower value) because the information revealed in a particular sample scenario may favor selecting some items outside this restricted set.

E.2. Information Relaxation Bounds for the Dynamic Assortment Problem

The challenge is to find penalties and information relaxations that make the bound on right side of (51) easy to compute and lead to reasonably tight bounds. For specificity, we will focus our discussion on the dynamic assortment example, though the ideas also apply in the applicant screening example and other dynamic selection problems. In the dynamic assortment example, the underlying uncertainties are the unknown (Poisson) demand rates for each product and the demand realizations for each item, in each period. In the original model, the demands are revealed for products when (and if) the products are selected; the demand rates are never revealed. We can consider a number of different relaxations, including:

(i) *Known rates*: The DM knows the demand rates for all products in advance, but demands are revealed sequentially only when the products are selected, as in the original model.

(ii) *Known demands*: The DM knows all demands for all products in all periods, in advance before making any selection decisions (i.e., the DM knows what demand would be if a product were to be selected); demands rates are never revealed.

(iii) *Perfect information*: The DM knows both demands and rates in advance.

(iv) *Uncensored demand*: Demands for all products are revealed sequentially (regardless of whether they are selected or not); demand rates are never revealed.

In the applicant screening example, we can consider analogous relaxations, where the applicants' quality and/or the signals are known in advance in the relaxed model.

In our discussion and numerical examples, we will focus on the known demands relaxation and consider a penalty based on the Lagrangian $L_{t+1}^{\boldsymbol{\lambda}}(\boldsymbol{x})$. Although we can use any $\boldsymbol{\lambda} \geq \mathbf{0}$, in our numerical examples we will take these to be optimal Lagrange multipliers $\boldsymbol{\lambda}^*$ given by solving the Lagrangian dual (7). We can estimate the known demands bound, $\max_{\pi' \in \Pi_{\mathbb{G}}} \mathbb{E}[\, r(\pi') - z(\pi')\,]$, by repeatedly:

(i) Drawing a demand rate γ_s for product s from the appropriate gamma distribution and then drawing demands for this product from a Poisson distribution with this rate. Let $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_T)$ where $\boldsymbol{d}_t = (d_{t,1}, \ldots, d_{t,S})$ denotes the randomly generated vector of product demands in period t.

(ii) Solving a deterministic *inner DP* (to be described shortly) to find the optimal value $\hat{V}_1(\boldsymbol{x}_1; \boldsymbol{d})$ given these demand realizations, incorporating the Lagrangian penalty.

We estimate the known demands bound by averaging the $\hat{V}_1(\boldsymbol{x}_1; \boldsymbol{d})$ for the different demand realizations $\boldsymbol{d}$.

Given a demand scenario $\boldsymbol{d}$, we can write the inner DP for this demand scenario as follows. Let $\hat{V}_{T+1}(\boldsymbol{x};\boldsymbol{d}) = 0$ and, for earlier t, we recursively define

$$\hat{V}_t(\boldsymbol{x};\boldsymbol{d}) = \max_{\boldsymbol{u}\in\mathcal{U}_t} \left\{ r_t(\boldsymbol{x},\boldsymbol{u}) - z_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t) + \hat{V}_{t+1}(\boldsymbol{\chi}_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t);\boldsymbol{d}) \right\} \tag{52}$$

where

$$z_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t) = L^{\boldsymbol{\lambda}}_{t+1}(\boldsymbol{\chi}_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t)) - \mathbb{E}\big[\, L^{\boldsymbol{\lambda}}_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x},\boldsymbol{u}))\,\big]\ . \tag{53}$$

Here the last term in (52) and the first term in (53) involve deterministic state transitions because the DM knows the demands: $\boldsymbol{\chi}_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t) = (\chi_{t,1}(x_1,u_1;d_{t,1}),\ldots,\chi_{t,S}(x_S,u_S;d_{t,S}))$ represents the state transitions with the given product demands for period t. The expectation in (53) is calculated using the same state-dependent negative-binomial distributions used in the original DP.

Using the law of iterated expectations, we know that $\mathbb{E}[\, z_t(\widetilde{\boldsymbol{x}}_t(\pi),\pi_t(\widetilde{\boldsymbol{x}}_t(\pi)))\,] = 0$ for any nonanticipative policy π. Thus the penalty $z(\pi) = \sum_t z_t(\widetilde{\boldsymbol{x}}_t(\pi),\pi_t(\widetilde{\boldsymbol{x}}_t(\pi)))$ is dual feasible and the known demands bound provides a performance bound, as in Proposition 8. This is an example of the general method for creating "good" dual feasible penalties described in BSS (2010). As discussed there, if we replace the Lagrangian $L^{\boldsymbol{\lambda}}_{t+1}$ in (53) with the optimal value function V^*_{t+1}, the information relaxation bound will be exactly equal to the optimal value. With this ideal penalty, the DM is exactly punished for using extra information: the benefit gained is exactly canceled by the penalty. With a penalty based on an approximate value function (such as the Lagrangian), the penalty approximately cancels this benefit. In general, to obtain good bounds, we want to choose generating functions that approximate the optimal value function well.

We now consider the DP (52) in more detail. First, note that that the penalty terms involving the Lagrangian $L^{\boldsymbol{\lambda}}_{t+1}$ decompose into the sum of item-specific values, as in (5). However, the inner DP (52) does not decompose into item-specific subproblems because the constraint on the total number of products selected ($\boldsymbol{u}\in\mathcal{U}_t$ where $\mathcal{U}_t$ is defined in (1)) links the decisions across items, as it did in the original DP (2). Thus, the inner DP – though deterministic – is still difficult to solve in problems with many items.

To decouple the inner DP (52), we relax the linking constraint in the same way that we relaxed the original DP (2). Consider Lagrange multipliers $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_T) \geq \mathbf{0}$ and let $\hat{L}^{\boldsymbol{\mu}}_{T+1}(\boldsymbol{x};\boldsymbol{d}) = 0$. The period-$t$ *inner Lagrangian* with demand realization $\boldsymbol{d}$ is then given recursively as

$$\hat{L}^{\boldsymbol{\mu}}_t(\boldsymbol{x};\boldsymbol{d}) = \max_{\boldsymbol{u}\in\{0,1\}^S} \left\{ r_t(\boldsymbol{x},\boldsymbol{u}) - z_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t) + \hat{L}^{\boldsymbol{\mu}}_{t+1}(\boldsymbol{\chi}_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t);\boldsymbol{d}) + \mu_t\left(N_t - \sum_{s=1}^S u_s\right)\right\}\ .$$

This can be decomposed into item-specific DPs as

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

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

$$\hat{V}^{\boldsymbol{\mu}}_{t,s}(x_s;\boldsymbol{d}_s) = \max\Bigg\{ r_{t,s}(x_s,1) - \mu_t - V^{\boldsymbol{\lambda}}_{s,t+1}(\chi_{t,s}(x_s,1,d_{t,s})) + \mathbb{E}\big[\, V^{\boldsymbol{\lambda}}_{s,t+1}(\widetilde{\chi}_{t,s}(x_s,1))\,\big] + \hat{V}^{\boldsymbol{\mu}}_{t+1,s}(\chi_{t,s}(x_s,1,d_{t,s})),$$
$$r_{t,s}(x_s,0) + \hat{V}^{\boldsymbol{\mu}}_{t+1,s}(\chi_{t,s}(x_s,0,d_{t,s}))\Bigg\}. \tag{54}$$

where $V^{\boldsymbol{\lambda}}_{t,s}$ is the value-function for the item-specific DP (6). Note that in the dynamic assortment model, the penalty term (53) is zero if a product is not selected because its state does not change.

These inner item-specific DPs and the Lagrangian satisfy properties like those of Propositions 1-4. In particular, the Lagrangian is an upper bound on the inner DP: $\hat{V}_t(\boldsymbol{x};\boldsymbol{d}) \leq \hat{L}^{\boldsymbol{\mu}}_t(\boldsymbol{x};\boldsymbol{d})$ for all $\boldsymbol{x}$, t, $\boldsymbol{d}$ and $\boldsymbol{\mu}\geq\mathbf{0}$.

To ensure we have the best possible bound for a given $\boldsymbol{d}$ and $\boldsymbol{x}$, we can solve the inner dual problem,

$$\min_{\boldsymbol{\mu} \geq \mathbf{0}} \ \hat{L}_1^{\boldsymbol{\mu}}(\boldsymbol{x}; \boldsymbol{d}) \ , \tag{55}$$

for an optimal $\boldsymbol{\mu}^*(\boldsymbol{x}, \boldsymbol{d})$. This is a convex optimization problem and can be solved using the cutting-plane method discussed in §A. Moreover, if we take the inner Lagrange multipliers $\boldsymbol{\mu}$ to be equal to the "outer" Lagrange multipliers $\boldsymbol{\lambda}$ used to define the penalty, we can use an induction argument to show that $\hat{L}_t^{\boldsymbol{\lambda}}(\boldsymbol{x}; \boldsymbol{d}) = L_t^{\boldsymbol{\lambda}}(\boldsymbol{x})$ for all t and $\boldsymbol{d}$.[11] Thus, since $\boldsymbol{\lambda}$ is feasible but not necessarily optimal for the inner Lagrangian dual problem (55), we have

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

Thus, for every demand scenario $\boldsymbol{d}$, the information relaxation bound $\hat{V}_1(\boldsymbol{x}; \boldsymbol{d})$ and its computable upper bound $\hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{x}, \boldsymbol{d})}(\boldsymbol{x}; \boldsymbol{d})$ will be at least as good as the Lagrangian bound $L_1^{\boldsymbol{\lambda}}(\boldsymbol{x})$.

We can also relate these bounds to the performance of a heuristic policy π in the same demand scenario. We focus on deterministic Markovian heuristic policies where the period-t selection decision π_t is chosen based on the current state $\boldsymbol{x}$. (When we are considering mixed policies, as in the optimal Lagrangian policy, let π be a particular realization of the mixed policy.) We assume that the actions selected by the heuristic are feasible, i.e., $\pi_t(\boldsymbol{x}) \in \mathcal{U}_t$. To facilitate comparison with those of the information relaxation, we will adjust the rewards using the penalty (53) as a control variate. Let $\hat{V}_t^{\pi}(\boldsymbol{x}; \boldsymbol{d})$ denote the value generated when following policy π, starting in state $\boldsymbol{x}$, given demand realization $\boldsymbol{d}$, adjusted by the control variate. We can write this value recursively in a form parallel to (52): let $\hat{V}_{T+1}^{\pi}(\boldsymbol{x}; \boldsymbol{d}) = 0$ and, for earlier t, we define

$$\hat{V}_t^{\pi}(\boldsymbol{x}; \boldsymbol{d}) = \left\{ r_t(\boldsymbol{x}, \pi_t(\boldsymbol{x})) - z_t(\boldsymbol{x}, \pi_t(\boldsymbol{x}); \boldsymbol{d}_t) + \hat{V}_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}, \pi_t(\boldsymbol{x}); \boldsymbol{d}_t); \boldsymbol{d}) \right\} . \tag{56}$$

Here this form exactly mimics the DP recursion (52), except the actions are chosen in accordance to the policy π rather than optimized. Thus we know that $\hat{V}_t^{\pi}(\boldsymbol{x}; \boldsymbol{d}) \leq \hat{V}_t(\boldsymbol{x}; \boldsymbol{d})$ for all t, $\boldsymbol{x}$, and $\boldsymbol{d}$. Moreover, because the penalty terms z_t have mean zero for all feasible policies, we know that the expected total reward when following policy π is $V_1^{\pi}(\boldsymbol{x}) = \mathbb{E}[\, \hat{V}_1^{\pi}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$, where the expectations are taken over the random demand scenarios. These control variates are helpful in reducing sampling error when estimating the expected values associated with a given policy and were used in the simulations of §6.2.

Combining these observations, we can say the following.

Theorem 2 (Ordered bounds)**.** *Consider any feasible and nonanticipative policy π, Lagrange multipliers $\boldsymbol{\lambda} \geq \mathbf{0}$ and initial state $\boldsymbol{x}$.*

(i) *For any demand realization $\boldsymbol{d}$, we have*

$$\hat{V}_1^{\pi}(\boldsymbol{x}; \boldsymbol{d}) \ \leq \ \hat{V}_1(\boldsymbol{x}; \boldsymbol{d}) \ \leq \ \hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{x}, \boldsymbol{d})}(\boldsymbol{x}; \boldsymbol{d}) \ \leq \ L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) \ . \tag{57}$$

(ii) *Taking expectations over random demand realizations $\tilde{\boldsymbol{d}}$, we have*

$$V_1^{\pi}(\boldsymbol{x}) \ = \ \mathbb{E}\Big[\hat{V}_1^{\pi}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \Big] \ \leq \ V_1^*(\boldsymbol{x}) \ \leq \ \mathbb{E}\Big[\hat{V}_1(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \Big] \ \leq \ \mathbb{E}\Big[\hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{x}, \tilde{\boldsymbol{d}})}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \Big] \ \leq \ L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) \ . \tag{58}$$

Working from the left in (58), we have the expected value with heuristic policy π ($V_1^{\pi}(\boldsymbol{x})$) is equal to the expected reward for this policy with the control variate included ($\mathbb{E}[\, \hat{V}_1^{\pi}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$). This value is less than or equal to the value with an optimal policy ($V_1^*(\boldsymbol{x})$), which is typically impossible to compute. This, in turn, is less than or equal to the known demands relaxation bound ($\mathbb{E}[\, \hat{V}_1(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$) which is also typically impossible to compute. However, the known demands bound is less than or equal to the Lagrangian relaxation of the known demands information relaxation bound with optimized Lagrange multipliers ($\mathbb{E}[\, \hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{x}, \boldsymbol{d})}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$), which is computable. Finally, all of these bounds are less than the ordinary Lagrangian bound ($L_1^{\boldsymbol{\lambda}}(\boldsymbol{x})$). The

[11] Note that the $V_{s,t+1}^{\boldsymbol{\lambda}}(\cdot)$ and $\hat{V}_{t+1,s}^{\boldsymbol{\mu}}(\cdot)$ terms in (54) cancel if $\boldsymbol{\mu} = \boldsymbol{\lambda}$ and we have the induction hypothesis that $V_{s,t+1}^{\boldsymbol{\lambda}}(\cdot) = \hat{V}_{s,t+1}^{\boldsymbol{\lambda}}(\cdot)$. Then (54) reduces to the the definition of $V_{s,t+1}^{\boldsymbol{\lambda}}(\cdot)$ in (6).

bounds in (57) show that the demand-dependent terms in (58) are ordered in every demand scenario $\boldsymbol{d}$ and less than or equal to the Lagrangian bound.

Though we have focused on the known demands relaxation in the dynamic assortment example, we can use the same approach and derive similar results with other relaxations and in other problems. In the applicant screening example, the information relaxation where all applicant signals are known in advance is exactly analogous to the known demands relaxation and we obtain the same results. If we consider the known rates relaxation instead of the known demands realization in the dynamic assortment example, we arrive at an inner DP similar to (52), but the deterministic demand transitions are replaced with Poisson distributions with (randomly drawn) known demand rates. This inner DP is also linked and we can use an inner Lagrangian relaxation to derive results analogous to those of Theorem 2.

E.3. Numerical Examples

The (a) panels of Figures 2-5 show information relaxation bounds for the dynamic assortment and applicant screening examples using the known demands and known signals relaxations. These bounds were evaluated with S equal to 4, 8, 16, 32, and 64 in the same 1000 sample scenarios (i.e., same demand and signal sequences) that were used to evaluate the heuristics. In all cases, we use penalties based on an optimal solution $\boldsymbol{\lambda}^*$ for the outer Lagrangian dual problem (7) and impose the policy restrictions discussed at the end of §E.1. These figures also show 95% confidence intervals for the estimated bounds; these confidence intervals are quite narrow, particularly for larger values of S.

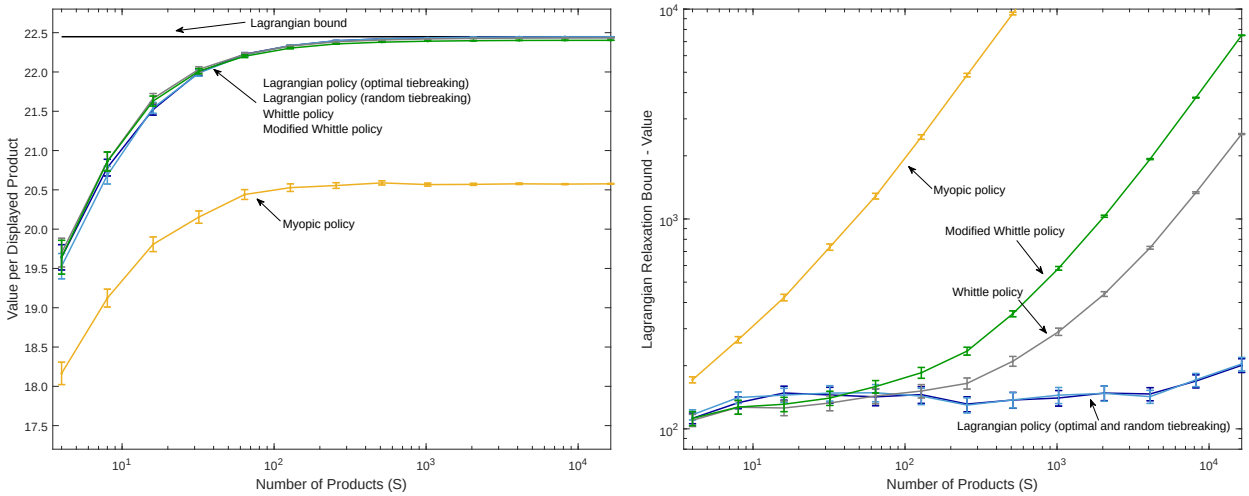
In the results, we see that the information relaxation bounds improve on the Lagrangian dual, particularly when S is small. The improvement is greatest in the dynamic assortment example with $T = 8$ and $S = 4$. In this case, the Lagrangian bound ensures that the Lagrangian index policy is within (approximately) \$0.88 per product displayed of the value given by an optimal solution. The information relaxation bound tells us that the Lagrangian index policy is in fact within \$0.16 per product displayed of an optimal solution. The improvements in bounds are less significant in the applicant screening example, particularly in the case with Bernoulli signals. Our intuition suggests that these information bounds are less effective when tiebreaking plays an important role: intuitively, the Lagrangian penalties "punish" the DM for using additional information in the selection decisions but do not punish for using this extra information to optimize tiebreaking. In all problems, the information relaxation bounds do not improve on the Lagrangian bound with large S: in these cases, the Lagrangian index policies are so close to the Lagrangian bound that there is very little room for the information relaxation bounds to improve upon the Lagrangian bound.

	Dynamic assortment example		Applicant screening example	
S	$T = 8$	$T = 20$	$n = 1$	$n = 5$
4	9.9	143	2.7	3.3
8	15.1	208	4.4	5.6
16	23.9	301	7.5	9.1
32	42.1	471	13.3	16.0
64	75.0	750	24.4	29.1

Table 3: Run times (seconds) for information relaxation bound calculations

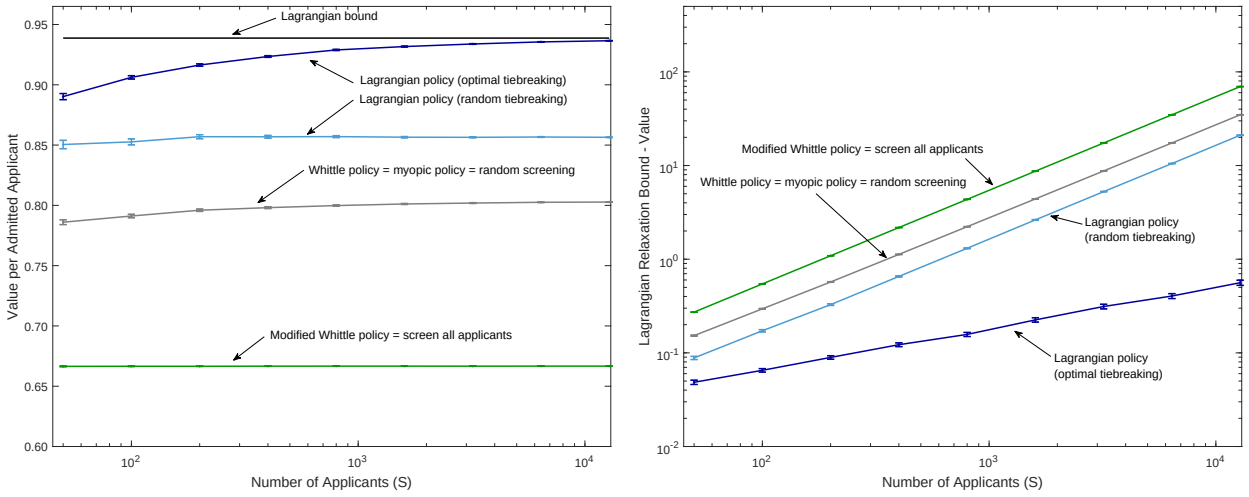
The run times are reported in Table 3. As discussed above, calculating these bounds requires solving the inner Lagrangian dual problems for each simulated demand (signal) sequence, for each product (applicant). This can be time consuming because the products (applicants) are not identical as each has its own demand (signal) sequence. We use the cutting-plane method in each case and start with $\boldsymbol{\mu} = \boldsymbol{\lambda}^*$, which yields the Lagrangian dual bound. If we cannot improve on this value, the cutting-plane algorithm typically stops after a few iterations. The run times grow roughly linearly with S, as one might expect, but not exactly because these no-improvement scenarios are more common with large S.

F. Numerical Results with Longer Horizons



(a) Relative performance (b) Performance gaps

Figure 7: Results for the dynamic assortment examples with horizon T=40



(a) Relative performance (b) Performance gaps

Figure 8: Results for the applicant screening examples with T=51 and Bernoulli signals ($n = 1$)

G. Additional Details on Weber and Weiss's Counterexample

Weber and Weiss (1990)'s example is useful for understanding dynamic selection problems with long or infinite horizons. Though Weber and Weiss considered a continuous-time, average-reward setting, their example can be adapted to discrete time with a long, but finite horizon. The example considers identical items, each having four states; the transition matrices and rewards are constant over time and are shown in Table 4. In each period, the DM must select exactly 83.5% of the items available. We assume that the system starts with 16%, 9%, 35% and 40% of the items in states one through four, respectively. We will consider a horizon T equal to 20,000 periods and focus on the dynamics of the Whittle and Lagrangian index policies in the deterministic mean field limit as the number of items S approaches infinity. The plots in Figure 9 show results for the first 3,000 of 20,000 periods; truncating these time series in this way makes the patterns easier to see.

	Probability Transition Matrices								Rewards	
	Selected				Not Selected					Not
State	1	2	3	4	1	2	3	4	Selected	Selected
1	0.9625	0.0075	0.0150	0.0150	0.9625	0.0075	0.0150	0.0150	0	10
2	0.0000375	0.9957625	0.0042	0.0000	0.0075	0.1525	0.8400	0.0000	10	10
3	0.0000	0.0000	0.9700	0.0300	0.0000	0.0000	0.9700	0.0300	10	1
4	0.0150	0.0000	0.0150	0.9700	0.0150	0.0000	0.0150	0.9700	10	0

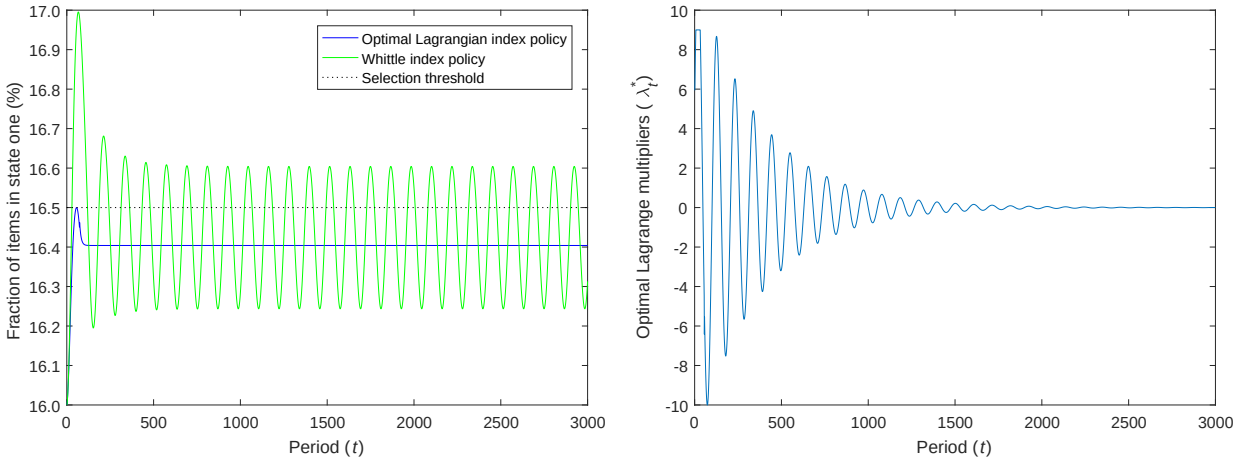
Table 4: Assumptions for Weber and Weiss (1990)'s example

First we consider the Whittle index policy. The Whittle indices may be calculated analytically and depend on an item's state (as usual) but not the period (a feature of this example). The Whittle indices are -10, 0, 9, and 10 for states one through four.[12] The ingenious feature of Weber and Weiss's example is that the fractions of items in each state cycles under the Whittle index policy. For example, Figure 9a shows the fraction of items in state one when following the Whittle index policy. The fraction of items in state one starts at 16%, rises to 17%, and ultimately settles into a cyclical pattern with fractions varying between 16.2% and 16.6%. The fractions in other states also vary cyclicly. In this example, the DM must select 83.5% of the items, so whenever the fraction in state one exceeds 16.5% (indicated with a dashed line in Figure 9a), the DM must select some items that are in state one. In periods where the Whittle index policy selects items in states two, three and four only, the policy generates a reward of 10. In periods where the policy selects some items in state one, the reward is less than 10, reflecting the zero reward when selecting items in state one.

Now consider the Lagrangian relaxation with a full set of T Lagrange multipliers. The optimal Lagrange multipliers $\boldsymbol{\lambda}^*$ (solving the dual problem (7)) are shown in Figure 9b and the state one fractions for the corresponding optimal Lagrangian index policy are shown in Figure 9a.[13] In Figure 9b we see that the optimal Lagrange multipliers λ_t^* cycle initially with dampening amplitude, approaching a steady state value of zero. The oscillations in the state fractions are less than those for the Whittle index policy and the fraction in state one remains at or below 16.5% in all periods, hitting 16.5% in period 56. How do the Lagrangian index and Whittle index policies differ? In the very early periods (1-6), the Lagrangian index policy prioritizes items in higher states, like the Whittle index policy. But in periods 8-31, the Lagrangian index policy prioritizes items in state two over state three, leaving some items in state three unselected. (Items in states two and three have the same index values in periods 7 and 32 and tiebreaking plays a role.) In most of the remaining periods, the Lagrangian index policy prioritizes items in the same way as the Whittle index. However, there is one later period (period 73) where the Lagrangian index policy is indifferent between selecting items in states one and two and the optimal Lagrangian index policy breaks ties so some items in state one are selected, earning zero reward. In this period, λ_t^* is -10 and the fraction of items in state one is strictly less than 16.5% so the DM is not forced to select items in state one in this period.

[12] For items in states one, three and four, the transition probabilities are identical in the active and passive states and the continuation values cancel in (15); it is easy to verify that (15) is satisfied with these index values. It is not hard to see with $\lambda_t = 0$ for all t, in every state the optimal value function is 10 times the number of periods remaining; (15) is thus satisfied in state two with $\lambda_t = 0$.

[13] This example took about 30 minutes to solve using an LP formulation of the Lagrangian dual (7).



(a) Fraction of items in state one

(b) Optimal Lagrange multipliers

Figure 9: Selected results for the Weber and Weiss example

These differences between the Whittle and Lagrangian index policies dampen the early oscillations seen in Figure 9a and guide the Lagrangian index policy to an equilibrium where the fraction of items in state one is approximately 16.4%; the fractions in other states also stabilize. In this equilibrium, the optimal Lagrange multipliers are zero and the Lagrangian and Whittle priority indices are equal: all items in states three and four are selected and approximately 99.0% of those in state two are selected. The rewards are 10 per period in this equilibrium. The optimal Lagrangian bound for the example, which is equal to the reward of the Lagrangian index policy, is slightly below 10 per period, reflecting the selection of some items in state one in period 73. The Whittle index policy performs worse because it regularly selects items in state one.

These numerical results depend on the initial fractions of items in each state, but the results are typical. For most initial conditions, the state fractions for the Whittle index policy settle into cycles as seen in Figure 9a where items in state one are routinely selected and the average reward is strictly less than 10. Similarly, for most initial conditions, the optimal Lagrange multipliers and state fractions for the Lagrangian index policy cycle initially, but approach an equilibrium distribution where the period reward is always 10. The exception to this typical behavior is that if we start the problem with initial conditions *exactly* equal to the equilibrium distribution, $\boldsymbol{\lambda} = 0$ is optimal for the Lagrangian dual problem and the Lagrangian and Whittle policies are equivalent and remain at this equilibrium distribution; however, this equilibrium is unstable and small deviations in initial conditions will lead the state distributions for Whittle index policies to oscillate.

H. Proofs for the Infinite-Horizon Extension

We begin our analysis of the infinite-horizon case by first considering how the result of Proposition 5 changes if we incorporate a discount factor $\delta \in [0,1)$ in the finite-horizon model of §2. We first briefly remark on how the results of the technical lemmas of §D are affected by discounting and then consider Proposition 5.

Lemma 1: Here the result is

$$L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) - V_1^{\pi}(\boldsymbol{x}) = \sum_{t=1}^{T} \delta^{t-1}\, \mathbb{E}[\, d_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t), \pi_t(\widetilde{\boldsymbol{x}}_t))\,]$$

where

$$\begin{aligned} d_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}, \boldsymbol{u}_t^{\pi}) = \lambda_t\big(N - n(\boldsymbol{u}_t^{\psi})\big) + r(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}) - r(\boldsymbol{x}_t, \boldsymbol{u}_t^{\pi}) \\ + \delta\,\mathbb{E}\Big[\, \bar{V}_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}))\Big] - \delta\,\mathbb{E}\big[\, \bar{V}_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}(\boldsymbol{x}_t, \boldsymbol{u}_t^{\pi}))\,\big]\ . \end{aligned}$$

The proof is analogous to the proof of Lemma 1.

Lemma 2: The result is exactly the same but discounting plays a role in the constants k_t. The inequality (46) is now

$$\big|\, V_t^{\pi}(\boldsymbol{x}') - V_t^{\pi}(\boldsymbol{x}'')\,\big| \le 2(\bar{r} - \underline{r})m \ + \ 2\delta k_{t+1}(\bar{r} - \underline{r})m$$

and we wind up with $k_t = 2(1 + \delta k_{t+1}) = \frac{2}{2\delta - 1}\big((2\delta)^{T-t+1} - 1\big)$.

Lemma 3: The result is the same but now $c_t = 1 + \delta k_{t+1} = 1/2k_t = \frac{1}{2\delta-1}\big((2\delta)^{T-t+1} - 1\big)$.

Lemma 4: The result and proof are unchanged.

Proposition 5: Using the analogs of Lemmas 1, 3, and 4 in the same way as before, we have:

$$\begin{aligned} L_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}) - V_1^{\widetilde{\pi}}(\boldsymbol{x}) &= \sum_{t=1}^{T} \delta^{t-1}\, \mathbb{E}[\, d_t(\widetilde{\boldsymbol{x}}_t, \widetilde{\psi}, \widetilde{\pi})\,] \\ &\le \sum_{t=1}^{T} \delta^{t-1} c_t (\bar{r} - \underline{r}) \sqrt{\bar{N}_t(1 - \bar{N}_t/S)}\ . \end{aligned}$$

Taking $\beta_t(T)$ (as claimed in equation (24)) to be

$$\beta_t(T) = \delta^{t-1} c_t = \frac{\delta^{t-1}}{2\delta - 1}\big((2\delta)^{T-t+1} - 1\big)\ ,$$

we obtain the result of Proposition 5. For future reference, we note that

$$\sum_{t=0}^{T} \beta_t(T) = \frac{1}{2\delta - 1}\left[2\delta^T(2^T - 1) - \frac{1 - \delta^T}{1 - \delta}\right]\ . \tag{59}$$

In preparation for the proof of Proposition 6, we note that the result of Proposition 5 can be extended to consider partial sums of cash flows, as claimed in (26). Specifically, consider two time horizons T and T' where $T' \le T$. Now suppose we define the optimal Lagrangian policy $\widetilde{\psi}$ and the corresponding optimal Lagrangian index policy $\widetilde{\pi}$ in the usual way for the longer time horizon T, with $\boldsymbol{\lambda}^*$ denoting the optimal Lagrange multipliers. Now consider the sum of the discounted expected cash flows for $\widetilde{\psi}$ and and $\widetilde{\pi}$ over the shorter horizon T':

$$\hat{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; T', T) \equiv \sum_{t=1}^{T'} \delta^{t-1}\, \mathbb{E}\big[\, \lambda_t\big(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))\big) + r_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t))\,\big]$$

$$\hat{V}_1^{\widetilde{\psi}}(\boldsymbol{x};T',T) \equiv \sum_{t=1}^{T'} \delta^{t-1}\mathbb{E}[\, r_t(\widetilde{\boldsymbol{x}}_t, \pi_t(\widetilde{\boldsymbol{x}}_t))\,]$$

Applying the argument of Proposition 5, but considering only the first T' periods, we obtain

$$\hat{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T',T) - \hat{V}_1^{\widetilde{\psi}}(\boldsymbol{x};T',T) \;\le\; \sum_{t=1}^{T'} \beta_t(T')(\bar{r}-\underline{r})\sqrt{\bar{N}_t(1-\bar{N}_t/S)} \;\le\; \sum_{t=1}^{T'} \beta_t(T')(\bar{r}-\underline{r})\sqrt{N}\;. \tag{60}$$

where $\beta_t(T')$ is given by (24). This then implies the result of (26).

Proof of Proposition 6. Consider two time horizons T and T' where $T' \le T$ and the optimal Lagrangian policy $\widetilde{\psi}$ and the corresponding optimal Lagrangian index policy $\widetilde{\pi}$ are based on the longer time horizon T. From (26) and (59), we have

$$\begin{aligned}\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T) - \bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T) &\le (\bar{r}-\underline{r})\left[\sum_{t=1}^{T'} \beta_t(T')\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right] \\ &= (\bar{r}-\underline{r})\left[\frac{1}{2\delta-1}\left(2\delta^{T'}(2^{T'}-1) - \frac{1-\delta^{T'}}{1-\delta}\right)\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right] \end{aligned} \tag{61}$$

Since we have assumed $\delta > 1/2$, $(2\delta-1) > 0$ and we can simplify the bracketed term in (61) by dropping terms:

$$\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T) - \bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T) \le (\bar{r}-\underline{r})\left[\frac{2(2\delta)^{T'}}{2\delta-1}\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right]. \tag{62}$$

Now consider the choice $T' = \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$. With this T', we have

$$\delta^{T'} = \delta^{\lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor} \le \frac{1}{\delta}\cdot \delta^{\log_2 \frac{S}{\sqrt{N}}} = \frac{1}{\delta}\cdot \left(2^{\log_2 \delta}\right)^{\log_2 \frac{S}{\sqrt{N}}} = \frac{1}{\delta}\cdot\left(\frac{\sqrt{N}}{S}\right)^{\log_2 \frac{1}{\delta}}, \tag{63}$$

where the inequality uses the fact that $\delta < 1$. Using the fact that $\delta > 1/2$ and hence $2\delta > 1$, we have

$$(2\delta)^{T'} = (2\delta)^{\lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor} \le (2\delta)^{\log_2 \frac{S}{\sqrt{N}}} = \frac{S}{\sqrt{N}}\cdot\left(\frac{\sqrt{N}}{S}\right)^{\log_2 \frac{1}{\delta}}. \tag{64}$$

Using (63) and (64), the bracketed term in (62) satisfies

$$\begin{aligned}\frac{2(2\delta)^{T'}}{2\delta-1}\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S \;&\le\; \frac{2}{2\delta-1}\sqrt{N}\cdot\left(\frac{S}{\sqrt{N}}\right)\cdot\left(\frac{\sqrt{N}}{S}\right)^{\log_2 \frac{1}{\delta}} + \frac{1}{\delta(1-\delta)}\cdot S\cdot\left(\frac{\sqrt{N}}{S}\right)^{\log_2 \frac{1}{\delta}} \\ &=\; \left(\frac{2}{2\delta-1} + \frac{1}{\delta(1-\delta)}\right)\cdot S\cdot\left(\frac{\sqrt{N}}{S}\right)^{\log_2 \frac{1}{\delta}},\end{aligned}$$

and the result of Proposition 6 then follows with $\gamma = \frac{2}{2\delta-1} + \frac{1}{\delta(1-\delta)}$. This choice of $T' = \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$ can be viewed as approximately minimizing the bound in (62). Specifically, this selection of T' differs from the minimizing T' by rounding down and dropping a constant term that complicates the resulting expressions. □

Proposition 6 when $\delta \in [0, 1/2]$: When $\delta < 1/2$, following a similar analysis, we obtain

$$\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; T) - \bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; T) \leq \left(\frac{1}{(1-\delta)(1-2\delta)} + \frac{2}{1-\delta} \right) \sqrt{N} \ .$$

Thus, in this case, we have $\sqrt{N}$ convergence as in the finite-horizon setting. When $\delta = 1/2$, we obtain

$$\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; T) - \bar{V}_1^{\tilde{\pi}}(\boldsymbol{x}; T) \leq \frac{2}{1-\delta} \sqrt{N} + 2 \log_2 \left(\frac{S}{\sqrt{N}} \right) \sqrt{N} \ .$$

This convergence is worse than $\sqrt{N}$ but not as slow as the case where $\delta > 1/2$.

Proof of Corollary 2. Using Proposition 6 and (28), we have

$$\begin{aligned}
\lim_{S \to \infty} \frac{L^{\boldsymbol{\lambda}^*}(\boldsymbol{x}; S) - V^{\tilde{\pi}}(\boldsymbol{x}; S)}{V^*(\boldsymbol{x}; S)} &\leq (\bar{r} - \underline{r}) \lim_{S \to \infty} \frac{\gamma S \left(\frac{\sqrt{N(S)}}{S} \right)^{\log_2 \frac{1}{\delta}}}{V^*(\boldsymbol{x}; S)} \\
&\leq (\bar{r} - \underline{r}) \lim_{S \to \infty} \frac{\gamma S \left(\frac{\sqrt{N(S)}}{S} \right)^{\log_2 \frac{1}{\delta}}}{\kappa S} \\
&\leq (\bar{r} - \underline{r}) \lim_{S \to \infty} \frac{\gamma}{\kappa} \left(\frac{\sqrt{N(S)}}{S} \right)^{\log_2 \frac{1}{\delta}} \\
&= 0 \ .
\end{aligned}$$

□

References

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

Bernstein, F., Li, Y. and Shang, K. (2015), 'A simple heuristic for joint inventory and pricing models with lead time and backorders', *Management Science* **62**(8), 2358–2373.

Bertsekas, D. P., Nedić, A. and Ozdaglar, A. E. (2003), *Convex Analysis and Optimization*, Athena Scientific.

Bertsimas, D. and Mišić, V. V. (2016), 'Decomposable markov decision processes: A fluid optimization approach', *Operations Research* **64**(6), 1537–1555.

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

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

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

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

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

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

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

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

Lai, G., Wang, M. X., Kekre, S., Scheller-Wolf, A. and Secomandi, N. (2011), 'Valuation of storage at a liquefied natural gas terminal', *Operations Research* **59**(3), 602–616.

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

Weber, R. R. and Weiss, G. (1990), 'On an index policy for restless bandits', *Journal of Applied Probability* **27**(3), 637–648.

Ye, F., Zhu, H. and Zhou, E. (2018), 'Weakly coupled dynamic program: Information and lagrangian relaxations', *IEEE Transactions on Automatic Control* **63**(3), 698–713.