Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Index Policies and Performance Bounds for Dynamic Selection Problems

David B. Brown, James E. Smith

To cite this article:
David B. Brown, James E. Smith (2020) Index Policies and Performance Bounds for Dynamic Selection Problems. Management Science 66(7):3029-3050. https://doi.org/10.1287/mnsc.2019.3342

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 66, No. 7, July 2020, pp. 3029–3050
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Index Policies and Performance Bounds for Dynamic Selection Problems

David B. Brown,[a] James E. Smith[b]

[a] Fuqua School of Business, Duke University, Durham, North Carolina 27708; [b] Tuck School of Business, Dartmouth College, Hanover, New Hampshire 03755

Contact: dbbrown@duke.edu, http://orcid.org/0000-0002-5458-9098 (DBB); jim.smith@dartmouth.edu, http://orcid.org/0000-0002-9429-7567 (JES)

Received: September 18, 2017
Revised: August 16, 2018; November 26, 2018
Accepted: March 5, 2019
Published Online in Articles in Advance: January 23, 2020

https://doi.org/10.1287/mnsc.2019.3342

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.

History: Accepted by Yinyu Ye, optimization.
Supplemental Material: The e-companion is available at https://doi.org/10.1287/mnsc.2019.3342.

Keywords: dynamic programming • restless bandits • Lagrangian relaxations • Gittins index • Whittle index

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 Section 3. 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, nonstationary 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 Section 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) grow 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]

We begin in Section 2 by defining the dynamic selection problem and introducing the dynamic assortment and applicant screening problems. In Section 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 the appendix. In Section 4, we define a number of heuristic policies including the Whittle index policy and the optimal Lagrangian index policy. In Section 5, we characterize the performance of the optimal Lagrangian index policy and present results on the asymptotic optimality of this policy. In Section 6, we simulate the heuristic policies of Section 4 in the context of the two example problems and evaluate their performance. In Section 7, we discuss the applicability of these methods in problems with long time horizons, considering the conjecture of Whittle (1988) 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 the electronic companion (EC) in Section EC4, 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 Section 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, that is, 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

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

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

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

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, that is, where the inequality constraint in Equation (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

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

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), 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 Caro and Gallien (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 Caro and Gallien (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, that is, $N_t = 0.25S$.

Caro and Gallien (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. Caro and Gallien (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 Caro and Gallien (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, for example, to a web page. For example, Rusmevichientong et al. (2010) consider a dynamic assortment model with capacity constraints (like constraint (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 five 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 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 Section 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, 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

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

Compared with 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

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

The first term in the maximization of Equation (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 (Equation (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 Section EC1.1. □

In Equation (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,

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

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*

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

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

(iii) Optimality conditions. The Lagrange multiplier vector $\boldsymbol{\lambda}^*$ *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 $\psi_{s,i} \in \Psi_s^*(\boldsymbol{\lambda}^*)$ $(n_s \le T+1)$ 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*

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

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

Proof. See Section EC1.1. □

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 $\bar{\psi} = (\bar{\psi}_1, \ldots, \bar{\psi}_S)$ where the item-specific mixed policies $\bar{\psi}_s$ 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 Section EC1.2. 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 $\psi_i \in \Psi^*(\boldsymbol{\lambda}^*)$ $(n \le T+1)$ and mixing weights $\{\gamma_i\}_{i=1}^{n}$ such that

$$S \sum_{i=1}^{n} \gamma_i\, p_t(\psi_i) = N_t \text{ for all } t \text{ such that } \lambda_t^* > 0 \text{ and}$$

$$S \sum_{i=1}^{n} \gamma_i\, p_t(\psi_i) \le N_t \text{ for all } t \text{ such that } \lambda_t^* = 0. \quad (12)$$

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 earlier, 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 2003, Adelman and Mersereau 2008, and Bertsimas and Mišić 2016). We discuss the LP formulation in more detail in Section EC1.3. In the appendix, 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 Section 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 formulation 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 two 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 Section 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 Section 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).

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 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).

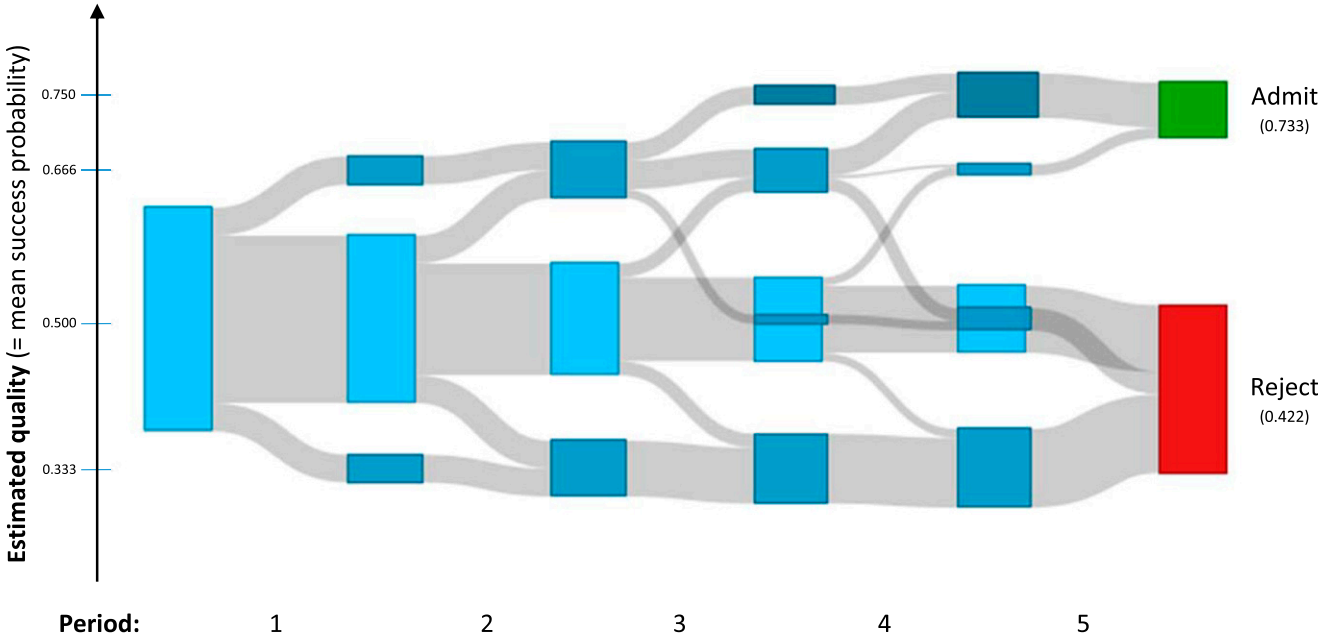
4. Heuristic Policies

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 Section 4.4) in Section 5 and evaluate the performance of these heuristics for the dynamic assortment and applicant screening problems in a simulation study in Section 6.

Table 1. Selection Probabilities for Policies Involved in an Optimal Mixture for the Applicant Screening Example

		Selection probabilities by period($p_t(\psi_i)$)					
Policy (ψ_i)		1	2	3	4	5	Mixing 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

Figure 1. Optimal Flows for the Lagrangian Relaxation of the Applicant Screening Example



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 Section 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,

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

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 Equation (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:

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

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 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 Equation (13),

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

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 Equation (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 Section EC2.1 for details), but this still essentially requires solving item-specific DPs once for each period and state. As mentioned in Section 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 Section 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 Equation (15) with a one-step lookahead value function and a normal distribution. In our numerical examples for the dynamic assortment problem in Section 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 Equation (15) is satisfied with $w = 0$. (See Proposition EC1 in Section EC2.2 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 finite horizons and/or 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, that is, $\boldsymbol{m} = (m_{t+1,s}(x_s), \ldots, m_{T,s}(x_s))$ for this same state x_s. We then take

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

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 Equation (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.

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 Section 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 Section 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 Section 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* $\tilde{\pi}$ as a Lagrangian index policy for $\boldsymbol{\lambda}^*$ that uses an optimal mixed policy $\tilde{\psi}$ 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 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 Equation (17).

We can generate a mixed policy $\tilde{\psi}$ 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 Equation (17) is randomly selected from $\{0, 1\}$ with state-dependent probabilities given in Section EC1.2.
- Proportional assignment: if some or all of the items are identical, we can sometimes construct a

nonrandom 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 noninteger remainders when necessary. For instance, in the applicant screening problem with the optimal policy mixture in Table 1, if $S = 1{,}000$, we assign (300, 25, 75, 250, 250, 200) applicants to the six 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 six policies, each 50% of the time. In Section 6, we use proportional assignments because it reduces the uncertainty in the model and seems to lead to slightly better performance (see Section 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 λ^* denote an optimal solution for the Lagrangian dual problem (7) with initial state $\boldsymbol{x}$. Let $\tilde{\psi}$ denote an optimal mixed policy for this Lagrangian and $\tilde{\pi}$ an optimal Lagrangian index policy that uses $\tilde{\psi}$ as a tiebreaker. Then*

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

where $\bar{N}_t$ is the expected number of items selected by $\tilde{\psi}$ 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 Section EC3.1. □

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

- The selection decisions made by the heuristic $\tilde{\pi}$ are based on priority indices that are aligned with the decisions made by $\tilde{\psi}$. Let n_t denote the number of items selected by the relaxed policy $\tilde{\psi}$ in period t in a particular state; this may be larger or smaller than N_t. From Equations (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 $\tilde{\psi}$. If $n_t < N_t$ items are selected by $\tilde{\psi}$, then these n_t items will be among the N_t items with the largest selection indices and will also be selected by $\tilde{\pi}$. If $n_t \ge N_t$ items are selected by $\tilde{\psi}$, then $\tilde{\pi}$ will select a subset of size N_t of those selected by $\tilde{\psi}$. In both cases, the number of items with different decisions is bounded by $|n_t - N_t|$. 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 $\tilde{\psi}$. With an optimal policy $\tilde{\psi}$ 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 $|\tilde{n}_t - N_t|$ 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 Equation (18) reflect the maximum possible number of changes in item states caused by the selection decision of $\tilde{\pi}$ deviating from $\tilde{\psi}$ 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 Equation (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 $\tilde{\pi}$ deviating from the state under $\tilde{\psi}$ 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 $L_1^{\lambda^*}(\boldsymbol{x}) - V_1^{\tilde{\pi}}(\boldsymbol{x})$ 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 Section EC3.2. 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,*

$$V_1^*(\boldsymbol{x}) - \Delta^{\tilde{\psi}}(\boldsymbol{x}) \le L_1^{\lambda^*}(\boldsymbol{x}) - \Delta^{\tilde{\psi}}(\boldsymbol{x}) \le V_1^{\tilde{\pi}}(\boldsymbol{x}) \le V_1^*(\boldsymbol{x}) \le L_1^{\lambda^*}(\boldsymbol{x}).$$

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 $\tilde{\pi}$ is feasible for the DP (2), that is, it satisfies the linking constraint (1). □

Since $\Delta^{\tilde{\psi}}(\boldsymbol{x})$ 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 Section 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 $V_t^*(\boldsymbol{x};S)$, $L_t^{\lambda^*}(\boldsymbol{x};S)$, and $V_t^{\tilde{\pi}}(\boldsymbol{x};S)$ denote the corresponding optimal value functions, values for the optimal Lagrangian, and value for the corresponding optimal Lagrangian index policy $\tilde{\pi}$. If the $V_1^*(\boldsymbol{x};S)$ are positive and satisfy*

$$\lim_{S\to\infty} \frac{V_1^*(\boldsymbol{x};S)}{\sqrt{N(S)}} = \infty, \tag{20}$$

then

$$\lim_{S\to\infty} \frac{L_1^{\lambda^*}(\boldsymbol{x};S) - V_1^{\tilde{\pi}}(\boldsymbol{x};S)}{V_1^*(\boldsymbol{x};S)} = 0. \tag{21}$$

Since $V_1^{\tilde{\pi}}(\boldsymbol{x}) \le V_1^*(\boldsymbol{x}) \le L_1^{\lambda^*}(\boldsymbol{x})$, Equation (21) implies

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

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

Proof. See Section EC3.1. □

This corollary implies that, when the growth condition (20) is satisfied, the gaps between $V_1^*(\boldsymbol{x};S)$, $L_1^{\lambda^*}(\boldsymbol{x};S)$, and $V_1^{\tilde{\pi}}(\boldsymbol{x};S)$, 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 (20) 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 growth condition (20) will be satisfied. We could normalize the ratios in Corollary 1 by the Lagrangian $L_1^{\lambda^*}(\boldsymbol{x};S)$ rather than $V_1^*(\boldsymbol{x};S)$ (because $V_1^*(\boldsymbol{x};S) \le L_1^{\lambda^*}(\boldsymbol{x};S)$) 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 Section 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 Section 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 λ^* 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 Section 4.4. As discussed in Sections 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. Runtimes

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

Table 2. Runtimes, Problem Sizes, and Related Statistics for Index Calculations

	Dynamic assortment			Applicant screening		
	$T = 8$	$T = 20$	$T = 40$	$n = 1$ $T = 5$	$n = 5$ $T = 5$	$n = 1$ $T = 51$
Runtimes (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 dynamic program	12,636	199,710	1,599,820	35	115	23,426
Cutting plane iterations	70	530	826	14	16	540

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 runtimes are dominated by the time required to solve the item-specific DPs (6). The time required to solve these DPs is primarily determined by the number of states that must be considered (also shown in Table 2). In problems with a fixed state space (such as the Weber and Weiss (1990) example discussed in Section 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 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 two 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.

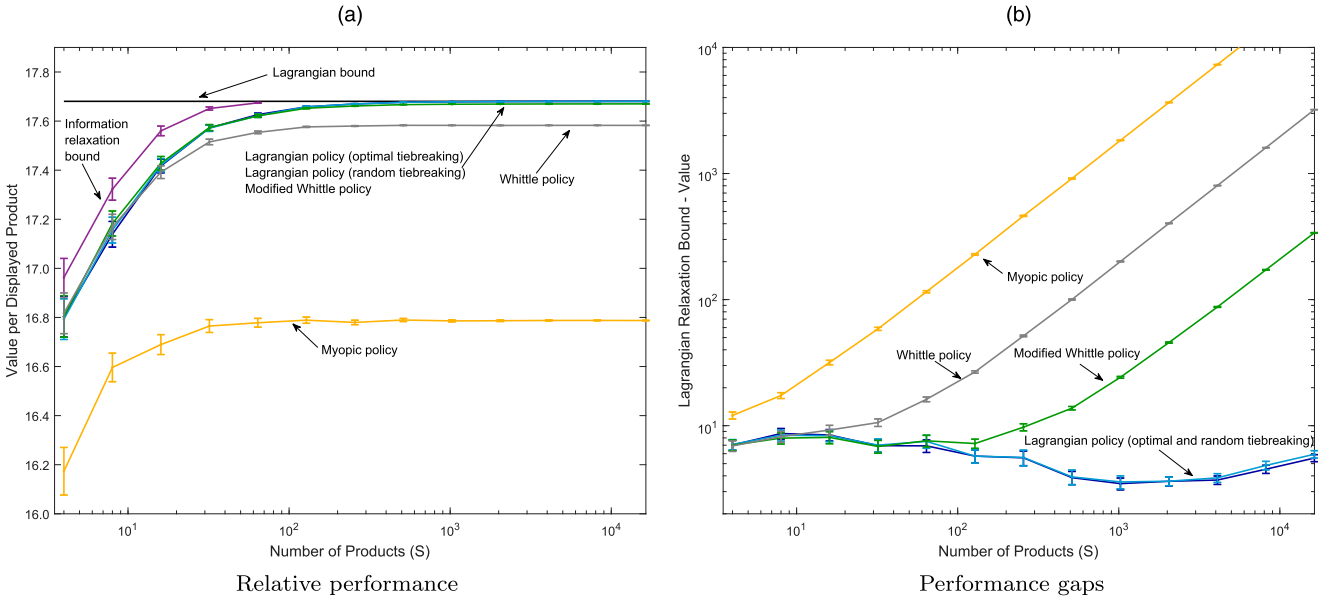
6.2. Simulation Results

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 1,000 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 Section EC4. The error bars in the figures represent 95% confidence intervals for these estimated values. The Lagrangian bounds $L_1^{\lambda^*}(\boldsymbol{x})$ 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 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.

6.2.1. 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

Figure 2. Results for the Dynamic Assortment Examples with Horizon $T = 8$



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!

Figure 3, (a) and (b), are like Figure 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

Figure 3. Results for the Dynamic Assortment Examples with Horizon $T = 20$

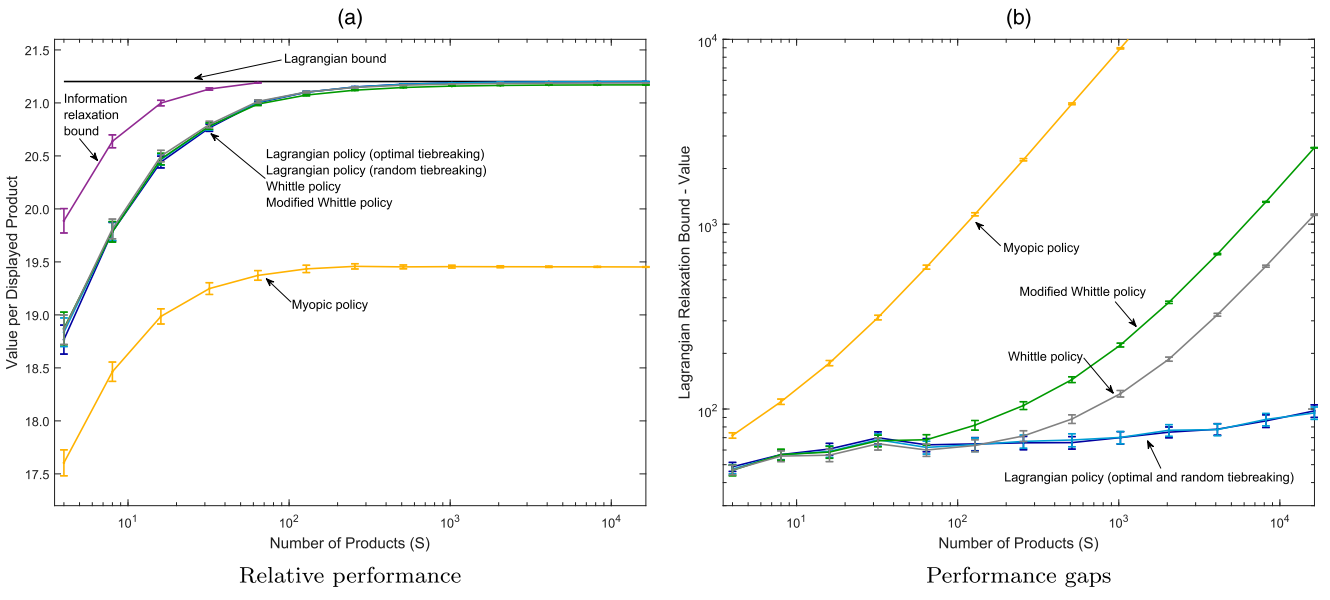
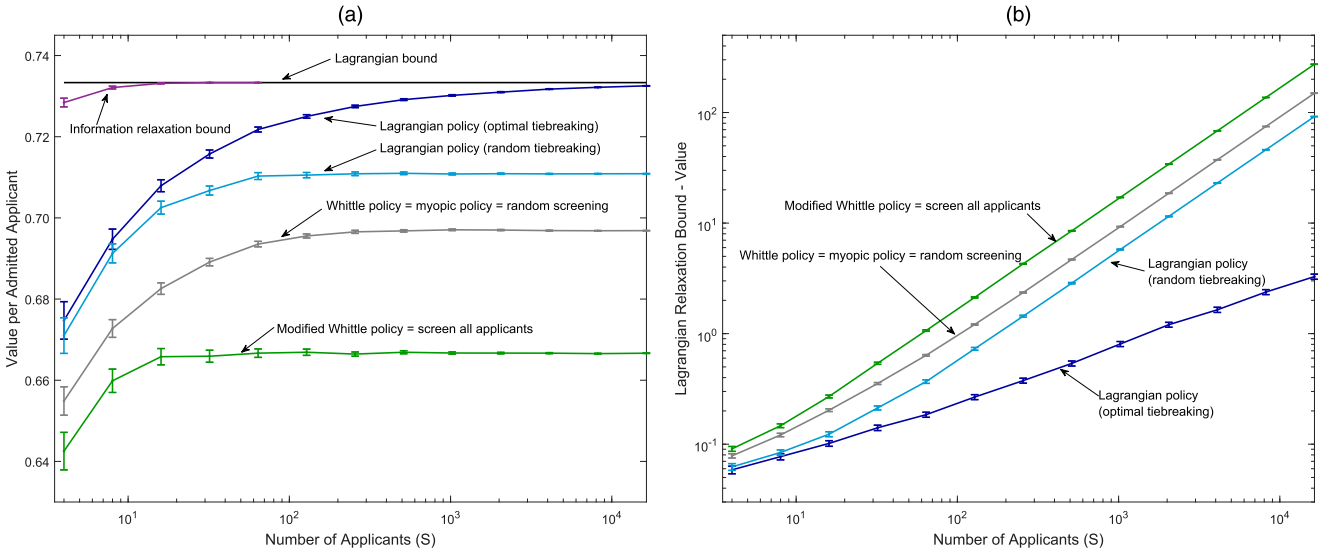


Figure 4. Results for the Applicant Screening Examples with $T = 5$ and Bernoulli Signals ($n = 1$)



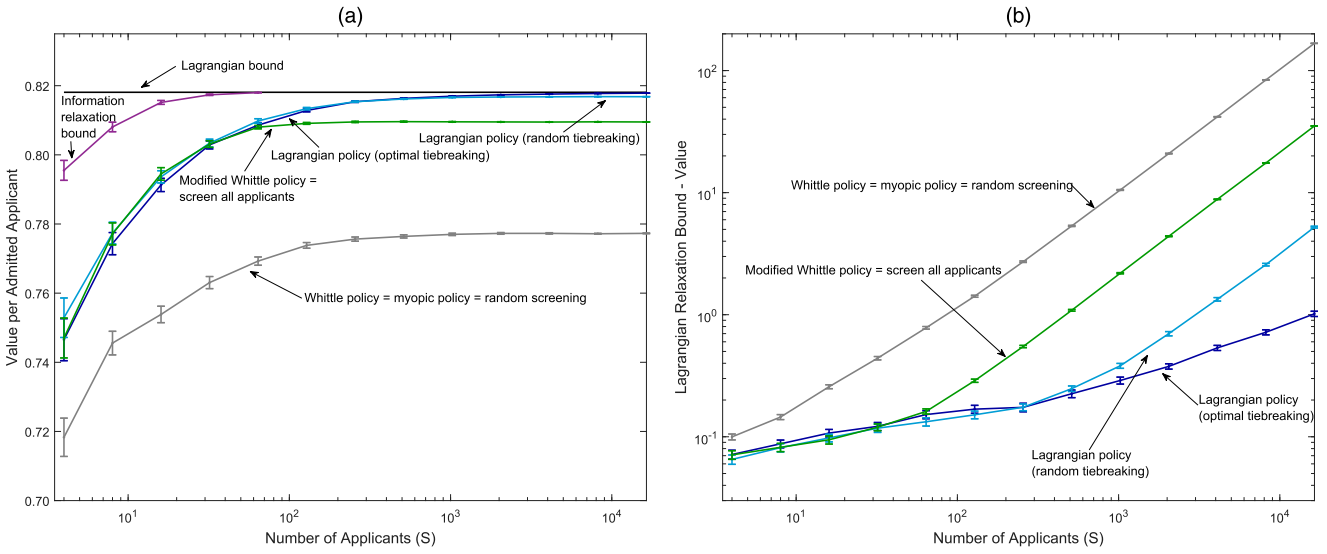
Relative performance Performance gaps

modified Whittle index policy for large S, but again both exhibit linear growth in the performance gap. 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 Section EC5.

Finally, although we do not show these results in Figures 2 and 3, we also simulated the one-period lookahead/normal approximation of the Whittle index developed by Caro and Gallien (2007) (see Section 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.

6.2.2. Applicant Screening Examples. The performance of the heuristics is more varied in the applicant

Figure 5. Results for the Applicant Screening Examples with $T = 5$ and Binomial Signals ($n = 5$)



Relative performance Performance gaps

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 Section 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 Section 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 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 Section 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 Section 5.

Figure 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 Section EC5. 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 Section EC4, we consider the use of information relaxations (e.g., Brown et al. 2010) with dynamic selection 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 described in the appendix play important roles in the analysis and computation.

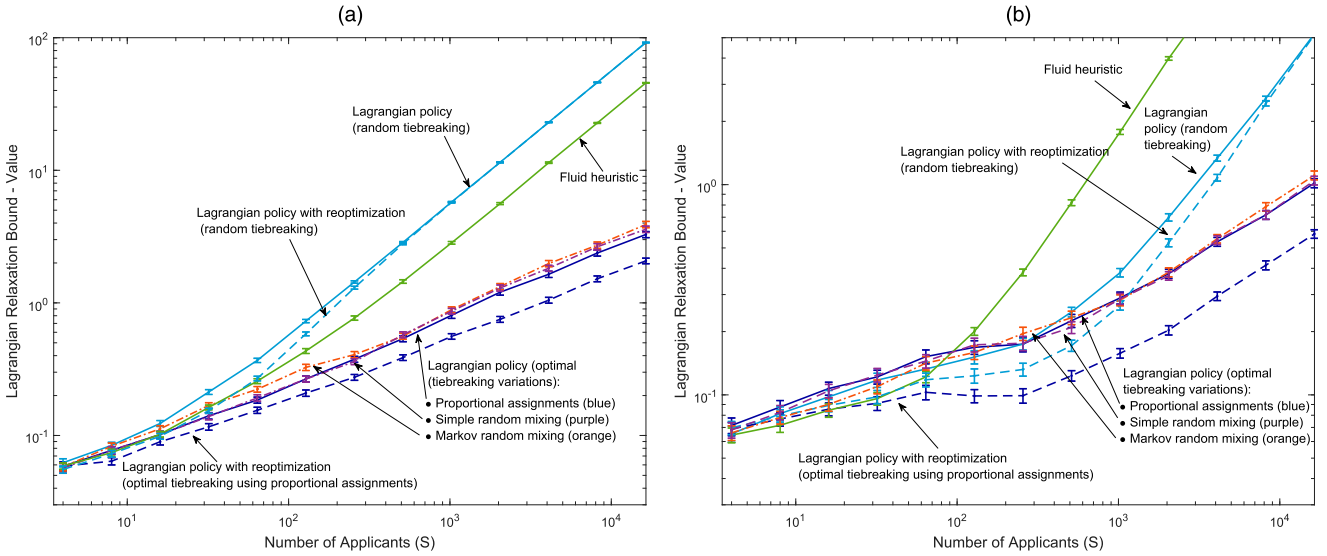
In our numerical examples, these information relaxation bounds are shown in the (a) panels of Figures 2–5. In these results, we see that the information relaxation bounds improve on the Lagrangian dual, particularly when S is small. The improvement is greatest in the dynamic assortment example with $T = 8$ 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 Section EC4.

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 reoptimization appears to outperform the optimal Lagrangian policy without reoptimization, but they both appear to exhibit $\sqrt{S}$ growth in the performance gap. These

Figure 6. Results for Applicant Screening Examples with Variations on the Heuristics



With Bernoulli signals (n=1) With binomial signals (n=5)

applicant screening examples are small enough to allow reoptimization (the runtimes range from nine 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 Section 4.4 for generating a mixed policy for tiebreaking with the optimal Lagrangian index policy, without reoptimization. As discussed in Section 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 (see Section EC1.3) 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 formulation; see Section EC1.3 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 the conjecture in Whittle (1988) on the asymptotic optimality of the Whittle index policy and the counterexample in Weber and Weiss (1990). We use this example to motivate the extension of the results of Section 5 to the infinite-horizon case with discounting, which we consider in Section 7.2.

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

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

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

where $\tilde{\pi}$ 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 $\tilde{\psi}$ 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 $\tilde{\psi}$ 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 $\tilde{\psi}$, whereas the optimal Lagrangian index policy is, by construction, aligned with $\tilde{\psi}$. Weber and Weiss (1990) showed that optimal policies asymptotically converge to the Lagrangian bound in the average reward setting [in the relative sense of Equation (22)] but provided an example that showed that the Whittle index policy need not be asymptotically optimal.

In Section EC6, 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 Section 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 the example from Weber and Weiss (1990), 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 (Equation (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

$$\beta_t(T) = \frac{\delta^{t-1}}{2\delta - 1}\left((2\delta)^{T-t+1} - 1\right). \tag{23}$$

(See Section EC7 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 Section EC7.)

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 $L_1^{\lambda^*}(\boldsymbol{x};T)$ denote the optimal Lagrangian with finite horizon T, defined as in Equation (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 $L_1^{\lambda^*}(\boldsymbol{x};\infty)$ 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

$$\underbrace{V_1^{\tilde{\pi}}(\boldsymbol{x};T) + \frac{\delta^T}{1-\delta}\underline{r}S}_{\equiv\, \bar{V}_1^{\tilde{\pi}}(\boldsymbol{x};T)} \le V_1^*(\boldsymbol{x}) \le L_1^{\lambda^*}(\boldsymbol{x};\infty) \le \underbrace{L_1^{\lambda^*}(\boldsymbol{x};T) + \frac{\delta^T}{1-\delta}\bar{r}S}_{\equiv\, \bar{L}_1^{\lambda^*}(\boldsymbol{x};T)}. \tag{24}$$

Here the term on the left, $\bar{V}_1^{\bar{\pi}}(x;T)$, 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 in Equation (24). The second inequality follows from Lagrangian duality, as in Proposition 1. The final term $\bar{L}_1^{\lambda^*}(x;T)$ 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 Section 5, we will show that, in relative terms, $\bar{V}_1^{\bar{\pi}}(x;T)$ approaches $\bar{L}_1^{\lambda^*}(x;T)$ as we increase S and T. Since the optimal value function $V_1^*(x)$ is bracketed by these terms in Equation (24), this will imply the desired asymptotic optimality result. In our analysis, we will consider sums of cash flows in the difference of $\bar{L}_1^{\lambda^*}(x;T) - \bar{V}_1^{\bar{\pi}}(x;T)$ over a horizon $T' \le T$ and obtain a bound of the form

$$\bar{L}_1^{\lambda^*}(x;T) - \bar{V}_1^{\bar{\pi}}(x;T) \le (\bar{r}-\underline{r})\left(\sum_{t=1}^{T'}\beta_t(T')\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right). \tag{25}$$

This follows from the argument underlying Proposition 5. We then choose T' to provide a good bound in Equation (25). 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 we 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 Equation (25), we have the following infinite-horizon analog of Proposition 5.

Proposition 6. *Let $\bar{L}_1^{\lambda^*}(x;T)$ and $\bar{V}_1^{\bar{\pi}}(x;T)$ be defined as in Equation (24) and let $\lfloor z \rfloor$ denote the largest integer less than or equal to z. For any $T \ge \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$,*

$$\bar{L}_1^{\lambda^*}(x;T) - \bar{V}_1^{\bar{\pi}}(x;T) \;\le\; \gamma(\bar{r}-\underline{r})S\left(\frac{\sqrt{N}}{S}\right)^{\log_2\frac{1}{\delta}}, \tag{26}$$

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 Equation (26) does not improve if we increase T beyond $T \ge \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$. 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 $T(S) \ge \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$. Let $\bar{L}_1^{\lambda^*}(x;S) = \bar{L}_1^{\lambda^*}(x;T(S))$ and $\bar{V}_1^{\bar{\pi}}(x;S) = \bar{V}_1^{\bar{\pi}}(x;T(S))$, as defined in Equation (24). If the optimal value functions $V_1^*(x;S)$ are positive and satisfy*

$$V_1^*(x;S) \ge \kappa S, \tag{27}$$

for some constant $\kappa > 0$, then

$$\lim_{S\to\infty} \frac{\bar{L}_1^{\lambda^*}(x;S) - \bar{V}_1^{\bar{\pi}}(x;S)}{V_1^*(x;S)} = 0. \tag{28}$$

Since the optimal value function $V_1^*(x;S)$ lies between $\bar{V}_1^{\bar{\pi}}(x;S)$ and $\bar{L}_1^{\lambda^*}(x;S)$, 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 (27) is stronger than that in Corollary 1, as we require $V_1^*(x;S)$ 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

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

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 λ^*) 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 Section 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 Section 5.

Acknowledgments

The authors thank conference and seminar participants at the INFORMS Annual Meeting (2016, 2017, and 2018); University of Maryland; Dartmouth College; University of Southern California; Northwestern University; the Advances in Decision Analysis Conference 2017 (Austin); University of California, Los Angeles; Princeton University; The University of Texas at Austin, Duke University; University of Illinois-Chicago; and the University of Virginia for helpful comments and questions. The authors also thank the anonymous associate editor and three anonymous referees for helpful comments and suggestions.

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

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

$$N^\top\lambda + \sum_{s=1}^{S} V_s(\lambda_{i_s}) + \nabla_{s,i_s}^\top(\lambda - \lambda_{i_s}), \tag{A.1}$$

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

$$\ell_k(\lambda) \equiv \max_{i_1,\dots,i_S \in \{1,\dots,k\}} \left\{ N^\top\lambda + \sum_{s=1}^{S} \left(V_s(\lambda_{i_s}) + \nabla_{s,i_s}^\top(\lambda - \lambda_{i_s}) \right) \right\}. \tag{A.2}$$

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

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

$$\lambda_{k+1} = \arg\min_{\lambda \geq 0} \ \ell_k(\lambda). \tag{A.3}$$

We then calculate the item-specific value functions $V_s(\lambda_{k+1})$ and subgradients $\nabla_{s,k+1} \in \partial V_s(\lambda_{k+1})$ for this new point, as well as the Lagrangian $L(\lambda_{k+1}) = N^\top\lambda_{k+1} + \sum_{s=1}^S V_s(\lambda_{k+1})$. The process continues until $\ell_k(\lambda_{k+1}) = L(\lambda_{k+1})$. In this terminal case, since λ_{k+1} minimizes $\ell_k(\lambda)$ and $\ell_k(\lambda) \leq L(\lambda)$ for all $\lambda \geq 0$, we know that λ_{k+1} is an optimal solution for the Lagrangian dual problem (7). If $\ell_k(\lambda_{k+1}) < L(\lambda_{k+1})$, we add the newly calculated values $V_s(\lambda_{k+1})$ and gradients $\nabla_{s,k+1}$ to form a new cutting-plane model $\ell_{k+1}(\lambda)$. Note that in this case, we will 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 λ_{k+1}, whereas $\min_{\lambda \geq 0} \ell_k(\lambda) = \ell_k(\lambda_{k+1}) < L(\lambda_{k+1})$. Since $L(\lambda)$ 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 (A.3) can be formulated as a linear program as

$$\begin{aligned} \min_{\lambda, v_s} \ & N^\top\lambda + \sum_{s=1}^{S} v_s \\ \text{s.t. } & v_s \geq V_s(\lambda_i) + \nabla_{s,i}^\top(\lambda - \lambda_i) \quad \forall i \in \{1,\dots,k\}, \\ & \qquad\qquad\qquad\qquad\quad \forall s \in \{1,\dots,S\}, \\ & \lambda \geq 0. \end{aligned} \tag{A.4}$$

As we proceed iteratively in the cutting-plane method, we add additional constraints for the new values $V_s(\lambda_{k+1})$ and

subgradients $\nabla_{s,k+1}$ at the new trial value $\boldsymbol{\lambda}_{k+1}$. We solve linear program (A.4) 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 linear program (A.4) as

$$\begin{aligned} \max_{\gamma_{s,i}} \ & \sum_{s=1}^{S}\sum_{i=1}^{k}(V_s(\boldsymbol{\lambda}_i) - \nabla_{s,i}^{\top}\boldsymbol{\lambda}_i)\gamma_{s,i} \\ \text{s.t.} \ & -\sum_{s=1}^{S}\sum_{i=1}^{k}\gamma_{s,i}\nabla_{s,i} \le N \\ & \sum_{i=1}^{k}\gamma_{s,i} = 1 \qquad \forall s \in \{1,\dots,S\}, \\ & \gamma_{s,i} \ge 0 \qquad \forall i \in \{1,\dots,k\}, \forall s \in \{1,\dots,S\}. \end{aligned} \tag{A.5}$$

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 Equation (A.5) 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 linear programs (A.4) and (A.5) similarly simplify. If we let S' denote the number of distinct items, the simplified version of the linear program (A.4) will have $S' + T$ decision variables and $k \times S'$ constraints. The basic solutions for the simplified version of the dual linear program (A.5) 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 linear programs (A.4) are typically easy to solve, even if the item-specific DPs have large state spaces.

Endnotes

[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 Section 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 Section 7.2.

[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 $\sum_{\tau=0}^{t-1}((\tau-1)\bar{d}+1))$ possible states, representing the values of (m, α) that could be obtained under some policy.

[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^*.

[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, for example, Le Boudec et al. (2007).

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

[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 $(m_{t+1,s}(x_s), \dots, m_{T,s}(x_s))$ for future periods for this same state x_s. This assumption is true in all of the examples that we consider.

[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 1 and Corollary 1 then follow with no additional changes.

[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 linear program (A.4) in the cutting-plane method when calculating Lagrangian indices.

[9] The standard cutting-plane method takes the maximum in Equation (A.2) 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 Equation (A.2) 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 (Equation (A.2)) and thereby accelerates convergence of the algorithm. This is particularly important when reoptimizing (as in Section 6.4) or calculating information relaxation bounds (Section EC4) where the items will necessarily be distinct.

References

Adelman D, Mersereau AJ (2008) Relaxations of weakly coupled stochastic dynamic programs. *Oper. Res.* 56(3):712–727.

Bernstein F, Kök AG, Xie L (2015) Dynamic assortment customization with limited inventories. *Manufacturing Service Oper. Management* 17(4):538–553.

Bertsekas DP, Nedić A, Ozdaglar AE (2003) *Convex Analysis and Optimization* (Athena Scientific, Belmont, MA).

Bertsimas D, Mersereau AJ (2007) A learning approach for interactive marketing to a customer segment. *Oper. Res.* 55(6):1120–1135.

Bertsimas D, Mišić VV (2016) Decomposable Markov decision processes: A fluid optimization approach. *Oper. Res.* 64(6):1537–1555.

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

Brown DB, Smith JE (2014) Information relaxations, duality, and convex stochastic dynamic programs. *Oper. Res.* 62(6):1394–1415.

Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 584(1):785–801.

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

Gittins J, Glazebrook K, Weber R (2011) *Multi-Armed Bandit Allocation Indices* (John Wiley & Sons, Chichester, UK).

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

Hodge DJ, Glazebrook KD (2015) On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Adv. Appl. Probab.* 47(3):652–667.

Hu W, Frazier P (2017), An asymptotically optimal index policy for finite-horizon restless bandits. Working paper, Cornell University, Ithaca, NY.

Kök AG, Fisher ML, Vaidyanathan R (2008) Assortment planning: Review of literature and industry practice. *Retail Supply Chain Management*, International Series in Operations Research & Management Science, vol. 223 (Springer, Boston), 99–153.

Le Boudec J-Y, McDonald D, Mundinger J (2007) A generic mean field convergence result for systems of interacting objects. *Proc. 4th Internat. Conf. Quantitative Evaluation of Systems* (Institute of Electrical and Electronics Engineers, Washington, DC), 3–18.

Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, Hoboken, NJ).

Rusmevichientong P, Shen Z-JM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

Topaloglu H (2009) Using Lagrangian relaxation to compute capacity-dependent bid prices in network revenue management. *Oper. Res.* 57(3):637–649.

Weber RR, Weiss G (1990) On an index policy for restless bandits. *J. Appl. Probab.* 27(3):637–648.

Whittle P (1988) Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.* 25(A):287–298.