Index Policies and Performance Bounds for Dynamic Selection Problems

David B. Brown and James E. Smith

Fuqua School of Business
Duke University
Durham, NC 27708
dbbrown@duke.edu, jes9@duke.edu

November 21, 2017

Abstract

We consider dynamic selection problems, where a decision maker repeatedly selects a set of items from a larger collection of available items. A classic example is the dynamic assortment problem, where a retailer chooses items to offer for sale subject to a display space constraint. The retailer may adjust the chosen 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 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 optimal Lagrangian index policy and bound and show that, under mild conditions, these policies and bounds are both asymptotically optimal for problems with many items; tiebreaking plays an essential role in the analysis of these index policies and has a surprising impact on performance. We also develop an efficient cutting-plane method for solving the Lagrangian dual problem and develop an information relaxation bound that improves on the standard Lagrangian bound. We demonstrate these policies and bounds in two large scale examples: a dynamic assortment problem with demand learning and an applicant screening problem.

Keywords: Models: dynamic programming, Restless bandits, Lagrangian relaxations, dynamic assortment problem

We would like to thank conference and seminar participants at INFORMS 2016 Nashville, University of Maryland, Dartmouth, University of Southern California, Northwestern University, and the Advances in Decision Analysis Conference 2017 (Austin) for helpful comments and questions.
1. Introduction

In the dynamic assortment problem, a decision maker (DM) – prototypically a retailer – chooses products to offer for sale, selecting from many possible products, but limited by display space. When product demands are uncertain, the retailer may want to update the assortment over the course of the selling season in response to demands observed in previous periods. Similar dynamic selection 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?). In these problems, the DM must balance exploitation (selecting those that appear to be the best) and exploration (learning about the performance of the available items) with limited time and limited resources.

These dynamic selection problems can be naturally formulated as finite-horizon stochastic dynamic programs (DPs) but are difficult to solve to optimality. Even when the state dynamics 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 these dynamic selection problems and provide upper bounds on the performance of an optimal policy. These performance bounds are useful for assessing the quality of a heuristic: if the average reward for a heuristic policy is close to this upper bound, we can be sure that it is a good heuristic.

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, provided we break ties appropriately, is asymptotically optimal: as we increase the number of items available and the number of items that can be selected, under mild conditions, the relative performance of the heuristic approaches the Lagrangian upper bound. We also develop information relaxation bounds that use the Lagrangian as a penalty and show that these bounds are at least as good as those provided by the Lagrangian itself.

We illustrate these results with two examples. The first is based on the dynamic assortment model with demand learning from ?, CG hereafter. The second is an applicant screening example 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 accepted or hired.

1.1. Literature Review

Our paper builds on and contributes to several related streams of literature. First, the dynamic selection problem can be viewed as a special case of a weakly coupled DP. For example, ?, ? and ? 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. All consider Lagrangian relaxations of weakly coupled DPs, similar to the Lagrangian relaxation in §3 below. Lagrangian relaxations of DPs have been used in a number of applications including revenue management (e.g., ?) and marketing (e.g., ? as well as ?).

The dynamic selection problem can also be viewed as a finite-horizon, non-stationary version of the restless bandit problem introduced in ?. 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 evolve when not selected. ? introduced an index policy where items are prioritized for selection according to an index that is essentially equivalent to the Gittins index. ? motivates this policy through a Lagrangian analysis, viewing the index as a breakeven Lagrange multiplier (see §4.2) and conjectured that in the infinite-horizon average reward setting these policies are asymptotically optimal for problems with many items. ? showed that this conjecture is true under certain conditions but need not be true in general. ? studied Whittle indices in the dynamic assortment problem. For a comprehensive discussion of the restless bandit problem, see ?. In §§4-5, we will evaluate the Whittle index policy and a modification of the Whittle index policy that may be more suitable for finite-horizon models with time-varying rewards or constraints.

Finally, we apply information relaxation bounds in the context of the dynamic selection problem, following ?, BSS hereafter. ? generalized earlier applications of information relaxations for valuing American options (see, e.g., ? and ?). Our application to dynamic selection problems can be viewed as a new application in a growing list of applications of information relaxation methods. In addition to the many applications to valuing options and other derivative securities, recent applications of information relaxations include managing natural gas storage (? and ?), dynamic portfolio optimization with transaction costs or taxes (? and ?), and inventory and pricing models with lead time and backorders (?). Our application of information relaxations to the dynamic selection problem combines information relaxations and Lagrangian relaxations. Information relaxations and Lagrangian relaxations were similarly combined in a network revenue management problem in ?, in a multiclass queueing problem in ?, and in ?.

1.2. Contributions and Outline

Our main contributions are:

(i) a detailed analysis of the Lagrangian relaxation of the DP and the development of an efficient cutting-plane method for solving the Lagrangian dual problem;

(ii) the development of an optimal Lagrangian index policy that performs well in examples and is proven to be asymptotically optimal; and

(iii) the development of an information relaxation bound based on the Lagrangian that is provably at least as good as the standard Lagrangian bound;

(iv) the application of these methods in two large scales examples, the classic dynamic assortment problem and a new applicant screening problem.

Contribution (i) is broadly applicable for weakly coupled DPs and improves on linear programming (LP)
formulations or subgradient methods that are commonly used in the literature. Contribution (iii) is helpful in assessing the performance of heuristics in applications. The examples mentioned in contribution (iv) demonstrate these methods and are of independent interest.

The most significant contribution is probably contribution (ii). In the related literature, the Whittle index policy (\? is often suggested as an ideal policy; for example, \? call the Whittle index an “exact desirability index.” Our numerical examples and theoretical results show that these Whittle index policies may not perform well and typically perform worse than the optimal Lagrangian policy, particularly for problems with many items. Similarly, the “fluid heuristic” developed in \? need not perform as well as the optimal Lagrangian policy. 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, e.g., $N$ 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 $N$ for the Whittle index policy and the fluid heuristic, whereas the performance gap grows with $\sqrt{N}$ for the optimal Lagrangian index. Tiebreaking plays a surprising and important role in this analysis and in the numerical examples; a Lagrangian index policy that breaks ties randomly may also exhibit linear growth in the performance gap.

We begin in §2 by defining the dynamic selection problem and introducing the dynamic assortment and applicant screening examples. In §3, we describe the Lagrangian relaxation, discuss its theoretical properties and introduce a cutting-plane method for efficiently solving the Lagrangian dual optimization problem; we also contrast this cutting-plane method with the LP formulation. In §4, we define a number of heuristic policies that we evaluate in the two examples in §5. In §6, we characterize the performance of the Lagrangian index policy and present results on the asymptotic optimality of these policies. Finally, in §7, we describe the Lagrangian-based information relaxation performance bounds and show how they improve on the standard Lagrangian bounds.

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 that item $s$ is (is not) selected. In each period, item $s$ generates a reward $r_t(x_s, u_s)$ 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 $\tilde{x}_t(x_s, u_s)$ with transitions also depending on the current state, the selection decision, and period. We let $x = (x_1, \ldots, x_S)$ denote a vector of item states, $u = (u_1, \ldots, u_S)$ a vector of selection decisions, and
\( \tilde{\chi}_t(x, u) = (\tilde{\chi}_{t,1}(x_1, u_1), \ldots, \tilde{\chi}_{t,S}(x_S, u_S)) \) 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., ?) imply there is an optimal policy that is deterministic and Markovian i.e., of the form \( \pi = (\pi_1, \ldots, \pi_T) \), where \( \pi_t(x) \) specifies a vector of selection decisions \( u \) given state vector \( x \), where \( u \) must be in

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

Taking the terminal value \( V_{T+1}(x) = 0 \), we can write the optimal value function for earlier periods as

\[
V_t^*(x) = \max_{u \in \mathcal{U}_t} \left\{ r_t(x, u) + \mathbb{E}[V_{t+1}^*(\tilde{\chi}_t(x, u))] \right\}
\]

where \( r_t(x, u) = \sum_{s=1}^S r_{t,s}(x_s, u_s) \) is the total reward for a given period.

For an arbitrary policy \( \pi \), we can write the corresponding value function \( V_t^\pi(x) \) recursively as

\[
V_t^\pi(x) = r_t(x, \pi_t(x)) + \mathbb{E}[V_{t+1}^\pi(\tilde{\chi}_t(x, \pi_t(x)))]
\]

where the terminal case is \( V_{T+1}^\pi(x) = 0 \) for all \( x \). A policy \( \pi \) is optimal for initial state \( x \) if it always satisfies the linking constraint and \( V_1^\pi(x) = V_1^*(x) \).

### 2.2. Dynamic Assortment Problem

As discussed in the introduction, in the dynamic assortment problem a retailer 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 ?, we assume the demand for product \( s \) follows a Poisson distribution with an unknown product-specific rate \( \gamma_s \). The demand rates are assumed to be independent across products and have a gamma prior with shape parameter \( m_s \) and inverse scale parameter \( \alpha_s \) \((m_s, \alpha_s > 0)\), which implies the mean and variance of \( \gamma_s \) are \( m_s/\alpha_s \) and \( m_s/\alpha_s^2 \). The state variable \( x_s \) for product \( s \) is the vector \((m_s, \alpha_s)\) 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 \( m_s/\alpha_s \); 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 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.

? considered several extensions of this basic model that also fit within the framework of dynamic selection problems. One 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. ? also considered an extension with switching costs, which requires keeping track of whether or not a product is currently displayed.

In our numerical examples, we will focus on the basic model considered in ? using parameters similar to those assumed there. We consider two horizons, \( T = 8 \) or 20. We assume that all products are \textit{a priori} identical with gamma distribution parameters \((m, \alpha) = (1.0, 0.1)\) (so the mean and standard deviation for the demand rate are both 10) and rewards are equal to the mean demand \( ms/\alpha_s \) (i.e., the profit margin is \$1 per unit). We will vary the number of products available \( S \) and assume that we can display 25\% of the products available in each period, i.e., \( N_t = 0.25S \).

### 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 \( \alpha, \beta > 0 \); 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 \( q \) on each trial. The number of trials \( n \) in the binomial distribution can be interpreted as a measure of the informativeness of the signals. For example, a binomial signal with \( n = 5 \) provides as much information as 5 signals from a Bernoulli signal (a binomial with \( n = 1 \)). After screening an applicant and observing a signal \( d_s \), the applicant’s state is updated using Bayes’ rule to \( (\alpha_s + d_s, \beta_s + n - d_s) \). 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 \textit{a priori} identical applicants with \((\alpha, \beta) = (1, 1)\). We will vary the number of applicants \( S \) and assume 25\% of the applicants can be accepted 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. In this problem, 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 candidates, a process which typically requires multiple screenings. With these parameters, the DM can screen candidates more than once only if some other candidates are not screened at all.
3. Lagrangian Relaxations

The DP (2) is difficult to solve because the constraint 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 will consider Lagrangian relaxations of this problem where we relax this linking constraint and decompose the value functions into computationally manageable subproblems. This Lagrangian relaxation can then be used to generate a heuristic selection policy (as described in §4) as well as an upper bound on the performance of an optimal policy. Propositions 1-3 are fairly standard in the literature on Lagrangian relaxations of DPs (e.g., ?, ?, and ?). Proposition 4 provides a detailed analysis of the gradient structure of the Lagrangian that is important in the cutting-plane algorithm developed in §3.3 and later analysis.

3.1. The Lagrangian

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

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

Compared to original 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 display as many items as desired (we now require \( u \in \{0,1\}^S \) rather than \( u \in \mathcal{U}_t \)). Both of these changes can only increase the optimal value so the Lagrangian \( L_t^\lambda(x) \) provides an upper bound on the true value function \( V_t^*(x) \).

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

Thus we can use the Lagrangian as a performance bound to assess the quality of a feasible policy.

The advantage of the Lagrangian relaxation is that, for any fixed \( \lambda \), we can decompose the Lagrangian into a sum of item-specific problems that can be solved independently.

Proposition 2 (Decomposition). For all \( x, t, \) and \( \lambda \geq 0 \),

\[
L_t^\lambda(x) = \sum_{t=t-1}^{T} \lambda_t N_t + \sum_{s=1}^{S} V_{t,s}^\lambda(x_s)
\]
where $V_{t,s}^\lambda(x_s)$ is the value function for an item-specific DP: $V_{t+1,s}^\lambda(x_s) = 0$ and

$$V_{t,s}^\lambda(x_s) = \max \left\{ r_{t,s}(x_s,1) - \lambda_t + \mathbb{E}[V_{t+1,s}^\lambda(\tilde{X}_{t,s}(x_s,1))] , \ r_{t,s}(x_s,0) + \mathbb{E}[V_{t+1,s}^\lambda(\tilde{X}_{t,s}(x_s,0))] \right\}. \quad (6)$$

The first term in the maximization of (6) is the value if the item is selected and the second term is the value if the item is not selected. Intuitively, the period-$t$ Lagrange multiplier $\lambda_t$ can be interpreted as a charge for using the constrained resource in period $t$. We will use $\psi$ to denote a deterministic Markovian policy for the Lagrangian relaxation (5) and $\psi_s$ to denote an item-specific policy for (6); we reserve $\pi$ for policies that respect the linking constraints (1).

### 3.2. The Lagrangian Dual Problem

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

$$\min_{\lambda \geq 0} L_1^\lambda(x). \quad (7)$$

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

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 $x$, $t$, and $\lambda \geq 0$, $L(\lambda)$ and $V_s(\lambda)$ are piecewise linear and convex in $\lambda$.

**Proof.** See Appendix A.1. \qed

In (6) we see that the Lagrange multipliers $\lambda_t$ appear as costs paid whenever an item is selected; thus the gradients of $V_s(\lambda)$ and $L(\lambda)$ will be related to the probability of selecting items under an optimal policy for the item-specific DPs (6) for the given $\lambda$. These selection probabilities are not difficult to compute when solving the DP. Since a convex function is differentiable almost everywhere, for “most” $\lambda$ these gradients will be unique. However, as piecewise linear functions, there may be places where $V_s(\lambda)$ and $L(\lambda)$ are not differentiable and the optimal solution for the Lagrangian dual minimization problem (7) is likely to be at such a “kink.” These kinks correspond to values of $\lambda$ 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 $\psi_s$ for the item-specific DP (6) and let $\Psi_1^s(\lambda)$ be the set of deterministic policies for the item-specific DP (6) with Lagrange multipliers $\lambda$ that are optimal for the initial state.

7
(i) **Subgradients for item-specific problems:** For any $\psi_s \in \Psi_s^*(\lambda)$,

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

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

$$
V_s(\lambda') \geq V_s(\lambda) + \nabla_s(\psi_s)(\lambda' - \lambda) \text{ for all } \lambda'.
$$

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

$$
\partial V_s(\lambda) = \text{conv}\{\nabla_s(\psi_s) : \psi_s \in \Psi_s^*(\lambda)\}
$$

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

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

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

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

(iii) **Optimality conditions.** $\mathbf{X}$ 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 $\psi_{s,i} \in \Psi_s^*(\mathbf{X})$ ($n_s \leq T + 1$) and mixing weights $\{\gamma_{s,i}\}_{i=1}^{n_s}$ (with $\gamma_{s,i} > 0$ and $\sum_{i=1}^{n_s} \gamma_{s,i} = 1$) such that

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

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

**Proof.** See Appendix A.1.

We can interpret the result of Proposition 4(iii) as saying that the optimal policies for the Lagrangian must satisfy the linking constraints (1) “on average” for a mixed policy $\tilde{\psi} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_S)$ where the item-specific mixed policies $\tilde{\psi}_s$ 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 must consider the uncertainty in the state evolution process when following a given item-specific policy $\psi_{s,i}$ (this determines $p_{t,s}(\psi_{s,i})$) and the probability of following policy $\psi_{s,i}$. We describe how to calculate the policies and mixing weights in the next subsection.

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 Appendix A.2. In both representations, we randomize independently across items.

In the special case where the all items are \textit{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: \( \lambda^* \) 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 \( \psi_i \in \Psi^*(\lambda^*) \) \((n \leq T + 1)\) and mixing weights \( \{ \gamma_i \}_{i=1}^n \) such that

\[
S \sum_{i=1}^n \gamma_i p_t(\psi_i) = N_t \text{ for all } t \text{ such that } \lambda_t > 0 \quad \text{and} \\
S \sum_{i=1}^n \gamma_i p_t(\psi_i) \leq N_t \text{ for all } t \text{ such that } \lambda_t = 0 .
\]

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

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

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 to solve the Lagrangian dual problem (7). Alternatively, one could use subgradient methods (as in, for example, ? and ?) or a Nelder-Mead method (as in ?) or an LP formulation that we discuss in the next subsection. Here we describe a cutting-plane method that exploits the structure of the subgradients described in Proposition 4. This cutting-plane method effectively exploits the separability (over items and time) in the Lagrangian dual problem and, unlike the subgradient or Nelder-Mead method, the cutting-plane method is guaranteed to terminate in a finite number of iterations with a provably optimal solution. In our setting, it is important to solve the Lagrangian dual problems efficiently and, as we will see in §5-6, asymptotic optimality of the Lagrangian index policy requires an optimal solution for the Lagrangian dual.

In this cutting-plane method, we proceed iteratively through a series of trial points \( \lambda_k \), 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; recall from Proposition 4, these subgradients correspond to selection probabilities for an optimal policy for the given \( \lambda_k \). By (9), we know \( V_s(\lambda) \geq V_s(\lambda_k) + \nabla_{s,k}^T(\lambda - \lambda_k) \) for each \( k \), i.e., the subgradients provide a linear approximation of \( V_s(\lambda) \) from below. We then approximate the Lagrangian \( L(\lambda) = N^T\lambda + \sum_{s=1}^S V_s(\lambda) \) as

\[
N^T\lambda + \sum_{s=1}^S V_s(\lambda_{i_s}) + \nabla_{s,i_s}^T(\lambda - \lambda_{i_s}) \quad \text{(13)}
\]
where we use the value and subgradient from iteration \(i_s, i_s \in \{1, \ldots, k\}\), to approximate \(V_s(\lambda)\). Taking the upper envelope of these linear approximations, we have the cutting-plane model of the

\[
\ell_k(\lambda) \equiv \max_{i_1, \ldots, i_S \in \{1, \ldots, k\}} \left\{ N^T \lambda + \sum_{s=1}^S (V_s(\lambda_{i_s}) + \nabla_{s,i_s}^T (\lambda - \lambda_{i_s})) \right\} .
\]  

\((14)\)

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

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

\[
\lambda_{k+1} = \arg \min_{\lambda \geq 0} \ell_k(\lambda) .
\]  

\((15)\)

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 \(L(\lambda_{k+1}) = N^T \lambda_{k+1} + \sum_{s=1}^S V_s(\lambda_{k+1})\). The process continues until \(\ell_k(\lambda_{k+1}) = L(\lambda_{k+1})\). In this terminal case, since \(\lambda_{k+1}\) minimizes \(\ell_k(\lambda)\) and \(\ell_k(\lambda) \leq L(\lambda)\), we know that \(\lambda_{k+1}\) is an optimal solution for \((7)\). 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 \(L\) since the new subgradient will support \(L\) at \(\lambda_{k+1}\) 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 in a finite number of iterations.

The cutting-plane optimization problem \((15)\) can be formulated as a linear program (LP) as

\[
\min_{\lambda, v_s} N^T \lambda + \sum_{s=1}^S v_s \\
\text{s.t. } v_s \geq V_s(\lambda_i) + \nabla_{s,i}^T (\lambda - \lambda_i) \quad \forall i \in \{1, \ldots, k\}, \forall s \in \{1, \ldots, S\} , \\
\lambda \geq 0 .
\]  

\((16)\)

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 \(\lambda_{k+1}\). We solve \((16)\) using the dual simplex method, using the optimal dual basis from one iteration as an initial dual basis for the next iteration.

---

\(^1\)The standard cutting-plane method takes the maximum in \((14)\) using values and subgradients of the objective function, here \(L(\lambda)\), at each stage. Effectively this requires using the values and gradients from the same iteration \(i_s\) for all items in \((14)\) 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 \((14)\) and thereby accelerates convergence of the algorithm.
We can write the dual of the LP (16) as
\[
\max_{\gamma_{s,i}} \sum_{s=1}^{S} \sum_{i=1}^{k} (V_s(\lambda_i) - \nabla s,i \lambda_i) \gamma_{s,i}
\]
\[
s.t. - \sum_{s=1}^{S} \sum_{i=1}^{k} \gamma_{s,i} \nabla s,i \leq N
\]
\[
\sum_{i=1}^{k} \gamma_{s,i} = 1 \quad \forall s \in \{1, \ldots, S\}
\]
\[
\gamma_{s,i} \geq 0 \quad \forall i \in \{1, \ldots, k\}, \forall s \in \{1, \ldots, S\}.
\]

In the final step of the cutting-plane method where \( \ell_k(\lambda_{k+1}) = L(\lambda^*) \), 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 (17) 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 \( \lambda_i \) and also optimal given \( \lambda^* \).

In our numerical examples, the computational bottleneck when solving the Lagrangian dual problem using the cutting-plane method is calculating the item-specific value functions (6) and their subgradients; the LP (16) is easy to solve. If some or all items are identical, the cutting-plane method can be simplified as the DP and its gradients need only be calculated once for the identical items; the LPS (16) and (17) similarly simplify. If we let \( S' \) denote the number of distinct items, the LP (16) has \( T + S' \) decision variables and \( k \times S' \) constraints, where \( k \) is the iteration in the cutting-plane method. Thus these LPS will be small, even if the item-specific DPs have large state spaces.

3.4. Linear Programming Formulations

We can also formulate the Lagrangian dual problem (7) as an LP; \( ?, ?, \) and \( ? \) considered similar LP formulations. First, following the standard LP formulation of a DP, we can write the item-specific DP (6) for item \( s \) with Lagrange multipliers \( \lambda \) as
\[
\min_{V^\lambda_{t,s}(x_s)} V^\lambda_{s,1}(x_s^0)
\]
\[
s.t. \quad V^\lambda_{t,s}(x_s) \geq r_{t,s}(x_s, u_s) - \lambda_t u_s + \sum_{\tilde{\chi}_{t,s}} p_t(\tilde{\chi}_{t,s} \mid x_s, u_s) \quad V^\lambda_{t+1,s}(\tilde{\chi}_{t,s}) \quad \forall t, x_s, u_s,
\]
where \( x_s^0 \) is the initial state of item \( s \) and \( p_t(\tilde{\chi}_{t,s} \mid x_s, u_s) \) is the conditional probability of state \( \tilde{\chi}_{t,s} \) occurring when starting in state \( x_s \) and taking action \( u_s \) (with \( u_s \in \{0, 1\} \)). The decision variables in this LP are the values \( V^\lambda_{t,s}(x_s) \) for each period \( t \) and state \( x_s \) and the constraints represent the Bellman equations (6). (We assume \( V^\lambda_{T+1,s}(x_s) = 0 \).) The value function constraints will be binding for optimal actions in states that are visited when following the optimal policy, but need not be binding for any action in states that are not visited by the optimal policy.

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

\[
\begin{align*}
\min_{\lambda, V_{1,s}^\lambda(x_s)} & \quad \sum_{i=1}^T \lambda_i N_t + \sum_{s=1}^S V_{1,s}^\lambda(x_s^0) \\
\text{s.t.} & \quad V_{t,s}^\lambda(x_s) \geq r_{t,s}(x_s, u_s) - \lambda_t u_s + \sum_{\chi_{t,s}} p_t(\chi_{t,s} | x_s, u_s) V_{t+1,s}^\lambda(\chi_{t,s}) \quad \forall s, t, x_s, u_s , \quad (19)
\end{align*}
\]

\( \lambda_t \geq 0 \quad \forall t . \)

If we let \(|X_s|\) be the size of the state space for item \( s \), this LP has \( T \times \left( 1 + \sum_{s=1}^S |X_s| \right) \) decision variables and \( 2 \times T \times \sum_{s=1}^S |X_s| \) constraints. (If some or all of the items are identical, this LP can be simplified.) Though this LP formulation delivers optimal values for \( \lambda \) and the initial values \( V_{1,s}^\lambda(x_s^0) \) for the item-specific DPs, it does not provide a full optimal value function for all periods and states because values for states that are not visited under the optimal policy do not affect the objective function. The Lagrangian heuristics defined in the next section require a full value function. To calculate these value functions using this LP formulation, we need to fix \( \lambda \) at the optimal value from (19) and solve LPs like (18) or (19) with an objective function that includes positive weights on the values \( V_{1,s}^\lambda(x_s) \) for all items, states, and periods.

Taking \( \nu_{t,s}(x_s, u_s) \) to be the dual variables for the constraints in (19), we can write the dual of (19) as:

\[
\begin{align*}
\max_{\nu_{t,s}(x_s, u_s)} & \quad \sum_s \sum_{x_s} \sum_{u_s} \sum_{x_{t,s}} r_{t,s}(x_s, u_s) \nu_{t,s}(x_s, u_s) \\
\text{s.t.} & \quad \sum_{u_s} \nu_{t,s}(x_s^0, u_s) = 1 \quad \forall s \quad (20) \\
& \quad \sum_{u_s} \nu_{t,s}(\chi_{t,s}, u_s) = \sum_{x_s} \sum_{u_s} p_t(\chi_{t,s} | x_s, u_s) \nu_{t-1,s}(x_s, u_s) \quad \forall s, t > 1, \chi_{t,s} , \\
& \quad \sum_{s} \sum_{x_s} \nu_{t,s}(x_s, 1) \leq N_t \quad \forall t \\
& \quad \nu_{t,s}(x_s, u_s) \geq 0 \quad \forall s, t, x_s, u_s .
\end{align*}
\]

The dual variables here have a natural interpretation as flows: \( \nu_{t,s}(x_s, u_s) \) can be interpreted as the probability of being in state \( x_s \) at time \( t \) and choosing action \( u_s \). The objective in (20) is the expected total reward. The first two constraints are flow conservation conditions: the total flow in the initial state \( x_s^0 \) for each item \( \sum_{u_s} \nu_{t,s}(x_s^0, u_s) \) is equal to 1 and the total flow into a later state \( \chi_{t,s} \) must have come from a transition from some previous state. The third constraint requires the linking constraint to hold “on average” and complementary slackness ensures that this linking constraint holds with equality in period \( t \) whenever \( \lambda_t > 0 \). This average linking constraint is thus equivalent to the necessary and sufficient conditions for optimality in the Lagrangian dual given in Proposition 4(iii). Complementary slackness also implies that if the optimal flow \( \nu_{t,s}(x_s, u_s) \) is positive, the corresponding value function inequality in (19) holds with equality: that is, the action \( u_s \) is optimal in state \( x_s \) in period \( t \). The optimal flows \( \nu_{t,s}(x_s, u_s) \) given by the LP (20) can also
Selection probabilities by period \( (p_t(\psi_i)) \) Mixing 
Policy \((\psi_i)\) | 1 | 2 | 3 | 4 | 5 | Mixing weight \((\gamma_i)\)
---|---|---|---|---|---|---
a: Never screen | 0 | 0 | 0 | 0 | 0 | 0.300
b: Screen once | 0 | 0 | 1 | 0 | 0.5 | 0.025
c: Screen once | 0 | 0 | 0 | 1 | 0.5 | 0.075
d: May screen twice | 1 | 0 | 0 | 0.5 | 0.333 | 0.250
e: May screen twice | 0 | 1 | 0.5 | 0 | 0.333 | 0.250
f: May screen twice | 0 | 0 | 1 | 0.5 | 0.333 | 0.100

Weighted average: 0.25 0.25 0.25 0.25 0.25 1.000

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

be calculated from the policies \( \psi_{s,i} \) and mixing weights \( \gamma_{s,i} \) given by the cutting-plane method of §3.3; see Appendix A.2.

Though this LP formulation gives some additional insight into the structure of the Lagrangian dual problem, the LP formulation does not appear to provide any computational advantages over the cutting-plane method of §3.3. Solving the LP (19) requires simultaneously considering all of the item-specific subproblems and all periods of the subproblems. The cutting-plane method more fully exploits the separability of the problem across items and across time, focusing on one item-specific DP (6) at a time when evaluating the Lagrangian and focusing on one period at a time when evaluating the item-specific DPs. For instance in the dynamic assortment example with horizon \( T = 20 \), solving the Lagrangian dual as an LP took about 16 hours, using a commercial LP solver (MOSEK) and exploiting the simplifications due to having identical items. In contrast, the cutting-plane method takes less than 2 minutes. Solving these Lagrangian dual problems efficiently is particularly important when calculating information relaxation bounds (§7) or when reoptimizing with the Lagrangian index policy (§5.3) because the dual problems must be solved repeatedly when calculating the bounds or simulating the policies.

3.5. Applicant Screening Example

To illustrate the Lagrangian model and the role of mixed policies, we consider the applicant screening example described in §2.3 in the case with Bernoulli signals. Here the horizon \( T \) is 5 and the DM can screen 25% of the applicants in the first four periods and can admit 25% in the final period. As discussed in §2.3, with these assumptions the DM must choose between screening each applicant once or screening some applicants more than once with the hope of identifying better candidates to admit. Using the cutting-plane method to solve the Lagrangian dual problem (7), we find an optimal solution with \( X = (0.0333, 0.0333, 0.0333, 0.0333, 0.60) \) and optimal policies \( \psi_i \) with selection probabilities \( p_t(\psi_i) \) and mixing weights \( \gamma_i \) shown in Table 1. This mixture of policies selects 25% of the applicants in each period, as required by the optimality condition (12).

Figure 1 portrays the evolution of the screening process with the optimal mixture of policies shown in Table 1. 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 the rectangles on the vertical axis represent 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 divided into the 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 time (from policy (c)) and screened a second time (from policies (d) and (f)).

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.

The policies and mixing weights in Table 1 are not unique, nor are the optimal Lagrange multipliers $\lambda^*$. 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 screened in each period, as required by the optimality condition (12).
4. Heuristic Policies

The optimal policies for the Lagrangian relaxation for a given $\lambda$ cannot be implemented because they regularly violate the linking constraints. For instance in the applicant screening example with an optimal $\lambda^*$ just discussed, the optimal Lagrangian policy screens and admits $N_t$ 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 constraints in every scenario and hence can be implemented. We will compare their performance in the dynamic assortment and applicant screening examples in the next section and analyze the performance of the optimal Lagrangian index policy (introduced in §4.4) in §6.

4.1. Index Policies

All of the heuristics we consider can be viewed as index policies. In an index policy, we calculate a priority index $i_{t,s}(x_s)$ (or desirability index) that indicates the relative attractiveness of selecting item $s$ in period $t$ when the item is in state $x_s$. Given the current 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. The linking constraints will thus be satisfied by construction and the resulting policies will be feasible for the dynamic selection problem (2). We will generally break ties among items with the same priority index randomly, with the exception of the optimal Lagrangian index policy described in §4.4.

The heuristics we consider all take the index to be an approximation of the value added by selecting item $s$ in period $t$ when the item is in state $x_s$,

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

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 $\lambda$ takes the approximate value function $W_{t+1,s}(\mathbf{x})$ to be the item-specific Lagrangian value function $V_{t+1,s}(\mathbf{x})$ given by (6). The myopic policy simply takes $W_{t+1,s}(\mathbf{x}) = 0$. The other heuristic policies we consider can be viewed as variations of the Lagrangian index policy.

Though we describe these heuristics as index policies, we can also view these heuristics as being greedy with respect to an approximate value function $W_t(\mathbf{x}) = \sum_{s=1}^{S} W_{t,s}(x_s)$. 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:

$$\max_{\mathbf{u} \in \mathcal{U}_t} \left\{ r_{t}(\mathbf{x}, \mathbf{u}) + \mathbb{E}[W_{t+1}(\tilde{\chi}_{t}(\mathbf{x}, \mathbf{u}))] \right\}. \tag{22}$$

 Ranking items by priority index and selecting $N_t$ items with the largest (nonnegative) indices solves the optimization problem (22) exactly. In the case of the Lagrangian index policy, the approximate value
function $W_{t+1}(x)$ differs from the Lagrangian $L_{t+1}^\lambda(x)$ by a constant. Thus a Lagrangian index policy can be viewed as using the Lagrangian as an approximate value function.

4.2. Whittle Index Policy

The Whittle index policy 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 $\lambda = w1$ where $1$ 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 an item and not selecting it,

$$r_{t,s}(x_s, 1) - w + E[V_{t+1,s}^w(\tilde{\chi}_{t,s}(x_s, 1))] = r_{t,s}(x_s, 0) + E[V_{t+1,s}^w(\tilde{\chi}_{t,s}(x_s, 0))]$$

or, equivalently, in the form of (21),

$$w = (r_{t,s}(x_s, 1) + E[V_{t+1,s}^w(\tilde{\chi}_{t,s}(x_s, 1))] - (r_{t,s}(x_s, 0) + E[V_{t+1,s}^w(\tilde{\chi}_{t,s}(x_s, 0))]). \tag{23}$$

It is important to note that these Whittle indices may not be well defined. For example, ? describes an example that is not “indexable” because there are multiple $w$ satisfying (23). 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 $\lambda = w1$ 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 Appendix B.1 for details), but this still essentially requires solving item-specific DPs once for each period and state.

? showed that Whittle indices are well defined in the dynamic assortment problem and noted that computing them is a “complicated task.” Rather than using this hard-to-compute Whittle index, they suggested using an approximate index that is based on approximating the expected continuation values in (23) with a one-step lookahead value function and a normal distribution.

The Whittle indices are well defined in the applicant screening example and are not difficult to compute, but turn out to be not very helpful. In period $T$, the Whittle index for any applicant is the expected reward associated with admitting the applicant. However, the Whittle indices for the earlier screening periods are all zero, regardless of the state $x_s$ of the applicant (see Appendix B.2 for a proof). Thus the Whittle index is not helpful for prioritizing applicants in the screening process. This failure of the Whittle index policy initially surprised us, but it perhaps should not be too 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 restless bandits that ? considered.

4.3. Modified Whittle Index Policy

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

\[
m_{t,s}(x_s) = (r_{t,s}(x_s, 1) + E[V_{t+1,s}^m(\tilde{\chi}_{t,s}(x_s, 1))] - (r_{t,s}(x_s, 0) + E[V_{t+1,s}^m(\tilde{\chi}_{t,s}(x_s, 0))]).
\] (24)

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 (24).

In our experiments with the 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 example, with our specific numerical assumption, the modified Whittle index policy prioritizes screening unscreened candidates, 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

As discussed in §3, for any \( \lambda \geq 0 \), the Lagrangian \( L^\lambda_1(x) \) provides an upper bound on the expected rewards of an optimal policy and hence can be used as a performance bound for any heuristic. Similarly, we can define a Lagrangian index policy for any \( \lambda \). Intuitively, one might expect Lagrange multipliers \( \lambda \) that lead to better performance bounds would be better approximate value functions and tend to generate better heuristics. We will show that optimal Lagrange multipliers \( \lambda^* \) do in fact generate an index policy that is asymptotically optimal (in a sense to be made precise in §6), but we need to take care when breaking ties if there are nonidentical items with equal priority indices.

To illustrate the issues with tiebreaking, consider implementing the Lagrangian index policy for \( \lambda^* \) in the applicant screening example with Bernoulli signals, as discussed in §3.5. 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 it turns out that 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.\(^2\) In all other (reachable)

\(^2\)Why are the priority indices equal? In the first period, in the item-specific DP (6), selecting and not selecting are both optimal in the initial state (see, e.g., the policies in Table 1) and the priority index (21) is equal to \( \lambda_1^* \). (The DP values of selecting and not selecting are equal with the Lagrange multiplier included, thus the index value is \( \lambda_1^* \).) In later periods, selecting and
states, the priority indices are smaller. For these later periods, tiebreaking can be important. For example, if the policy consistently breaks this tie in favor of screening unscreened applicants, all applicants will be screened once, leading the DM in the final period to choose applicants to admit from the many applicants with a single positive signal. Consistently breaking ties the other way (in favor of rescreening applicants with a positive signal) is also not ideal. We can do better by breaking ties in a more sophisticated manner.

Given an index policy \( \pi \) defined by indices \( i_{t,s}(x_s) \), we can define a new index policy \( \pi' \) that uses a policy \( \psi = (\psi_1, \ldots, \psi_S) \) as a tiebreaker by defining a new index

\[
i'_{t,s}(x_s) = i_{t,s}(x_s) - \epsilon \cdot (1 - \psi_{t,s}(x_s)),
\]

for a small \( \epsilon > 0 \). Here \( \epsilon \) is chosen to be small enough so that it 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 \( \psi_s \); items not selected by \( \psi_s \) in a given period/state are penalized slightly, so they will “lose” on this tiebreaker. Also note that items with a priority index of zero will not be selected with this new index policy if \( \psi_s \) does not select the item. We break any remaining ties randomly.

We define an optimal Lagrangian index policy \( \bar{\pi} \) as a Lagrangian index policy for \( \lambda^* \) that uses an optimal mixed policy \( \bar{\psi} \) for the Lagrangian dual problem (7) as a tiebreaker. We can generate a mixed policy \( \bar{\psi} \) for tiebreaking using any of the three methods discussed after Proposition 4(iii).

- Simple random mixing: independently randomly assign each item \( s \) a policy \( \psi_s \) according to the mixing weights of Proposition 4(iii) in each scenario.
- Markov random mixing: \( \psi_{t,s}(x_s) \) in (25) is randomly selected from \( \{0, 1\} \) with state-dependent probabilities given in Appendix A.2.
- Proportional assignment: if some or all of the items are identical, we can sometimes construct a non-random tiebreaking policy \( \psi \) where items are assigned different policies with proportions reflecting the desired mixing weights.

In our numerical examples, we will generate tiebreaking policies \( \psi_s \) using proportional assignments, using random mixing to allocate non-integer remainders when necessary. For example in the applicant screening example with the optimal policy mixture in Table 1, if \( S = 1000 \), we assign \((300, 25, 75, 250, 250, 200)\) applicants to the 6 policies listed in Table 1. If \( S = 100 \), the desired proportions are not integers, so we randomize, assigning \((30, 3, 7, 25, 25, 20)\) or \((30, 2, 8, 25, 25, 20)\) to these policies, each 50% of the time. We use proportional assignments because it reduces the uncertainty in the model and seems to lead to slightly better performance. If we want to eliminate uncertainty in the policies altogether, we can use the so-called “method of conditional expectations” (see, e.g., 7) to convert a mixed optimal Lagrangian index policy into a deterministic policy that performs at least as well as the mixed policy.

By taking the tiebreaking rule to be an optimal policy \( \bar{\psi} \) for the Lagrangian, we can ensure asymptotic
optimality of the heuristic policy (see §6). The use of an optimal tiebreaking method also make a significant difference in the numerical results in the applicant screening examples that are presented in the next section.

5. Numerical Examples

In this section, we compare the performance for the heuristics considered in the previous section in the context of the dynamic assortment and applicant screening examples. 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 $\lambda^*$ to the Lagrangian dual (7) which randomly breaks ties among items with the same priority index, and (v) an optimal Lagrangian index policy, which breaks ties as discussed in §4.4. As discussed in §2.2-2.3, we consider two versions of the dynamic assortment example (with horizon $T$ equal to 8 and 20) and two versions of the applicant screening example (with the informativeness of the binomial signal $n$ equal to 1 and 5). We will vary the number of items considered ($S$) in both examples.

5.1. Run Times

To implement the Whittle, modified Whittle and Lagrangian index policies, we must first calculate their respective indices; Table 2 reports the times required to calculate these indices for all states for each example. All calculations were performed using Matlab on a personal computer; we used MOSEK within Matlab to solve the LP (16) appearing in the cutting-plane method. In these examples, the items are identical so we need only calculate indices for a single item, regardless of the number of items $S$ considered.

<table>
<thead>
<tr>
<th>Index</th>
<th>Dynamic assortment example</th>
<th>Applicant screening example</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T = 8$</td>
<td>$T = 20$</td>
</tr>
<tr>
<td>Whittle</td>
<td>24.0</td>
<td>7,039</td>
</tr>
<tr>
<td>Modified Whittle</td>
<td>8.8</td>
<td>982</td>
</tr>
<tr>
<td>Lagrangian</td>
<td>0.9</td>
<td>126</td>
</tr>
</tbody>
</table>

Table 2: Run times (seconds) for index calculations

In these index calculations, the run times are dominated by the time required to solve the item-specific DPs (6). The time to required to solve these DPs is primarily determined by the number of states that must be considered (shown at the bottom of Table 2). In the dynamic assortment examples, the size of the DP grows rapidly with $T$ and the run times increase accordingly: with $T = 20$, the Whittle indices require about 2 hours to compute, the modified Whittle indices require about 16 minutes, and the Lagrangian indices require about two minutes. In the applicant screening examples, the DPs are much simpler and the index calculations all take less than a few hundredths of a second.

---

3Detailed 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, MOSEK Version 7.1.0.60.

4We truncate the demand distributions at $\bar{d} = 150$ (thereby including 99.9999% of the possible demand outcomes). In period $t$, there are $\sum_{r=0}^{T-1}(r-1)d+1$ possible states, representing the values of $(m, \alpha)$ that could be obtained under some policy.
5.2. Simulation Results

Figures 2-5 describe the performance of the heuristics with numbers of items $S$ (products or applicants) equal to 4, 8, 16, . . . , 16,384. In all cases, we scale $N_t$ (the number of products displayed or applicants screened/admitted) with $S$, taking $N_t = 0.25S$. Note the horizontal axes in the figures showing $S$ are plotted on a log scale. The heuristics are evaluated using simulation, with a sample of 1000 trials. The samples are common across heuristics, meaning, for any given $S$, the products have the same randomly generated demands (and applicants have the same signals) for all heuristics. The expected total rewards $V_1^N(x)$ for the heuristics are estimated from these simulations using a control variate based on the Lagrangian that will be discussed later; see (43) below. The error bars in the figures represent 95% confidence intervals for these estimated values. The Lagrangian bounds $L_N^N(x)$ 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. (We will discuss the information relaxation bound later.) The (b) panels of these figures show estimates of the performance gap for the heuristics, $L_N^N(x) - V_1^N(x)$, where these estimates of the gaps are plotted on a log scale.

**Dynamic Assortment Example.** In the dynamic assortment example 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 more closely at the performance gaps in Figure 2(b) in absolute terms rather than relative terms, we see that the gaps for both Whittle index policies grow linearly in $S$ (linear growth corresponds to a slope of one in the log-log plot). In contrast, the performance gaps for the Lagrangian index policies grow sublinearly. This implies that in Figure 2(a), the modified Whittle index policy approaches an asymptote below the Lagrangian bound, whereas the two Lagrangian index policies truly approach the Lagrangian bound. In this example, there is no difference between the two Lagrangian index policies because there are no scenarios where products in different states have the same priority indices, so the tiebreaking rules do not matter.

Note that these 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 modified Whittle index policy for large $S$, but again both exhibit linear growth in the performance gap.
Figure 2: Results for the dynamic assortment example with horizon $T=8$

Figure 3: Results for the dynamic assortment example with horizon $T=20$
The performance gaps for the Lagrangian index policies again grow sublinearly. With $S = 16,384$, the total reward for the optimal Lagrangian index policy is approximately $1,736,761$ (with a mean standard error of $4$) and the Lagrangian bound is $1,736,858$, so the optimal Lagrangian index policy is within $98$ of the optimal solution.

In the (a) panels of Figures 2 and 3 (as well as Figures 4 and 5 below), we see that the relative performance is increasing in $S$ for all of the heuristics we considered. Intuitively and informally, the heuristics benefit from some economies of scale. As discussed in §4.1, these heuristics can be viewed as selecting items to solve an optimization problem (22) with an approximate value function. As we move from $S$ to $2S$ items and increase $N_t$ to $2N_t$, one could separately optimize two batches of $S$ items (choosing $N_t$ items from each batch) and obtain an expected value of that is twice the value with $S$ on this approximate objective. However, by combining these two batches into a single large batch, we can perhaps do better: in some scenarios, we might select more than $N_t$ from one batch and less than $N_t$ from the other batch to obtain $2N_t$ total. Thus we would expect the performance of the heuristic for the combined batch to be at least twice the value with $S$ items. This suggests that we might expect performance to be increasing in relative terms and intuitively we would expect the magnitude of this effect to decrease with increasing $S$. However, this argument is informal as there is no direct link between improving performance on the approximate objective in (22) with improving the total expected reward for the heuristic.

**Applicant Screening Example.** The performance of the heuristics is more varied in the applicant screening example. We first consider the case with Bernoulli signals ($n = 1$). In Figure 4(a), we see that all of the heuristic policies other than the optimal Lagrangian index policy approach an asymptote below the Lagrangian bound. As discussed in §4.3, 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; this has an expected value of 0.666 per applicant admitted. (With small $S$, there is some chance that fewer than 25% of the applicants will receive a positive signal so the expected value is less than 0.666 per applicant admitted.) As discussed in §4.2, the Whittle indices during the screening stages are all zero, so the Whittle index policy reduces to randomly selecting applicants to screen. Since the rewards are zero during the screening periods, the myopic policy also reduces to random screening. This random screening policy outperforms “screen all applicants” (as suggested by the modified Whittle index policy) in this case because it generates some candidates with two or more positive signals who will be preferred to those with a single positive signal. The difference between the Lagrangian index policies with optimal and random tiebreaking highlights the importance of tiebreaking, as discussed in §4.4. In Figure 4(b), we see that the performance gaps grow linearly in $S$ for all of the heuristics other than the optimal Lagrangian index policy, as we would expect given the results in Figure 4(a). The performance gap for the optimal Lagrangian index policy appears to grow with $\sqrt{S}$ (the line has slope 0.5 in the log-log plot) which is consistent with our theoretical analysis in §6.

Figures 5(a) and (b) show the same results for the case with binomial signals where $n = 5$. Here the
(a) Relative performance

(b) Performance gaps

Figure 4: Results for the applicant screening example with Bernoulli signals (n=1)

(a) Relative performance

(b) Performance gaps

Figure 5: Results for the applicant screening example with binomial signals (n=5)
results are similar but the policy that screens all applicants (as suggested by the modified Whittle indices) outperforms random screening (as suggested by the standard Whittle indices). 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. “Screen all applicants” is still worse than the Lagrangian index policies, for large \( S \). The difference between the two tiebreaking methods in the Lagrangian index policy is also less 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 \( S \).

5.3. Variations

Figure 6 shows results for several variations on the heuristics discussed above, focusing on the applicant screening example. 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 applicant screening examples are small enough to allow reoptimization (the run times range from 9 to 46 seconds for the results reported in the figure), but reoptimization would be very time consuming in the dynamic assortment examples. With reoptimization, we have to solve the Lagrangian dual problem in every period in every simulated scenario and these dual 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.

Figure 6: Results for applicant screening example with variations on the heuristics.
and the errors grow linearly in $S$. Thus reoptimization is not a substitute for being smart about tiebreaking.

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

Finally, we show results for the fluid heuristic described in ?. The fluid heuristic is based on reoptimization of the Lagrangian dual problem and recommends choosing actions (here selecting items) in period $t$ and state $x_s$ to maximize the total flow,

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

where the $\nu_{t,s}(x_s, u_s)$ are the optimal flows given by the solution to the LP (20) for the given period and state. The intuition behind this heuristic is that flows are positive for items that would be selected in the Lagrangian (this follows from complementary slackness as discussed after (20)). Intuitively, the fluid heuristic is similar to the Lagrangian heuristic with reoptimization, but the fluid heuristic breaks ties to maximize flows rather than using an optimal tiebreaking approach. 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.

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

**Proposition 5.** Let $X$ denote an optimal solution for the Lagrangian dual problem (7) with initial state $x$. 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

$$L^X_1(x) - \Delta^{\tilde{\psi}}(x) \leq V^\tilde{\pi}_1(x),$$

where

$$\Delta^{\tilde{\psi}}(x) = (\bar{r} - r) \sum_{t=1}^T \beta_t \sqrt{\bar{N}_t(1 - \bar{N}_t/S)},$$

$\bar{N}_t$ is the expected number of items selected by $\tilde{\psi}$ in period $t$ ($\bar{N}_t = N_t$ if $\lambda_t > 0$, and $\bar{N}_t \leq N_t$ if $\lambda_t = 0$), and the $\beta_t$ are nonnegative constants that depend only on $t$ and $T$. Moreover,

$$\Delta^{\tilde{\psi}}(x) \leq (\bar{r} - r) \left( \sum_{t=1}^T \beta_t \right) \sqrt{N}.$$
Proof. See Appendix C.1.

The proof of Proposition 5 considers the states $\bar{x}_t$ visited using the policy $\bar{\pi}$ that is optimal for the Lagrangian relaxation and characterizes the differences in rewards generated by $\bar{\pi}$ and those generated by the corresponding optimal Lagrangian index policy $\bar{\pi}$. The key observation is that $\bar{\pi}$ and $\bar{\pi}$ make similar selection decisions and the expected number of items with different decisions is bounded by $|n(\bar{\psi}_t(\bar{x}_t)) - N_i|$, where $n(\bar{\psi}_t(\bar{x}_t))$ denotes the number of items selected by the Lagrangian policy $\bar{\psi}$ in period $t$ and state $\bar{x}_t$. The use of $\bar{\psi}$ as a tiebreaker for $\bar{\pi}$ ensures the two policies make similar selections when there are ties. With an optimal policy $\bar{\psi}$ for the Lagrangian and $\lambda^*_t > 0$, the difference $n(\bar{\psi}_t(\bar{x}_t)) - N_t$ has zero mean (by Proposition 4(iii)) and the expectation of $|n(\bar{\psi}_t(\bar{x}_t)) - N_t|$ is bounded by a standard deviation term of the form $\sqrt{N_t(1 - N_t/S)}$. The assumption that the state transitions and mixing of policies are independent across items is important here as it ensures that the standard deviations grow with $\sqrt{N}$ rather than $N$.

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 $V^*_t(x)$.

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

$$V^*_t(x) - \Delta^{\bar{\pi}}(x) \leq L^N_t(x) - \Delta^\pi(x) \leq V^\bar{\pi}_t(x) \leq V^*_t(x) \leq L^X_t(x).$$

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

Since $\Delta^\pi(x)$ is bounded by a term that grows at a rate less than $\sqrt{N}$, Proposition 5 and Theorem 1 provide insight into the asymptotic performance of the Lagrangian index policy and bound for large problems. The results of our numerical examples in §5.2 are consistent with this: in the (b) panels of Figures 2-5, we increase $S$ by adding identical items, increasing $N_t$ proportionally. We saw that the performance gap for the optimal Lagrangian heuristic, $L^X_t(x) - V^\bar{\pi}_t(x)$, grows at a rate less than or equal to $\sqrt{S}$ or, equivalently, less than or equal to $\sqrt{N}$. This implies the relative performance gap, $(L^X_t(x) - V^\bar{\pi}_t(x))/N_t$, approaches zero as we increase $S$ or $N$, as observed in the (a) panels of Figures 2-5.

The next result establishes this kind of asymptotic optimality for large problems in a more general sense. 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 and let $V^*_t(x; S)$, $L^X_t(x; S)$ and $V^\bar{\pi}_t(x; S)$, denote the corresponding optimal value functions, values for the optimal Lagrangian, and value for the optimal Lagrangian index policy $\bar{\pi}$. If the $V^*_t(x; S)$ are positive and satisfy

$$\lim_{S \to \infty} \frac{V^*_t(x; S)}{\sqrt{N(S)}} = \infty,$$

(31)
then
\[
\lim_{S \to \infty} \frac{L^X_1(x;S) - \tilde{V}_1^\pi(x;S)}{V^*_1(x;S)} = 0.
\] (32)

Since \(V_1^\pi(x) \leq V^*_1(x) \leq L^X_1(x;S)\), this implies
\[
\lim_{S \to \infty} \frac{V^*_1(x;S) - \tilde{V}_1^\pi(x;S)}{V^*_1(x;S)} = 0 \quad \text{and} \quad \lim_{S \to \infty} \frac{L^X_1(x;S) - V^*_1(x;S)}{V^*_1(x;S)} = 0.
\]

Proof. See Appendix C.2.

The proposition implies that, when the growth condition (31) is satisfied, the gaps between \(V^*_1(x;S)\), \(L^X_1(x;S)\) and \(V_1^\pi(x;S)\), when normalized by \(V^*_1(x;S)\), will 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 is mild. For example, if the expected reward associated with selecting each item is bounded away from zero and \(N_t(S)\) approaches infinity as \(S\) approaches infinity, then (31) will be satisfied.

Note that we could normalize the ratios in Corollary 1 by the Lagrangian \(L^X_1(x;S)\) rather than \(V^*_1(x;S)\) (because \(V^*_1(x;S) \leq L^X_1(x;S)\)) 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 did in §5.2, 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.

Our numerical examples demonstrate that asymptotic optimality need not hold for the Whittle index policy or the modified Whittle index policy. Asymptotic optimality also need not hold for the Lagrangian index policies for arbitrary Lagrange multipliers \(\lambda\) or even for Lagrangian index policies based on the optimal Lagrange multipliers \(\lambda^*\) if we do not break ties appropriately. In the dynamic assortment examples, the performance gaps appeared to increase at a rate less than \(\sqrt{N}\), but in the applicant screening examples, the growth rate appeared to grow with \(\sqrt{N}\). We have developed simple analytic examples where the gap between the Lagrangian and optimal Lagrangian policies asymptotically grows with \(\sqrt{N}\); see Appendix C.3. Thus \(\sqrt{N}\) is the best possible growth rate for these performance gaps for general dynamic selection problems.

7. Information Relaxation Bounds

In the numerical examples of §5, the gaps between the optimal Lagrangian index policy and Lagrangian bound were very small for large \(S\), but were more substantial for small \(S\). One might wonder whether these gaps are due to the policies being suboptimal or due to slack in the Lagrangian bound. In this section, we consider the use of information relaxations to improve the Lagrangian bounds. We will briefly and informally review the theory of information relaxation bounds as developed in ?; discuss the application to our examples, and discuss numerical results.

7.1. Information Relaxation Bounds

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

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

$$V_1^*(x) = \max_{\pi \in \Pi_F} \mathbb{E}[r(\pi)]$$

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

If we let $\Pi_G$ denote a larger set of policies ($\Pi_F \subseteq \Pi_G$) that can use additional information,\footnote{To formalize the definitions of these sets of policies, a policy can be defined as a mapping from the underlying outcome space to selection decisions ($u_1, \ldots, u_T$) for each product and each period (with $u_t \in U_t$). Policies in the DP (2) that make selections as a function of the current state of the system can be viewed as imposing measurability restrictions on this more general set of policies. The relaxed model imposes a weaker set of measurability restrictions. See \cite{note1} for more discussion.} we can solve a relaxed version of the DP to obtain an upper bound on the primal DP:

$$V_1^*(x) = \max_{\pi \in \Pi_F} \mathbb{E}[r(\pi)] \leq \max_{\pi \in \Pi_G} \mathbb{E}[r(\pi)] .$$

Unfortunately, the bounds given by (34) will be weak if the extra information provided in the relaxation is valuable. To counter this, we incorporate a penalty that “punishes” the DM for using information that would not actually be available when making decisions. The penalty $z(\pi)$ is a policy-dependent random variable, like the rewards, i.e., $z(\pi) = \sum_t z_t(\tilde{x}_t(\pi), \pi_t(\tilde{x}_t(\pi)))$ for some set of period-$t$ penalty terms $z_t(x, u)$. A penalty $z(\pi)$ is dual feasible if $\mathbb{E}[z(\pi)] \leq 0$ for all $\pi \in \Pi_F$; that is, if the expected penalty is nonpositive for all nonanticipative policies.

The following weak duality result from \cite{note1} is the key tool for generating performance bounds using information relaxations.

**Proposition 6** (Weak duality). Suppose $\Pi_F \subseteq \Pi_G$. If policy $\pi$ is nonanticipative (i.e., $\pi \in \Pi_F$) and penalty $z$ is dual feasible then

$$\mathbb{E}[r(\pi)] \leq \max_{\pi' \in \Pi_G} \mathbb{E}[r(\pi') - z(\pi')] .$$

**Proof.** We have:

$$\mathbb{E}[r(\pi)] \leq \mathbb{E}[r(\pi) - z(\pi)] \leq \max_{\pi' \in \Pi_G} \mathbb{E}[r(\pi') - z(\pi')] .$$

Given $\pi \in \Pi_F$, the first inequality follows from the definition of dual feasibility ($\mathbb{E}[z(\pi)] \leq 0$) and the second inequality follows from the fact that $\Pi_F \subseteq \Pi_G$. \qed
provide a strong duality result that shows that there is a penalty such that the value for the relaxed model is exactly equal to the optimal value for the original, but these penalties require knowledge of the optimal value function (more on this in the next subsection).

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

7.2. Information Relaxation Bounds for the Dynamic Assortment Problem

The challenge is to find penalties and information relaxations that make the bound on right side of (35) easy to compute and lead to reasonably tight bounds. For specificity, we will focus our discussion on the dynamic assortment example, though the ideas also apply in the applicant screening example and other dynamic selection problems. In the dynamic assortment example, the underlying uncertainties are the unknown (Poisson) demand rates for each product and the demand realizations for each item, in each period. In the original model, the demands are revealed for products when (and if) the products are selected; the demand rates are never revealed. We can consider a number of different relaxations, including:

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

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

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

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

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

In our discussion and numerical examples, we will focus on the known demands relaxation and consider a penalty based on the Lagrangian \( L_{i+1}^{\lambda}(x) \). Although we can use any \( \lambda \geq 0 \), in our numerical examples we will take these to be optimal Lagrange multipliers \( \lambda' \) given by solving the Lagrangian dual (7). We can estimate the known demands bound, \( \max_{\pi' \in \Pi_G} \mathbb{E}[r(\pi') - z(\pi')] \), by repeatedly:

(i) Drawing a demand rate \( \gamma_s \) for product \( s \) from the appropriate gamma distribution and then drawing demands for this product from a Poisson distribution with this rate. Let \( d = (d_1, \ldots, d_T) \) where
\[ d_t = (d_{t,1}, \ldots, d_{t,S}) \] denotes the randomly generated vector of product demands in period \( t \).

(ii) Solving a deterministic inner DP (to be described shortly) to find the optimal value \( \hat{V}_1(x_1; d) \) given these demand realizations, incorporating the Lagrangian penalty.

We estimate the known demands bound by averaging the \( \hat{V}_1(x_1; \hat{d}) \) for the different demand realizations \( \hat{d} \).

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

\[
\hat{V}_t(x; \hat{d}) = \max_{u \in U_t} \left\{ v_t(x, u) - z_t(x, u; d_t) + \hat{V}_{t+1}(\chi_t(x, u; d_t); \hat{d}) \right\}
\]

(36)

where

\[
z_t(x, u; d_t) = L_{t+1}^\lambda(\chi_t(x, u; d_t)) - \mathbb{E}[L_{t+1}(\hat{\chi}_t(x, u))].
\]

(37)

Here the last term in (36) and the first term in (37) involve deterministic state transitions because the DM knows the demands: \( \chi_t(x, u; d_t) = (\chi_{t,1}(x_1, u_1; d_{t,1}), \ldots, \chi_{t,S}(x_S, u_S; d_{t,S})) \) represents the state transitions with the given product demands for period \( t \). The expectation in (37) is calculated using the same state-dependent negative-binomial distributions used in the original DP.

Using the law of iterated expectations, we know that \( \mathbb{E}[z_t(\bar{x}_t(\pi), \pi_t(\bar{x}_t(\pi)))] = 0 \) for any nonanticipative policy \( \pi \). Thus the penalty \( z(\pi) = \sum_t z_t(\bar{x}_t(\pi), \pi_t(\bar{x}_t(\pi))) \) is dual feasible and the known demands bound provides a performance bound, as in Proposition 6. This is an example of the general method for creating "good" dual feasible penalties described in \( \xi \). As discussed there, if we replace the Lagrangian \( L_{t+1}^\lambda \) in (37) with the optimal value function \( V_{t+1}^\star \), the information relaxation bound will be exactly equal to the optimal value. With this ideal penalty, the DM is exactly punished for using extra information: the benefit gained is exactly canceled by the penalty. With a penalty based on an approximate value function (such as the Lagrangian), the penalty approximately cancels this benefit. In general, to obtain good bounds, we want to choose generating functions that approximate the optimal value function well.

We now consider the DP (36) in more detail. First, note that that the penalty terms involving the Lagrangian \( L_{t+1}^\lambda \) decompose into the sum of item-specific values, as in (5). However, the inner DP (36) does not decompose into item-specific subproblems because the constraint on the total number of products selected \( \{u \in U_t \} \) links the decisions across items, as it did in the original DP (2). Thus, the inner DP – though deterministic – is still difficult to solve in problems with many items.

To decouple the inner DP (36), we relax the linking constraint in the same way that we relaxed the original DP (2). Consider Lagrange multipliers \( \mu = (\mu_1, \ldots, \mu_T) \geq 0 \) and let \( \hat{\lambda}_{t+1}(x; d) = 0 \). The period-\( t \) inner Lagrangian with demand realization \( d \) is then given recursively as

\[
\hat{\lambda}_t^\mu(x; d) = \max_{u \in \{0,1\}^S} \left\{ v_t(x, u) - z_t(x, u; d_t) + \hat{\lambda}_{t+1}^\mu(\chi_t(x, u; d_t); d) + \mu_t \left( N_t - \sum_{s=1}^S u_s \right) \right\}
\]

(38)
This can be decomposed into item-specific DPs as

\[ \hat{L}_t^\mu(x; d) = N_t \sum_{t=1}^T \mu_t + \sum_{s=1}^S \hat{V}_{t,s}^\mu(x_s; d_s) \]  

(39)

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

\[ \hat{V}_{t,s}^\mu(x_s; d_s) = \max \left\{ r_{t,s}(x_s, 1) - \mu_t - V_{s,t+1}^\lambda(\chi_{t,s}(x_s, d_{t,s})) + \mathbb{E} \left[ V_{s,t+1}^\lambda(\chi_{t,s}(x_s, 1)) \right] + \hat{V}_{t+1,s}^\mu(\chi_{t,s}(x_s, d_{t,s})), \right. \]

\[ \left. r_{t,s}(x_s, 0) + \hat{V}_{t+1,s}^\mu(\chi_{t,s}(x_s, 0, d_{t,s})) \right\}. \]

(40)

where \( V_{t,s}^\lambda \) is the value-function for the item-specific DP (6). Note that in the dynamic assortment model, the penalty term (37) is zero if a product is not selected because its state does not change.

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

To ensure we have the best possible bound for a given \( d \) and \( x \), we can solve the inner dual problem,

\[ \min_{\mu \geq 0} \hat{L}_1^\mu(x; d) , \]

(41)

for an optimal \( \mu^*(x, d) \). This is a convex optimization problem and can be solved using the cutting-plane method discussed in §3.3. Moreover, if we take the inner Lagrange multipliers \( \mu \) to be equal to the “outer” Lagrange multipliers \( \lambda \) used to define the penalty, we can use an induction argument to show that \( \hat{L}_t^\mu(x; d) = L_t^\lambda(x) \) for all \( t \) and \( d \). Thus, since \( \lambda \) is feasible but not necessarily optimal for the inner Lagrangian dual problem (41), we have

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

(42)

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

We can also relate these bounds to the performance of a heuristic policy \( \pi \) in the same demand scenario. We focus on deterministic Markovian heuristic policies where the period-\( t \) selection decision \( \pi_t \) is chosen based on the current state \( x \). (When we are considering mixed policies, as in the optimal Lagrangian policy, let \( \pi \) be a particular realization of the mixed policy.) We assume that the actions selected by the heuristic are feasible, i.e., \( \pi_t(x) \in \mathcal{U}_t \). To facilitate comparison with those of the information relaxation, we will adjust the rewards using the penalty (37) as a control variate. Let \( \hat{V}_t^\pi(x; d) \) denote the value generated when following policy \( \pi \), starting in state \( x \), given demand realization \( d \), adjusted by the control variate. We can write this
value recursively in a form parallel to (36): let 
\[ \hat{V}_{t+1}^\pi(x; \tilde{d}) = 0 \] and, for earlier \( t \), we define
\[
\hat{V}_t^\pi(x; d) = \left\{ r_t(x, \pi_t(x)) - z_t(x, \pi_t(x); d_t) + \hat{V}_{t+1}^\pi(\tilde{X}_t(x, \pi_t(x); d_t); d) \right\}. \tag{43}
\]

Here this form exactly mimics the DP recursion (36), except the actions are chosen in accordance to the policy \( \pi \) rather than optimized. Thus we know that 
\[ \hat{V}_t^\pi(x; d) \leq \hat{V}_t(x; d) \] for all \( t, x, \) and \( d \). Moreover, because the penalty terms \( z_t \) have mean zero for all feasible policies, we know that the expected total reward when following policy \( \pi \) is 
\[ V_1^\pi(x) = \mathbb{E}[\hat{V}_1^\pi(x; \tilde{d})], \] where the expectations are taken over the random demand scenarios. These control variates are very helpful in reducing sampling error when estimating the expected values associated with a given policy and were used in the simulations of §5.2.

Combining these observations, we can say the following.

**Theorem 2 (Ordered bounds).** Consider any feasible and nonanticipative policy \( \pi \), Lagrange multipliers \( \lambda \geq 0 \) and initial state \( x \).

(i) For any demand realization \( d \), we have
\[
\hat{V}_1^\pi(x; d) \leq \hat{V}_1(x; d) \leq L_1^\mu(x, d) \leq L_1^\lambda(x). \tag{44}
\]

(ii) Taking expectations over random demand realizations \( \tilde{d} \), we have
\[
V_1^\pi(x) = \mathbb{E}[\hat{V}_1^\pi(x; \tilde{d})] \leq V_1^*(x) \leq \mathbb{E}[\hat{V}_1(x; \tilde{d})] \leq \mathbb{E}[L_1^\mu(x, \tilde{d})] \leq L_1^\lambda(x). \tag{45}
\]

Working from the left in (45), we have the expected value with heuristic policy \( \pi \) \( (V_1^\pi(x)) \) is equal to the expected reward for this policy with the control variate included \( (\mathbb{E}[\hat{V}_1^\pi(x; \tilde{d})]) \). This value is less than or equal to the value with an optimal policy \( (V_1^*(x)) \), which is typically impossible to compute. This, in turn, is less than or equal to the known demands relaxation bound \( (\mathbb{E}[\hat{V}_1(x; \tilde{d})]) \) which is also typically impossible to compute. However, the known demands bound is less than or equal to the Lagrangian relaxation of the known demands information relaxation bound with optimized Lagrange multipliers \( (\mathbb{E}[L_1^\mu(x, \tilde{d})]) \), which is computable. Finally, all of these bounds are less than the ordinary Lagrangian bound \( (L_1^\lambda(x)) \). The bounds in (44) show that the demand-dependent terms in (45) are ordered in every demand scenario \( d \) and less than or equal to the Lagrangian bound.

Though we have focused on the known demands relaxation in the dynamic assortment example, we can use the same approach and derive similar results with other relaxations and in other problems. In the applicant screening example, the information relaxation where all applicant signals are known in advance is exactly analogous to the known demands relaxation and we obtain the same results. If we consider the known rates relaxation instead of the known demands realization in the dynamic assortment example, we arrive at an inner DP similar to (36), but the deterministic demand transitions are replaced with Poisson
distributions with (randomly drawn) known demand rates. This inner DP is also linked and we can use an inner Lagrangian relaxation to derive results analogous to those of Theorem 2.

7.3. Numerical Examples

The (a) panels of Figures 2-5 show information relaxation bounds for the dynamic assortment and applicant screening examples using the known demands and known signals relaxations. These bounds were evaluated with $S$ equal to 4, 8, 16, 32, and 64 in the same 1000 sample scenarios (i.e., same demand and signal sequences) that were used to evaluate the heuristics. In all cases, we use penalties based on an optimal solution $X^*$ for the outer Lagrangian dual problem (7) and impose the policy restrictions discussed at the end of §7.1. These figures also show 95% confidence intervals for the estimated bounds; these confidence intervals are quite narrow, particularly for larger values of $S$.

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

<table>
<thead>
<tr>
<th>$S$</th>
<th>$T = 8$</th>
<th>$T = 20$</th>
<th>$n = 1$</th>
<th>$n = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9.9</td>
<td>143</td>
<td>2.7</td>
<td>3.3</td>
</tr>
<tr>
<td>8</td>
<td>15.1</td>
<td>208</td>
<td>4.4</td>
<td>5.6</td>
</tr>
<tr>
<td>16</td>
<td>23.9</td>
<td>301</td>
<td>7.5</td>
<td>9.1</td>
</tr>
<tr>
<td>32</td>
<td>42.1</td>
<td>471</td>
<td>13.3</td>
<td>16.0</td>
</tr>
<tr>
<td>64</td>
<td>75.0</td>
<td>750</td>
<td>24.4</td>
<td>29.1</td>
</tr>
</tbody>
</table>

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

The run times are reported in Table 3. As discussed above, calculating these bounds requires solving the inner Lagrangian dual problems for each simulated demand (signal) sequence, for each product (applicant). This can be time consuming because the products (applicants) are not identical as each has its own demand (signal) sequence. We use the cutting-plane method in each case and start with $\mu = X^*$, which yields the Lagrangian dual bound. If we cannot improve on this value, the cutting-plane algorithm typically stops after a few iterations. The run times grow roughly linearly with $S$, as one might expect, but not exactly because these no-improvement scenarios are more common with large $S$. 33
8. Conclusions

The methods and results developed here for dynamic selection problems could be adapted to more general weakly-coupled dynamic programs. For example, one might consider a budget constraint with items of varying costs rather than cardinality constraints. The analysis of the Lagrangian in §3 and the cutting-plane method generalize directly; the information relaxation bounds of §7 also generalize directly. The Lagrangian heuristics could also be generalized, but with general linking constraints, we would solve the approximate DP optimization problems (22) as discrete (binary) optimization problems rather than ranking by index values. It is not immediately clear how to adapt the performance analysis of §6 to this more general setting. A key part of the analysis is ensuring that the optimal policies for the Lagrangian relaxation and the Lagrangian index heuristic make similar selections. Establishing this similarity in selections may require more delicate arguments if the heuristics are defined in terms of solutions of a general discrete optimization problem.

The methods and results of the paper could also be generalized to infinite-horizon problems with discounting. For numerical reasons (e.g., the solution of the item-specific DPs and for the cutting-plane method), we would not have an infinite sequence of period-specific Lagrange multipliers, but we could, for example, allow the Lagrange multipliers to vary for a finite number of early periods and then be constant thereafter (as in ?). The information relaxation bounds would also have to be adapted to consider the infinite horizon; see ?. The optimal Lagrangian policies would no longer ensure that \( N_t \) or fewer items are selected on average in periods without period-specific Lagrange multipliers and the performance analysis of §6 would have to be adapted accordingly.

The numerical and theoretical results of this paper suggests 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 easier to apply 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 \( \lambda^* \)) that lead to the correct 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. Selected Proofs

A.1. Proofs for §3: Lagrangian Relaxations

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

\[
V_s(\lambda) = \max_{\psi_s} \sum_{t=1}^{T} \mathbb{E}[r_{t,s}(\mathbf{x}_{t,s}(x_{1,s}; \psi_s)), \psi_{t,s}(\mathbf{x}_{t,s}(x_{1,s}; \psi_s)) - \lambda_t \psi_{t,s}(\mathbf{x}_{t,s}(x_{1,s}; \psi_s))] \tag{46}
\]

where \( \mathbf{x}_{t,s}(x_{1,s}; \psi_s) \) is the random state for item \( s \) in period \( t \) when starting in state \( x_{1,s} \) and following policy \( \psi_s \). For a fixed policy \( \psi_s \), the objective in (46) is linear in \( \lambda \). The pointwise maximum over these linear functions yields a piecewise linear and convex function. The Lagrangian \( L(\lambda) \), as a finite sum of piecewise linear convex functions \( V_s(\lambda) \) (plus additional linear terms), is also piecewise linear and convex.

Proof of Proposition 4. (i): Consider the representation of the item-specific DP given in equation (46) in the proof of Proposition 3. There, for a fixed policy \( \psi_s \), the objective in (46) is linear in \( \lambda \) and the \( t^{th} \) element of the gradient \( \nabla_s(\psi_s) \) with policy \( \psi_s \) is \( -\mathbb{E}[\psi_{t,s}(\mathbf{x}_{t,s}(x_{1,s}; \psi_s))] \), which is \(-p_{t,s}(\psi_s)\). The subdifferential result (10) then follows from Danskin’s Theorem (see, e.g., Proposition 4.5.1, p. 245). This subdifferential result implies \( \nabla_s(\psi_s) \) is a subgradient of \( V_s(\lambda) \) for any \( \psi_s \in \Psi_s(\lambda) \).

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

(iii) A necessary and sufficient condition for \( X \) to be optimal for the Lagrangian dual problem (7) is

\[
0 \in \partial L(X) + \mathcal{N}_{\{\lambda \geq 0\}}(X)
\]

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

A.2. Constructing a Markov Random Policy.

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

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

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

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

Thus the probability of being in state \( x_s \) with this mixed policy is \( \nu_{t,s}(x_s, 0; \tilde{\psi}) + \nu_{t,s}(x_s, 1; \tilde{\psi}) \). If \( \tilde{\psi} \) is an optimal mixed policy for the Lagrangian dual problem, \( \nu_{t,s}(x_s, u_s; \tilde{\psi}) \) is an optimal solution for the LP (20).
For a Markov random policy that corresponds to the mixed distribution \( \tilde{\psi} \), we can take the probability of selecting an item \( s \) in state \( x \) in period \( t \) to be:

\[
\frac{\nu_{t,s}(x,1; \tilde{\psi})}{\nu_{t,s}(x,0; \tilde{\psi}) + \nu_{t,s}(x,1; \tilde{\psi})}
\]  

(47)

By construction, this will generate the same state-action probabilities as \( \tilde{\psi} \), will select the same number of items on average in each period as \( \tilde{\psi} \), and will have the same expected total reward as \( \tilde{\psi} \). Note that these selection probabilities will be undefined when the probability of being in state \( x \) in period \( t \) (in the denominator of (47)) is zero. These undefined selection probabilities are irrelevant for evaluating the Lagrangian policies, but may be relevant when we use the Lagrangian policy as a tiebreaker as discussed in §4.4. In our numerical examples, we take these undefined probabilities to be 0.5; one could experiment with other choices if desired.

B. Notes on Whittle Indices

B.1. Calculating Whittle Indices

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

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

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

\[
\Delta_{t,s}^{w}(x) = (r_{t,s}(x,1) + E[V_{t+1,s}^{w}(x_{t,s}(x,1))] - (r_{t,s}(x,0) + E[V_{t+1,s}^{w}(x_{t,s}(x,0))])
\]

\[
\sigma_{t,s}^{w}(x) = E[\eta_{t+1,s}^{w}(x_{t,s}(x,1))] - E[\eta_{t+1,s}^{w}(x_{t,s}(x,0))].
\]

Here \( \Delta_{t,s}^{w}(x) \) is the difference on the right side of (23) and \( \sigma_{t,s}^{w}(x) \) is the partial derivative of \( \Delta_{t,s}^{w}(x) \) with respect to \( w \).

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

\[
\delta^{*} = \min_{t,s} \left\{ \Delta_{t,s}^{w}(x) - w : \psi_{t,s}(x; w) = 1 \right\}.
\]  

(48)

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

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

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

The breakpoint calculation in (48) can be understood as follows: for any states and periods satisfying \( \psi_{t,s}(x; w) = 1 \), selection is strictly optimal at the current \( w \), and hence \( \Delta_{t,s}^{w}(x) > w \) in such states. Since \( \sigma_{t,s}^{w}(x) \) represents the partial derivative of \( \Delta_{t,s}^{w}(x) \) with respect to \( w \), we seek a value \( \delta \) such that \( w + \delta \) is a new Whittle index, i.e., \( \delta \) satisfies

\[
\Delta_{t,s}^{w}(x) + \sigma_{t,s}^{w}(x) \cdot \delta = w + \delta.
\]

The ratio in (48) represents the largest increase to \( w \) such that the policy \( \psi_{s} \) remains optimal. For times and states attaining this value in (48), we are indifferent between selecting and not selecting the item at \( w + \delta^{*} \).
The efficiency of this procedure is improved by noting some properties of the value functions and derivatives when updating in step (ii), i.e., as \( w \) is replaced with \( w' = w + \delta \). First, we need only update \( \eta_{t,s}^w(x_s) \) and \( \sigma_{t,s}^w(x_s) \) in time periods up to \( t^* \), where \( t^* \) is the earliest time period attaining the minimum in (iv). The partial derivatives for later periods are unchanged because no decisions change after period \( t^* \). Second, we can update the differences as \( \Delta_{t,s}^w(x_s) = \Delta_{t,s}^w(x_s) + \sigma_{t,s}^w(x_s) \cdot \delta \). This follows from the fact that the policy \( \psi_s \) is optimal from \( w \) to \( w + \delta \) and thus the value functions are linear functions of \( w \) in this range.

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

### B.2. Whittle Indices for the Applicant Screening Example

Here we show that in the applicant screening example, the Whittle indices have a particularly simple form. We let \( \mu(x_s) \) denote an applicant’s mean quality in state \( x_s \), i.e., \( \mu(x_s) = \alpha_s / (\alpha_s + \beta_s) \). The item-specific DP (6) with \( \lambda = w1 \) is given recursively as

\[
V_{t,s}^w(x_s) = \max \{-w + \mathbb{E}[V_{t+1,s}^w(\tilde{x}_{t,s}(x_s,1))], V_{t+1,s}^w(x_s)\}.
\]

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

\[
-w + \mu(x_s) = 0 \quad \text{for } t = T, \quad \text{and} \quad -w + \mathbb{E}[V_{t+1,s}^w(\tilde{x}_{t,s}(x_s,1))] = V_{t+1,s}^w(x_s) \quad \text{for } t < T.
\]

We show the following.

**Proposition 7.** In the applicant screening example, for all \( s, t, \) and \( x_s \), the Whittle index is unique.

(i) In the final period \( (t = T) \), the Whittle index is \( \mu(x_s) \).

(ii) In screening periods \( (t < T) \), the Whittle index is zero.

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

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

(ii) We first show that \( w = 0 \) is a Whittle index for \( t < T \). In this case, \( V_{T,s}^w(x_s) = \mu(x_s) \), since \( \mu(x_s) > 0 \). By induction and using the fact that the posterior mean is equal to the prior mean, for \( t < T \), we have \( V_{t,s}^w(x_s) = \mathbb{E}[V_{t+1,s}^w(\tilde{x}_{t,s}(x_s,1))] = \mathbb{E}[\mu(\tilde{x}_{t,s}(x_s,1))] = \mu(x_s) \). Thus (50) holds for \( w = 0 \).

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

\[
-w + \mathbb{E}[V_{t+1,s}^w(\tilde{x}_{t,s}(x_s,1))] = -w + \mathbb{E}[\mu(\tilde{x}_{t,s}(x_s,1)) + (T - t)w] = \mu(x_s) - (T - t + 1)w
\]

where the first inequality follows from the induction hypothesis and the second from the fact that the expected posterior mean is equal to the prior mean. From the induction hypothesis, not screening in period \( t \) yields

\[
V_{t+1,s}^w(x_s) = \mu(x_s) - (T - t)w
\]

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

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

\[
V_{t+1,s}^w(x_s) = \max\{\mu(x_s) - w, 0\}
\]

37
and screening yields:

\[-w + \mathbb{E}[V_{t+1, s}(\tilde{X}_{t, s}(x_s, 1))] = -w + \mathbb{E}[\max\{\mu(\tilde{X}_{t, s}(x_s, 1)) - w, 0\}]\]

\[< -w + \mathbb{E}[\mu(\tilde{X}_{t, s}(x_s, 1))] = -w + \mu(x_s) \leq \max\{\mu(x_s) - w, 0\}\]

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

C. Proofs for §6: Analysis of the Optimal Lagrangian Index Policy

C.1. Proof of Proposition 5

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

**Lemma 1.** For any \(\lambda \geq 0\) and initial state \(x\), let \(\psi\) be an optimal policy for the Lagrangian (5), and let \(\bar{x}_t\) denote the state transition process generated by \(\psi\). Then, for any policy \(\pi\),

\[
L_1^\lambda(x) - V_1^\pi(x) = \sum_{t=1}^{T} \mathbb{E}[d_t(\bar{x}_t, \psi_t(\bar{x}_t), \pi_t(\bar{x}_t))] \tag{51}
\]

where

\[
d_t(x_t, u_t^\psi, u_t^\pi) = \lambda_t(N_t - n(u_t^\psi)) + r_t(x_t, u_t^\psi) - r_t(x_t, u_t^\pi) + \mathbb{E}[V_{t+1}^\pi(\tilde{X}_t(x_t, u_t^\psi))] - \mathbb{E}[V_{t+1}^\pi(\tilde{X}_t(x_t, u_t^\pi))]. \tag{52}
\]

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

**Proof.** Since \(\psi\) is an optimal policy for the Lagrangian \(L_1^\lambda\) starting in state \(x\), we have

\[
L_1^\lambda(x) = \sum_{t=1}^{T} \mathbb{E}[\lambda_t(N_t - n(\psi_t(\bar{x}_t))) + r_t(\bar{x}_t, \psi_t(\bar{x}_t))]. \tag{53}
\]

We also have

\[
V_1^\pi(x) = V_1^\pi(x) + \sum_{t=2}^{T} \mathbb{E}[V_t^\pi(\bar{x}_t)] - \sum_{t=2}^{T} \mathbb{E}[V_t^\pi(\bar{x}_t)]
\]

\[= \sum_{t=1}^{T} \mathbb{E}[V_t^\pi(\bar{x}_t)] - \sum_{t=1}^{T} \mathbb{E}[\tilde{X}_t(\bar{x}_t, \psi_t(\bar{x}_t))]
\]

\[= \sum_{t=1}^{T} \mathbb{E}[r_t(\bar{x}_t, \pi_t(\bar{x}_t))] + \mathbb{E}[V_{t+1}^\pi(\tilde{X}_t(\bar{x}_t, \pi_t(\bar{x}_t)))] - \mathbb{E}[V_{t+1}^\pi(\tilde{X}_t(\bar{x}_t, \psi_t(\bar{x}_t)))] \]

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

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

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

$$\left| V_{t}^{\pi}(x') - V_{t}^{\pi}(x'') \right| \leq k_t \cdot (\bar{r} - r) m .$$

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

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

$$\left| V_{t}^{\pi}(x') - V_{t}^{\pi}(x'') \right| = \left| r_t(x', \pi(x')) - r_t(x'', \pi(x'')) + E[V_{t+1}^{\pi}(\tilde{X}_t(x', \pi(x')))] - E[V_{t+1}^{\pi}(\tilde{X}_t(x'', \pi(x'')))] \right|$$

$$\leq \left| r_t(x', \pi(x')) - r_t(x'', \pi(x'')) \right| + \left| E[V_{t+1}^{\pi}(\tilde{X}_t(x', \pi(x')))] - E[V_{t+1}^{\pi}(\tilde{X}_t(x'', \pi(x'')))] \right|$$

$$\leq 2(\bar{r} - r)m + 2k_{t+1}(\bar{r} - r)m .$$

(54)

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

$$\left| r_t(x', \pi(x')) - r_t(x'', \pi(x'')) \right| \leq 2(\bar{r} - r)m .$$

With differences for at most $2m$ item decisions and state transitions that are independent across items, the random continuation states $\tilde{X}_t(x', \pi(x'))$ and $\tilde{X}_t(x'', \pi(x''))$ will differ for at most $2m$ items. Thus, using the induction hypothesis, we have

$$\left| E[V_{t+1}^{\pi}(\tilde{X}_t(x', \pi(x')))] - E[V_{t+1}^{\pi}(\tilde{X}_t(x'', \pi(x'')))] \right| \leq 2k_{t+1}(\bar{r} - r)m ,$$

completing the proof of the inequality (54). Then taking $k_t = 2(1 + k_{t+1}) = 2^{T-t+2} - 2$, we obtain the result of the lemma.

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

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

$$d_t(\tilde{x}_t, \psi_t(\tilde{x}_t), \pi_t(\tilde{x}_t)) \leq (\lambda_t + c_t(\bar{r} - r)) \cdot \left| n(\psi_t(\tilde{x}_t)) - N_i \right| .$$

(55)

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

$$d_t(\tilde{x}_t, \psi_t(\tilde{x}_t), \pi_t(\tilde{x}_t)) \leq c_t(\bar{r} - r) \cdot \max\{n(\psi_t(\tilde{x}_t)) - N_i, 0\} .$$

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

39
We consider two cases. Case (i): Suppose the Lagrangian policy \( \psi \) selects \( n(\psi_t(\vec{x}_t)) < N_t \) items. Those items selected by \( \psi \) with \( \hat{i}_{t,s}(x_{t,s}) > \lambda_t \) will be included in the top \( N_t \) items as ranked by the priority index and will thus also be selected by the heuristic \( \pi \). The tiebreaking rules ensure that any other items selected by \( \psi \) with \( \hat{i}_{t,s}(x_{t,s}) = \lambda_t \) will also be selected by \( \pi \). \( \pi \) may also select up to \( N_t - n(\psi_t(\vec{x}_t)) \) additional items with nonnegative priority indices that were not selected by \( \psi \). (We note for future reference that if \( \lambda_t = 0 \), then in this case \( \psi \) and \( \pi \) will select exactly the same items.)

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

The desired result (55) can now be established as follows:

\[
\begin{align*}
\theta_t(\vec{x}_t, \psi_t(\vec{x}_t), \pi_t(\vec{x}_t)) &= \lambda_t (N_t - n(\psi_t(\vec{x}_t))) + r_t(\vec{x}_t, \psi_t(\vec{x}_t)) - r_t(\vec{x}_t, \pi_t(\vec{x}_t)) \\
&\quad + \mathbb{E}[V^\pi_{t+1}(\vec{x}_t, \psi_t(\vec{x}_t)) - V^\pi_{t+1}(\vec{x}_t, \pi_t(\vec{x}_t))] \\
&\quad \leq \lambda_t |n(\psi_t(\vec{x}_t)) - N_t| + (\bar{r} - r) |n(\psi_t(\vec{x}_t)) - N_t| \\
&\quad + 2(\bar{r} - r)k_{t+1} |n(\psi_t(\vec{x}_t)) - N_t| \\
&= \left( \lambda_t + (\bar{r} - r)(1 + 2k_{t+1}) \right) |n(\psi_t(\vec{x}_t)) - N_t|.
\end{align*}
\]

The inequality above follows term by term. (a) \( \leq (a') \): Because \((N_t - n(\psi_t(\vec{x}_t))) \leq |n(\psi_t(\vec{x}_t)) - N_t|\), we have

\[
\lambda_t (N_t - n(\psi_t(\vec{x}_t))) \leq \lambda_t |n(\psi_t(\vec{x}_t)) - N_t|.
\]

(b) \( \leq (b') \): Because \( \psi \) and \( \pi \) will select no more than \( |n(\psi_t(\vec{x}_t)) - N_t| \) different items, we have

\[
r_t(\vec{x}_t, \psi_t(\vec{x}_t)) - r_t(\vec{x}_t, \pi_t(\vec{x}_t)) \leq (\bar{r} - r) |n(\psi_t(\vec{x}_t)) - N_t|.
\]

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

\[
\mathbb{E}[V^\pi_{t+1}(\vec{x}_t, \psi_t(\vec{x}_t))] - \mathbb{E}[V^\pi_{t+1}(\vec{x}_t, \pi_t(\vec{x}_t))] \leq 2(\bar{r} - r)k_{t+1} |n(\psi_t(\vec{x}_t)) - N_t|.
\]

where \( k_t \) is as defined in Lemma 2. The desired result then follows by taking \( c = (1 + 2k_{t+1}) \).

In the case where \( \lambda_t = 0 \), as discussed above, in Case (i) \( \psi \) and \( \pi \) will select the same items, so combining Cases (i) and (ii), \( \psi \) and \( \pi \) will select no more than \( \max\{n(\psi_t(\vec{x}_t)) - N_t, 0\} \) different items. The proof then proceeds as before, noting that (56) holds because \( \lambda_t = 0 \).

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

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

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

\[
\mathbb{E}[|\tilde{n}_t(\vec{\psi}) - N_t|] \leq \sqrt{N_t(1 - N_t/S)}.
\]

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

\[
\mathbb{E}[\max\{\tilde{n}_t(\vec{\psi}) - N_t, 0\}] \leq \sqrt{N_t(1 - N_t/S)},
\]

\[
40
\]
where \( \tilde{N}_t = \mathbb{E}[\tilde{n}_t(\tilde{\psi})] \leq N_t \).

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

\[
\text{Var}[\tilde{n}_t(\tilde{\psi})] = \sum_{s=1}^{S} p_{t,s}(1 - p_{t,s})
= \sum_{s=1}^{S} p_{t,s} - \sum_{s=1}^{S} p_{t,s}^2
= \mathbb{E}[\tilde{n}_t(\tilde{\psi})] - \sum_{s=1}^{S} p_{t,s}^2
\leq \mathbb{E}[\tilde{n}_t(\tilde{\psi})] - \mathbb{E}[\tilde{n}_t(\tilde{\psi})]^2/S
= \mathbb{E}[\tilde{n}_t(\tilde{\psi})](1 - \mathbb{E}[\tilde{n}_t(\tilde{\psi})]/S)
\]

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

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

\[
\mathbb{E}\left[\left|\tilde{n}_t(\tilde{\psi}) - N_t\right|^2\right] \leq \text{Var}\left[\tilde{n}_t(\tilde{\psi}) - N_t\right]
= \text{Var}\left[\tilde{n}_t(\tilde{\psi})\right]
\leq \mathbb{E}[\tilde{n}_t(\tilde{\psi})](1 - \mathbb{E}[\tilde{n}_t(\tilde{\psi})]/S)
= N_t(1 - \bar{N}_t/S)
\]

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

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

\[
\mathbb{E}\left[\max\{|\tilde{n}_t(\tilde{\psi}) - N_t|, 0\}\right]^2 \leq \mathbb{E}\left[\max\{|\tilde{n}_t(\tilde{\psi}) - \bar{N}_t|, 0\}\right]^2
\leq \mathbb{E}\left[\left|\tilde{n}_t(\tilde{\psi}) - \bar{N}_t\right|^2\right]
= \text{Var}\left[\tilde{n}_t(\tilde{\psi}) - \bar{N}_t\right]
\leq \mathbb{E}[\tilde{n}_t(\tilde{\psi})](1 - \mathbb{E}[\tilde{n}_t(\tilde{\psi})]/S)
= \bar{N}_t(1 - \bar{N}_t/S)
\]

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

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

\[
V_1^\pi(x) = L_1^X(x) - \sum_{t=1}^{T} \mathbb{E}\left[d_t(x_t, \tilde{\psi}, \tilde{\pi})\right]
\geq L_1^X(x) - \sum_{t=1}^{T} \left\{ (\lambda_t^* + c_t(\bar{r} - r)) \cdot \mathbb{E}\left[|n(\tilde{\psi}(x_t)) - \bar{N}_t|\right] \right\} \\
\quad \text{if } \lambda_t^* > 0
\left\{ c_t(\bar{r} - r) \cdot \mathbb{E}\left[\max\{n(\psi(x_t)) - N_t, 0\}\right] \right\} \\
\quad \text{if } \lambda_t^* = 0
\]
The growth assumption implies 
\[ L_1^N(x) - \sum_{t=1}^{T} (\lambda_t^* + c_t(\bar{r} - r)) \sqrt{\tilde{N}_t(1 - \tilde{N}_t/S)} \]
where \( \tilde{N}_t = N_t \) if \( \lambda_t^* > 0 \) and \( \tilde{N}_t = \mathbb{E}[\tilde{n}_t(\bar{\psi})] \leq N_t \) if \( \lambda_t^* = 0 \). When considering expectations involving the mixed policies, we assume that the realizations of \( \bar{\psi} \) and \( \bar{r} \) are coordinated so the realized \( \bar{r} \) is the Lagrangian index policy with the realized \( \bar{\psi} \) as tiebreaker: this is necessary when applying Lemma 3 in the second line above.

Finally, we note that an optimal \( \lambda_t^* \) must satisfy \( \lambda_t^* \leq (T - t)(\bar{r} - r) \). If this were not the case, it would not be optimal to select any item in any state at time \( t \) under the optimal Lagrangian policy \( \bar{\psi} \), contradicting the fact that \( N_t \) items are selected on average using \( \bar{\psi} \) if \( \lambda_t > 0 \) (see Proposition 4(iii)). Thus we take \( \beta_t = T - t + c_t \) to obtain the result of the proposition.

The inequality (29) then follows from the fact that \( \sqrt{N_t(1 - N_t/S)} \leq \sqrt{\tilde{N}_t} \leq \sqrt{N} \).

\[ \square \]

C.2. Proof of Corollary 1

Proof of Corollary 1. Theorem 1 implies
\[ 0 \leq \frac{L_2^N(x; S) - V_2^\psi(x; S)}{V_1^\psi(x; S)} \leq (\bar{r} - r) \sum_{t=1}^{T} \beta_t \frac{\sqrt{N_t(S)(1 - N_t(S)/S)}}{V_1^\psi(x; S)} \leq (\bar{r} - r) \sum_{t=1}^{T} \beta_t \frac{\sqrt{N(S)}}{V_1^\psi(x; S)}. \]
The growth assumption implies \( \lim_{S \to \infty} \sqrt{N(S)/V_1^\psi(x; S)} = 0 \), which gives the desired result (32).

\[ \square \]

C.3. Example Showing the Lagrangian Performance Gap of \( \sqrt{N} \) is Tight

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

(i) \( S/2 \) items are a priori identical and yield rewards \( r_{1,s}(x_{s}^0, 1) = 1 \) in their initial state \( x_{s}^0 \). If selected in period one, in period two these items transition to state \( \bar{x} \) with probability 1/2 and to state \( \bar{z} \) with probability 1/2, with \( r_{2,s}(\bar{x}, 1) = 2 \) and \( r_{2,s}(\bar{z}, 1) = 0 \). If not selected, these items do not change state. Let \( S_1 \) denote this set of items.

(ii) \( S/4 \) items are identical and yield deterministic rewards \( r_{1,s}(x_{s}^0, 1) = 1/2 \) if selected in either period, and never transition from their initial state \( x_{s}^0 \), whether selected or not. Let \( S_2 \) denote this set of items.

(iii) The remaining \( S/4 \) items are identical and yield deterministic rewards \( r_{1,s}(x_{s}^0, 1) = 1/4 \) if selected in either period, and never transition from their initial state \( x_{s}^0 \), whether selected or not. Let \( S_3 \) denote this set of items.

All items yield zero reward when not selected.

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

(i) For \( s \in S_1 \): In period two, \( V_2^X(\bar{x}) = 7/4 \), \( V_2^X(\bar{z}) = 0 \), \( \mathbb{E}[V_2^X(\bar{x}_{1,s}(x_{s}^0, 1))] = 7/8 \), and it is strictly optimal to select in state \( \bar{x} \) and not select in state \( \bar{z} \) in period one. In period one, it is strictly optimal to select: the value of selecting is \( r_{1,s}(x_{s}^0, 1) - \lambda_1^* + \mathbb{E}[V_2^X(\bar{x}_{1,s}(x_{s}^0, 1))] = 11/8 \) and the value of not selecting is \( 0 + V_2^X(x_{s}^0) = 1 - \lambda_2^* = 3/4 \). Thus, for \( s \in S_1 \), there is a single optimal policy \( \psi_s \) for \( s \in S_1 \).

(ii) For \( s \in S_2 \): In period two, \( V_2^X(x_{s}^0) = 1/4 \) and it is strictly optimal to select. In period one, selecting or not selecting are both optimal: the value for selecting is \( r_{1,s}(x_{s}^0, 1) - \lambda_1^* + V_2^X(x_{s}^0) = 1/4 \) and the value for not selecting is \( V_2^X(x_{s}^0) = 1/4 \). For all \( s \in S_2 \), we take \( \psi_s \) to be the optimal policy that does not select these items in period one.
(iii) For \( s \in S_3 \): In period two, \( V_{2, s}^\pi(x_s^0) = 0 \) and selecting and not selecting are both optimal. In period one, not selecting is strictly optimal. For all \( s \in S_4 \), we take \( \psi_s \) to be the optimal policy that does not select these items in period two.

With these optimal policies, we select exactly \( N = S/2 \) items (all items in \( S_1 \)) in period one. In period two, we select those items in \( S_1 \) that transition to \( \pi \) (expected number equal to \( S/4 \)) and select all \( S/4 \) items in \( S_2 \), for a total of \( S/2 \) items in expectation. By Proposition 4(iii), this implies that \( X = (1/2, 1/4) \) is optimal.

**Total Reward with the Optimal Policy for the Lagrangian Relaxation.** In the Lagrangian relaxation, it is optimal to select all items in \( S_1 \) in the first period. We let \( Y \) denote the random variable corresponding to the number of items in \( S_1 \) that transition to \( \pi \) in period two. The distribution of \( Y \) is binomial with \( S/2 \) trials and probability \( 1/2 \).

The first period rewards are simply \( S/2 \), as exactly \( N = S/2 \) items with reward 1 are selected. In the second period, all \( Y \) items in \( S_1 \) are selected and yield reward 2, and all \( S/4 \) items in \( S_2 \), each yielding reward 1/2, are selected. The Lagrangian penalty in period two is \( \lambda Y (S/2 - Y - S/4) = S/16 - Y/4 \). Putting this together, the total reward in the Lagrangian relaxation given \( Y \) is \((7/4)Y + (11/16)S\).

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

\[
\begin{align*}
\text{for } s \in S_1 : \ & i_1,s(x_s^0) = (r_1,s(x_s^0,1) + \mathbb{E}[V_{2,s}^X(\chi_{1,s}(x_s^0,1))]) - (r_1,s(x_s^0,0) + V_{2,s}^X(x_s^0)) = (1 + 7/8) - (0 + 3/4) = 9/8, \\
\text{for } s \in S_2 : \ & i_1,s(x_s^0) = (r_1,s(x_s^0,1) + V_{2,s}^X(x_s^0)) - (r_1,s(x_s^0,0) + V_{2,s}^X(x_s^0)) = (1/2 + 1/4) - (0 + 1/4) = 1/2, \\
\text{for } s \in S_3 : \ & i_1,s(x_s^0) = (r_1,s(x_s^0,1) + V_{2,s}^X(x_s^0)) - (r_1,s(x_s^0,0) + V_{2,s}^X(x_s^0)) = (1/4 + 0) - (0 + 0) = 1/4,
\end{align*}
\]

and thus all items in \( S_1 \) are selected in the first period by the optimal Lagrangian index policy.

In the second period, the selection indices in the optimal Lagrangian index policy equal the item’s rewards in their current state. Thus, in period two, the optimal Lagrangian index policy selects all \( Y \) items in \( S_1 \) that yield reward 2, possibly in addition to some other items, which differ in two cases:

(a) If \( Y < S/4 \), then all \( S/4 \) items in \( S_2 \) are also selected, each yielding reward 1/2, as well as \( S/2 - (Y + S/4) = S/4 - Y \) items in \( S_1 \) are selected, each yielding reward 1/4. The total reward (including period one) in this case is \((7/4)Y + (11/16)S\), equal to the Lagrangian relaxation value.

(b) If \( Y \geq S/4 \), then \( S/2 - Y \leq S/4 \) items from \( S_2 \) are also selected, yielding a total reward (including period one) of \((3/2)Y + (3/4)S\).

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

\[
L_1^X(x) - V_1^\pi(x) = \mathbb{E} \left[ \mathbb{I}\{Y \geq S/4\} \left( \frac{7}{4} Y + \frac{11}{16} S - \frac{3}{2} Y - \frac{3}{4} S \right) \right]
\]

\[
= \mathbb{E} \left[ \mathbb{I}\{Y \geq S/4\} \left( \frac{Y}{4} - \frac{S}{16} \right) \right]
\]

\[
= \frac{1}{4} \mathbb{E} \left[ \mathbb{I}\{Y \geq S/4\} \left( Y - \frac{S}{4} \right) \right]
\]

\[
= \frac{1}{4} \mathbb{E} \left[ \left( Y - \frac{S}{4} \right)^+ \right].
\]

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

\[
L_1^X(x; S) - V_1^\pi(x; S) = \frac{1}{4} \mathbb{E} \left[ (Y - S/4)^+ \right] = \frac{1}{8} \mathbb{E} \left[ |Y - S/4| \right] = \frac{\sqrt{2S}}{8\sqrt{8\pi}} = \frac{\sqrt{N}}{8\sqrt{2\pi}}.
\]
References


