Electronic Companion for

Index Policies and Performance Bounds for Dynamic Selection Problems

by David B. Brown and James E. Smith

EC1. Selected Proofs and Details for §3

EC1.1. Proofs for §3: Lagrangian Relaxations

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

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

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

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

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

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

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

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

EC1.2. Constructing a Markov Random Policy

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

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

x_s and choosing action u_s with the simple mixed policy $\widetilde{\psi}$. This is given by:

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

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

Thus the probability of being in state x_s with this mixed policy is $\nu_{t,s}(x_s, 0; \widetilde{\psi}) + \nu_{t,s}(x_s, 1; \widetilde{\psi})$. If $\widetilde{\psi}$ is an optimal mixed policy for the Lagrangian dual problem, $\nu_{t,s}(x_s, u_s; \widetilde{\psi})$ is an optimal solution for the LP (EC-5).

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

$$\frac{\nu_{t,s}(x_s, 1; \widetilde{\psi})}{\nu_{t,s}(x_s, 0; \widetilde{\psi}) + \nu_{t,s}(x_s, 1; \widetilde{\psi})} \tag{EC-2}$$

By construction, this will generate the same state-action probabilities as $\widetilde{\psi}$, will select the same number of items on average in each period as $\widetilde{\psi}$, and will have the same expected total reward as $\widetilde{\psi}$. Note that these selection probabilities will be undefined when the probability of being in state x_s in period t (in the denominator of (EC-2)) is zero. These undefined selection probabilities are irrelevant for evaluating policies for the Lagrangian relaxation, but may be relevant when we use the policy for the Lagrangian relaxation as a tiebreaker for the optimal Lagrangian index policy (as discussed in §4.4P. In our numerical examples, we take these undefined probabilities to be 0.5.

EC1.3. Linear Programming Formulation of the Lagrangian Dual Problem

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

$$\begin{aligned} \min_{V^{\boldsymbol{\lambda}}_{t,s}(x_s)} \quad & V^{\boldsymbol{\lambda}}_{s,1}(x^0_s) \\ \text{s.t.} \quad & V^{\boldsymbol{\lambda}}_{t,s}(x_s) \geq \ r_{t,s}(x_s, u_s) - \lambda_t u_s + \sum_{\widetilde{\chi}_{t,s}} p_t(\widetilde{\chi}_{t,s} \,|\, x_s, u_s) \ V^{\boldsymbol{\lambda}}_{t+1,s}(\widetilde{\chi}_{t,s}) \quad \forall \ t, x_s, u_s \ , \end{aligned} \tag{EC-3}$$

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

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

$$\begin{aligned} \min_{\boldsymbol{\lambda}, \ V^{\boldsymbol{\lambda}}_{t,s}(x_s)} \quad & \sum_{t=1}^{T} \lambda_t N_t + \sum_{s=1}^{S} V^{\boldsymbol{\lambda}}_{1,s}(x^0_s) \\ \text{s.t.} \quad & V^{\boldsymbol{\lambda}}_{t,s}(x_s) \geq \ r_{t,s}(x_s, u_s) - \lambda_t u_s + \sum_{\widetilde{\chi}_{t,s}} p_t(\widetilde{\chi}_{t,s} \,|\, x_s, u_s) \ V^{\boldsymbol{\lambda}}_{t+1,s}(\widetilde{\chi}_{t,s}) \qquad \forall \ s, t, x_s, u_s \ , \\ & \lambda_t \geq \ 0 \qquad\qquad \forall \ t \ . \end{aligned} \tag{EC-4}$$

If we let $|X_s|$ be the size of the state space for item s, this LP has $T \times \left(1 + \sum_{s=1}^{S} |X_s|\right)$ decision variables and $2 \times T \times \sum_{s=1}^{S} |X_s|$ constraints. (If some or all of the items are identical, this LP can be simplified.) Though this LP formulation delivers optimal values for $\boldsymbol{\lambda}$ and the initial values $V^{\boldsymbol{\lambda}}_{1,s}(x^0_s)$ for the item-specific

DPs, it does not provide a full optimal value function for all periods and states because values for states that are not visited under the optimal policy do not affect the objective function. The Lagrangian index policy defined in §4 requires a full value function. To calculate these value functions using this LP formulation, we need to fix $\boldsymbol{\lambda}$ at the optimal value from (EC-4) and solve LPs like (EC-3) with an objective function that includes positive weights on the values $V^{\boldsymbol{\lambda}}_{t,s}(x_s)$ for all items, states, and periods.

Taking $\nu_{t,s}(x_s, u_s)$ to be the dual variables for the constraints in (EC-4), we can write the dual of (EC-4) as:

$$
\begin{aligned}
\max_{\nu_{t,s}(x_s,u_s)} \quad & \sum_t \sum_s \sum_{x_s} \sum_{u_s} r_{t,s}(x_s, u_s)\nu_{t,s}(x_s, u_s) & & \\
\text{s.t.} \quad & \sum_{u_s} \nu_{1,s}(x^0_s, u_s) = 1 & \forall\ s\ , & \qquad \text{(EC-5)} \\
& \sum_{u_s} \nu_{t,s}(\widetilde{\chi}_{t,s}, u_s) = \sum_{x_s} \sum_{u_s} p_t(\widetilde{\chi}_{t,s} \mid x_s, u_s)\ \nu_{t-1,s}(x_s, u_s) & \forall\ s, t > 1, \widetilde{\chi}_{t,s}\ , & \\
& \sum_s \sum_{x_s} \nu_{t,s}(x_s, 1) \le N_t & \forall\ t\ , & \\
& \nu_{t,s}(x_s, u_s) \ge 0 & \forall\ s, t, x_s, u_s\ . &
\end{aligned}
$$

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 (EC-5) is the expected total reward. The first two constraints are flow conservation conditions: the total flow in the initial state x^0_s for each item ($\sum_{u_s} \nu_{1,s}(x^0_s, u_s)$) is equal to 1 and the total flow into a later state $\widetilde{\chi}_{t,s}$ must have come from a transition from some previous state. The third constraint requires the linking constraint to hold "on average" and complementary slackness ensures that this linking constraint holds with equality in period t whenever $\lambda_t > 0$. This average linking constraint is thus equivalent to the necessary and sufficient conditions for optimality in the Lagrangian dual given in Proposition 4(iii). Complementary slackness also implies that if the optimal flow $\nu_{t,s}(x_s, u_s)$ is positive, the corresponding value function inequality in (EC-4) 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 (EC-5) can also be calculated from the policies $\psi_{s,i}$ and mixing weights $\gamma_{s,i}$ given by the cutting-plane method of Appendix A; see Appendix EC1.2.

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

In our numerical examples of §6.4, we consider a decomposed version of the fluid heuristic where we solve the Lagrangian dual problem (EC-5) in each period and select items to maximize the total flow,

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

where the $\nu_{t,s}(x_s, u_s)$ are the optimal flows for the given period and state given by the solution to (EC-5). Any ties are broken randomly. As noted after (EC-5), complementary slackness implies that if the optimal flow $\nu_{t,s}(x_s, u_s)$ in the LP is positive, the action u_s is optimal in state x_s in period t. The heuristic chooses items to maximize this flow.

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

EC2. Notes on Whittle Indices

EC2.1. Calculating Whittle Indices

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

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

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

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

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

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

$$\delta^\star = \min_{t,x_s}\left\{ \frac{\Delta^w_{t,s}(x_s) - w}{1 - \sigma^w_{t,s}(x_s)} : \psi_{t,s}(x_s; w) = 1 \right\}\ . \tag{EC-6}$$

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

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

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

The breakpoint calculation in (EC-6) can be understood as follows: for any states and periods satisfying $\psi_{t,s}(x_s; w) = 1$, selection is strictly optimal at the current w, and hence $\Delta^w_{t,s}(x_s) > w$ in such states. Since $\sigma^w_{t,s}(x_s)$ represents the partial derivative of $\Delta^w_{t,s}(x_s)$ with respect to w, we seek a value δ such that $w + \delta$ is a new Whittle index, i.e., δ satisfies

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

The ratio in (EC-6) represents the largest increase to w such that the policy ψ_s remains optimal. For times and states attaining this value in (EC-6), we are indifferent between selecting and not selecting the item at $w + \delta^\star$.

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

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

EC2.2. Whittle Indices for the Applicant Screening Example

Here we show that in the applicant screening example, the Whittle indices have a particularly simple form. We let $\mu(x_s)$ denote an applicant's mean quality in state x_s, which we assume to be positive. For example,

with a beta prior $\mu(x_s) = \alpha_s/(\alpha_s + \beta_s)$. The item-specific DP (6) with $\boldsymbol{\lambda} = w\mathbf{1}$ is given recursively as $V^{w\mathbf{1}}_{T,s}(x_s) = \max\{\mu(x_s) - w, 0\}$ and, for $t < T$,

$$V^{w\mathbf{1}}_{t,s}(x_s) = \max\{-w + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big],\ V^{w\mathbf{1}}_{t+1,s}(x_s)\}\ . \tag{EC-7}$$

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

$$\begin{aligned} -w + \mu(x_s) &= 0 && \text{for } t = T, \text{ and} \\ -w + \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big] &= V^{w\mathbf{1}}_{t+1,s}(x_s) && \text{for } t < T. \end{aligned} \tag{EC-8}$$

We show the following.

Proposition EC1. *In the applicant screening example, for all s, t, and x_s, the Whittle index is unique.*
(i) *In the final period ($t = T$), the Whittle index is $\mu(x_s)$.*
(ii) *In screening periods ($t < T$), the Whittle index is zero.*

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

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

(ii) We first show that $w = 0$ is a Whittle index for $t < T$. In this case, $V^{w\mathbf{1}}_{T,s}(x_s) = \mu(x_s)$, since $\mu(x_s) > 0$. By induction and using the fact that the posterior mean is equal to the prior mean, for $t < T$, we have $V^{w\mathbf{1}}_{t,s}(x_s) = \mathbb{E}\big[\, V^{w\mathbf{1}}_{t+1,s}(\widetilde{\chi}_{t,s}(x_s, 1))\,\big] = \mathbb{E}[\,\mu(\widetilde{\chi}_{t,s}(x_s, 1))\,] = \mu(x_s)$. Thus (EC-8) holds for $w = 0$.

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

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

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

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

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

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

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

and screening yields:

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

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

EC3. Proofs for §5: Analysis of the Optimal Lagrangian Index Policy

EC3.1. Proof of Proposition 5

The proof of Proposition 5 relies on three key steps which we state in Lemmas EC1, EC3, and EC4 below. Lemma EC2 supports Lemma EC3. In this discussion, we let $n(\boldsymbol{u}_t) = \sum_{s=1}^{S} u_{t,s}$ denote the number of items selected with action vector $\boldsymbol{u}_t$.

Lemma EC1. *For any $\boldsymbol{\lambda} \geq \mathbf{0}$ and initial state $\boldsymbol{x}$, let ψ be an optimal policy for the Lagrangian* (5)*, and let $\widetilde{\boldsymbol{x}}_t$ denote the state transition process generated by ψ. Then, for any policy π,*

$$L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) - V_1^{\pi}(\boldsymbol{x}) = \sum_{t=1}^{T} \mathbb{E}\big[\, d_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t), \pi_t(\widetilde{\boldsymbol{x}}_t))\,\big] \tag{EC-9}$$

where

$$\begin{aligned} d_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}, \boldsymbol{u}_t^{\pi}) = \lambda_t\big(N_t - n(\boldsymbol{u}_t^{\psi})\big) + r_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}) - r_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\pi}) \\ + \mathbb{E}\Big[\, V_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\psi}))\Big] - \mathbb{E}\big[\, V_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}_t, \boldsymbol{u}_t^{\pi}))\,\big]\ . \end{aligned} \tag{EC-10}$$

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

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

$$L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{t=1}^{T} \mathbb{E}\big[\, \lambda_t\big(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))\big) + r_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t))\,\big]\ . \tag{EC-11}$$

We also have

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

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

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

Lemma EC2. *Let π be an index policy and suppose states $\boldsymbol{x}'$ and $\boldsymbol{x}''$ differ for m or fewer items. Then, for any t, there exists a nonnegative constant k_t (that depends only on t and T) such that:*

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

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

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

$$\begin{aligned}\big|V_t^\pi(\boldsymbol{x}') - V_t^\pi(\boldsymbol{x}'')\big| &= \big|r_t(\boldsymbol{x}', \pi(\boldsymbol{x}')) - r_t(\boldsymbol{x}'', \pi(\boldsymbol{x}'')) + \mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}', \pi(\boldsymbol{x}')))\,\big] - \mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}'', \pi(\boldsymbol{x}')))\,\big]\big| \\ &\le \big|r_t(\boldsymbol{x}', \pi(\boldsymbol{x}')) - r_t(\boldsymbol{x}'', \pi(\boldsymbol{x}''))\big| + \big|\,\mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}', \pi(\boldsymbol{x}')))\,\big] - \mathbb{E}\big[\, V_{t+1}^\pi(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}'', \pi(\boldsymbol{x}')))\,\big]\big| \\ &\le 2(\bar{r} - \underline{r})m \;+\; 2k_{t+1}(\bar{r} - \underline{r})m \end{aligned} \tag{EC-12}$$

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

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

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

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

completing the proof of the inequality (EC-12). 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 EC1 in the case where the policy π is a Lagrangian index policy with a tiebreaker that is an optimal policy ψ for the Lagrangian for any $\boldsymbol{\lambda}$. The key observation in the proof is to note that though ψ and π may select different numbers of items in a given state, the choices will differ for at most $\big|n(\psi_t(\boldsymbol{x}_t)) - N_t\big|$ items.

Lemma EC3. *For any $\boldsymbol{\lambda} \ge \mathbf{0}$ and initial state $\boldsymbol{x}$, let ψ be an optimal policy for the Lagrangian* (5)*, and let π be the Lagrangian index policy for $\boldsymbol{\lambda}$ with ψ as a tiebreaker. For each t, there exists a nonnegative constant c_t (depending only on t and T), such that for all $\widetilde{\boldsymbol{x}}_t$ that may be realized when following policy ψ,*

$$d_t(\widetilde{\boldsymbol{x}}_t, \psi_t(\widetilde{\boldsymbol{x}}_t), \pi_t(\widetilde{\boldsymbol{x}}_t)) \le \lambda_t\big(N_t - n(\psi_t(\widetilde{\boldsymbol{x}}_t))\big) + c_t(\bar{r} - \underline{r})\big|n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\big|\ . \tag{EC-13}$$

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

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

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

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

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

The desired result (EC-13) can now be established as follows:

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

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

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

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

- $(c) \le (c')$: Because ψ and π will select no more than $\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$ different items and state transitions are independent across items, the random continuation states $\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}', \pi(\widetilde{\boldsymbol{x}}'))$ and $\widetilde{\boldsymbol{\chi}}_t(\widetilde{\boldsymbol{x}}'', \pi(\widetilde{\boldsymbol{x}}''))$ will differ for at most $\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$ items. Lemma EC2 then implies

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

  where k_t is as defined in Lemma EC2.

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

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

The final lemma provides a bound on the expectations of the $\big|\, n(\psi_t(\widetilde{\boldsymbol{x}}_t)) - N_t\,\big|$ terms appearing in Lemma EC3 by calculating the variance of these quantities.

Lemma EC4. *Let $\boldsymbol{\lambda}^*$ denote an optimal solution for the Lagrangian dual problem (7) with initial state $\boldsymbol{x}$ and let $\widetilde{\psi}$ denote an optimal mixed policy. Let $\tilde{n}_t(\widetilde{\psi}) = n(\widetilde{\psi}_t(\widetilde{\boldsymbol{x}}_t(\boldsymbol{x}, \widetilde{\psi})))$.*

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

$$\mathbb{E}\big[\, \big|\, \tilde{n}_t(\widetilde{\psi}) - N_t\,\big|\,\big] \le \sqrt{N_t(1 - N_t/S)}\;. \tag{EC-14}$$

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

$$\mathbb{E}[\max\{\tilde{n}_t(\widetilde{\psi}) - N_t, 0\}\,] \le \sqrt{\bar{N}_t(1 - \bar{N}_t/S)}\;, \tag{EC-15}$$

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

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

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

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

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

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

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

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

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

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

□

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

Proof of Proposition 5. Using the notation of Lemmas EC1, EC3, and EC4 and applying these results in that order, we have:

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

where $\bar{N}_t = N_t$ if $\lambda_t^* > 0$ and $\bar{N}_t = \mathbb{E}[\,\tilde{n}_t(\widetilde{\psi})\,] \le N_t$ if $\lambda_t^* = 0$. In the final step above, we also use the fact that $\mathbb{E}[\,N_t - n(\widetilde{\psi}(\widetilde{\boldsymbol{x}}_t))\,] = 0$ when $\lambda_t^* > 0$; see Proposition 4(iii). When considering expectations involving the mixed policies, we assume that the realizations of $\widetilde{\psi}$ and $\widetilde{\pi}$ are coordinated so the realized π is the Lagrangian index policy with the realized ψ as tiebreaker: this is necessary when applying Lemma EC3 in the second line above. Taking $\beta_t = c_t = 2^{T-t+1} - 1$, we obtain the result of the proposition.

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

Proof of Corollary 1. Theorem 1 implies

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

The growth assumption implies $\lim_{S\to\infty}\sqrt{N(S)}/V_1^*(\boldsymbol{x};S)=0$, which gives the desired result (21). □

EC3.2. Example Showing the Lagrangian Performance Gap of $\sqrt{N}$ is Tight

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

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

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

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

All items yield zero reward when not selected.

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

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

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

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

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

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

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

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

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

$$s \in \mathcal{S}_2 : i_{1,s}(x_s^0) = (r_{1,s}(x_s^0, 1) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) - (r_{1,s}(x_s^0, 0) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) = (1/2 + 1/4) - (0 + 1/4) = 1/2,$$

$$s \in \mathcal{S}_3 : i_{1,s}(x_s^0) = (r_{1,s}(x_s^0, 1) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) - (r_{1,s}(x_s^0, 0) + V_{2,s}^{\boldsymbol{\lambda}^*}(x_s^0)) = (1/4 + 0) - (0 + 0) = 1/4,$$

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

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

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

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

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

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

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

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

EC4. Information Relaxation Bounds

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

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

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

EC4.1. Information Relaxation Bounds

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

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

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

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

If we let $\Pi_{\mathbb{G}}$ denote a larger set of policies ($\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$) that can use additional information,[EC1] we can solve a relaxed version of the DP to obtain an upper bound on the primal DP:

$$V_1^*(\boldsymbol{x}) = \max_{\pi \in \Pi_{\mathbb{F}}} \mathbb{E}[\, r(\pi)\,] \le \max_{\pi \in \Pi_{\mathbb{G}}} \mathbb{E}[\, r(\pi)\,]\ . \tag{EC-16}$$

Unfortunately, the bounds given by (EC-16) will be weak if the extra information provided in the relaxation is valuable. To counter this, we incorporate a penalty that "punishes" the DM for using information that would not actually be available when making decisions. The penalty $z(\pi)$ is a policy-dependent random variable, like the rewards, i.e., $z(\pi) = \sum_t z_t(\widetilde{\boldsymbol{x}}_t(\pi), \pi_t(\widetilde{\boldsymbol{x}}_t(\pi)))$ for some set of period-t penalty terms $z_t(x_t, u_t)$. A penalty $z(\pi)$ is *dual feasible* if $\mathbb{E}[\, z(\pi)\,] \le 0$ for all $\pi \in \Pi_{\mathbb{F}}$; that is, if the expected penalty is nonpositive for all nonanticipative policies.

[EC1] To formalize the definitions of these sets of policies, a policy can be defined as a mapping from the underlying outcome space to selection decisions $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ for each product and each period (with $\boldsymbol{u}_t \in \mathcal{U}_t$). Policies in the DP (2) that make selections as a function of the current state of the system can be viewed as imposing measurability restrictions on this more general set of policies. The relaxed model imposes a weaker set of measurability restrictions. See BSS (2010) for more discussion.

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

Proposition EC2 (Weak duality)**.** *Suppose $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. If policy π is nonanticipative (i.e., $\pi \in \Pi_{\mathbb{F}}$) and penalty z is dual feasible then*

$$\mathbb{E}[\, r(\pi)\,] \;\leq\; \max_{\pi' \in \Pi_{\mathbb{G}}} \mathbb{E}[\, r(\pi') - z(\pi')\,] \;. \tag{EC-17}$$

Proof. We have:

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

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

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

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

EC4.2. 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 (EC-17) easy to compute and lead to reasonably tight bounds. For specificity, we will focus our discussion on the dynamic assortment example, though the ideas also apply in the applicant screening example and other dynamic selection problems. In the dynamic assortment example, the underlying uncertainties are the unknown (Poisson) demand rates for each product and the demand realizations for each item, in each period. In the original model, the demands are revealed for products when (and if) the products are selected; the demand rates are never revealed. We can consider a number of different relaxations, including:

(i) *Known rates*: The DM knows the demand rates for all products in advance, but demands are revealed sequentially only when the products are selected, as in the original model.
(ii) *Known demands*: The DM knows all demands for all products in all periods, in advance before making any selection decisions (i.e., the DM knows what demand would be if a product were to be selected); demands rates are never revealed.
(iii) *Perfect information*: The DM knows both demands and rates in advance.
(iv) *Uncensored demand*: Demands for all products are revealed sequentially (regardless of whether they are selected or not); demand rates are never revealed.

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

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

(i) Drawing a demand rate γ_s for product s from the appropriate gamma distribution and then drawing demands for this product from a Poisson distribution with this rate. Let $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_T)$ where $\boldsymbol{d}_t = (d_{t,1}, \ldots, d_{t,S})$ denotes the randomly generated vector of product demands in period t.
(ii) Solving a deterministic *inner DP* (to be described shortly) to find the optimal value $\hat{V}_1(\boldsymbol{x}_1; \boldsymbol{d})$ given these demand realizations, incorporating the Lagrangian penalty.

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

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

$$\hat{V}_t(\boldsymbol{x};\boldsymbol{d}) = \max_{\boldsymbol{u}\in\mathcal{U}_t} \left\{ r_t(\boldsymbol{x},\boldsymbol{u}) - z_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t) + \hat{V}_{t+1}(\boldsymbol{\chi}_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t);\boldsymbol{d}) \right\} \tag{EC-18}$$

where

$$z_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t) = L^{\boldsymbol{\lambda}}_{t+1}(\boldsymbol{\chi}_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t)) - \mathbb{E}\big[\, L^{\boldsymbol{\lambda}}_{t+1}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x},\boldsymbol{u}))\,\big]\,. \tag{EC-19}$$

Here the last term in (EC-18) and the first term in (EC-19) involve deterministic state transitions because the DM knows the demands: $\boldsymbol{\chi}_t(\boldsymbol{x},\boldsymbol{u};\boldsymbol{d}_t) = (\chi_{t,1}(x_1,u_1;d_{t,1}),\ldots,\chi_{t,S}(x_S,u_S;d_{t,S}))$ represents the state transitions with the given product demands for period t. The expectation in (EC-19) is calculated using the same state-dependent negative-binomial distributions used in the original DP.

Using the law of iterated expectations, we know that $\mathbb{E}[\, z_t(\widetilde{\boldsymbol{x}}_t(\pi),\pi_t(\widetilde{\boldsymbol{x}}_t(\pi)))\,] = 0$ for any nonanticipative policy π. Thus the penalty $z(\pi) = \sum_t z_t(\widetilde{\boldsymbol{x}}_t(\pi),\pi_t(\widetilde{\boldsymbol{x}}_t(\pi)))$ is dual feasible and the known demands bound provides a performance bound, as in Proposition EC2. This is an example of the general method for creating "good" dual feasible penalties described in BSS (2010). As discussed there, if we replace the Lagrangian $L^{\boldsymbol{\lambda}}_{t+1}$ in (EC-19) with the optimal value function V^*_{t+1}, the information relaxation bound will be exactly equal to the optimal value. With this ideal penalty, the DM is exactly punished for using extra information: the benefit gained is exactly canceled by the penalty. With a penalty based on an approximate value function (such as the Lagrangian), the penalty approximately cancels this benefit. In general, to obtain good bounds, we want to choose generating functions that approximate the optimal value function well.

We now consider the DP (EC-18) in more detail. First, note that that the penalty terms involving the Lagrangian $L^{\boldsymbol{\lambda}}_{t+1}$ decompose into the sum of item-specific values, as in (5). However, the inner DP (EC-18) does not decompose into item-specific subproblems because the constraint on the total number of products selected ($\boldsymbol{u}\in\mathcal{U}_t$ where $\mathcal{U}_t$ is defined in (1)) links the decisions across items, as it did in the original DP (2). Thus, the inner DP – though deterministic – is still difficult to solve in problems with many items.

To decouple the inner DP (EC-18), we relax the linking constraint in the same way that we relaxed the original DP (2). Consider Lagrange multipliers $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_T) \geq \mathbf{0}$ and let $\hat{L}^{\boldsymbol{\mu}}_{T+1}(\boldsymbol{x};\boldsymbol{d}) = 0$. The period-$t$ *inner Lagrangian* with demand realization $\boldsymbol{d}$ is then given recursively as

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

This can be decomposed into item-specific DPs as

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

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

$$\hat{V}^{\boldsymbol{\mu}}_{t,s}(x_s;\boldsymbol{d}_s) = \max\Bigg\{ r_{t,s}(x_s,1) - \mu_t - V^{\boldsymbol{\lambda}}_{s,t+1}(\chi_{t,s}(x_s,1,d_{t,s})) + \mathbb{E}\big[\, V^{\boldsymbol{\lambda}}_{s,t+1}(\widetilde{\chi}_{t,s}(x_s,1))\,\big] + \hat{V}^{\boldsymbol{\mu}}_{t+1,s}(\chi_{t,s}(x_s,1,d_{t,s})),$$
$$r_{t,s}(x_s,0) + \hat{V}^{\boldsymbol{\mu}}_{t+1,s}(\chi_{t,s}(x_s,0,d_{t,s})) \Bigg\}. \tag{EC-20}$$

where $V^{\boldsymbol{\lambda}}_{t,s}$ is the value-function for the item-specific DP (6). Note that in the dynamic assortment model, the penalty term (EC-19) is zero if a product is not selected because its state does not change.

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

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

$$\min_{\boldsymbol{\mu} \geq \mathbf{0}} \ \hat{L}_1^{\boldsymbol{\mu}}(\boldsymbol{x}; \boldsymbol{d}) \ , \tag{EC-21}$$

for an optimal $\boldsymbol{\mu}^*(\boldsymbol{x}, \boldsymbol{d})$. This is a convex optimization problem and can be solved using the cutting-plane method discussed in §A. Moreover, if we take the inner Lagrange multipliers $\boldsymbol{\mu}$ to be equal to the "outer" Lagrange multipliers $\boldsymbol{\lambda}$ used to define the penalty, we can use an induction argument to show that $\hat{L}_t^{\boldsymbol{\lambda}}(\boldsymbol{x}; \boldsymbol{d}) = L_t^{\boldsymbol{\lambda}}(\boldsymbol{x})$ for all t and $\boldsymbol{d}$.[EC2] Thus, since $\boldsymbol{\lambda}$ is feasible but not necessarily optimal for the inner Lagrangian dual problem (EC-21), we have

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

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

We can also relate these bounds to the performance of a heuristic policy π in the same demand scenario. We focus on deterministic Markovian heuristic policies where the period-t selection decision π_t is chosen based on the current state $\boldsymbol{x}$. (When we are considering mixed policies, as in the optimal Lagrangian policy, let π be a particular realization of the mixed policy.) We assume that the actions selected by the heuristic are feasible, i.e., $\pi_t(\boldsymbol{x}) \in \mathcal{U}_t$. To facilitate comparison with those of the information relaxation, we will adjust the rewards using the penalty (EC-19) as a control variate. Let $\hat{V}_t^{\pi}(\boldsymbol{x}; \boldsymbol{d})$ denote the value generated when following policy π, starting in state $\boldsymbol{x}$, given demand realization $\boldsymbol{d}$, adjusted by the control variate. We can write this value recursively in a form parallel to (EC-18): let $\hat{V}_{T+1}^{\pi}(\boldsymbol{x}; \boldsymbol{d}) = 0$ and, for earlier t, we define

$$\hat{V}_t^{\pi}(\boldsymbol{x}; \boldsymbol{d}) = \left\{ r_t(\boldsymbol{x}, \pi_t(\boldsymbol{x})) - z_t(\boldsymbol{x}, \pi_t(\boldsymbol{x}); \boldsymbol{d}_t) + \hat{V}_{t+1}^{\pi}(\widetilde{\boldsymbol{\chi}}_t(\boldsymbol{x}, \pi_t(\boldsymbol{x}); \boldsymbol{d}_t); \boldsymbol{d}) \right\} . \tag{EC-22}$$

Here this form exactly mimics the DP recursion (EC-18), except the actions are chosen in accordance to the policy π rather than optimized. Thus we know that $\hat{V}_t^{\pi}(\boldsymbol{x}; \boldsymbol{d}) \leq \hat{V}_t(\boldsymbol{x}; \boldsymbol{d})$ for all t, $\boldsymbol{x}$, and $\boldsymbol{d}$. Moreover, because the penalty terms z_t have mean zero for all feasible policies, we know that the expected total reward when following policy π is $V_1^{\pi}(\boldsymbol{x}) = \mathbb{E}[\, \hat{V}_1^{\pi}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$, where the expectations are taken over the random demand scenarios. These control variates are helpful in reducing sampling error when estimating the expected values associated with a given policy and were used in the simulations of §6.2.

Combining these observations, we can say the following.

Theorem EC1 (Ordered bounds)**.** *Consider any feasible and nonanticipative policy π, Lagrange multipliers $\boldsymbol{\lambda} \geq \mathbf{0}$ and initial state $\boldsymbol{x}$.*

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

$$\hat{V}_1^{\pi}(\boldsymbol{x}; \boldsymbol{d}) \ \leq \ \hat{V}_1(\boldsymbol{x}; \boldsymbol{d}) \ \leq \ \hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{x}, \boldsymbol{d})}(\boldsymbol{x}; \boldsymbol{d}) \ \leq \ L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) \ . \tag{EC-23}$$

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

$$V_1^{\pi}(\boldsymbol{x}) \ = \ \mathbb{E}\Big[\hat{V}_1^{\pi}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \Big] \ \leq \ V_1^*(\boldsymbol{x}) \ \leq \ \mathbb{E}\Big[\hat{V}_1(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \Big] \ \leq \ \mathbb{E}\Big[\hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{x}, \tilde{\boldsymbol{d}})}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \Big] \ \leq \ L_1^{\boldsymbol{\lambda}}(\boldsymbol{x}) \ . \tag{EC-24}$$

Working from the left in (EC-24), we have the expected value with heuristic policy π ($V_1^{\pi}(\boldsymbol{x})$) is equal to the expected reward for this policy with the control variate included ($\mathbb{E}[\, \hat{V}_1^{\pi}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$). This value is less than or equal to the value with an optimal policy ($V_1^*(\boldsymbol{x})$), which is typically impossible to compute. This, in turn, is less than or equal to the known demands relaxation bound ($\mathbb{E}[\, \hat{V}_1(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$) which is also typically impossible to compute. However, the known demands bound is less than or equal to the Lagrangian relaxation of the known demands information relaxation bound with optimized Lagrange multipliers ($\mathbb{E}[\, \hat{L}_1^{\boldsymbol{\mu}^*(\boldsymbol{x}, \boldsymbol{d})}(\boldsymbol{x}; \tilde{\boldsymbol{d}}) \,]$), which is computable. Finally, all of these bounds are less than the ordinary Lagrangian bound ($L_1^{\boldsymbol{\lambda}}(\boldsymbol{x})$). The

[EC2] Note that the $V_{s,t+1}^{\boldsymbol{\lambda}}(\cdot)$ and $\hat{V}_{t+1,s}^{\boldsymbol{\mu}}(\cdot)$ terms in (EC-20) cancel if $\boldsymbol{\mu} = \boldsymbol{\lambda}$ and we have the induction hypothesis that $V_{s,t+1}^{\boldsymbol{\lambda}}(\cdot) = \hat{V}_{s,t+1}^{\boldsymbol{\lambda}}(\cdot)$. Then (EC-20) reduces to the the definition of $V_{s,t+1}^{\boldsymbol{\lambda}}(\cdot)$ in (6).

bounds in (EC-23) show that the demand-dependent terms in (EC-24) are ordered in every demand scenario $\boldsymbol{d}$ and less than or equal to the Lagrangian bound.

Though we have focused on the known demands relaxation in the dynamic assortment example, we can use the same approach and derive similar results with other relaxations and in other problems. In the applicant screening example, the information relaxation where all applicant signals are known in advance is exactly analogous to the known demands relaxation and we obtain the same results. If we consider the known rates relaxation instead of the known demands realization in the dynamic assortment example, we arrive at an inner DP similar to (EC-18), 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 EC1.

EC4.3. Numerical Examples

The (a) panels of Figures 2-5 show information relaxation bounds for the dynamic assortment and applicant screening examples using the known demands and known signals relaxations. These bounds were evaluated with S equal to 4, 8, 16, 32, and 64 in the same 1000 sample scenarios (i.e., same demand and signal sequences) that were used to evaluate the heuristics. In all cases, we use penalties based on an optimal solution $\boldsymbol{\lambda}^*$ for the outer Lagrangian dual problem (7) and impose the policy restrictions discussed at the end of §EC4.1. These figures also show 95% confidence intervals for the estimated bounds; these confidence intervals are quite narrow, particularly for larger values of S.

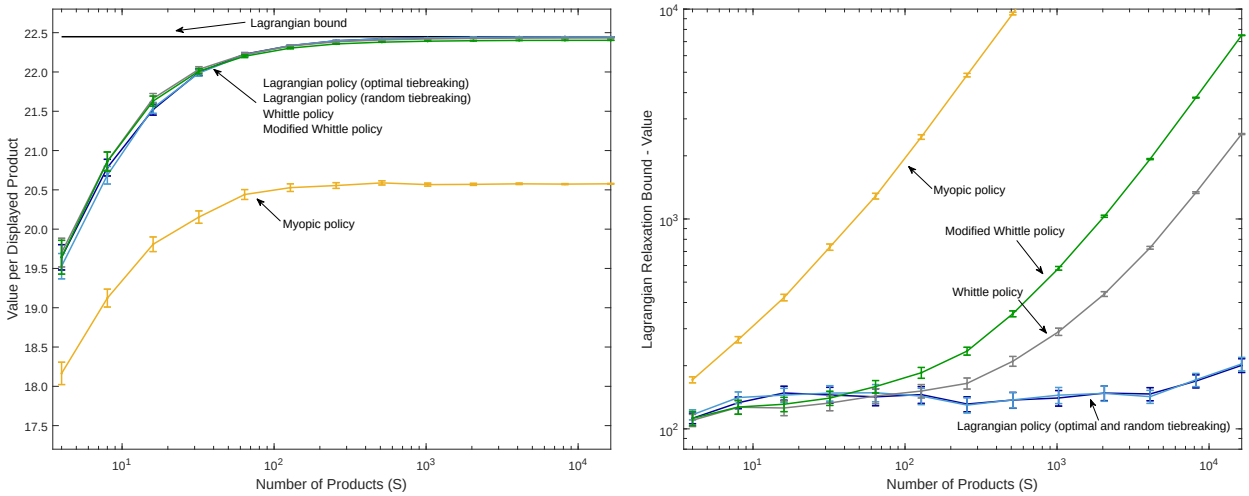
In the results, we see that the information relaxation bounds improve on the Lagrangian dual, particularly when S is small. The improvement is greatest in the dynamic assortment example with $T = 8$ and $S = 4$. In this case, the Lagrangian bound ensures that the Lagrangian index policy is within (approximately)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

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

Table EC-1: Run times (seconds) for information relaxation bound calculations

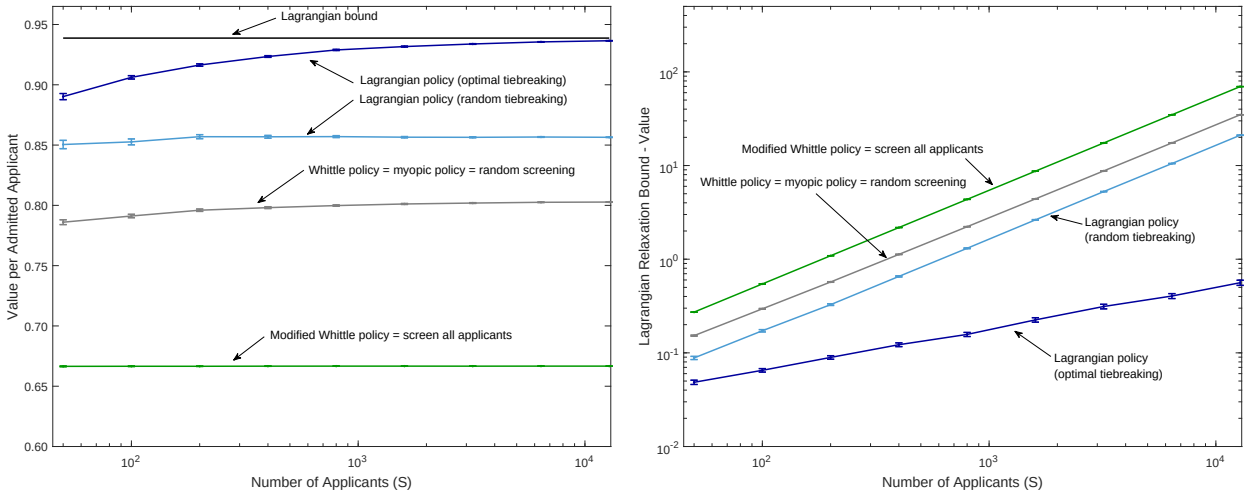
The run times are reported in Table EC-1. As discussed above, calculating these bounds requires solving the inner Lagrangian dual problems for each simulated demand (signal) sequence, for each product (applicant). This can be time consuming because the products (applicants) are not identical as each has its own demand (signal) sequence. We use the cutting-plane method in each case and start with $\boldsymbol{\mu} = \boldsymbol{\lambda}^*$, which yields the Lagrangian dual bound. If we cannot improve on this value, the cutting-plane algorithm typically stops after a few iterations. The run times grow roughly linearly with S, as one might expect, but not exactly because these no-improvement scenarios are more common with large S.

EC5. Numerical Results with Longer Horizons



(a) Relative performance (b) Performance gaps

Figure EC-1: Results for the dynamic assortment examples with horizon T=40



(a) Relative performance (b) Performance gaps

Figure EC-2: Results for the applicant screening examples with T=51 and Bernoulli signals ($n = 1$)

EC6. Additional Details on Weber and Weiss's Counterexample

Weber and Weiss (1990)'s example is useful for understanding dynamic selection problems with long or infinite horizons. Though Weber and Weiss considered a continuous-time, average-reward setting, their example can be adapted to discrete time with a long, but finite horizon. The example considers identical items, each having four states; the transition matrices and rewards are constant over time and are shown in Table EC-2. In each period, the DM must select exactly 83.5% of the items available. We assume that the system starts with 16%, 9%, 35% and 40% of the items in states one through four, respectively. We will consider a horizon T equal to 20,000 periods and focus on the dynamics of the Whittle and Lagrangian index policies in the deterministic mean field limit as the number of items S approaches infinity. The plots in Figure EC-3 show results for the first 3,000 of 20,000 periods; truncating these time series in this way makes the patterns easier to see.

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

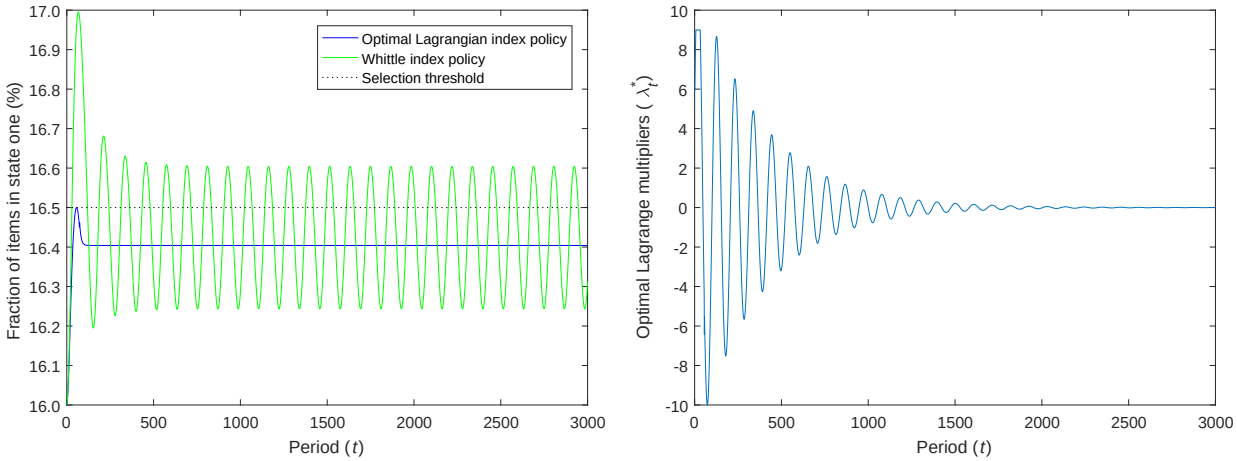
Table EC-2: Assumptions for Weber and Weiss (1990)'s example

First we consider the Whittle index policy. The Whittle indices may be calculated analytically and depend on an item's state (as usual) but not the period (a feature of this example). The Whittle indices are -10, 0, 9, and 10 for states one through four.[EC3] The ingenious feature of Weber and Weiss's example is that the fractions of items in each state cycles under the Whittle index policy. For example, Figure EC-3a shows the fraction of items in state one when following the Whittle index policy. The fraction of items in state one starts at 16%, rises to 17%, and ultimately settles into a cyclical pattern with fractions varying between 16.2% and 16.6%. The fractions in other states also vary cyclicly. In this example, the DM must select 83.5% of the items, so whenever the fraction in state one exceeds 16.5% (indicated with a dashed line in Figure EC-3a), the DM must select some items that are in state one. In periods where the Whittle index policy selects items in states two, three and four only, the policy generates a reward of 10. In periods where the policy selects some items in state one, the reward is less than 10, reflecting the zero reward when selecting items in state one.

Now consider the Lagrangian relaxation with a full set of T Lagrange multipliers. The optimal Lagrange multipliers $\boldsymbol{\lambda}^*$ (solving the dual problem (7)) are shown in Figure EC-3b and the state one fractions for the corresponding optimal Lagrangian index policy are shown in Figure EC-3a.[EC4] In Figure EC-3b we see that the optimal Lagrange multipliers λ_t^* cycle initially with dampening amplitude, approaching a steady state value of zero. The oscillations in the state fractions are less than those for the Whittle index policy and the fraction in state one remains at or below 16.5% in all periods, hitting 16.5% in period 56. How do the Lagrangian index and Whittle index policies differ? In the very early periods (1-6), the Lagrangian index policy prioritizes items in higher states, like the Whittle index policy. But in periods 8-31, the Lagrangian index policy prioritizes items in state two over state three, leaving some items in state three unselected. (Items in states two and three have the same index values in periods 7 and 32 and tiebreaking plays a role.) In most of the remaining periods, the Lagrangian index policy prioritizes items in the same way as the Whittle index. However, there is one later period (period 73) where the Lagrangian index policy is indifferent between selecting items in states one and two and the optimal Lagrangian index policy breaks ties so some items in state one are selected, earning zero reward. In this period, λ_t^* is -10 and the fraction of items in state one is strictly less than 16.5% so the DM is not forced to select items in state one in this period.

[EC3] For items in states one, three and four, the transition probabilities are identical in the active and passive states and the continuation values cancel in (15); it is easy to verify that (15) is satisfied with these index values. It is not hard to see with $\lambda_t = 0$ for all t, in every state the optimal value function is 10 times the number of periods remaining; (15) is thus satisfied in state two with $\lambda_t = 0$.

[EC4] This example took about 30 minutes to solve using an LP formulation of the Lagrangian dual (7).



(a) Fraction of items in state one

(b) Optimal Lagrange multipliers

Figure EC-3: Selected results for the Weber and Weiss example

These differences between the Whittle and Lagrangian index policies dampen the early oscillations seen in Figure EC-3a and guide the Lagrangian index policy to an equilibrium where the fraction of items in state one is approximately 16.4%; the fractions in other states also stabilize. In this equilibrium, the optimal Lagrange multipliers are zero and the Lagrangian and Whittle priority indices are equal: all items in states three and four are selected and approximately 99.0% of those in state two are selected. The rewards are 10 per period in this equilibrium. The optimal Lagrangian bound for the example, which is equal to the reward of the Lagrangian index policy, is slightly below 10 per period, reflecting the selection of some items in state one in period 73. The Whittle index policy performs worse because it regularly selects items in state one.

These numerical results depend on the initial fractions of items in each state, but the results are typical. For most initial conditions, the state fractions for the Whittle index policy settle into cycles as seen in Figure EC-3a where items in state one are routinely selected and the average reward is strictly less than 10. Similarly, for most initial conditions, the optimal Lagrange multipliers and state fractions for the Lagrangian index policy cycle initially, but approach an equilibrium distribution where the period reward is always 10. The exception to this typical behavior is that if we start the problem with initial conditions *exactly* equal to the equilibrium distribution, $\boldsymbol{\lambda} = 0$ is optimal for the Lagrangian dual problem and the Lagrangian and Whittle policies are equivalent and remain at this equilibrium distribution; however, this equilibrium is unstable and small deviations in initial conditions will lead the state distributions for Whittle index policies to oscillate.

EC7. Proofs for the Infinite-Horizon Extension

We begin our analysis of the infinite-horizon case by first considering how the result of Proposition 5 changes if we incorporate a discount factor $\delta \in [0,1)$ in the finite-horizon model of §2. We first briefly remark on how the results of the technical lemmas of §EC3 are affected by discounting and then consider Proposition 5.

Lemma EC1: Here the result is

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

where

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

The proof is analogous to the proof of Lemma EC1.

Lemma EC2: The result is exactly the same but discounting plays a role in the constants k_t. The inequality (EC-12) is now

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

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

Lemma EC3: The result is the same but now $c_t = 1 + \delta k_{t+1} = 1/2k_t = \frac{1}{2\delta-1}\big((2\delta)^{T-t+1} - 1\big)$.

Lemma EC4: The result and proof are unchanged.

Proposition 5: Using the analogs of Lemmas EC1, EC3, and EC4 in the same way as before, we have:

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

Taking $\beta_t(T)$ (as claimed in equation (23)) to be

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

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

$$\sum_{t=0}^{T} \beta_t(T) = \frac{1}{2\delta - 1}\left[2\delta^T(2^T - 1) - \frac{1 - \delta^T}{1 - \delta}\right]\ . \tag{EC-25}$$

In preparation for the proof of Proposition 6, we note that the result of Proposition 5 can be extended to consider partial sums of cash flows, as claimed in (25). Specifically, consider two time horizons T and T' where $T' \le T$. Now suppose we define the optimal Lagrangian policy $\widetilde{\psi}$ and the corresponding optimal Lagrangian index policy $\widetilde{\pi}$ in the usual way for the longer time horizon T, with $\boldsymbol{\lambda}^*$ denoting the optimal Lagrange multipliers. Now consider the sum of the discounted expected cash flows for $\widetilde{\psi}$ and and $\widetilde{\pi}$ over the shorter horizon T':

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

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

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

$$\hat{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T',T) - \hat{V}_1^{\widetilde{\psi}}(\boldsymbol{x};T',T) \;\le\; \sum_{t=1}^{T'} \beta_t(T')(\bar{r}-\underline{r})\sqrt{\bar{N}_t(1-\bar{N}_t/S)} \;\le\; \sum_{t=1}^{T'} \beta_t(T')(\bar{r}-\underline{r})\sqrt{N}\,. \qquad \text{(EC-26)}$$

where $\beta_t(T')$ is given by (23). This then implies the result of (25).

Proof of Proposition 6. Consider two time horizons T and T' where $T' \le T$ and the optimal Lagrangian policy $\widetilde{\psi}$ and the corresponding optimal Lagrangian index policy $\widetilde{\pi}$ are based on the longer time horizon T. From (25) and (EC-25), we have

$$\begin{aligned}\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T) - \bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T) &\le (\bar{r}-\underline{r})\left[\sum_{t=1}^{T'} \beta_t(T')\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right] \\ &= (\bar{r}-\underline{r})\left[\frac{1}{2\delta-1}\left(2\delta^{T'}(2^{T'}-1) - \frac{1-\delta^{T'}}{1-\delta}\right)\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right] \qquad \text{(EC-27)}\end{aligned}$$

Since we have assumed $\delta > 1/2$, $(2\delta-1) > 0$ and we can simplify the bracketed term in (EC-27) by dropping terms:

$$\bar{L}_1^{\boldsymbol{\lambda}^*}(\boldsymbol{x};T) - \bar{V}_1^{\widetilde{\pi}}(\boldsymbol{x};T) \le (\bar{r}-\underline{r})\left[\frac{2(2\delta)^{T'}}{2\delta-1}\sqrt{N} + \frac{\delta^{T'}}{1-\delta}S\right]. \qquad \text{(EC-28)}$$

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

$$\delta^{T'} = \delta^{\lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor} \le \frac{1}{\delta}\cdot \delta^{\log_2 \frac{S}{\sqrt{N}}} = \frac{1}{\delta}\cdot \left(2^{\log_2 \delta}\right)^{\log_2 \frac{S}{\sqrt{N}}} = \frac{1}{\delta}\cdot\left(\frac{\sqrt{N}}{S}\right)^{\log_2 \frac{1}{\delta}}, \qquad \text{(EC-29)}$$

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

$$(2\delta)^{T'} = (2\delta)^{\lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor} \le (2\delta)^{\log_2 \frac{S}{\sqrt{N}}} = \frac{S}{\sqrt{N}}\cdot\left(\frac{\sqrt{N}}{S}\right)^{\log_2 \frac{1}{\delta}}. \qquad \text{(EC-30)}$$

Using (EC-29) and (EC-30), the bracketed term in (EC-28) satisfies

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

and the result of Proposition 6 then follows with $\gamma = \frac{2}{2\delta-1} + \frac{1}{\delta(1-\delta)}$. This choice of $T' = \lfloor \log_2 \frac{S}{\sqrt{N}} \rfloor$ can be viewed as approximately minimizing the bound in (EC-28). Specifically, this selection of T' differs from the minimizing T' by rounding down and dropping a constant term that complicates the resulting expressions. □

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

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

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

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

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

Proof of Corollary 2. Using Proposition 6 and (27), we have

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

□

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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