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

OPERATIONS RESEARCH
Vol. 67, No. 2, March–April 2019, pp. 577–597
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Approximations to Stochastic Dynamic Programs via Information Relaxation Duality

Santiago R. Balseiro,[a] David B. Brown[b]

[a] Graduate School of Business, Columbia University, New York, New York 10027; [b] Fuqua School of Business, Duke University, Durham, North Carolina 27708

Contact: srb2155@columbia.edu, https://orcid.org/0000-0002-0012-3292 (SRB); dbbrown@duke.edu, https://orcid.org/0000-0002-5458-9098 (DBB)

Received: January 18, 2016
Revised: November 22, 2017
Accepted: May 21, 2018
Published Online in Articles in Advance: March 21, 2019

Subject Classifications: dynamic programming; analysis of algorithms: suboptimal algorithms
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2018.1782

Abstract. In the analysis of complex stochastic dynamic programs, we often seek strong theoretical guarantees on the suboptimality of heuristic policies. One technique for obtaining performance bounds is perfect information analysis: this approach provides bounds on the performance of an optimal policy by considering a decision maker who has access to the outcomes of all future uncertainties before making decisions, that is, fully relaxed nonanticipativity constraints. A limitation of this approach is that in many problems perfect information about uncertainties is quite valuable, and thus, the resulting bound is weak. In this paper, we use an information relaxation duality approach, which includes a penalty that punishes violations of the nonanticipativity constraints, to derive stronger analytical bounds on the suboptimality of heuristic policies in stochastic dynamic programs that are too difficult to solve. The general framework we develop ties the heuristic policy and the performance bound together explicitly through the use of an approximate value function: heuristic policies are greedy with respect to this approximation, and penalties are also generated in a specific way using this approximation. We apply this approach to three challenging problems: stochastic knapsack problems, stochastic scheduling on parallel machines, and sequential search problems. In each of these problems, we consider a greedy heuristic policy generated by an approximate value function and a corresponding penalized perfect information bound. We then characterize the gap between the performance of the policy and the information relaxation bound in each problem; the results imply asymptotic optimality of the heuristic policy for specific "large" regimes of interest.

Supplemental Material: The online appendices are available at https://doi.org/10.1287/opre.2018.1782.

Keywords: dynamic programming • greedy heuristic policies • information relaxation duality • asymptotic optimality • stochastic knapsack problems • stochastic scheduling • sequential search problems

1. Introduction

Dynamic programming (DP) is a powerful and widely used framework for studying sequential decision making in the face of uncertainty. Unfortunately, stochastic dynamic programs are often far too difficult to solve as the number of states that need to be considered typically grows exponentially with the problem size. As a result, we are often relegated to consider suboptimal, heuristic policies. In specific problem instances, a variety of methods, often employing Monte Carlo simulation, may be used to assess the quality of heuristic policies. More broadly, we may also seek strong analytical guarantees on the performance of heuristic policies. Ideally, this analysis will allow us to conclude that a heuristic policy provides a good approximation to the optimal policy on a broad set of instances or at least allow us to understand on what instances the heuristic policy is nearly optimal.

One technique for obtaining performance bounds is "perfect information analysis." This approach provides bounds by considering a decision maker (DM) who has advance access to the outcomes of all future uncertainties, that is, a problem with fully relaxed nonanticipativity constraints. We refer to this as the *perfect information problem*. For each sample path, the DM then solves a deterministic optimization problem, which is often much easier to analyze than the original, stochastic DP. The typical analysis compares the expected performance of the heuristic policy under consideration with the expected performance of the perfect information problem, or the *perfect information bound*. This approach has been used successfully to analyze heuristic policies in some applications; we survey a few in Section 1.1. In many problems, however, perfect information about uncertainties is quite valuable and leads to a weak bound as a result; this limits the applicability of the approach.

In this paper, we study the information relaxation duality approach developed in Brown et al. (2010) (BSS hereafter). The framework in Brown et al. (2010) involves "information relaxations" in which some (i.e., imperfect information) or all (i.e., perfect information) of the uncertainties are revealed in advance as well as a penalty that punishes violations of the nonanticipativity constraints. Brown et al. (2010) show both weak duality and strong duality. Weak duality ensures that any penalty that is *dual feasible*—in that it does not impose a positive, expected penalty on any nonanticipative policy—leads to an upper bound on the expected reward with any feasible policy, including an optimal policy. Strong duality ensures the existence of a dual feasible penalty such that the upper bound equals the expected reward with an optimal policy. Thus, by including a dual feasible penalty, we may be able to improve the perfect information bound. We refer to the optimization problem in which we relax all nonanticipativity constraints and include a dual feasible penalty as the *penalized perfect information problem* and the associated bound as the *penalized perfect information bound*.

Although the general theory we use closely follows Brown et al. (2010), the primary application of information relaxation duality to this point has been as a computational method for evaluating heuristic policies in a variety of applications. In contrast, our objective here is to use the approach to derive theoretical guarantees on the performance of heuristic policies in complex DPs. The general setup we study considers a given approximation to the optimal continuation value and a heuristic that selects actions "greedily" with respect to this approximate continuation value. We then generate a dual feasible penalty using this same approximate continuation value by taking the penalty to be the sum of the realized approximate continuation values minus their expectations; this construction follows the recipe for generating "good" penalties from Brown et al. (2010). We then obtain a bound on the suboptimality of the heuristic policy by analyzing the difference between the performance of the policy and the penalized perfect information problem in every sample path.

This basic recipe is broadly applicable: every approximate continuation value leads to a corresponding greedy heuristic policy and a "paired" penalized perfect information bound. It is not hard to show that with an optimal continuation value (and, hence, an optimal policy), the penalized perfect information bound and policy coincide in every sample path; this suggests that, with a good approximation to the optimal continuation value, the penalized perfect information problem will be nearly "aligned" with the heuristic policy in every sample path and, thus, close in value, thereby proving that the policy is nearly optimal. We apply the approach to three challenging problems: (1) stochastic knapsack problems, (2) stochastic scheduling on parallel machines, and (3) sequential search problems. In each problem, the method leads to analytical bounds on the suboptimality of a particular greedy heuristic policy and imply asymptotic optimality of the policy in specific regimes of interest.

1. *Stochastic knapsack.* In this problem (Dean et al. 2008), there is a set of items available to be inserted into a knapsack of finite capacity. Each item has a deterministic value and a stochastic size that is independently distributed. The actual size of an item is unknown until insertion of that item is attempted. The DM repeatedly selects items for insertion until the capacity overflows, and at that moment, the problem ends. The goal is to maximize the expected value of all items successfully inserted into the knapsack. In this problem, we need to balance choosing items with high values against items that tend to use little capacity (i.e., have small sizes). We consider an approximate continuation value that values remaining capacity in a simple linear fashion; the resulting greedy heuristic policy simply ranks items according to their value per expected size. Using this approximate continuation value in the penalty, it is nearly optimal to select items according to the greedy policy in the penalized perfect information problem in every sample path. The resulting gap implies the greedy policy is asymptotically optimal with many items, provided capacity scales at a particular rate.

2. *Stochastic scheduling on parallel machines.* This is the problem of scheduling a set of jobs on identical parallel machines when no preemptions are allowed. Job processing times are stochastic and independently distributed, and the processing times of each job are known only after a job is completed. The goal is to minimize the total expected weighted completion time. We consider an approximate continuation value from an identical scheduling problem but with a single machine that is M times faster, where M is the number of machines in the original problem. The resulting greedy heuristic policy schedules jobs in decreasing order of the ratio of a job's weight to its expected processing time (the weighted shortest expected processing time (WSEPT) first order). Using this approximate continuation value in the penalty, it is not hard to show that the greedy policy is nearly optimal in the penalized perfect information problem in every sample path. The resulting bound implies that the greedy policy is asymptotically optimal as the number of jobs grows large. The specific performance result complements existing performance guarantees on the greedy (WSEPT) policy in the scheduling literature, for example, in Weiss (1990) and Möhring et al. (1999).

3. *Sequential search for the best alternatives.* We consider a variation of the sequential search problem studied in

Weitzman (1979). In this problem, a DM sequentially searches a given set of alternatives with unknown rewards, drawn from independent distributions. Search is costly but reveals the rewards of an alternative. The DM can select previously revealed alternatives and collect the associated rewards. Weitzman (1979) studies the case when at most one alternative can be selected and characterizes the optimal policy in terms of a simple reservation price rule. We study a variation in which the DM has the capacity to select multiple alternatives, which significantly complicates the problem. We use an approximate continuation value based on a model with an infinite number of alternatives; the resulting greedy policy is a simple reservation price rule with reservation prices depending only on the remaining capacity. Using this approximate continuation value in the penalty, we can show that in the penalized perfect information problem the value of each unit of capacity is no larger than the corresponding reservation price in each sample path. This allows us to conclude that the greedy policy is asymptotically optimal as the number of alternatives grows large.

As we demonstrate in all three problems, the penalty is essential for obtaining a good bound. In each problem, we provide a simple example that is easy to solve in closed form but the perfect information bound is weak. For example, in the stochastic knapsack problem, the perfect information problem involves revealing all item sizes prior to any item selection decisions, and the DM can avoid inserting items with large realized sizes. In Section 3, we reproduce a convincing example from Dean et al. (2008) for which these bounds can be arbitrarily weak. With the inclusion of the penalty we consider, however, we recover a tight bound.

The approach does not circumvent the need for problem-specific insights and some problem-specific analysis (e.g., in determining the approximate continuation value). We view this as typical of most approximation schemes used to obtain policies and bounds. For example, in the approximate linear programming approach to dynamic programming (e.g., de Farias and Roy 2003), one needs to specify appropriate basis functions. Similarly, in applying Lagrangian duality to analyze DPs (e.g., Adelman and Mersereau 2008), one needs to specify the Lagrange multipliers. A potentially attractive feature of our approach, however, is that the penalized perfect information problem and the heuristic policy are explicitly linked in a way that can be directly applied to many other problems.

The rest of the paper is organized as follows. Section 1.1 reviews some related work. Section 2 develops the general theory related to information relaxation duality and provides an overview of the approach. We study the stochastic knapsack problem in Section 3, the stochastic scheduling problem in Section 4, and the sequential search problem in Section 5. Each of these sections follows the general theory in Section 2 in self-contained fashion: in each section, we begin by describing the problem and the perfect information bound, and we then discuss the penalized perfect information bound and present our performance analysis. Section 6 concludes. Most proofs are in the appendix. Online Appendices B, C, and D present some extensions.

1.1. Literature Review

In this section, we discuss the connection of our work to several streams of literature. First, our paper naturally relates to the literature on information relaxations. Brown et al. (2010) draw inspiration from a stream of papers on "martingale duality methods" aimed at calculating upper bounds on the price of high-dimensional, American options, tracing back to independent developments by Haugh and Kogan (2004) and Rogers (2002). Rogers (2007) independently developed similar ideas as in Brown et al. (2010) for perfect information relaxations of Markov decision process (MDPs) using change-of-measure techniques.

In terms of applications of information relaxations, there are many other applications to option pricing problems (Andersen and Broadie 2004, Brown et al. 2010, Desai et al. 2012); inventory management problems (Brown et al. 2010, Brown and Smith 2014); valuation of natural gas storage (Lai et al. 2010, Nadarajah et al. 2015); integrated models of procurement, processing, and commodity trading (Devalkar et al. 2011); dynamic portfolio optimization (Brown and Smith 2011, Haugh et al. 2016); linear-quadratic control with constraints (Haugh and Lim 2012); network revenue management problems (Brown and Smith 2014); and multiclass queueing systems (Brown and Haugh 2017). Of central concern in these papers is computational tractability: the goal is to use an information relaxation and penalty that render the upper bounds sufficiently easy to compute. A recurring theme in these papers is that relatively easy-to-compute policies are often nearly optimal, and the bounds computed from information relaxations are essential in showing this. This line of work has focused on computing numerical bounds for specific problem instances; our focus is on using information relaxations to derive analytical guarantees on the performance of heuristic policies.

Perfect information bounds (without penalty) have been successfully used in theoretically analyzing heuristic policies in several applications in operations research and computer science and are often referred to as "hindsight bounds" or "offline optimal bounds." Talluri and van Ryzin (1998) show that static bid–price policies are asymptotically optimal in network revenue management when capacities and the length of the horizon are large; they provide various upper bounds on the performance of the optimal policy, including perfect information bounds. Feldman et al. (2010) study

the online stochastic packing problem in the setting in which the underlying probabilistic model is unknown and show that a training-based primal-dual heuristic is asymptotically optimal when the number of items and capacities are large; they use the perfect information bound as a benchmark. Manshadi et al. (2012) study the same problem when the underlying probability distributions are known by the decision maker and present an algorithm that achieves at least 0.702 of the perfect information bound. Garg et al. (2008) study the stochastic Steiner tree problem in which each demand vertex is drawn independently from some distribution and show that greedy policy is nearly optimal relative to the perfect information bound. Similarly, Grandoni et al. (2008) study stochastic variants of set cover and facility location problems and show that suitably defined greedy policies perform well with respect to the expected cost with perfect information. Finally, in computer science, there is a large body of work on *competitive analysis*, which revolves around studying the performance of online algorithms relative to the performance of an optimal "offline" algorithm that knows the entire input in advance. In this line of work, there is no underlying probabilistic model for the inputs, and instead, performance is measured relative to the offline optimum in the worst case (see, e.g., Borodin and El-Yaniv 1998 for a comprehensive review).

The knapsack problem has a rich history in operations research with various applications in areas such as finance, advertising, transportation, revenue management, and scheduling (Martello and Toth 1990). The version of the stochastic knapsack problem we study was introduced by Dean et al. (2008) although variants of this problem have been studied earlier. For example, Papstavrou et al. (1996) study a version in which items arrive stochastically and in take-it-or-leave-it fashion; Dean et al. (2008) provide an overview of earlier work. Dean et al. (2008) study both nonadaptive policies and adaptive policies for the problem and show that the loss for restricting attention to nonadaptive policies is at most a factor of four. In addition, they provide sophisticated linear programming bounds based on polymatroid optimization. The nonadaptive policy they consider is a greedy policy that inserts items in decreasing order of the ratio of value to expected size. When item sizes are small relative to capacity, Dean et al. (2008) show that the greedy policy performs within a factor of two of the optimal policy. Derman et al. (1978) show that the greedy policy is optimal in the case of exponentially distributed sizes for a "covering" variation of the problem (we discuss this variation in Section 3.5). Blado et al. (2016) develop approximate dynamic programming bounds for this problem and in extensive numerical experiments find that the greedy policy often performs well, especially for examples with many items. In subsequent work, Blado and Toriello (2016) establish some asymptotic optimality results for the greedy policy using a different approach involving approximate linear programming relaxations.

Stochastic scheduling is a fundamental problem in operations research with a vast literature, which we do not attempt to review here; see Pinedo (2012) for a comprehensive review. Weiss (1990) originally established the optimality gap of the WSEPT policy for scheduling on parallel machines and proved that this policy is asymptotically optimal under mild conditions. Möhring et al. (1999) study polyhedral relaxations of the performance space of stochastic parallel machine scheduling and provide bounds on the performance of the WSEPT policy. The result we present in Section 4 is closest to a result in Weiss (1990); in Online Appendix C, we provide an alternative and novel proof of the result in Möhring et al. (1999) using penalized perfect information bounds.

Finally, there is long literature on sequential search problems with many applications. The model we study is similar to that in Weitzman (1979), who describes the problem in terms of opening "boxes" with a priori unknown rewards and derives an optimal policy. Weitzman (1979) calls this reservation price policy "Pandora's Rule." The optimal reservation price rule in Weitzman (1979) is essentially equivalent to a Gittins index (Gittins and Jones 1974) policy. Many variations of this problem have been studied, but it appears that "Pandora's Rule does not ready generalize" (Weitzman 1979, p. 650) in many natural extensions of the problem. With the ability to select only one alternative ("box"), the problem reduces to a stopping problem; when the DM can select multiple alternatives, the DM may intermittently recall past alternatives as opportunities for finding better alternatives dwindle, and optimal policies may need to track previous search results as part the problem state. The ability to select multiple alternatives also appears to significantly complicate other versions of sequential search problems; for example, Vanderbei (1980) studies a version of the secretary problem with the goal of selecting a best subset of the search population.

2. General Theory

In describing the basic results related to information relaxations, we work with a general MDP formulation. We consider a discrete-time, finite-horizon MDP with T periods and decision periods denoted by $t = 1, \ldots, T$. We let x_t denote the MDP state at time t. In each period, a DM must choose an action a from a set $A_t(x_t)$. After an action $a_t \in A_t(x_t)$ is selected, a random variable ε_t is realized, the DM collects reward $r_t(x_t, a_t, \varepsilon_t)$, and the state transitions according to $x_{t+1} = \chi_t(x_t, a_t, \varepsilon_t)$, where the ε_t are independent random variables. Feasible policies in this primal DP must be *nonanticipative*, meaning the actions selected in period t must be measurable

functions of the past actions and states. We restrict feasible policies to be deterministic and Markovian, and we use $\alpha = (\alpha_t)_{t=1}^T$ to denote a feasible policy; that is, α is a sequence of deterministic functions α_t that map from a state x_t to an action in $A_t(x_t)$ for all t and x_t. We let $\mathcal{A}$ denote the set of feasible Markovian policies. For any policy $\alpha \in \mathcal{A}$, we denote the expected performance of the policy by $V^\alpha \triangleq \mathbb{E}[r(\alpha)]$, where $r(\alpha) \triangleq \sum_{t=1}^T r_t(x_t, \alpha_t(x_t), \varepsilon_t)$. The DM's goal is to select a feasible policy that maximizes the expected value of the total reward.

We let V^* denote the optimal value and assume there exists at least one feasible policy $\alpha \in \mathcal{A}$ attaining V^*. We can write the primal DP succinctly as

$$V^* = \max_{\alpha \in \mathcal{A}} \mathbb{E}[r(\alpha)]. \tag{1}$$

Under a variety of standard conditions (see, e.g., Puterman 1994), it is well known that we can equivalently write Equation (1) as a recursion over value functions; we take $V_{T+1}^* = 0$, and

$$V_t^*(x_t) = \max_{a \in A_t(x_t)} \mathbb{E}_{\varepsilon_t}\left[r_t(x_t, a, \varepsilon_t) + V_{t+1}^*(\chi_t(x_t, a, \varepsilon_t))\right], \tag{2}$$

with $V^* = V_1^*(x_1)$ in the initial state x_1. We let $Q_t^*(x, a, \varepsilon) \triangleq r_t(x, a, \varepsilon) + V_{t+1}^*(\chi_t(x, a, \varepsilon))$ denote the *optimal continuation value* in period t when the current state is x, the action taken is a, and the ensuing realized uncertainty is ε. Equation (2) implies that an optimal action $\alpha_t^*(x_t)$ should be "greedy" with respect to the expected value of Q_t; specifically, an optimal policy selects actions according to

$$\alpha_t^*(x_t) \in \arg\max_{a_t \in A_t(x_t)} \mathbb{E}_{\varepsilon_t}\left[Q_t^*(x_t, a_t, \varepsilon_t)\right].$$

The function $\mathbb{E}_{\varepsilon_t}\left[Q_t^*(x_t, a_t, \varepsilon_t)\right]$ depends on the state–action pair (x_t, a_t) and is often referred to as the "Q-factor" in the approximate dynamic programming and reinforcement learning literature (see, e.g., Bertsekas 2000, section 6.4). We modify this definition to also include the uncertainties ε_t as these values are revealed in the information relaxations.

All of the heuristic policies we study can be interpreted as policies that are "greedy" with respect to a given approximation of the optimal continuation value. In particular, given an approximation $Q_t(x, a, \varepsilon)$ of the optimal continuation value, the heuristic policies select actions according to

$$\alpha_t(x_t) \in \arg\max_{a_t \in A_t(x_t)} \mathbb{E}_{\varepsilon_t}\left[Q_t(x_t, a_t, \varepsilon_t)\right]. \tag{3}$$

2.1. Duality Results

We work exclusively with the *perfect information relaxation*, in which all uncertainties are revealed prior to making decisions. This relaxation can be viewed as one in which all nonanticipativity constraints are removed; Brown et al. (2010) consider more general information relaxations that correspond to partial relaxations of the nonanticipativity constraints.

Formally, we consider a relaxed DP in which the sequence $(\varepsilon_t)_{t=1}^T$ is revealed to the DM prior to making any decisions. Letting $\boldsymbol{\varepsilon} \triangleq (\varepsilon_1, \ldots, \varepsilon_T)$, $\mathbf{a} \triangleq (a_1, \ldots, a_T)$, and $A(\boldsymbol{\varepsilon}) \triangleq \{\mathbf{a} : a_t \in A_t(x_t) \text{ for all } t\}$, we can write the perfect information problem given $\boldsymbol{\varepsilon}$ as a deterministic optimization problem:

$$V^{\mathrm{P}}(\boldsymbol{\varepsilon}) = \sup_{\mathbf{a} \in A(\boldsymbol{\varepsilon})} \sum_{t=1}^T r_t(x_t, a_t, \varepsilon_t), \tag{4}$$

where states evolve as $x_{t+1} = \chi_t(x_t, a_t, \varepsilon_t)$. Because selecting actions according to any feasible, nonanticipative policy is feasible in Equation (4), the perfect information problem provides an upper bound on V^* in expectation; that is, $V^* \le \mathbb{E}_{\boldsymbol{\varepsilon}}\left[V^{\mathrm{P}}(\boldsymbol{\varepsilon})\right]$.

Unfortunately, this perfect information upper bound is often quite weak: absent any penalty for the additional information, the DM with perfect information may obtain an expected reward much larger than V^*. In the problems we study, our goal is to show that a simple, heuristic policy performs well in some asymptotic limit, and we show with explicit examples that the perfect information bound is not an asymptotically tight bound in each of these problems.

Following Brown et al. (2010), we, therefore, consider including a penalty that attempts to compensate for the benefit of the additional information. As with rewards, the penalty is an action-dependent random variable $z(\mathbf{a}) \triangleq \sum_{t=1}^T z_t(x_t, a_t, \varepsilon_t)$ for some sequence of period-t penalty functions $(z_t)_{t=1}^T$. We say a penalty is dual feasible if $\mathbb{E}[z(\alpha)] = 0$ for all $\alpha \in \mathcal{A}$; that is, a dual feasible penalty "charges" zero expected penalty for any nonanticipative policy. The penalty may, however, charge a positive expected penalty to policies that "cheat" by violating the nonanticipativity constraints.

We use the following version of the "weak duality" result from Brown et al. (2010) to generate performance bounds on the primal DP. In what follows, we let $V_z^\alpha(\boldsymbol{\varepsilon}) = \sum_{t=1}^T r_t(x_t, \alpha_t(x_t), \varepsilon_t) - z_t(x_t, \alpha_t(x_t), \varepsilon_t)$, that is, the penalized performance of a given feasible policy $\alpha \in \mathcal{A}$; for a dual feasible penalty z, we have $\mathbb{E}[z(\alpha)] = 0$, and hence, $\mathbb{E}_\varepsilon[V_z^\alpha(\boldsymbol{\varepsilon})] = V^\alpha$.

Proposition 1 (Weak Duality). *For any $\alpha \in \mathcal{A}$ and any dual feasible penalty z,*

$$V^\alpha \le \mathbb{E}_{\boldsymbol{\varepsilon}}\left[V_z^{\mathrm{P}}(\boldsymbol{\varepsilon})\right] \triangleq V_z^{\mathrm{P}}, \tag{5}$$

where $V_z^{\mathrm{P}}(\boldsymbol{\varepsilon}) \triangleq \sup_{\mathbf{a} \in A(\boldsymbol{\varepsilon})}\{r(\mathbf{a}) - z(\mathbf{a})\}$. In particular, we have $V^* - V^\alpha \le \mathbb{E}_{\boldsymbol{\varepsilon}}[V_z^{\mathrm{P}}(\boldsymbol{\varepsilon}) - V_z^\alpha(\boldsymbol{\varepsilon})]$.

Proof. We have $V^\alpha = \mathbb{E}[r(\alpha)] = \mathbb{E}[r(\alpha) - z(\alpha)] \le \mathbb{E}_\varepsilon\left[\sup_{\mathbf{a} \in A(\varepsilon)}\{r(\mathbf{a}) - z(\mathbf{a})\}\right] = V_z^{\mathrm{P}}$, where the second equality

follows by feasibility of α and dual feasibility of the penalty and the inequality follows by the fact that the actions selected by α are feasible in every sample path $\boldsymbol{\varepsilon}$. The last statement follows because $V^* \le V_z^{\mathrm{P}}$ (take α to be an optimal policy in $\mathcal{A}$ with $V^\alpha = V^*$), and $\mathbb{E}_{\boldsymbol{\varepsilon}}[V_z^\alpha(\boldsymbol{\varepsilon})] = V^\alpha$ again by dual feasibility of z. □

Taking α to be an optimal, nonanticipative policy, Proposition 1 implies that $V^* \le V_z^{\mathrm{P}}$ for any dual feasible penalty z. We can think of the upper bound V_z^{P} in terms of Monte Carlo simulation: we randomly simulate sample paths ε, solve the deterministic "inner problem" $V_z^{\mathrm{P}}(\boldsymbol{\varepsilon}) = \max_{\mathbf{a}\in A(\varepsilon)}\{r(\mathbf{a}) - z(\mathbf{a})\}$ given $\boldsymbol{\varepsilon}$, and average the resulting optimal values. (In everything that follows, we assume the sup in Equation (5) is almost surely attained, justifying the use max in place of sup; this holds in all of our examples.)

In our analysis of heuristic policies, we compare the value $V_z^{\mathrm{P}}(\boldsymbol{\varepsilon})$ of the penalized perfect information problem to the value of the heuristic policy in every sample path. To facilitate this comparison, we include the penalty in the performance of the policy; this does not affect the expected reward of the policy. The last part of Proposition 1 states that the expected value of $\mathbb{E}_{\boldsymbol{\varepsilon}}[V_z^{\mathrm{P}}(\boldsymbol{\varepsilon}) - V_z^\alpha(\boldsymbol{\varepsilon})]$ is an upper bound on the suboptimality of the policy. The goal in a given problem is to show that this upper bound is sufficiently small.

We use the approximate continuation value Q_t that generates the greedy heuristic policy to generate penalties as well. Specifically, we use penalties of the form

$$z_t(x_t, a_t, \varepsilon_t) = Q_t(x_t, a_t, \varepsilon_t) - \mathbb{E}_{\tilde{\varepsilon}_t}[Q_t(x_t, a_t, \tilde{\varepsilon}_t)], \quad (6)$$

where Q_t is the approximation used in Equation (3) for selecting actions in the heuristic policy. By applying the law of iterated expectations, we see for any $\alpha \in \mathcal{A}$, that $\mathbb{E}_{\varepsilon_t}[z_t(x_t, \alpha_t(x_t), \varepsilon_t)] = 0$ for each time period t and state x_t, and hence, the penalty from Equation (6) is dual feasible.

Intuitively, the goal of the penalty is to cancel the benefit of perfect information. By taking $Q_t = Q_t^*$, where $Q_t^*(x, a, \varepsilon) = r_t(x, a, \varepsilon) + V_{t+1}^*(\chi_t(x, a, \varepsilon))$, the penalty in Equation (6) perfectly cancels the benefit of perfect information, and the optimal value of the penalized perfect information problem in Equation (5) equals V^* in every sample path. To see this, fix a sample path $\boldsymbol{\varepsilon}$ and fix the actions $\mathbf{a} \in A(\boldsymbol{\varepsilon})$ and states to be optimal actions and states for the penalized perfect information problem given the sample path. We have

$$\begin{aligned} V_z^{\mathrm{P}}(\boldsymbol{\varepsilon}) &= r(\mathbf{a}) - z(\mathbf{a}) \\ &= \sum_{t=1}^{T} r_t(x_t, a_t, \varepsilon_t) + \mathbb{E}_{\tilde{\varepsilon}_t}\left[Q_t^*(x_t, a_t, \tilde{\varepsilon}_t)\right] - Q_t^*(x_t, a_t, \varepsilon_t) \\ &\le \sum_{t=1}^{T} V_t^*(x_t) - V_{t+1}^*(x_{t+1}) = V_1^*(x_1) - V_{T+1}^* = V^*, \end{aligned}$$

where the inequality follows from the definition of Q_t^*, using that $V_t^*(x_t) = \max_{a\in A_t(x_t)} \mathbb{E}_{\tilde{\varepsilon}_t}\left[Q_t^*(x_t, a, \tilde{\varepsilon}_t)\right]$ from Equation (2), and $x_{t+1} = \chi_t(x_t, a_t, \varepsilon_t)$. Because the actions of an optimal policy $\alpha_t^*(x_t^*)$, where x_t^* is the state induced by α_t^*, are feasible for the penalized perfect information, by a similar argument it follows that $V^* \le V_z^{\mathrm{P}}(\boldsymbol{\varepsilon})$. This implies that strong duality holds and that the optimal value of the penalized perfect information problem equals V^* in every sample path.

Overview of Approach. As just discussed, we can in principle obtain a tight bound using the "ideal penalty" based on the optimal continuation value Q_t^*. This does not directly help us in problems with large state spaces, where Q_t^* is difficult to determine. Nonetheless, the strong duality result just discussed is suggestive: a good approximation to the optimal continuation value Q_t^* will lead to both a good policy and a good penalty, and the penalized performance of the policy will be close to the penalized perfect information value in every sample path.

Our approach works as follows:

1. Design an approximation $Q_t(x, a, \varepsilon)$ to the optimal continuation value. This determines a heuristic policy α using Equation (3) and a penalty z using Equation (6).

2. Establish an error bound $\Delta(\boldsymbol{\varepsilon})$ such that for every sample path $\boldsymbol{\varepsilon}$:

$$V_z^{\mathrm{P}}(\boldsymbol{\varepsilon}) - V_z^\alpha(\boldsymbol{\varepsilon}) \le \Delta(\boldsymbol{\varepsilon}). \quad (7)$$

3. Taking expectations, using Proposition 1, and denoting $\Delta = \mathbb{E}_{\boldsymbol{\varepsilon}}[\Delta(\boldsymbol{\varepsilon})]$, we obtain

$$V^* - V^\alpha \le V_z^{\mathrm{P}} - V^\alpha \le \Delta\,.$$

A given approximation Q_t leads to both a policy and a performance bound, and the task then reduces to deterministic analysis to obtain a bound on the gap in Equation (7) in each sample path. What constitutes a "good" value of Δ depends on the goals of the analysis; in the problems we study, we seek an error bound Δ that is (relatively) small in some asymptotic regime of interest, which allows us to conclude that α is asymptotically optimal in this limit.

3. Stochastic Knapsack Problem

We consider the version of the stochastic knapsack problem studied in Dean et al. (2008). There is a set of items, indexed by $i \in \mathcal{I} = \{1, \ldots, I\}$, available to be inserted into a knapsack of initial capacity κ. Item $i \in \mathcal{I}$ has a deterministic value denoted by $v_i \ge 0$ and a stochastic size denoted by $s_i \ge 0$. The sizes are independent random variables with known, arbitrary distributions. The actual size of an item is unknown until the item is selected for insertion. Random values can be easily

accommodated, provided values are independent and independent of sizes, by replacing each random value with its expectation.

At each decision epoch, the DM selects an item i and attempts to insert it into the knapsack. After that, the size s_i of item i is revealed, and a value of v_i is obtained if i is successfully inserted, that is, if s_i is no larger than the remaining capacity. The DM repeatedly selects items for insertion until the capacity overflows. At that moment, the problem ends, and the value of the overflowing item is not collected. The goal is to maximize the expected value of all items successfully inserted into the knapsack.

A feasible policy $\alpha \in \mathcal{A}$ is a mapping that determines the next item $\alpha(\mathcal{S}, c)$ to attempt to insert into the knapsack given the set of remaining items $\mathcal{S} \subseteq \mathcal{I}$ and the remaining knapsack capacity $c \in [0, \kappa]$. We denote the decision epochs by $t = 1, \ldots, I$. For a given $\alpha \in \mathcal{A}$, we let $\mathcal{S}_t^\alpha$ denote the items available for insertion at the beginning of time t and c_t^α denote the knapsack's remaining capacity. To simplify the notation, we let $\alpha_t = \alpha(\mathcal{S}_t^\alpha, c_t^\alpha)$ denote the item to be inserted at time t under policy α. The initial conditions are $\mathcal{S}_1^\alpha = \mathcal{I}$ and $c_1^\alpha = \kappa$. At time t, item $\alpha_t = \alpha(\mathcal{S}_t, c_t)$ is selected for insertion, and the state is updated as $\mathcal{S}_{t+1}^\alpha = \mathcal{S}_t^\alpha \setminus \{\alpha_t\}$ and $c_{t+1}^\alpha = c_t^\alpha - s_{\alpha_t}$. If the item fits into the knapsack, that is, if $c_{t+1}^\alpha \geq 0$, the value of v_{α_t} is collected. Otherwise, the problem ends.

We let $\tau^\alpha = \inf\{t \geq 1 : c_{t+1}^\alpha < 0\}$ denote the stopping time corresponding to the first time capacity overflows. We can then write the problem as

$$V^* = \max_{\alpha \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{I \wedge (\tau^\alpha - 1)} v_{\alpha_t}\right].$$

3.1. Approximate Continuation Value and Greedy Policy

We let $Q_t(c, i, s_i)$ denote the approximate continuation value in period t when the remaining knapsack capacity is c and item i with unknown size s_i is selected for insertion. We use the approximation

$$Q_t(c, i, s_i) = v_i + r_i(c - s_i). \tag{8}$$

In Equation (8), we use a simple linear approximation for the value of remaining capacity. We consider the specific choice $r_i \triangleq v_i/\mathbb{E}[s_i]$, that is, the ratio of the selected item's value to its expected size; this approximation naively assumes the currently selected item i appropriately captures the marginal value of remaining capacity irrespective of the subset $\mathcal{S}$ of remaining items. Because the expected continuation value is $\mathbb{E}_{s_i}[Q_t(c, i, s_i)] = v_i/(\mathbb{E}[s_i])c$, Equation (3) implies that the greedy policy induced by Equation (8) sorts the items in decreasing order of r_i and inserts items in this order until the knapsack overflows or no items remain.[1] Without loss of generality, we assume that items are sorted in decreasing order of this ratio; that is, $r_1 \geq r_2 \geq \ldots \geq r_I$. The expected performance of the greedy policy is given by

$$V^{\mathrm{G}} = \mathbb{E}\left[\sum_{t=1}^{I \wedge (\tau^{\mathrm{G}} - 1)} v_t\right],$$

where τ^{G} is the first time that capacity overflows under the greedy policy.

Dean et al. (2008) show that a randomized variant of the greedy policy performs within a factor of 7/32 of the optimal value. It is possible to find simple examples in which the greedy policy can perform arbitrarily poorly (see, e.g., a deterministic example of this with $I = 2$ in §4 of Dean et al. 2008). In general, the greedy policy performs poorly, even in deterministic examples, when sizes are large relative to capacity: because we cannot add fractional amounts of items, the ratio of value to size may not be a good proxy for the marginal value of adding an item.

On the other hand, we might expect the greedy policy to perform well when sizes are small relative to capacity: with many small items, the problem "smoothes" in a certain sense. We can gain intuition for this from the deterministic case by considering the linear programming (LP) relaxation of the problem that allows the DM to insert fractional items. Because the greedy ordering is optimal in the LP relaxation, in the deterministic case we have

$$V^* - V^{\mathrm{G}} \leq V^{\mathrm{LP}} - V^{\mathrm{G}} \leq \max_{i=1,\ldots,I} v_i, \tag{9}$$

where V^{LP} is the optimal objective value of the LP relaxation. In Equation (9), the gap in the last inequality arises from potential lost value of an overflowing item, which can be included fractionally in the LP relaxation but cannot be included by the greedy policy. If we then consider scaling the problem so that capacity increases by an integer factor $\theta \geq 1$ and we make θ copies of all items, we conclude from Equation (9) that the relative suboptimality of the greedy policy goes to zero as θ gets larger. In this sense, in the deterministic problem, the greedy policy performs well as we consider problems with many items that are small relative to capacity.

We derive a result analogous to Equation (9) for the stochastic version of the problem in which the decision maker optimizes over all possible nonanticipative policies. The result then allows us to analyze the performance of the greedy policy as the number of items grows large—and the problem is, thus, increasingly difficult to solve—under certain conditions on the capacity, values, and the distributions of sizes.

3.2. Perfect Information Bound

Consider a clairvoyant with access to all future realizations of the sizes $\mathbf{s} = (s_i)_{i=1}^I$ before selecting any items. Given a sample path $\mathbf{s} \in \mathbb{R}_+^I$, we let $V^{\mathrm{P}}(\mathbf{s})$ denote the optimal (deterministic) value for sample path $\mathbf{s}$ with perfect information about sizes. The expected value $V^{\mathrm{P}} = \mathbb{E}_{\mathbf{s}}[V^{\mathrm{P}}(\mathbf{s})]$ is an upper bound for the optimal performance; that is, $V^* \le V^{\mathrm{P}}$. The perfect information problem is equivalent to the deterministic knapsack problem

$$V^{\mathrm{P}}(\mathbf{s}) = \max_{\mathbf{x}\in\{0,1\}^I} \left\{ \sum_{i=1}^I v_i x_i \ : \ \sum_{i=1}^I s_i x_i \le \kappa \right\},$$

where $x_i \in \{0,1\}$ indicates whether item i is included in the knapsack.

Unfortunately, the perfect information bound may be quite loose: by knowing the realizations of all sizes in advance, the DM can avoid inserting potentially large items. Dean et al. (2008) demonstrate this with the following example: consider the case when all items are symmetric with value one, and each item's size is either zero or $\kappa + \varepsilon$ (for some $\varepsilon > 0$) with probability 1/2. Because the items are symmetric, the problem is trivial, and it is easy to show that $V^{\mathrm{G}} = V^* = 1 - (1/2)^I \le 1$. On the other hand, in the perfect information problem, it is optimal to select every item with a realized size of zero. Because this occurs with probability 1/2 for each item and items are independent, this leads to the very poor upper bound of $V^{\mathrm{P}} = I/2$.

3.3. Penalized Perfect Information Bound

To improve the upper bound, we impose a penalty that punishes violations of the nonanticipativity constraints. Equation (6) implies that the period-t penalty induced by Equation (8) is $z_t(c, i, s_i) = r_i(\mathbb{E}[s_i] - s_i)$. Denoting by a_t the item to be inserted at time t, we obtain that the total penalty is given by

$$z(\mathbf{a}) = \sum_{t=1}^{I\wedge\tau^\alpha} r_{a_t}(\mathbb{E}[s_{a_t}] - s_{a_t}).$$

Note that we include the penalty terms until τ^α, that is, until it is known that an overflow has occurred. This is required to ensure dual feasibility of the penalty: for any feasible policy $\alpha \in \mathcal{A}$, τ^α is a stopping time, but $\tau^\alpha - 1$ is not.

We let $V_z^{\mathrm{P}}(\mathbf{s})$ denote the optimal (deterministic) value of the penalized perfect information problem for sample path $\mathbf{s} \in \mathbb{R}_+^I$. From Proposition 1, we obtain an upper bound $V^* \le V_z^{\mathrm{P}}$, where $V_z^{\mathrm{P}} = \mathbb{E}_{\mathbf{s}}[V_z^{\mathrm{P}}(\mathbf{s})]$ denotes the penalized perfect information bound. In Appendix A.1, we discuss how to calculate $V_z^{\mathrm{P}}(\mathbf{s})$ by solving an integer program with additional variables representing which item, if any, overflows the knapsack.

Recall that the DM with perfect information may "cheat" by selecting items with low realized sizes. The penalty creates an incentive for the DM with perfect information to resist selecting items with realized sizes that are small relative to their expected sizes: the penalized value of selecting item i becomes $v_i + r_i(s_i - \mathbb{E}[s_i]) = r_i s_i$. Thus, in the penalized perfect information problem, the DM may "cheat" and select items with low realized sizes but will also receive less value for doing so.

It is instructive to see how this works on the example from Dean et al. (2008) as discussed in Section 3.2 with I symmetric items of value one and sizes that are either zero or $\kappa + \epsilon$ with probability 1/2. Recall that a greedy policy is (trivially) optimal, and the optimal value is $V^* = 1 - (1/2)^I$, but the perfect information problem leads to the poor bound of $V^{\mathrm{P}} = I/2$. In the penalized perfect information problem, the value for selecting an item is $r_i s_i$, which is zero if $s_i = 0$ and two if $s_i = \kappa + \epsilon$: in particular, any items with realized sizes of zero provide zero value as well. Moreover, we can select at most one item with realized positive size of $\kappa + \epsilon$—in particular, an item that overflows the knapsack. Because the penalty for the item that overflows the knapsack equals one, we obtain $V_z^{\mathrm{P}}(\mathbf{s}) = 1$ if $s_i > 0$ for any i and $V_z^{\mathrm{P}}(\mathbf{s}) = 0$ otherwise. Because $\mathbb{P}\{s_i = 0 \ \forall \ i\} = (1/2)^I$, the penalized perfect information bound then is

$$V_z^{\mathrm{P}} \ = \ 1 - (1/2)^I \ = \ V^*,$$

that is, we recover a tight bound for all values of I.

In general, the penalty aligns the perfect information problem with the greedy policy in that it is "nearly" optimal for the perfect information policy to select items according to the greedy ordering. The "nearly" involves quantifying the slack in the upper bound because of value collected from an overflowing item, analogous to the analysis of LP relaxations in the deterministic case in Equation (9).

3.4. Performance Analysis

We now formalize this discussion in the general case. We show that the greedy policy incurs a small loss in value compared with the optimal policy when the scale of the problem increases and, in particular, that the greedy policy is asymptotically optimal under conditions that we make precise. Let $V_z^{\mathrm{G}}(\mathbf{s}) = \sum_{i=1}^{I\wedge(\tau^{\mathrm{G}}-1)} v_i + \sum_{i=1}^{I\wedge\tau^{\mathrm{G}}} r_i(s_i - \mathbb{E}[s_i])$ denote the penalized performance of the greedy policy under sample path $\mathbf{s}$ and $x^+ = \max(x, 0)$ for $x \in \mathbb{R}$. The following result compares, for every sample path, the penalized performance of the greedy policy to the performance of the penalized perfect information problem.

Proposition 2. *For every sample path* $\mathbf{s}$*, the penalized performance of the greedy policy satisfies*

$$V_z^{\mathrm{P}}(\mathbf{s}) - V_z^{\mathrm{G}}(\mathbf{s}) \le \max_{i\in\mathcal{J}} v_i + \max_{i\in\mathcal{J}} r_i(s_i - \mathbb{E}[s_i])^+. \qquad (10)$$

We prove Equation (10) by relating the optimal value of the penalized perfect information problem $V_z^{\mathrm{P}}(\mathbf{s})$ to the penalized performance of the greedy policy $V_z^{\mathrm{G}}(\mathbf{s})$. Recall that with the penalty, the values for selecting items with low realized sizes are adjusted downward, and thus, the DM with perfect information has less incentive to "cheat" by selecting items with low realized sizes. The DM with perfect information, however, can still "cheat" by choosing a large item to overflow the knapsack because it receives the value of the overflowing item. We handle this issue by decomposing the penalized perfect information problem into (1) a traditional deterministic knapsack problem and (2) another problem in which the DM can choose any item as a candidate to overflow the knapsack regardless of whether this item actually leads to the overflow. In the LP relaxation of the first problem, the greedy policy is optimal, and a loss of at most $\max_i v_i$ is incurred because the last item can be included fractionally in the LP relaxation but cannot be included by the greedy policy. This leads to the first loss term in Equation (10). In the second problem, the DM simply chooses the item with largest penalized value $r_i(s_i - \mathbb{E}[s_i])$ whenever this value is nonnegative as a candidate to overflow the knapsack, which leads to the second loss term in Equation (10).

Taking expectations in Equation (10) and using the duality results of Section 2, we obtain the following guarantees on the performance of the greedy policy.

Corollary 1. *The performance of the greedy policy satisfies:*

1. *Performance guarantee.*

$$V^* - V^{\mathrm{G}} \le V_z^{\mathrm{P}} - V^{\mathrm{G}} \le \max_i v_i + \mathbb{E}\left[\max_{i\in\mathcal{I}} r_i(s_i - \mathbb{E}[s_i])^+\right].$$

2. *Asymptotic optimality. If* $\lim_{I\to\infty} \frac{1}{\kappa}\mathbb{E}[\max_i r_i s_i] = 0$, *then*

$$\lim_{I\to\infty} \frac{1}{\kappa}(V^* - V^{\mathrm{G}}) = 0\,. \tag{11}$$

Corollary 1 shows that the greedy policy is asymptotically optimal when the expected maximum penalized value $r_i s_i$ grows more slowly than the capacity of the knapsack; this, in turn, limits the value the penalized perfect information DM can obtain by overflowing the knapsack (the asymptotic regime allows the capacity κ to scale with the number of items I).[2] Note that asymptotic optimality requires $\frac{1}{\kappa}\mathbb{E}[\max_i r_i s_i] \to 0$: this follows from the fact that $\max_i v_i \le \mathbb{E}[\max_i r_i s_i]$ and $(s_i - \mathbb{E}[s_i])^+ \le s_i$ as we show in the proof of the result. The condition given in Corollary 1, part 2, although general, may be difficult to verify directly. The next result provides more easily verifiable sufficient conditions for asymptotic optimality. We say a_n is *little omega* of b_n or $a_n = \omega(b_n)$ if $\lim_{n\to\infty} a_n/b_n = \infty$, that is, a_n grows asymptotically faster than b_n.

Corollary 2. *Suppose that* $r_i \le \bar{r}$ *for some* $\bar{r} < \infty$ *independent of* I. *Then Equation* (11) *holds if*

1. *Sizes are uniformly bounded; that is,* $s_i \le \bar{s} < \infty$, *and capacity scales as* $\kappa = \omega(1)$.
2. *Sizes have uniformly bounded* $p > 1$ *moments, that is,* $\mathbb{E}[s_i^p] \le m < \infty$, *and capacity scales as* $\kappa = \omega(I^{1/p})$.
3. *Sizes are uniformly sub-Gaussian, that is, there exist* $a, b > 0$ *such that* $\mathbb{P}\{s_i > x\} \le a\exp(-bx^2)$ *for all* $x \ge 0$, *and capacity scales as* $\kappa = \omega(\sqrt{\log I})$.

Intuitively, the growth of the maximum penalized value $\max_i r_i s_i$ is governed to a large extent by the tails of the distributions of sizes. When items are symmetric, roughly speaking, we have that $\mathbb{E}[\max_i s_i] \approx F^{-1}(I/(I+1))$, where F is the cumulative distribution function of sizes, and thus, we need capacity to grow at least as $F^{-1}(I/(I+1))$ for Equation (11) to hold. Corollary 2 makes this intuition precise and provides the necessary growth rate of capacity for different families of distributions. When sizes are uniformly bounded (e.g., uniform or Bernoulli), the penalized values are trivially bounded, and it suffices that capacity grow unbounded at any rate. When sizes have p-moments, it suffices that capacity grow at a power-law rate. When sizes are sub-Gaussian, it suffices that capacity grow at a logarithmic rate.

3.5. Stochastic Covering Variation

We consider a variation of the stochastic knapsack problem in which a DM needs to select components to cover certain requirement at minimum cost.[3] Following Derman et al. (1978), we consider an application to equipment maintenance in which a system, requiring a certain component to function, must operate for a fixed amount of time. The component needs to be replaced each time it fails. There is a set of *spare* components, indexed by $i \in \mathcal{I} = \{1,\dots,I\}$, that can be used to operate the system. Component i has a deterministic cost v_i and a random operating life s_i. The operating lives are independent random variables with known, arbitrary distributions. The actual operating life of a component is unknown at the time of replacement. The goal is to minimize the expected cost of operating the system for κ units of time. As in Derman et al. (1978), we guarantee that the problem is feasible by assuming that there is an infinite supply of one type of component. These components are indexed by $i > I$ and assumed to have deterministic cost $\bar{c}$ and operating life $\bar{s}$. We let $\mathcal{I}^+ = \{1,\dots\}$ denote the indices of all components with indices $1,\dots,I$ corresponding to the I spare components.

As before, a policy $\alpha \in \mathcal{A}$ is a mapping that determines the next component $\alpha(\mathcal{S}, c)$ to be selected given the set of remaining spare components $\mathcal{S} \subseteq \mathcal{I}$ and the remaining time c. We denote the decision epochs by $t \ge 1$. To simplify the notation, we let α_t denote the tth

component used under policy α. We let $\tau^\alpha = \inf\{t \geq 1 : \sum_{j=1}^{t} s_{\alpha_j} \geq \kappa\}$ be the total number of components required by policy α to operate the system until the finish date. We can then write the problem as[4]

$$V^* = \min_{\alpha \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{\tau^\alpha} v_{\alpha_t}\right].$$

Consider the approximation to the continuation value given in Equation (8). The greedy policy induced by this approximation sorts the components in *increasing* order of cost per expected life, $r_i \triangleq v_i/\mathbb{E}[s_i]$, and selects components for replacement until the target time is reached. Without loss of generality, we assume that components are sorted in increasing order of this ratio. The expected performance of the greedy policy is denoted by $V^{\mathrm{G}} = \mathbb{E}\left[\sum_{t=1}^{\tau^{\mathrm{G}}} v_t\right]$, where τ^{G} is the total number of components used by the greedy policy to operate the system until the finish date.

As in the knapsack problem, the perfect information bound may be quite loose: by knowing the realizations of all operating lives in advance, the DM with perfect information can avoid selecting components with short realized operating lives. To improve the lower bound, we impose a penalty that punishes violations of the nonanticipativity constraints. Equation (6) implies that the period-t penalty induced by Equation (8) is $z_t(c, i, s_i) = r_i(\mathbb{E}[s_i] - s_i)$. We let $V_z^{\mathrm{P}}(\mathbf{s})$ denote the optimal (deterministic) value of the penalized perfect information problem for sample path $\mathbf{s} \in \mathbb{R}_+^I$. By Proposition 1, we obtain a lower bound $V_z^{\mathrm{P}} \leq V^*$, where $V_z^{\mathrm{P}} = \mathbb{E}_{\mathbf{s}}[V_z^{\mathrm{P}}(\mathbf{s})]$ denotes the penalized perfect information bound.

Using a similar analysis as in the knapsack problem, we can compare the penalized performance of the greedy policy, denoted by $V_z^{\mathrm{G}}(\mathbf{s})$, with the performance of the penalized perfect information problem in every sample path. Unlike Proposition 2, the bound has a single loss term because the cost of the component used at termination counts toward the objective.

Proposition 3. *For every sample path* $\mathbf{s}$*, the penalized performance of the greedy policy satisfies*

$$V_z^{\mathrm{G}}(\mathbf{s}) - V_z^{\mathrm{P}}(\mathbf{s}) \leq \max_{i \in \mathcal{I}^+} r_i s_i \,. \tag{12}$$

Taking expectations in Equation (12) and using the duality results of Section 2, we obtain the following guarantees on the performance of the greedy policy:

$$V^{\mathrm{G}} - V^* \leq V^{\mathrm{G}} - V_z^{\mathrm{P}} \leq \mathbb{E}\left[\max_{i \in \mathcal{I}^+} r_i s_i\right].$$

Thus, the greedy policy is asymptotically optimal in the sense that $(V^{\mathrm{G}} - V^*)/\kappa \to 0$ as $\kappa \to \infty$ when the expected maximum penalized cost grows more slowly than the finishing date. Because the components $i > I$ have deterministic operating lives, the conditions provided in Corollary 2 on the distributions of s_i for the I spare components are sufficient to guarantee asymptotic optimality.

4. Stochastic Scheduling on Parallel Machines

We consider the problem of scheduling a set of jobs on identical parallel machines with the objective of minimizing the total weighted completion time when no preemptions are allowed. Job processing times are stochastic, and the processing time of each job is not fully known until a job is completed. Formally, consider a set of jobs, indexed by $j \in \mathcal{J} = \{1, \ldots, J\}$, to be scheduled on M identical parallel machines. The processing time of job $j \in \mathcal{J}$ is independent of the machine and denoted by the random variable p_j. Job processing times are assumed to be independent (but not necessarily identical) with finite means. We let C_j denote the completion time of job $j \in \mathcal{J}$; that is, C_j equals the waiting time until processing of j starts plus the processing time p_j of j. Each job has an associated weight w_j, and the objective is to minimize the expected total weighted completion time $\mathbb{E}[\sum_{j=1}^{J} w_j C_j]$. Using Graham's notation, the problem can be written as $PM//\mathbb{E}[\sum_j w_j C_j]$ (see Pinedo 2012).

A policy $\alpha \in \mathcal{A}$ is a sequence of mappings that determine the job to be processed, denoted by $\alpha(t, \mathcal{W}, \mathcal{P}, \mathbf{s})$ at time t given the set of waiting jobs $\mathcal{W} \subseteq \mathcal{J}$, the set of jobs $\mathcal{P} \subseteq \mathcal{J}$ currently in process, and the amount of processing currently elapsed $\mathbf{s} \in \mathbb{R}_+^J$ on each job in process. We let $\mathcal{W}_t^\alpha \subseteq \mathcal{J}$ denote the subset of jobs waiting for service at time t and $\mathcal{P}_t^\alpha \subseteq \mathcal{J}$ denote the subset of jobs under process at time t using policy α. Denoting the time when all jobs are completed by $\tau^\alpha = \inf\{t \geq 0 : \mathcal{W}_t^\alpha \cup \mathcal{P}_t^\alpha = \emptyset\}$, the completion time of job $j \in \mathcal{J}$ is given by $C_j^\alpha = \int_0^{\tau^\alpha} \mathbf{1}\{j \in \mathcal{W}_t^\alpha \cup \mathcal{P}_t^\alpha\}\, \mathrm{d}t$. We restrict attention to policies satisfying $\mathbb{E}\tau^\alpha < \infty$. The problem can be written as

$$V^* = \min_{\alpha \in \mathcal{A}} \mathbb{E}\left[\sum_{j=1}^{J} w_j C_j^\alpha\right].$$

Although this problem is perhaps most naturally described as a continuous-time DP, we can equivalently write the problem as a discrete-time DP by discretizing all job processing time distributions; decision epochs then correspond to possible completion times of some job. Discretization can be done with arbitrary precision and is standard in the scheduling literature (see, e.g., Skutella et al. 2016). Nonetheless, weak duality (Proposition 1) relies primarily on using relaxed sets of policies (by relaxing the nonanticipativity constraints) and still leads to valid performance bounds for continuous-time problems.

4.1. WSEPT Policy and Single Machine Approximation

It is well known (Rothkopf 1966) that the WSEPT policy is optimal in the case of a single machine. This policy sorts the jobs in decreasing order of weight per *expected* processing time $r_j \triangleq w_j/\mathbb{E}[p_j]$ and then schedules the jobs in this order without idling. With multiple machines, this is no longer true, and it may even be optimal to idle in some states (for examples, see Uetz 2003). When there are many jobs relative to machines, we may nonetheless expect the policy that schedules jobs in WSEPT order in nonidling fashion to perform well (we simply refer to this as the WSEPT policy). In this regime, the system is heavily loaded, machines will nearly always be processing, and we may expect with an optimal policy that the total processing time will be shared approximately equally across all M machines. Thus, when there are many jobs, the system behavior resembles that of the same problem but with a single machine that is M times faster.

We consider approximations of the optimal continuation value inspired by this intuition. Assume that jobs are sorted by decreasing order of weight per expected processing time; that is, $r_1 \geq r_2 \geq \ldots \geq r_J$. We consider the following approximate continuation value when the set $\mathcal{W} \subseteq \mathcal{J}$ of jobs remains to be processed, and we assign job $j \in \mathcal{W}$ at time t, and its realized processing time is p_j:

$$Q_t(\mathcal{W}, j, p_j) = \frac{1}{M} w_j p_j + \frac{1}{M} \sum_{\ell \in \mathcal{W}\setminus\{j\}} w_\ell \left(p_j + \sum_{k \in \mathcal{W}\setminus\{j\},\, k \leq \ell} \mathbb{E}[p_k] \right). \tag{13}$$

The first term in Equation (13) captures the weighted completion time remaining for the currently selected job j. The second term captures the expected weighted completion time of the remaining jobs $\mathcal{W} \setminus \{j\}$ under the approximation that these jobs are processed optimally by a single machine (hence, the processing of these remaining jobs by label order, which corresponds to WSEPT). We scale the processing times by $1/M$ to reflect the fact that the single machine in this approximation is M times faster.

Because Equation (13) is the optimal continuation value for a single machine problem, the heuristic policy that is greedy with respect to Equation (13) corresponds to the WSEPT policy; in other words, with jobs indexed according to their WSEPT order, we have $\min\{j \in \mathcal{W}\} \in \arg\min_{j \in \mathcal{W}} \mathbb{E}_{\tilde{p}_j}[Q_t(\mathcal{W}, j, \tilde{p}_j)]$ (note that the greedy policy does not consider the option of idling an available machine, which may be optimal). We let V^{G} denote the expected performance of this policy; that is, $V^{\mathrm{G}} = \mathbb{E}[\sum_{j=1}^{J} w_j C_j^{\mathrm{G}}]$, where C_j^{G} denotes the completion time of job $j \in \mathcal{J}$ under the WSEPT policy. The goal is to compare the performance V^{G} using the WSEPT policy to the performance V^* using an optimal policy.

4.2. Perfect Information Bound

Consider a clairvoyant with access to all future realizations of the processing times $\mathbf{p} = (p_j)_{j=1}^{J}$. Given a sample path $\mathbf{p} \in \mathbb{R}_+^J$, we let $V^{\mathrm{P}}(\mathbf{p})$ denote the optimal (deterministic) total weighted completion time with perfect information. The expected value $V^{\mathrm{P}} = \mathbb{E}_{\mathbf{p}}[V^{\mathrm{P}}(\mathbf{p})]$ is the perfect information bound, which in this problem is a lower bound for the optimal performance; that is, $V^{\mathrm{P}} \leq V^*$.

The perfect information bound may be loose in general because there can be substantial benefit to knowing the realized processing times in advance. To illustrate this, we consider a simple example with one machine and J jobs with weight one, and each jobs' processing time is either ϵ or one, each with probability 1/2. Because the jobs are a priori identical, the problem is trivial, and it is easy to show that $V^{\mathrm{G}} = V^* = (1+\epsilon)J(J+1)/4$. On the other hand, in the perfect information problem, it is optimal to first schedule every short job with a realized processing time of ϵ. Let $I_t = \sum_{j=1}^{J} \mathbf{1}\{p_j = t\}$ denote the number of jobs with processing time $t \in \{\epsilon, 1\}$, respectively. The total completion time of the short jobs is $\epsilon I_\epsilon(I_\epsilon+1)/2$, and the total completion time of the long jobs is $\epsilon I_\epsilon I_1 + I_1(I_1+1)/2$. Taking expectations and using the fact that $I_\epsilon + I_1 = J$ together with the fact that I_1 is binomially distributed with J trials and success probability 1/2 because jobs are independent leads us to the poor lower bound of $V^{\mathrm{P}} = J(J+3+\epsilon(3J+1))/8$. For J large, this lower bound is off from V^* by nearly a factor of two.

4.3. Penalized Perfect Information Bound

With the single machine approximate continuation value given in Equation (13), the resulting penalty according to Equation (6) for assigning job j at time t is given by $z_t(\mathcal{W}, j, p_j) = \frac{1}{M}\sum_{k \in \mathcal{W}} w_k(\mathbb{E}[p_j] - p_j)$. For a given sample path $\mathbf{p}$, we let $A(\mathbf{p})$ denote the set of feasible deterministic scheduling policies in the perfect information problem. For a given feasible schedule $\mathbf{a} \in A(\mathbf{p})$, after rearranging sums, we can write the total penalty as

$$z(\mathbf{a}) = \frac{1}{M} \sum_{j \in \mathcal{J}} w_j \left(\sum_{i \preceq_{\mathbf{a}} j} \mathbb{E}[p_i] - p_i \right),$$

where we use $i \prec_{\mathbf{a}} j$ to denote all jobs i preceding j using $\mathbf{a}$ (and $i \preceq_{\mathbf{a}} j$ to also include j itself in a summation). We can then write the penalized perfect information problem as

$$\begin{aligned} V_z^{\mathrm{P}}(\mathbf{p}) &= \min_{\mathbf{a} \in A(\mathbf{p})} \sum_{j \in \mathcal{J}} w_j C_j^{\mathbf{a}} + z(\mathbf{a}) \\ &= \min_{\mathbf{a} \in A(\mathbf{p})} \frac{1}{M} \sum_{j \in \mathcal{J}} w_j \sum_{i \preceq_{\mathbf{a}} j} \mathbb{E}[p_i] + \sum_{j \in \mathcal{J}} w_j \left(C_j^{\mathbf{a}} - \frac{1}{M} \sum_{i \preceq_{\mathbf{a}} j} p_i \right). \end{aligned} \tag{14}$$

The first term represents the expected weighted completion time when jobs are processed in the order given by $\mathbf{a}$ by a single machine that is M times faster. The second term is an error term, which can be interpreted as the difference between the weighted completion time in the given sample path in the actual problem (i.e., with M machines) and the weighted completion time in the single machine approximation. If only the first term were present in this objective, the WSEPT policy would be optimal in the penalized perfect information problem: this follows from the fact that this term corresponds to the expected value for a single machine–scheduling problem. In fact, if we return to the one-machine example in Section 4.2, the penalized perfect information bound is tight in every sample path. In our analysis, we effectively bound the second (i.e., error) term, which allows us to conclude under mild conditions that the WSEPT policy is asymptotically optimal in the regime of many jobs.

4.4. Performance Analysis

We let $V_z^{\mathrm{G}}(\mathbf{p})$ denote the penalized value of the greedy policy given sample path $\mathbf{p}$; following the discussion in Section 2.1, we have $\mathbb{E}_{\mathbf{p}}[V_z^{\mathrm{G}}(\mathbf{p})] = V^{\mathrm{G}}$. Using the penalty described previously leads us to the following result.

Proposition 4. *For every sample path* $\mathbf{p}$*, the WSEPT policy satisfies*

$$V_z^{\mathrm{G}}(\mathbf{p}) - V_z^{\mathrm{P}}(\mathbf{p}) \le \left(\frac{M-1}{M}\right) \sum_{j \in \mathcal{J}} w_j \max_{i \in \mathcal{J}} p_i. \tag{15}$$

We show Equation (15) by bounding from below the second term in the penalized perfect information objective in Equation (14). We do this by using a known fact from deterministic scheduling: for any perfect information-scheduling policy $\mathbf{a} \in A(\mathbf{p})$, the start time $S_j^{\mathbf{a}}$ for any job j must satisfy

$$S_j^{\mathbf{a}} \ge \frac{1}{M} \sum_{i \prec_{\mathbf{a}} j} p_i - \frac{M-1}{M} \cdot \max_{i \in \mathcal{J}} p_i.$$

This leads to a lower bound on the second term in Equation (14) that is independent of $\mathbf{a}$; the result then follows from the fact that the WSEPT policy is optimal for the first term in Equation (14) in isolation.

If we consider a scaling of the number of jobs J, where job weights and expected processing times are bounded away from zero, then the optimal cost V^* scales quadratically with J. Under mild conditions, the gap implied by Proposition 4 grows slower than quadratically in J; for example, if job weights are uniformly bounded and processing times are bounded, then the gap implied by Proposition 4 is linear in J, which implies that the WSEPT policy is asymptotically optimal as J grows large.

Taking expectations of Equation (15) and using Proposition 1 leads to the following:

Corollary 3. *The performance of the WSEPT policy satisfies*:

1. *Performance guarantee.*

$$V^{\mathrm{G}} - V^* \le V^{\mathrm{G}} - V_z^{\mathrm{P}} \le \left(\frac{M-1}{M}\right) \sum_{j \in \mathcal{J}} w_j \mathbb{E}\left[\max_{i \in \mathcal{J}} p_i\right]. \tag{16}$$

2. *Asymptotic optimality. If weights are uniformly bounded, that is,* $w_j \le \bar{w}$ *for some* $\bar{w} < \infty$ *independent of the number of jobs* J *and* $\lim_{J\to\infty} \frac{1}{J}\mathbb{E}[\max_{j\in\mathcal{J}} p_j] = 0$*, then*

$$\lim_{J \to \infty} \frac{1}{J^2}(V^{\mathrm{G}} - V^*) = 0\,. \tag{17}$$

As with the stochastic knapsack problem, we can state conditions on the processing time distributions similar to those in Corollary 2 that would ensure that $\lim_{J\to\infty} \frac{1}{J}\mathbb{E}[\max_{j\in\mathcal{J}} p_j] = 0$ (e.g., this condition holds provided the mean and variance of the processing times are uniformly bounded). Note that, in the case of $M = 1$, Corollary 3, part 1, shows that WSEPT is optimal and the penalized perfect information bound is tight (in every sample path, according to Proposition 4). Following the discussion on strong duality in Section 2.1, this is to be expected: with $M = 1$ the continuation value Equation (13) is optimal, and hence, the resulting penalty is ideal.

Proposition 3.1 of Weiss (1990) is similar to Equation (16), but Equation (16) is stronger by a factor of two. Using a different approach involving valid inequalities for the performance space of all feasible scheduling policies, Möhring et al. (1999) show that the WSEPT policy satisfies

$$V^{\mathrm{G}} - V^* \le \frac{(M-1)(1+\rho)}{2M} \sum_{j \in \mathcal{J}} w_j \mathbb{E}[p_j], \tag{18}$$

where ρ is an upper bound on the squared coefficient of variation of job processing times. We show in Online Appendix C that we can also establish Equation (18) using penalized perfect information analysis, albeit with a different penalty. Although Equation (18) is a tighter bound than Equation (16) in many cases, this is not always true. For example, when jobs are identical and processing times are Bernoulli with probability q, Equation (16) is tighter than Equation (18) whenever $q \le 1 - (1/2)^{1/J}$.

5. Sequential Search for the Best Alternatives

We consider a variation of the sequential search problem studied by Weitzman (1979). In this problem, a DM—for example, a firm wanting to hire new employees or an investor considering real estate options—sequentially explores a given set of alternatives in an effort to find the most valuable ones. Exploration is costly but provides information about the value of

an alternative. The DM can either stop the search and select a previously explored alternative or continue searching. The model considered in Weitzman (1979) assumes the DM can only accept a single alternative. In many applications of this basic model, the DM may, in fact, have the ability to accept multiple alternatives. For example, firms may wish to hire several employees from a batch of applicants; we consider this variation.

We formally describe the model with a finite number N of a priori identical alternatives. Associated with each alternative is an unknown reward; we let r_n denote the random reward associated with alternative n and assume that rewards are nonnegative, independent, and identically distributed with finite mean. At the beginning of each time period, the DM can pay a search cost $s \geq 0$ to explore an alternative, which reveals the alternative's reward. At the end of each time period, the DM can select any available (i.e., previously explored but not yet selected) alternatives, thereby collecting their associated rewards. Rewards are discounted in each period according to a discount factor $\delta \in (0, 1]$. The DM may explore as many alternatives as desired but, because of capacity limitations, can select at most K alternatives in total. The problem ends when K alternatives have been selected or the DM stops exploring (because of choice or because no capacity remains or all N alternatives have been explored). To avoid trivialities, we assume that $\delta \mathbb{E}_r[r] \geq s$ as, otherwise, the DM has no incentive to explore any alternatives. The goal is to maximize the expected discounted reward, net of search costs, associated with selecting alternatives.

We denote decision epochs by $t = 1, \ldots, N$. There are two decisions per time period: (1) at the beginning of the period, whether to explore another alternative or stop the search and (2) at the end of the period, which available alternatives, if any, to select. A feasible policy $\alpha \in \mathcal{A}$ is given by a sequence of functions $(\alpha_t^{\mathrm{E}})_{t=1}^N$ and $(\alpha_t^{\mathrm{S}})_{t=1}^N$, each taking as input the remaining capacity k and the set of rewards $\mathcal{R}$ associated with available alternatives. The function α_t^{E} maps to $\{EXPLORE, STOP\}$, indicating whether the DM explores another alternative or stops the search at the beginning of period t. The function α_t^{S} maps to a set $\mathcal{S} \subseteq \mathcal{R}$, where $|\mathcal{S}| \leq k$, of available rewards to select at the end of period $t < N$; we assume α_N^{S} maps to a set $\mathcal{S}$ such that $|\mathcal{S}| = k$ because any optimal policy would use any remaining capacity to select the highest remaining rewards at the end of the horizon.

For a given $\alpha \in \mathcal{A}$, we let k_t^α denote the remaining selection capacity and let $\mathcal{S}_t^\alpha$ denote the set of rewards selected at time t. Capacity evolves according to $k_{t+1}^\alpha = k_t^\alpha - |\mathcal{S}_t^\alpha|$ with $k_1^\alpha = K$. We let τ^α denote the stopping time corresponding to the first time capacity runs out or the DM stops the search. We can then write the problem as

$$V^* = \max_{\alpha \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{N \wedge \tau^\alpha} \left(\delta^t \sum_{r \in \mathcal{S}_t^\alpha} r - \delta^{t-1} s\right)\right].$$

In this problem, alternatives are a priori identical, so which alternative to explore next is irrelevant. In contrast, selection decisions in each period present a complex trade-off between (1) immediately selecting available alternatives and (2) preserving capacity to select potentially more valuable alternatives at the expense of further discounting and additional search costs.

5.1. Approximate Continuation Value and Greedy Policy

In the case of $K = 1$, Weitzman (1979) shows that a threshold policy can be used to determine whether to select the best available alternative; specifically, Weitzman (1979) shows that when $t < N$, it is optimal to select the highest available reward (and, hence, stop before searching all alternatives) if and only if this reward is above a *reservation price* v^* given by the solution to $v^* = \delta \mathbb{E}_r[\max(r, v^*)] - s$. The value v^* represents the indifference point between immediate selection and continuing the search one more period and is equivalent to the Gittins index for an unexplored alternative. The problem is more complicated when $K > 1$: optimal policies may deplete capacity over time by intermittently recalling previously explored alternatives, and optimal selection thresholds in general may depend on remaining capacity k, the largest k available rewards, and time.

We instead consider a greedy policy based on reservation prices that only depend on remaining capacity; these reservation prices will be nonincreasing with capacity, reflecting a DM who is more discriminating as capacity diminishes. With this assumption, the greedy policy will either select an alternative immediately after it is explored or (possibly) at the end of the horizon. We let $v_j \geq 0$ denote the reservation price for the jth last unit of capacity (i.e., v_1 corresponds to the final unit of capacity) and interpret v_j as an approximation of the marginal value of this unit of capacity with an optimal policy.

Because the DM has two decisions per time period, we use approximate continuation values Q_t^S and Q_t^E for exploration and selection, respectively. It is easiest to write our approximation denoting a selection decision $a_t \in \{0, 1\}$, which indicates whether the most recently explored alternative with reward r_t is selected at time t. Letting k denote the capacity remaining prior to this selection, we approximate the continuation value of selection as

$$Q_t^{\mathrm{S}}(k, r_t, a_t) = r_t a_t + (1 - a_t) v_k + \sum_{j=1}^{k-1} v_j. \tag{19}$$

With Equation (19), the greedy policy $a_t \in \arg\max_{a \in \{0,1\}} \cdot Q_t^{\mathrm{S}}(k, r_t, a)$ selects the current alternative if and only

if its reward r_t is no smaller than the marginal value (or reservation price) v_k.[5] The remaining term $\sum_{j=1}^{k-1} v_j$ captures the approximate net present value of all remaining capacity. Including the search cost and discounting, we can thus approximate the continuation value of exploration as

$$Q_t^{\mathrm{E}}(k, r_t) = \delta \max(r_t, v_k) + \delta \sum_{j=1}^{k-1} v_j - s\,. \tag{20}$$

Because $v_j \geq 0$, the assumption $\delta \mathbb{E}_r[r] \geq s$ implies that $\mathbb{E}_r[Q_t^{\mathrm{E}}(k, r)] \geq 0$. Thus, taking the value of stopping search equal to zero, the greedy policy will continue exploring as long as some positive capacity remains.

Choice of Reservation Prices. Although we can consider a greedy policy based on any reservation prices, we show that with a particular choice of reservation prices, the greedy policy is asymptotically optimal when the number N of alternatives grows large.

Our main focus is on the case when capacity grows more slowly than the number of alternatives; that is, $K = o(N)$. In this regime, capacity is precious, and good selection policies must balance selecting an alternative against the net present value of many possible future search opportunities. (In Section 5.5, we consider the regime in which capacity grows proportionally with the number of alternatives.) Motivated by this intuition, we consider an approximation in which the number of alternatives is infinite and the DM never recalls previously explored alternatives. Optimal policies in this approximation use reservation prices v_k that only depend on remaining capacity k and are defined recursively from $k = 1$ to K as

$$v_k + \sum_{j=1}^{k-1} v_j = \delta \mathbb{E}_r[\max(r, v_k)] - s + \delta \sum_{j=1}^{k-1} v_j\,. \tag{21}$$

The value $\sum_{j=1}^{k} v_j$ represents the optimal net present value with k units of capacity in the infinite alternative model. Given values $v_1, \ldots, v_{k-1}$, we can thus interpret Equation (21) as describing the value v_k that makes the DM indifferent between (1) immediate selection of an alternative with reward v_k (and using the kth last unit of capacity) and (2) exploring another alternative (and preserving the kth last unit of capacity for at least one more period). The term $\sum_{j=1}^{k-1} v_j$ is discounted on the right because preserving the kth last unit of capacity delays use of all remaining capacity by an additional period. Because $\delta \mathbb{E}_r[r] \geq s$, it is not hard to see there exist reservation prices v_k satisfying Equation (21) that are nonincreasing in k. In addition, because reservation prices are (weakly) increasing as remaining capacity decreases, it can never be optimal to recall previously explored alternatives using this approximation. The reservation price v_1 for the last alternative to select corresponds to the reservation price v^* in Weitzman (1979). When $\delta = 1$, the reservation prices satisfy $v_k = v^*$ for every k, where $v^* = \mathbb{E}_r[\max(r, v^*)] - s$, that is, again the optimal reservation price in Weitzman (1979) but with $\delta = 1$.

We let V^{G} denote the expected performance of the greedy policy with v_k chosen as in Equation (21).

5.2. Perfect Information Bound

Consider a clairvoyant with access to all future realizations of the rewards $\mathbf{r} = (r_n)_{n=1}^N$ before exploring or selecting any alternatives. Given a sample path $\mathbf{r} \in \mathbb{R}_+^N$, we let $V^{\mathrm{P}}(\mathbf{r})$ denote the optimal (deterministic) value for sample path $\mathbf{r}$ with perfect information about rewards. The expected value $V^{\mathrm{P}} = \mathbb{E}_{\mathbf{r}}[V^{\mathrm{P}}(\mathbf{r})]$ is an upper bound for the optimal performance; that is, $V^* \leq V^{\mathrm{P}}$.

With perfect information, unexplored alternatives are no longer a priori identical and the DM can "explore" and then select alternatives known to have large rewards.[6] The DM with perfect information would always select an alternative immediately after exploring it, and the perfect information problem is equivalent to the deterministic assignment problem

$$\begin{aligned} V^{\mathrm{P}}(\mathbf{r}) = \max_{\mathbf{x} \in \{0,1\}^N} & \sum_{t=1}^{N} \sum_{n=1}^{N} \delta^{t-1}(\delta r_n - s) x_{n,t} \\ \text{s.t.} \quad & \sum_{n=1}^{N} x_{n,t} \leq 1 \qquad \forall t, && \text{(22a)} \\ & \sum_{t=1}^{N} x_{n,t} \leq 1 \qquad \forall n, && \text{(22b)} \\ & \sum_{t=1}^{N} \sum_{n=1}^{N} x_{n,t} \leq K, && \text{(22c)} \end{aligned}$$

where $x_{n,t} \in \{0, 1\}$ indicates whether alternative n is selected in period t. Equation (22a) ensures that at most one alternative can be selected per time period, (22b) ensures that each alternative can be selected at most once, and (22c) ensures that the DM can select at most K alternatives.

The perfect information bound may be quite loose as the DM is free to only select alternatives with large rewards in every sample path. For example, consider the problem when $K = N$ and $\delta = 1$. The problem is trivial, and $V^* = N(\mathbb{E}_r[r] - s)$ because the optimal policy explores and selects all alternatives. On the other hand, in the perfect information problem, it is optimal to select only alternatives with realized rewards greater than the search cost. This leads to the weak upper bound of $V^{\mathrm{P}} = N\mathbb{E}_r[(r - s)^+]$.

5.3. Penalized Perfect Information Bound

To improve the upper bound, we impose a penalty that punishes violations of the nonanticipativity constraints. We construct penalties according to Equation (6), using the approximate continuation values Q_t^{S} and Q_t^{E} inducing the greedy policy. Selection decisions are induced by

$Q_t^S(k, r_t, a_t)$, which is measurable with respect to past actions and states and, thus, leads to zero penalty. The approximate continuation value for exploration, however, does lead to a useful penalty: Equation (6) implies that the period-t penalty induced by Equation (20) is $z_t(k, r_t) = \delta \max(r_t, v_k) - \delta\mathbb{E}_{\tilde{r}}[\max(\tilde{r}, v_k)] = \delta(r_t - v_k)^+ - \delta\mathbb{E}_{\tilde{r}}[(\tilde{r} - v_k)^+]$. Using $\mathbf{a}$ to represent the vector of exploration and selection decisions and $\tau^{\mathbf{a}}$ to denote the time the search ends with $\mathbf{a}$, the total penalty is

$$z(\mathbf{a}) = \sum_{t=1}^{N\wedge\tau^{\mathbf{a}}} \delta^t((r_t - v_{k_t^{\mathbf{a}}})^+ - \mathbb{E}_{\tilde{r}}[(\tilde{r} - v_{k_t^{\mathbf{a}}})^+]).$$

We let $V_z^{\mathrm{P}}(\mathbf{r})$ denote the optimal (deterministic) value of the penalized perfect information problem for sample path $\mathbf{r} \in \mathbb{R}_+^N$. From Proposition 1, we have $V^* \le V_z^{\mathrm{P}}$.

As discussed, the DM with perfect information may "cheat" by selecting alternatives known to have large rewards. This penalty helps to align the perfect information DM with the greedy policy by "punishing" the perfect information DM for selecting alternatives with realized rewards (net of reservation prices) that are large relative to their expected value. Thus, although the DM in the penalized perfect information problem may still "cheat" and select alternatives with large realized rewards, the penalty reduces the benefit of such cheating.

It is instructive to see how this works on the example from Section 5.2. Recall that the optimal value is $V^* = N(\mathbb{E}_r[r] - s)$, but the perfect information bound without penalty provides the bound of $V^{\mathrm{P}} = N\mathbb{E}_r[(r - s)^+]$. With $\delta = 1$, the reservation prices in Equation (21) reduce to those as in Weitzman (1979); that is, $v_k = v^*$ with $\mathbb{E}_r[(r - v^*)^+] = s$. In the penalized problem, the penalized reward for selecting alternative n is $r_n - s - (r_n - v^*)^+ + \mathbb{E}_r[(r - v^*)^+] = \min(r_n, v^*)$, and the payoff for exploring but not selecting the alternative is $-s - (r_n - v^*)^+ + \mathbb{E}_r[(r - v^*)^+] = -(r_n - v^*)^+$. Moreover, because $v^* \ge 0$, we conclude that the DM with perfect information selects all alternatives in every sample path. Because these actions coincide with the optimal policy in every sample path, we conclude that $V_z^{\mathrm{P}} = V^*$; that is, we recover a tight bound.

5.4. Performance Analysis

We now formalize this discussion in the general case. We show that the greedy policy incurs a small loss in value compared with the optimal policy when the number of alternatives is large and, in particular, that the greedy policy is asymptotically optimal under conditions that we make precise. We let $V_z^{\mathrm{G}}(\mathbf{r})$ denote the penalized value of the greedy policy for sample path $\mathbf{r}$.

The following result compares the penalized performance of the greedy policy with the performance of the penalized perfect information problem along each sample path. We let $\bar{k}^{\mathrm{G}}$ denote the number of alternatives, if any, "recalled" by the greedy policy at $t = N$; formally, we have $\bar{k}^{\mathrm{G}} = K - \sum_{t=1}^{N\wedge\tau^{\mathrm{G}}} \mathbf{1}\{r_t \ge v_{k_t^{\mathrm{G}}}\}$, where k_t^{G} denotes the capacity remaining at the start of period t using the greedy policy.

Proposition 5. *For every sample path* $\mathbf{r}$*, the penalized performance of the greedy policy satisfies*

$$V_z^{\mathrm{P}}(\mathbf{r}) - V_z^{\mathrm{G}}(\mathbf{r}) \le \delta^N \sum_{j=1}^{\bar{k}^{\mathrm{G}}} v_j. \tag{23}$$

Moreover, when $\delta = 1$*, we have* $V_z^{\mathrm{P}}(\mathbf{r}) = V_z^{\mathrm{G}}(\mathbf{r})$ *for every sample path* $\mathbf{r}$*.*

In the proof of Proposition 5, we show that, in the penalized perfect information problem, the value of the jth last unit of capacity is no larger than v_j in Equation (21) in every sample path regardless of the specifics (i.e., which alternatives are selected and when) of the selection decisions. Intuitively, even if an alternative with a large realized reward r_t is selected for this unit of capacity, the penalty $-\delta(r_t - v_j)^+ + \delta\mathbb{E}_r[(r - v_j)^+]$ compensates for this and prevents the selection value for this unit of capacity from exceeding v_j. The penalized performance of the greedy policy for the jth last unit of capacity also reduces to be v_j unless this capacity is fulfilled by an alternative that is recalled by the greedy policy at $t = N$. Combining these results leads us to Equation (23). We note that Equation (23) applies to the same problem when the DM cannot recall past alternatives because in proving Equation (23) we drop the value of any alternatives recalled by the greedy policy at $t = N$.

Another consequence of the proof of Proposition 5 is that $V_z^{\mathrm{P}}(\mathbf{r}) \le \sum_{j=1}^{K} v_j$; that is, the infinite alternative model without recall provides an upper bound, and $V_z^{\mathrm{P}}(\mathbf{r})$ provides a tighter upper bound in every sample path. Finally, as noted in Proposition 5, in the case of $\delta = 1$, these arguments can be strengthened to show that the optimal actions in the penalized perfect information problem coincide with those of the greedy policy, thus implying the greedy policy is, in fact, optimal in this case.

Using Proposition 1, we then obtain the following guarantees on the performance of the greedy policy.

Corollary 4. *The performance of the greedy policy satisfies*:

1. *Performance guarantee. When* $\delta = 1$, $V^{\mathrm{G}} = V^* = V_z^{\mathrm{P}}$*, and in general, we have*

$$V^* - V^{\mathrm{G}} \le V_z^{\mathrm{P}} - V^{\mathrm{G}} \le \delta^N \mathbb{E}\left[\sum_{j=1}^{\bar{k}^{\mathrm{G}}} v_j\right]. \tag{24}$$

2. *Asymptotic optimality. If either* (a) δ *is fixed or* (b) $K = o(N)$ *and* $\limsup_{N\to\infty} (1 - \delta)N < \infty$*, then*

$$\lim_{N\to\infty} \frac{1}{K}(V^* - V^{\mathrm{G}}) = 0.$$

Corollary 4 states that the greedy policy is asymptotically optimal when the discount factor is fixed (for any scaling of capacity) or if capacity grows at a slower rate than N and $\delta \to 1$. This latter scaling is arguably more interesting and can be interpreted as a setting with a fixed time horizon in which a large number of decisions needs to be made and the time between decisions is small. Intuitively, when the number of alternatives grows faster than capacity, the greedy policy will tend to exhaust capacity relatively early in the horizon, and the infinite alternative model provides a good approximation even if the DM is becoming increasingly patient. In the proof we argue, using a concentration inequality, that the probability that the greedy policy reaches the end of the horizon with positive capacity converges to zero as $N \to \infty$, and thus, the expected value of the gap in Equation (23) becomes vanishingly small relative to capacity. As a result, with either condition (a) or (b), the convergence rate to optimality is exponential in the number of alternatives.

5.5. Extensions

In this section, we discuss a generalization in which alternatives are not a priori identical and then discuss how our results extend to a regime in which initial capacity grows proportionally with the number of alternatives.

Multiple Types. In some applications, alternatives may not be a priori identical (e.g., an employer may expect performance levels to vary across applicants with different experiences and backgrounds). We consider a variant in which there is a set $\mathcal{I} = \{1, \dots, I\}$ of types for the alternatives. An alternative of type $i \in \mathcal{I}$ has reward drawn independent and identically distributed from a type-dependent distribution $F_i(\cdot)$ and search cost s_i. The DM has N opportunities to explore alternatives, and each type has an infinite supply of alternatives. In each period $t = 1, \dots, N$, the DM needs to determine which type of alternative to explore or stop the search.

As before, we consider an approximation of the problem in which the number of search opportunities is infinite and the DM never recalls previously explored alternatives. Reservation prices v_k only depend on remaining capacity k and are defined recursively from $k = 1$ to K as

$$v_k + \sum_{j=1}^{k-1} v_j = \max_{i \in \mathcal{I}} \left\{ \delta \mathbb{E}_{r_i} [\max(r_i, v_k)] - s_i \right\} + \delta \sum_{j=1}^{k-1} v_j , \quad (25)$$

where $\mathbb{E}_{r_i}$ denotes the expectation with respect to the reward distribution F_i for type i. Similar to Equation (21), the value v_k makes the DM indifferent between immediately selecting an alternative of reward v_k (and, hence, using the kth unit of capacity) and exploring another alternative (and preserving the kth unit of capacity for at least one more period). At a given level of capacity k, the greedy policy induced by Equation (25) will explore the same type until an alternative is selected; at that point, capacity is reduced by one, and the DM may choose to explore a different type of alternative.

It is not hard to show that the performance guarantees of Proposition 5 and Corollary 4 extend to this setting. In this variation, the greedy policy is again asymptotically optimal when the discount factor is fixed (for any scaling of capacity) or if capacity grows at a slower rate than N and $\delta \to 1$.

Large Capacity. Our analysis can be extended to a regime in which initial capacity grows proportionally with the number of alternatives; that is, $K = \Theta(N)$. In this regime, the greedy policy induced by the reservation prices given in Equation (21) is not asymptotically optimal. To see this, consider an example with $\delta = 1 - 1/N$ and $K = N$. Here, the optimal policy explores and selects all alternatives, implying that a greedy policy with $v = 0$ is optimal. It is not hard to see that the reservation prices given in Equation (21) are strictly positive during a nonvanishing fraction of the horizon. Thus, the greedy policy will postpone selecting a significant number of alternatives until the end of the horizon, which introduces a loss because of discounting; it can be shown that this loss does not go to zero even though $\delta \to 1$ in this case.

In Online Appendix D, we consider reservation prices that depend on time but not capacity; these reservation prices are based on an approximation in which the capacity constraint is satisfied in expectation. This approximation is predicated on the fact that, when the number of alternatives is large, the state trajectories of an optimal policy tend to concentrate around the expected trajectory. We use reservation prices of the form $v_t = \lambda \delta^{-t}$, where $\lambda \geq 0$ can be interpreted as the Lagrange multiplier of the capacity constraint. Because reservation prices are increasing with time, the resulting greedy policy will only select an alternative immediately after it is explored or at the end of the horizon.

Using these reservation prices within the penalties, we can bound from above the gap between the penalized perfect information problem and the penalized performance of the greedy policy and again conclude that the greedy policy is asymptotically as the number of alternatives grows.

6. Conclusion

The general recipe proposed here is broadly applicable to problems in which we have an underlying approximate value function of interest. Although penalized perfect information analysis weaves a common thread in the problems we study here, problem-specific analysis remains: we need to derive a bound between the

penalized perfect information problem and the (penalized) value of the greedy heuristic policy. The upside is that the analysis for this key step is deterministic.

For the key step just described, we leveraged results from the deterministic version of the problem. In the stochastic knapsack problem, we applied LP relaxations to the penalized perfect information problem. In the stochastic scheduling problem, we applied deterministic scheduling bounds on the job completion times in the penalized perfect information problem. In the sequential search problem, we viewed the penalized perfect information problem as a deterministic assignment problem and bounded the value of each unit of capacity.

Our main goal was to demonstrate the method on different problems. Although the problems we studied have some similarities, there are important differences. For example, total rewards in the knapsack or sequential search problems roughly scale linearly with capacity, whereas total costs scale quadratically with the number of jobs in the scheduling problem. The budget constraints in the knapsack and search problems are somewhat different: in the knapsack problem, values are deterministic, but capacity is consumed by stochastic sizes, whereas in the search problem, capacity consumption is deterministic, but rewards are stochastic.

The approximations used in each of the problems were relatively simple as were the resulting policies. It would be interesting to see if approximate continuation values with more complex state dependence could lead to better policies, tighter penalized information relaxation bounds, and perhaps faster rates of convergence to optimality. Variations of these specific models with more realistic and complex features would also be interesting to consider. Finally, and most importantly, we are hopeful that this method can be applied successfully and more broadly in the analysis of heuristic policies for many other stochastic DPs.

Appendix. Proofs

A.1. Proof of Proposition 2

We bound the penalized perfect information value from above in terms of the penalized performance of the greedy policy. Herein we write the penalized perfect information problem as an integer program; relaxing constraints (A.1b) and (A.1d), we obtain that problem (A.1) decouples in terms of the decision variables $\mathbf{x}$ and $\mathbf{y}$. Thus, we obtain the upper bound

$$\bar{V}_z^{\mathrm{P}}(\mathbf{s}) \le \underbrace{\max_{\mathbf{x}\in\{0,1\}^I} \ \sum_{i=1}^I r_i s_i x_i \quad \text{s.t.} \ \sum_{i=1}^I s_i x_i \le 1}_{\clubsuit} + \underbrace{\max_{\mathbf{y}\in\{0,1\}^I} \ \sum_{i=1}^I r_i(s_i - \mathbb{E}[s_i]) y_i \quad \text{s.t.} \ \sum_{i=1}^I y_i \le 1}_{\spadesuit}.$$

We conclude the proof by bounding each term at a time.

For the first problem, note that the ratio of value to size of each item is r_i as in the greedy policy. By considering the continuous relaxation to $x_i \in [0,1]$, we obtain that $x_i = 1$ for $i < \tau^{\mathrm{G}}$, and $x_i \in (0,1]$ for $i = \tau^{\mathrm{G}}$ whenever $\tau^{\mathrm{G}} \le I$ (and $x_i = 1$ for all i when $\tau^{\mathrm{G}} > I$). Rounding up to one the last fractional item and using that $r_i s_i = v_i + r_i(s_i - \mathbb{E}[s_i])$, we obtain the upper bound

$$\clubsuit \le \sum_{i=1}^{I\wedge\tau^{\mathrm{G}}} r_i s_i = \sum_{i=1}^{I\wedge(\tau^{\mathrm{G}}-1)} v_i + \sum_{i=1}^{I\wedge\tau^{\mathrm{G}}} r_i(s_i - \mathbb{E}[s_i]) + v_{\tau^{\mathrm{G}}} \mathbf{1}\{\tau^{\mathrm{G}} \le I\}$$
$$= V_z^{\mathrm{G}}(\mathbf{s}) + v_{\tau^{\mathrm{G}}} \mathbf{1}\{\tau^{\mathrm{G}} \le I\} \le V_z^{\mathrm{G}}(\mathbf{s}) + \max_{i\in\mathcal{I}} v_i .$$

For the second problem, note that the optimal solution selects the item with maximum objective, and thus,

$$\spadesuit = \max_{i\in\mathcal{I}} r_i(s_i - \mathbb{E}[s_i])^+ .$$

The result then follows.

Integer programming formulation for $\mathbf{V}_{\mathbf{z}}^{\mathrm{P}}(\mathbf{s})$: To calculate $V_z^{\mathrm{P}}(\mathbf{s})$, because the item that overflows the knapsack now counts toward the objective, we need to explicitly account for the overflowing item whenever it exists. We obtain an upper bound on the penalized perfect information problem for a fixed sample path $\mathbf{s}$ by solving the integer programming problem

$$\bar{V}_z^{\mathrm{P}}(\mathbf{s}) \triangleq \max_{\mathbf{x},\mathbf{y}\in\{0,1\}^I} \sum_{i=1}^I (v_i + r_i(s_i - \mathbb{E}[s_i]))x_i + \sum_{i=1}^I r_i(s_i - \mathbb{E}[s_i]) y_i$$
$$\text{s.t.} \sum_{i=1}^I s_i x_i \le \kappa , \tag{A.1a}$$
$$x_i + y_i \le 1 , \quad \forall i \in \mathcal{I} , \tag{A.1b}$$
$$\sum_{i=1}^I y_i \le 1 , \tag{A.1c}$$
$$\sum_{i=1}^I s_i(x_i + y_i) \ge \kappa(1 - x_i) , \quad \forall i \in \mathcal{I} , \tag{A.1d}$$

where $x_i \in \{0,1\}$ indicates whether the item is selected and fits the knapsack, and $y_i \in \{0,1\}$ indicates if the item overflows the knapsack. Constraint (A.1b) imposes that an item either fits the knapsack or overflows it. Constraint (A.1c) guarantees that there is at most one overflowing item. Constraint (A.1d) requires that the overflowing item, if one exists, causes the selected capacity to exceed the capacity of the knapsack. This constraint is vacuous when all items fit the knapsack, that is, if there is no overflow. Note that we can only be sure that $V_z^{\mathrm{P}}(\mathbf{s}) \le \bar{V}_z^{\mathrm{P}}(\mathbf{s})$ because the overflowing item y_i can be chosen to exactly match the capacity of the knapsack. For item y_i to actually overflow the knapsack, we need to make inequality (A.1d) strict. When the distribution of item sizes are absolutely continuous or lattice (i.e., there exists some $h > 0$ such that $\mathbb{P}\{s_i \in \{0, h, 2h, \ldots\}\} = 1$ for all $i \in \mathcal{I}$), replacing constraint (A.1d) by $\sum_{i=1}^I s_i(x_i + y_i) \ge (\kappa + \epsilon)(1 - x_i)$ for some $\epsilon > 0$ in problem (A.1) gives that $\bar{V}_z^{\mathrm{P}}(\mathbf{s}) = V_z^{\mathrm{P}}(\mathbf{s})$. The bound given by $\bar{V}_z^{\mathrm{P}}(\mathbf{s})$, however, suffices for our analysis.

A.2. Proof of Corollary 1

Part 1 of the result follows by taking expectations of Equation (10) and applying Proposition 1. Equation (11) follows because $\max_i v_i = \max_i \mathbb{E}[r_i s_i] \le \mathbb{E}[\max_i r_i s_i]$ from Jensen's inequality together with the fact that $(s_i - \mathbb{E}[s_i])^+ \le s_i$.

A.3. Proof of Corollary 2

Because the value-to-size ratios are bounded, it suffices to show that $E_I \triangleq \frac{1}{\kappa}\mathbb{E}[\max_{i=1,\dots,I} s_i] \to 0$ as $I \to \infty$. We prove each case separately.

Case (a). This case follows trivially because $E_I \le \bar{s}/\kappa$ and $\kappa = \omega(1)$.

Case (b). Using Lyapunov's inequality and linearity of expectations, we obtain that

$$E_I \le \frac{1}{\kappa}\mathbb{E}[\max_i |s_i|^p]^{1/p} \le \frac{1}{\kappa}\left(\sum_{i=1}^{I}\mathbb{E}[|s_i|^p]\right)^{1/p} \le \frac{(mI)^{1/p}}{\kappa}, \quad \text{(A.2)}$$

which converges to zero because $\kappa = \omega(I^{1/p})$.

Case (c). Because s_i is sub-Gaussian, using the tail formula for expectations, we obtain that

$$\begin{aligned}\mathbb{E}[|s_i|^p] &= \int_0^\infty px^{p-1}\mathbb{P}\{s_i > x\}\mathrm{d}x \le a\int_0^\infty px^{p-1}\exp(-bx^2)\mathrm{d}x \\ &= \frac{p}{2b^{p/2}}\Gamma(p/2).\end{aligned}$$

Stirling's approximation implies that there exists a constant $c>0$ such that $\mathbb{E}[|s_i|^p]^{1/p} \le c\sqrt{p}$. Using Equation (A.2) and setting $p = \log(I)$, we obtain that $E_I \to 0$ as $I \to \infty$ when $\kappa = \omega(\sqrt{\log(I)})$.

A.4. Proof of Proposition 3

We bound the penalized perfect information value from below in terms of the penalized performance of the greedy policy for a fixed sample path **s**. The penalized perfect information problem can be written as

$$\begin{aligned}V_z^{\mathrm{P}}(\mathbf{s}) = \min_{\mathbf{x}\in\{0,1\}^\infty} &\sum_{i=1}^{\infty}(v_i + r_i(s_i - \mathbb{E}[s_i]))x_i \\ \text{s.t.} &\sum_{i=1}^{\infty} s_i x_i \ge \kappa,\end{aligned}$$

where $x_i \in \{0,1\}$ indicates whether the component is selected. Note that the ratio of cost to operating life of each component is r_i as in the greedy policy. By considering the continuous relaxation to $x_i \in [0,1]$, we obtain that $x_i = 1$ for $i < \tau^{\mathrm{G}}$, and $x_i \in (0,1]$ for $i = \tau^{\mathrm{G}}$. Rounding down to zero the last fractional item and using that $r_i s_i = v_i + r_i(s_i - \mathbb{E}[s_i])$, we obtain the lower bound

$$V_z^{\mathrm{P}}(\mathbf{s}) \ge \sum_{i=1}^{\tau^{\mathrm{G}}-1} r_i s_i = V_z^{\mathrm{G}}(\mathbf{s}) - r_{\tau^{\mathrm{G}}}s_{\tau^{\mathrm{G}}} \ge V_z^{\mathrm{G}}(\mathbf{s}) - \max_{i\in\mathcal{J}^+} r_i s_i,$$

where the first equality follows because the penalized performance of the greedy policy is $V_z^{\mathrm{G}}(\mathbf{s}) = \sum_{i=1}^{\tau^{\mathrm{G}}} r_i s_i$. The result then follows.

A.5. Proof of Proposition 4

Fix a sample path **p** and consider any feasible action $\boldsymbol{a} \in A(\mathbf{p})$ in the perfect information problem. Letting S_j^a denote the start time of a given job j using $\boldsymbol{a}$, we first note that

$$S_j^a \ge \frac{1}{M}\sum_{i \prec_a j} p_i - \frac{M-1}{M}\max_{i\in\mathcal{J}} p_i, \quad \text{(A.3)}$$

which is a known result from deterministic scheduling (see, e.g., the proof of proposition 3.1 in Weiss 1990). From Equation (14), we can then bound the penalized perfect information objective from below:

$$\begin{aligned}V_z^{\mathrm{P}}(\mathbf{p}) &= \min_{\boldsymbol{a}\in A(\mathbf{p})}\frac{1}{M}\sum_{j\in\mathcal{J}} w_j\sum_{i\preceq_a j}\mathbb{E}[p_i] + \sum_{j\in\mathcal{J}} w_j\left(C_j^{\boldsymbol{a}} - \frac{1}{M}\sum_{i\preceq_a j} p_i\right) \\ &\ge \min_{\boldsymbol{a}\in A(\mathbf{p})}\frac{1}{M}\sum_{j\in\mathcal{J}} w_j\sum_{i\preceq_a j}\mathbb{E}[p_i] + \frac{M-1}{M}\sum_{j\in\mathcal{J}} w_j\left(p_j - \max_{i\in\mathcal{J}} p_i\right) \\ &= \frac{1}{M}\sum_{j=1}^{J} w_j\sum_{i\preceq j}\mathbb{E}[p_i] + \frac{M-1}{M}\sum_{j\in\mathcal{J}} w_j\left(p_j - \max_{i\in\mathcal{J}} p_i\right),\end{aligned}$$

where the inequality follows from Equation (A.3) and the fact that $C_j^{\mathbf{a}} = S_j^{\mathbf{a}} + p_j$. The second equality follows from the fact that the WSEPT policy is optimal for the remaining objective (the only term that depends on **a** in the second line is the objective for a single machine–scheduling problem) and the fact that the jobs are indexed according to WSEPT. The penalized objective of the WSEPT policy $V_z^{\mathrm{G}}(\mathbf{p})$ corresponds to Equation (14) with the policy **a** corresponding to the WSEPT policy. We note that the completion time C_j^{G} using WSEPT for any job j satisfies $C_j^{\mathrm{G}} \le (1/M)\sum_{i<j} p_i + p_j$ (see, e.g., lemma 3.3 of Hall et al. 1997). We then have

$$\begin{aligned}V_z^{\mathrm{G}}(\mathbf{p}) &= \frac{1}{M}\sum_{j=1}^{J} w_j\sum_{i\le j}\mathbb{E}[p_i] + \sum_{j\in\mathcal{J}} w_j\left(C_j^{\mathrm{G}} - \frac{1}{M}\sum_{i\le j} p_i\right) \\ &\le \frac{1}{M}\sum_{j=1}^{J} w_j\sum_{i\le j}\mathbb{E}[p_i] + \frac{M-1}{M}\sum_{j\in\mathcal{J}} w_j p_j.\end{aligned}$$

Equation (15) then follows from the above lower bound on $V_z^{\mathrm{P}}(\mathbf{p})$ and this upper bound on $V_z^{\mathrm{G}}(\mathbf{p})$.

A.6. Proof of Corollary 3

Equation (16) follows from taking expectations in Equation (15) and applying Proposition 1. Equation (17) then follows directly given the assumption that $w_j \le \bar{w}$.

A.7. Proof of Proposition 5

We first analyze the penalized perfect information problem. In the perfect information problem, the DM must select exactly K alternatives, and any selected alternative must have been previously explored. Assume that the rewards are labeled so that they would be optimally explored in the penalized perfect information problem (i.e., r_t corresponds to the t^{th} alternative that is explored). There is no penalty for selecting alternatives, and hence, the penalized reward for selecting reward r_t at time $\tau \ge t$ is given simply by $\delta^\tau r_t$. The penalized cost for exploring this alternative, however, is given by

$$\begin{aligned}\delta^{t-1}(-s - z_t) &= \delta^{t-1}(-s + \delta\mathbb{E}_{\tilde{r}}[(\tilde{r} - v_{k_t})^+] - \delta(r_t - v_{k_t})^+) \\ &= \delta^t\min(v_{k_t} - r_t, 0) + (1-\delta)\delta^{t-1}\sum_{j=1}^{k_t} v_j, \quad \text{(A.4)}\end{aligned}$$

where we use k_t to denote the capacity remaining at the start of time t and that

$$(1-\delta)\sum_{j=1}^{k} v_j = \delta\mathbb{E}_{\tilde{r}}[(\tilde{r} - v_k)^+] - s$$

from Equation (21). Note that any term v_j in the sum $\sum_{j=1}^{k_t} v_j$ will continue to be accrued in the penalized perfect information problem up to and including the period when the jth unit of capacity (counting down from K to one) is consumed by some selection. Let τ_j denote the time when the jth

unit of capacity is consumed or, equivalently, when the $(K-j+1)^{\text{th}}$ alternative is selected, and let $t_j \le \tau_j$ denote the time at which the alternative used for this selection is explored. Decomposing the penalized rewards across units of capacity consumption, we then have

$$\begin{aligned} V_z^{\mathrm{P}}(\mathbf{r}) &\le \sum_{j=1}^{K}\left((1-\delta)\sum_{t=1}^{\tau_j}\delta^{t-1}v_j + \delta^{t_j}\min(v_{k_{t_j}} - r_{t_j}, 0) + \delta^{\tau_j} r_{t_j}\right) \\ &= \sum_{j=1}^{K}\left((1-\delta^{\tau_j})v_j + \delta^{t_j}\min(v_{k_{t_j}} - r_{t_j}, 0) + \delta^{\tau_j} r_{t_j}\right) \\ &\le \sum_{j=1}^{K}\left((1-\delta^{\tau_j})v_j + \delta^{\tau_j} v_j\right) = \sum_{j=1}^{K} v_j\,. \end{aligned}$$

The first inequality follows because we exclude only terms of the form $\delta^t \min(v_{k_t} - r_t, 0) \le 0$ for alternatives that are explored but never selected. The first equality follows by evaluating the geometric series for the first term. The second inequality follows from the fact that $\delta^{t_j}\min(v_{k_{t_j}} - r_{t_j}, 0) + \delta^{\tau_j} r_{t_j} \le \delta^{\tau_j} v_j$, which we argue as follows:

$$\begin{aligned} &\delta^{t_j}\min(v_{k_{t_j}} - r_{t_j}, 0) + \delta^{\tau_j} r_{t_j} \le \delta^{t_j}\min(v_j - r_{t_j}, 0) + \delta^{\tau_j} r_{t_j} \\ &\quad\le \delta^{\tau_j}\min(v_j - r_{t_j}, 0) + \delta^{\tau_j} r_{t_j} = \delta^{\tau_j}\min(v_j, r_{t_j}) \le \delta^{\tau_j} v_j\,, \end{aligned}$$

where the first inequality follows from the fact that v_j is nonincreasing in j, and $k_{t_j} \ge j$ for every j. The second inequality follows because $\tau_j \ge t_j$, $\delta \in (0,1]$, and the term in the minimization is no larger than zero.

We next analyze the penalized performance of the greedy policy. Notice that the reservation prices v_k are nonincreasing in k. Because capacity decreases monotonically, the greedy policy will only select an alternative immediately after it is explored or at $t = N$. Because the greedy policy selects an alternative immediately whenever $r_t \ge v_{k_t}$, using Equation (A.4) we obtain that the penalized payoff at time $t < N$ is given by

$$\delta^{t-1}\left(\delta r_t \mathbf{1}\{r_t \ge v_{k_t}\} - s - z_t\right) = \delta^t v_{k_t}\mathbf{1}\{r_t \ge v_{k_t}\} + (1-\delta)\delta^{t-1}\sum_{j=1}^{k_t} v_j\,.$$

In period N, if any positive capacity remains, the greedy policy will explore the final alternative and then select the k_N highest rewards. Because rewards are nonnegative, the total reward selected in period N is no smaller than $\delta^N v_{k_N}\mathbf{1}\{r_N \ge v_{k_N}\} + (1-\delta)\delta^{N-1}\sum_{j=1}^{k_N} v_j$.

As in the analysis above, let τ_j denote the time when the $(K-j+1)^{\text{th}}$ alternative is selected by the greedy policy. We can then bound the penalized performance of the greedy policy as

$$\begin{aligned} V_z^{\mathrm{G}}(\mathbf{r}) &\ge \sum_{t=1}^{N\wedge\tau^{\mathrm{G}}}\left(\delta^t v_{k_t}\mathbf{1}\{r_t \ge v_{k_t}\} + (1-\delta)\delta^{t-1}\sum_{j=1}^{k_t} v_j\right) \\ &= \sum_{j=\bar{k}^{\mathrm{G}}+1}^{K}\delta^{\tau_j} v_j + (1-\delta)\sum_{j=1}^{K} v_j \sum_{t=1}^{N\wedge\tau_j}\delta^{t-1} \\ &= \sum_{j=\bar{k}^{\mathrm{G}}+1}^{K}\delta^{\tau_j} v_j + \sum_{j=1}^{K} v_j(1-\delta^{N\wedge\tau_j}) = \sum_{j=1}^{K} v_j - \delta^N\sum_{j=1}^{\bar{k}^{\mathrm{G}}} v_j\,, \end{aligned}$$

where the inequality follows by the observation that the total reward selected in period N is no smaller than $v_{k_N}\mathbf{1}\{r_N \ge v_{k_N}\}$, the first equality follows because the penalized reward of the alternative selected at time τ_j is v_j, the second equality from the formula for geometric series, and the last because $\tau_j = N$ for $j \le \bar{k}^{\mathrm{G}}$. The result then follows because $V_z^{\mathrm{P}}(\mathbf{r}) \le \sum_{j=1}^{K} v_j$ as argued previously.

When $\delta = 1$, the reservation prices v_k equal a constant value v^* for each $k = 1,\dots,K$ with $\mathbb{E}_{\tilde{r}}[(\tilde{r} - v^*)^+] = s$. In this case, in the penalized perfect information problem, the value for exploring and selecting alternative t is given by $\min(r_t, v^*)$, and the value for exploring but not selecting an alternative is given by $\min(0, r_t - v^*) \le 0$. Thus, in the perfect information problem, it is optimal to explore and select the K alternatives with the largest realized rewards and the optimal value is given by $V_z^{\mathrm{P}}(\mathbf{r}) = \sum_{k=1}^{K}\min(r_{(k)}, v^*)$, where $r_{(k)}$ denotes the kth largest value of $\mathbf{r}$; that is, $r_{(1)} \ge r_{(2)} \ge \cdots \ge r_{(N)}$. For the greedy policy, the penalized reward for selecting the alternative explored at time t is given by $v^*\mathbf{1}\{r_t \ge v^*\}$. If the greedy policy explores all N alternatives and capacity $\bar{k}^{\mathrm{G}} > 0$ remains, then the largest $\bar{k}^{\mathrm{G}}$ remaining alternatives are selected. We can then write the penalized performance of the greedy policy as

$$V_z^{\mathrm{G}}(\mathbf{r}) = (K - \bar{k}^{\mathrm{G}})v^* + \sum_{k=K-\bar{k}^{\mathrm{G}}+1}^{K} r_{(k)} = \sum_{k=1}^{K}\min(r_{(k)}, v^*) = V_z^{\mathrm{P}}(\mathbf{r})\,.$$

A.8. Proof of Corollary 4

Equation (24) follows from taking expectations in Equation (23) and applying Proposition 1.

We now prove the second part of the result. We first assume condition (a); that is, the discount factor is fixed. It is not hard to see that the reservation prices are nondecreasing in δ. Let $\hat{v}$ be the reservation price when $\delta = 1$, that is, the solution of the equation $\mathbb{E}_{\tilde{r}}[(\tilde{r} - \hat{v})^+] = s$. Using the fact that the reservation prices v_k are nonincreasing in k, it follows that the loss term in Equation (24) is bounded above by $\delta^N K\hat{v}$. Then we have

$$\frac{1}{K}(V^* - V^{\mathrm{G}}) \le \frac{\delta^N K\hat{v}}{K} = \delta^N\hat{v} \overset{N\to\infty}{\longrightarrow} 0.$$

We now consider condition (b). Using the fact that $\delta \in (0,1)$ and the fact that $v_k \le \hat{v}$ as before, it follows that the loss term in Equation (24) is bounded from above by $\hat{v}\mathbb{E}[\bar{k}^{\mathrm{G}}]$. We next bound $\bar{k}^{\mathrm{G}}$ from above. Because reservation prices satisfy $v_k \le \hat{v}$, we obtain that $\bar{k}^{\mathrm{G}} = K - \sum_{t=1}^{N\wedge\tau^{\mathrm{G}}}\mathbf{1}\{r_t \ge v_{k_t}\} \le \max(0, K - \sum_{t=1}^{N}\mathbf{1}\{r_t \ge \hat{v}\})$. Because rewards are independent and identically distributed, we have $\sum_{t=1}^{N}\mathbf{1}\{r_t \ge \hat{v}\} \sim \mathrm{Bin}(N,\hat{p})$; that is, the sum is distributed as a binomial random variable with N trials and probability $\hat{p} = \mathbb{P}\{\tilde{r} \ge \hat{v}\}$. We next argue that $\hat{p} > 0$. Because $\mathbb{E}[\tilde{r}] < \infty$, for every $\epsilon > 0$ there exists some $\bar{r} > 0$ such that $\mathbb{E}[\tilde{r}\mathbf{1}\{\tilde{r} \ge \bar{r}\}] < \epsilon$. Without loss we can assume that $\bar{r} \ge v^*$. Using that $(r - v^*)^+ \le (\bar{r} - v^*)\mathbf{1}\{r \ge v^*\} + (r - \bar{r})^+$, we obtain by taking expectations over $\tilde{r}$ that $s \le (\bar{r} - v^*)\hat{p} + \mathbb{E}[(\tilde{r} - \bar{r})^+] \le \bar{r}\hat{p} + \epsilon$ because $\bar{r}, \bar{v}^* \ge 0$. Therefore, $\hat{p} \ge (s - \epsilon)/\bar{r}$, and the result follows because ϵ is arbitrary.

Because $K = o(N)$, we know that for any $\epsilon \in [0,\hat{p})$ it holds that $K \le (\hat{p} - \epsilon)N$ for sufficiently large N. This leads us to

$$\begin{aligned} \mathbb{E}[\bar{k}^{\mathrm{G}}] &\le \mathbb{E}\left[\max(0, K - \mathrm{Bin}(N,\hat{p}))\right] \le K\mathbb{P}\left\{\mathrm{Bin}(N,\hat{p}) \le K\right\} \\ &\le K\exp\left(-2\epsilon^2 N\right), \end{aligned}$$

where the second inequality follows because $\max(0, K-a) \le K\mathbf{1}\{a \le K\}$ for $a, K \ge 0$ and the second inequality from Hoeffding's inequality.

Putting this together with Equation (24), we conclude

$$\frac{1}{K}(V^* - V^{\mathrm{G}}) \le \frac{1}{K}\hat{v}\,\mathbb{E}[\tilde{k}^{\mathrm{G}}] \le \hat{v}\exp(-2\epsilon^2 N) \overset{N\to\infty}{\longrightarrow} 0\ .$$

Endnotes

[1] Dean et al. (2008) considers an alternate policy that sorts items according to w_i/μ_i, where $w_i = v_i\mathbb{P}\{s_i \le \kappa\}$ denotes the effective value of item i and $\mu_i = \mathbb{E}[\min\{s_i, \kappa\}]$ denotes the mean truncated size of item i. This takes into account the fact that in the event that $s_i > \kappa$, the actual realization of the size is irrelevant because item i certainly overflows the knapsack, and the DM will never collect the item's value in this case. In Online Appendix B, we show that our results extend to this effective value formulation.

[2] We present the asymptotic optimality result in absolute form as is customary in the operations research and regret-based learning literature. It is not hard to derive similar results in relative form by providing a lower bound on the performance of the greedy policy in terms of the problem parameters. Relative bounds are common in the approximation algorithm literature.

[3] We thank an anonymous referee for suggesting this variation.

[4] In contrast to the knapsack problem, the cost of the component used at termination counts toward the objective. Derman et al. (1978) considers the case with rebates; that is, the cost of the last component used is returned. Our results can be extended to accommodate rebates.

[5] To simplify notation, we do not include in Equation (19) the effect of previously explored alternatives that are recalled by the greedy policy at $t = N$. Our primary analysis of the greedy policy still applies without including these possible rewards.

[6] In the primal DP, we implicitly impose the constraint that the DM must explore an alternative before selecting it. We can impose this constraint in the perfect information relaxation as well. Thus, the DM with perfect information still has to pay a search cost to select an alternative even if the rewards are all known in advance. Note also that the perfect information relaxation gives the DM freedom to explore the alternatives in any order; in the case of a priori identical items, we could obtain tighter upper bounds by fixing the exploration sequence to be the same in every sample path. This variation appears more difficult to analyze, however, and would not extend to problems with alternatives that are not a priori identical (e.g., as in Section 5.5).

References

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

Andersen LM, Broadie M (2004) Primal-dual simulation algorithm for pricing multidimensional American options. *Management Sci.* 50(9):1222–1234.

Bertsekas DP (2000) *Dynamic Programming and Optimal Control* (Athena Scientific, Belmont, MA).

Blado D, Toriello A (2016) Relaxation analysis for the dynamic knapsack problem with stochastic item sizes. Working paper, Georgia Institute of Technology, Atlanta.

Blado D, Hu W, Toriello A (2016) Semi-infinite relaxations for the dynamic knapsack problem with stochastic item sizes. *SIAM J. Optim.* 26(3):1625–1648.

Borodin A, El-Yaniv R (1998) *Online Computation and Competitive Analysis* (Cambridge University Press, New York).

Brown DB, Haugh MB (2017) Information relaxation bounds for infinite horizon Markov decision processes. *Oper. Res.* 65(5): 1355–1379.

Brown DB, Smith JE (2011) Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds. *Management Sci.* 57(10):1752–1770.

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

Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 58(4-part-1):785–801.

Dean BC, Goemans MX, Vondrák J (2008) Approximating the stochastic knapsack problem: The benefit of adaptivity. *Math. Oper. Res.* 33(4):945–964.

de Farias DP, Roy BV (2003) The linear programming approach to approximate dynamic programming. *Oper. Res.* 51(6): 850–865.

Derman C, Lieberman C, Ross S (1978) A renewal decision problem. *Management Sci.* 24(5):554–561.

Desai V, Farias VF, Moallemi CC (2012) Pathwise optimization for optimal stopping problems. *Management Sci.* 58(12):2292–2308.

Devalkar S, Anupindi R, Sinha A (2011) Integrated optimization of procurement, processing, and trade of commodities. *Oper. Res.* 59(6):1369–1381.

Feldman J, Henzinger M, Korula N, Mirrokni VS, Stein C (2010) Online stochastic packing applied to display ad allocation. *Proc. 18th Annual Eur. Conf. Algorithms: Part I, ESA'10* (Springer-Verlag, Berlin), 182–194.

Garg N, Gupta A, Leonardi S, Sankowski P (2008) Stochastic analyses for online combinatorial optimization problems. *Proc. 19th Annual ACM-SIAM Sympos. Discrete Algorithms, SODA '08* (Society for Industrial and Applied Mathematics, Philadelphia), 942–951.

Gittins J, Jones D (1974) A dynamic allocation index for the sequential design of experiments. Gani J, ed. *Progress in Statistics* (North-Holland, Amsterdam), 241–266.

Grandoni F, Gupta A, Leonardi S, Miettinen P, Sankowski P, Singh M (2008) Set covering with our eyes closed. *Proc. Foundations of Comput. Sci. 2008. FOCS '08. IEEE 49th Annual IEEE Sympos.* (IEEE Computer Society, Los Alamitos, CA), 347–356.

Hall LA, Schulz AS, Shmoys DB, Wein J (1997) Scheduling to minimize average completion time: Off-line and on-line approximation algorithms. *Math. Oper. Res.* 22(3):513–544.

Haugh MB, Kogan L (2004) Pricing American options: A duality approach. *Oper. Res.* 52(2):258–270.

Haugh MB, Lim AE (2012) Linear-quadratic control and information relaxations. *Oper. Res. Lett.* 40(6):521–528.

Haugh MB, Iyengar G, Wang C (2016) Tax-aware dynamic asset allocation. *Oper. Res.* 64(4):849–866.

Lai G, Margot F, Secomandi N (2010) An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Oper. Res.* 58(3):564–582.

Manshadi VH, Gharan SO, Saberi A (2012) Online stochastic matching: Online actions based on offline statistics. *Math. Oper. Res.* 37(4):559–573.

Martello S, Toth P (1990) *Knapsack Problems: Algorithms and Computer Implementations* (John Wiley & Sons, New York).

Möhring RH, Schulz AS, Uetz M (1999) Approximation in stochastic scheduling: The power of LP-based priority policies. *J. ACM* 46(6):924–942.

Nadarajah S, Margot F, Secomandi N (2015) Relaxations of approximate linear programs for the real option management of commodity storage. *Management Sci.* 61(12):3054–3076.

Papstavrou JD, Rajagopalan S, Kleywegt AJ (1996) The dynamic and stochastic knapsack problem with deadlines. *Management Sci.* 42(12):1706–1718.

Pinedo ML (2012) *Scheduling* (Springer-Verlag, New York).

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

Rogers L (2002) Monte Carlo valuation of American options. *Math. Finance* 12(3):271–286.

Rogers L (2007) Pathwise stochastic optimal control. *SIAM J. Control Optim.* 46(3):1116–1132.

Rothkopf M (1966) Scheduling with random service times. *Management Sci.* 12(9):703–713.

Skutella M, Sviridenko M, Uetz M (2016) Unrelated machine scheduling with stochastic processing times. *Math. Oper. Res.* 41(3):851–864.

Talluri K, van Ryzin G (1998) An analysis of bid-price controls for network revenue management. *Management Sci.* 44(11):1577–1593.

Uetz M (2003) When greediness fails: Examples from stochastic scheduling. *Oper. Res. Lett.* 31(6):413–419.

Vanderbei R (1980) The optimal choice of a subset of a population. *Math. Oper. Res.* 5(4):481–486.

Weiss G (1990) Approximation results in parallel machines stochastic scheduling. *Ann. Oper. Res.* 26(1):195–242.

Weitzman ML (1979) Optimal search for the best alternative. *Econometrica* 47(3):641–654.

Santiago R. Balseiro is an assistant professor in the decision, risk, and operations division at the Graduate School of Business, Columbia University. His primary research interests are in the area of dynamic optimization, stochastic systems, and game theory with applications in revenue management and internet advertising.

David B. Brown is an associate professor in the decision sciences area at the Fuqua School of Business at Duke University. His research focuses on developing and analyzing approximation methods for large-scale stochastic optimization problems with an emphasis on applications in operations and finance.