Approximations to Stochastic Dynamic Programs via Information Relaxation Duality

Santiago R. Balseiro and David B. Brown

Fuqua School of Business
Duke University
srb43@duke.edu, dbbrown@duke.edu

January 18, 2016

Abstract

In the analysis of complex stochastic dynamic programs (DPs), we often seek strong theoretical guarantees on the suboptimality of heuristic policies: a common technique for obtaining such guarantees 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, i.e., fully relaxed non-anticipativity constraints. A limitation of this approach is that in many problems perfect information conveys excessive power to the decision maker, which leads to weak bounds. In this paper we leverage the information relaxation duality approach of Brown, Smith, and Sun (2010) to show that by including a penalty that punishes violations of these non-anticipativity constraints, we can derive stronger bounds and *analytically characterize* the suboptimality of heuristic policies in stochastic dynamic programs that are too difficult to solve. We study three challenging problems: stochastic scheduling on parallel machines, a stochastic knapsack problem, and a stochastic project completion problem. For each problem, we use this approach to derive analytical bounds on the suboptimality gap of a simple policy. In each case, these bounds imply asymptotic optimality of the policy for a particular scaling that renders the problem increasingly difficult to solve. As we discuss, the penalty is crucial for obtaining good bounds, and must be chosen carefully in order to link the bounds to the performance of the policy in question. Finally, for the stochastic knapsack and stochastic project completion problems, we find in numerical examples that this approach performs strikingly well.

Subject classifications: Dynamic programming, information relaxation duality, asymptotic optimality, stochastic scheduling, stochastic knapsack problems, stochastic project completion.

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 all instances, or at least allow us to understand on what types of instances the heuristic policy will perform well.

A common technique in the analysis of heuristic policies is "perfect information analysis." This approach provides bounds by considering a decision maker who has advance access to the outcomes of all future uncertainties, i.e., a problem with fully relaxed non-anticipativity constraints. We refer to this as the *perfect information problem*. For each sample path, the decision maker then solves a deterministic optimization problem, which is often 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 in the analysis of heuristic policies in a number of applications; we survey a few in Section 1.1. In many problems, however, perfect information may convey excessive power to the decision maker and lead to weak bounds as a result: this limits the applicability of this approach.

Can we improve the quality of these bounds while retaining their amenability to theoretical analysis? In this paper we provide a positive answer to this - in the context of three challenging DPs - by leveraging the approach of Brown, Smith and Sun (2010) (BSS hereafter). The framework in BSS 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 non-anticipativity constraints. BSS 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 non-anticipative policy - leads to an upper bound on the expected reward with any policy, including an optimal policy. Strong duality ensures the existence of a dual feasible penalty (albeit one that may be hard to compute) 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 bounds. When we include a penalty, we refer to the optimization problem in which we relax the non-anticipativity constraints as the *penalized perfect information problem*,

and the associated bound as the *penalized perfect information bound*.

To our knowledge, the use of information relaxations with penalties heretofore has been exclusively as a computational method for evaluating heuristic policies in applications. Our objective in this paper is different: we wish to use the approach to derive theoretical guarantees on the performance of heuristic policies in complex DPs. Although our results provide analytical bounds that hold for all problem instances, we are especially interested in identifying asymptotic regimes for which the heuristic policies in consideration approach optimality. We illustrate the approach in a study of three specific problems: (i) stochastic scheduling on parallel machines, (ii) a stochastic knapsack problem, and (iii) a stochastic project completion problem. In order to obtain theoretical guarantees, we consider simple heuristic policies that are amenable to analysis and show that these simple policies approach optimality in a particular asymptotic regime. We study:

(i) *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 study the performance of a greedy policy that schedules jobs in a fixed order based on weights and expected processing times, and show that this policy is asymptotically optimal as the number of jobs grows large relative to the number of machines. Although the result we derive is already known from Möhring et al. (1999) (and a similar result is in Weiss (1990)), we provide an alternate and novel proof using a penalized perfect information bound that allows us to exploit well-known results from deterministic scheduling.

(ii) *Stochastic knapsack.* In this problem (due to Dean et al., 2008), there is a set of items available to be inserted to 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 decision maker 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. We study the performance of a greedy policy that orders items based on their values and expected sizes, and show that this policy is asymptotically optimal when the number items grows and capacity is commensurately scaled.

(iii) *Stochastic project completion.* We consider a model of a firm working on a project that, upon completion, generates reward for the firm. A motivating example is a firm developing a new product, with the reward being the expected NPV from launching the product. The firm is concurrently working on different *alternatives*, and finishing *any* of these alternatives completes the project. In each period, the firm can choose to accelerate one of the alternatives, at a cost. The goal is to maximize the total expected discounted reward over an infinite horizon. We consider static policies that commit exclusively to accelerating the same alternative in every period, and show that these policies are asymptotically optimal as the number of initial steps to completion 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 performs quite poorly. For example, in the stochastic knapsack problem, the perfect information problem involves revealing all item sizes prior to any item selection decisions. When the decision maker knows the realizations of all sizes in advance, she can avoid inserting potentially large items, which can result in weak bounds. 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.

In each problem, we choose penalties carefully in order that they be: (a) dual feasible, so that we obtain a bound; (b) simple enough so that we can analyze the bound; and (c) somehow connected to the heuristic policy so that we can relate the bound to the performance of the heuristic policy. In all three problems, the penalties cancel out information "locally" for each possible action at any point in time. For example, in the stochastic knapsack problem, if we select an item in the penalized perfect information problem, we deduct a penalty that is proportional to the difference between the expected size of the item and the realized size of the item. This tilts incentives in the perfect information problem towards selecting items with large realized sizes, which, absent the penalty, would be better to avoid in the perfect information problem. Such a penalty does not completely counteract the benefit of perfect information, but by properly tuning the constants of proportionality on the penalty, we can characterize the gap between the penalized perfect information bound and the greedy policy and show that the greedy policy approaches optimality as the number of items (and capacity) is scaled.

Finally, although our original goal in this paper was to develop these techniques for the sake of theoretical analysis, we were surprised to also discover that the use of these "simple" penalties often led to strikingly good performance in numerical examples. Specifically, we compute the bounds on many randomly generated examples for the stochastic knapsack problem and the stochastic project completion problem and find small gaps, particularly in comparison to other state-of-the-art approaches.

The rest of the paper is organized as follows. Section 1.1 reviews some related papers. Section 2 demonstrates the use of the approach on the stochastic scheduling problem, Section 3 discusses the stochastic knapsack problem, and Section 4 discusses the stochastic project completion problem. These sections are self-contained: 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. We conclude Sections 3 and 4 with some numerical experiments that demonstrate the bounds. Section 5 concludes with some guidance on how to apply these ideas to other problems. All proofs are available in Appendix B.

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. BSS 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 BSS for perfect information relaxations of MDPs using change of measure techniques. In Appendix A we provide a review of the key definitions and results of information relaxation duality.

In terms of applications of information relaxations, there are many other applications to options pricing problems (Andersen and Broadie, 2004; BSS; Desai et al., 2012); inventory management problems (BSS; 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., 2014); linear-quadratic control with constraints (Haugh and Lim, 2012); network revenue management problems (Brown and Smith, 2014); and multiclass queueing systems (Brown and Haugh, 2014). 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. Again, this line of work focuses on numerically computing bounds for specific problem instances, while our objective in this paper is to derive analytical guarantees on the performance of heuristic policies for a large class of problem instances.

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 where 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 where 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 locations 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).

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. In a key paper, Weiss (1990) originally established the optimality gap of the WSEPT (Weighted Shortest Expected Processing Time first) 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 new sharp bounds on the performance of the WSEPT policy. We provide an alternative and novel proof of the result in Möhring et al. (1999) using penalized perfect information bounds. We carefully choose the penalty in a way that allows us to use the work of Hall et al. (1997) on valid inequalities for deterministic scheduling.

The version of the stochastic knapsack problem we study was introduced in the influential paper 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 item 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. Blado et al. (2015) develop approximate dynamic programming style bounds for this problem and in extensive numerical experiments find that the greedy policy often performs well, especially for examples with many items. In this paper we show that the greedy policy approaches optimality as the number of items grows and capacity is commensurately scaled.

Finally, the stochastic project completion is a new problem we developed and is motivated by a basic tension faced by firms considering multiple alternatives in product development: how should firms optimally invest in these alternatives, balancing costs with the pressure to develop a viable product quickly? This basic

tradeoff can arise in R&D settings in a host of industries (e.g., pharmaceutical, manufacturing, high tech, etc.). Other papers have studied similar models. Childs and Triantis (1999) study dynamic R&D investment policies with a variety of alternatives that may be accelerated or abandoned. Ding and Eliashberg (2002) study decision tree models for a "pipeline problem" faced, e.g., by pharmaceutical companies. Santiago and Vakili (2005) study firm R&D investment in which a single project can be managed "actively" (accelerated to more favorable states through costly investment) or "passively" (continued in a baseline fashion). The model we study involves a firm managing a potentially large number of alternatives over an infinite horizon. This problem bears some resemblance to a restless bandit problem (Whittle, 1988) in that the states of each alternative may change in each period, but differs in that the problem ends upon completion of any alternative. The resulting DP can be quite challenging, but by using penalized perfect information analysis we show that a static policy approaches optimality when the number of steps to completion grows large.

2 Illustrative example: stochastic scheduling on parallel machines

In this section, we demonstrate the use of the approach on a stochastic scheduling problem. Although the main result in this section is already known from Möhring et al. (1999), we provide an alternative and novel proof that uses penalized perfect information bounds.

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 times of each job are known only after a job is completed. Formally, let $\mathcal{N} = \{1, \ldots, n\}$ denote a set of jobs to be scheduled on m identical parallel machines. The processing time of job $i \in \mathcal{N}$ is independent of the machine and given by the random variable p_i. Processing times are assumed to be independently distributed (but not necessarily identical) with finite second moments. Let C_i be the completion time of job $i \in \mathcal{N}$, that is, the sum of the waiting time until service and the processing time p_i. Each job has a weight w_i and the objective is to minimize the expected total weighted completion time $\mathbb{E}[\sum_{i=1}^{n} w_i C_i]$. Using "Graham's notation," the problem can be written as $PM//\mathbb{E}[\sum_i w_i C_i]$ (see Pinedo, 2012).

We let Π denote the set of non-anticipative, adaptive policies. A policy $\pi \in \Pi$ is a mapping $2^{\mathcal{N}} \times 2^{\mathcal{N}} \times \mathbb{R}^n \to \mathcal{N}$ that determines the next job to be processed, denoted by $\pi(W, P, \mathbf{s})$, given the set of waiting jobs $W \subseteq \mathcal{N}$, the set of jobs $P \in \mathcal{N}$ currently in process, and the amount of work done $\mathbf{s} \in \mathbb{R}^n$ on each job in process. In this model time is continuous and decision epochs correspond to the time when a job is completed and a machine becomes available. We let $W_t^\pi \subseteq \mathcal{N}$ denote the subset of jobs waiting for service at time t and $P_t^\pi \subseteq \mathcal{N}$ denote the subset of jobs under process at time t under policy π. Denoting the time when all jobs are completed

by $\tau^\pi = \inf\{t \geq 0 : W_t^\pi = \emptyset\}$, the completion time of job $i \in \mathcal{N}$ is given by $C_i^\pi = \int_0^{\tau^\pi} \mathbb{1}\{i \in W_t^\pi \cup P_t^\pi\}\, \mathrm{d}t$. We restrict attention to policies satisfying $\mathbb{E}\tau^\pi < \infty$. The problem can be written as

$$J^* = \min_{\pi \in \Pi} \mathbb{E} \sum_{i=1}^n w_i C_i^\pi \,.$$

2.1 WSEPT policy

We consider the WSEPT policy (Weighted Shortest Expected Processing Time first), which sorts the jobs in decreasing order of weight per *expected* processing time $r_i := w_i/\mathbb{E}[p_i]$ and then schedules the jobs in this order. Without loss of generality, we assume that items are sorted by decreasing order of weight per expected processing time; that is, $r_1 \geq r_2 \geq \ldots \geq r_n$. The expected performance of this policy is

$$J^{\mathrm{G}} = \mathbb{E} \sum_{i=1}^n w_i C_i^{\mathrm{G}} \,,$$

where C_i^{G} denotes the completion time of job $i \in \mathcal{N}$ under the WSEPT policy. We aim to analytically compare the expected performance of the WSEPT policy to that of the optimal adaptive, non-anticipative policy.

2.2 Perfect information bound

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

Unfortunately, the perfect information bound may be poor 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 n jobs with weight one, and the processing times are a two-point distribution supported on $\{\epsilon, 1\}$ with equal probability for some $\epsilon \in (0, 1)$, i.e., each jobs's processing time is either ϵ or 1, each with probability 1/2. Since the jobs are identical, the problem is trivial and it is easy to show that $J^{\mathrm{G}} = J^* = (1+\epsilon)n(n+1)/4$. 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 - this can be seen by invoking the well-known result due to Smith (1956) showing that a WSPT policy (Weighted Shortest Processing Time first) is optimal with a single deterministic machine. Let $I_t = \sum_{i=1}^n \mathbb{1}\{p_i = t\}$ denote the number of jobs with processing time $t \in \{\epsilon, 1\}$, respectively. The total completion time of the long jobs is $\epsilon I_\epsilon(I_\epsilon + 1)/2$ and the total completion time of the short jobs is $\epsilon I_\epsilon I_1 + I_1(I_1+1)/2$. Taking expectations and using the fact that $I_\epsilon + I_1 = n$ together with the

fact that I_1 is binomially distributed with n trials and success probability $1/2$ since jobs are independent, leads us to the poor lower bound of $J^{\mathrm{P}} = n(n+3+\epsilon(3n+1))/8$. Note that for n large, this lower bound is off from J^* by nearly a factor of two.

2.3 Penalized perfect information bound

Given a constant vector $\mathbf{z} = (z_i)_{i=1}^n \in \mathbb{R}^n$, we consider the integral $M_t^\pi = \int_0^t \sum_{i \in W_s^\pi} z_i(p_i - \mathbb{E}[p_i])\, \mathrm{d}s$. We will use $M_{\tau^\pi}^\pi$ as a penalty. We claim $M_{\tau^\pi}^\pi$ is dual feasible in the sense of BSS, that is, $M_{\tau^\pi}^\pi$ does not penalize in expectation any non-anticipative policy $\pi \in \Pi$:

$$\mathbb{E}\left[M_{\tau^\pi}^\pi\right] = 0\,. \tag{1}$$

Condition (1) follows because, for every non-anticipative policy $\pi \in \Pi$, τ^π is a stopping time and $M_{\tau^\pi}^\pi$ is a martingale with respect to the natural filtration (i.e., the filtration that describes the decision maker's state of information at the beginning of the decision period). Hence the Optional Stopping Theorem implies that $\mathbb{E}\left[M_{\tau^\pi}^\pi\right] = M_0^\pi = 0$ because $\mathbb{E}\tau^\pi < \infty$. We can express $M_{\tau^\pi}^\pi$ in terms of completion times as

$$M_{\tau^\pi}^\pi = \sum_{i=1}^n z_i(p_i - \mathbb{E}[p_i])C_i^\pi - p_i z_i(p_i - \mathbb{E}[p_i])\,,$$

by adding and subtracting $\int_0^{\tau^\pi} \sum_{i \in P_s^\pi} z_i(p_i - \mathbb{E}[p_i])\, \mathrm{d}s$ and using the fact that $p_i = \int_0^{\tau^\pi} \mathbb{1}\{i \in P_s^\pi\}\, \mathrm{d}s$.

Including the penalty, the scheduling problem becomes

$$J_z^* = \min_{\pi \in \Pi} \left\{ \mathbb{E} \sum_{i=1}^n \left(w_i + z_i(p_i - \mathbb{E}[p_i])\right) C_i^\pi \right\} - \mathbb{E} \sum_{i=1}^n p_i z_i (p_i - \mathbb{E}[p_i])\,,$$

where we remove the terms that are independent of the scheduling policy from the minimization.

With perfect information and a given sample path $\mathbf{p} \in \mathbb{R}_+^n$, let $J_z^{\mathrm{P}}(\mathbf{p})$ be the optimal total weighted completion time including the penalty. We claim that $J_z^{\mathrm{P}} = \mathbb{E}_{\mathbf{p}}[J_z^{\mathrm{P}}(\mathbf{p})]$ is a lower bound on J^*. To see this, for any $\pi \in \Pi$, we denote by J^π and J_z^π the expected weighted completion time of π with and without the penalty, respectively; we then have

$$J^\pi \;=\; J_z^\pi \;\geq\; J_z^{\mathrm{P}},$$

where the equality follows from (1) since the penalty is zero mean for any $\pi \in \Pi$ and the inequality follows from the fact that π is feasible for the problem with perfect information. This inequality holds for all $\pi \in \Pi$,

and in particular it holds for the optimal policy, which implies $J^* \geq J_z^{\mathrm{P}}$. This is an example of the weak duality result from BSS; we refer the reader to Appendix A and in particular Lemma A.1 for a more detailed discussion of this general result.

The penalized perfect information problem is a deterministic scheduling problem with weights that depend on the actual realized processing times. Let Π^{P} denote the set of all policies in the perfect information problem given $\mathbf{p}$; for notational ease we suppress the dependence of Π^{P} on $\mathbf{p}$. Then, for sample path $\mathbf{p}$, we have

$$J_z^{\mathrm{P}}(\mathbf{p}) = \min_{\pi \in \Pi^{\mathrm{P}}} \left\{ \sum_{i=1}^{n} \left(w_i + z_i(p_i - \mathbb{E}[p_i]) \right) C_i^{\pi} \right\} - \sum_{i=1}^{n} p_i z_i (p_i - \mathbb{E}[p_i]) \,, \tag{2}$$

which can be determined by solving the deterministic scheduling problem $PM//\sum_i w_i^z C_i$ with weights $w_i^z = w_i + z_i(p_i - \mathbb{E}[p_i])$ and known processing times as given by $\mathbf{p}$.

This procedure provides a lower bound for any $\mathbf{z} \in \mathbb{R}^n$. Recall that the perfect information bound may "cheat" by scheduling first the jobs with higher ratio of weight to realized processing time. We seek to align the perfect information policy with the WSEPT policy by penalizing the decision maker with perfect information for scheduling early the jobs with realized processing times that are small relative to their expected values. A natural choice that accomplishes this well is $z_i = r_i = w_i/\mathbb{E}[p_i]$. With this choice of penalty, we obtain $w_i^z = r_i p_i$ and the perfect information policy effectively places less weight on jobs with shorter completion times. Since in every sample path the penalized perfect information problem is a deterministic scheduling problem with weights $w_i^z = r_i p_i$, the ratio of weights to processing times in the penalized perfect information problem is simply r_i, which is exactly the ratio used by the WSEPT policy in ranking jobs. In fact, if we return to the one machine example in Section 2.2, the penalized perfect information bound now becomes perfectly tight. This follows from the fact that for every sample path a WSPT policy ranking jobs by $w_i^z/p_i = r_i$ is optimal in the penalized perfect information problem, and this corresponds to the feasible WSEPT policy that ranks jobs by r_i (in this example, all the r_i are equal). Thus, a policy in Π minimizes the penalized costs over all perfect information policies Π^{P} in every sample path, which implies $J_z^{\mathrm{P}} = J_z^* = J^*$. In this example, the penalty entirely wipes out any benefit of advance information about job processing times. (Note that this argument holds for any distribution of job times; thus this argument provides an alternate proof of the well-known result (Rothkopf, 1966) that WSEPT is optimal for stochastic scheduling on a single machine.)

2.4 Performance analysis

In the general case, we can use the approach just discussed to show the following result.

Proposition 2.1 (Corollary 4.1 of Möhring et al. (1999)). *Suppose that processing times satisfy that* $Var[p_i]/\mathbb{E}[p_i]^2 \leq \Delta$ *uniformly over all jobs* $i \in \mathcal{N}$. *Then the WSEPT policy satisfies*

$$J^{\mathrm{G}} \geq J^* \geq J_z^{\mathrm{P}} \geq J^{\mathrm{G}} - \frac{m-1}{2m}(\Delta+1)\sum_{i=1}^{n} w_i \mathbb{E}[p_i]\,.$$

If we consider a scaling of the number of jobs n, the optimal cost J^* scales quadratically with n. The gap between the WSEPT policy and J^*, however, only scales as $O(n)$, provided job weights and mean processing times are uniformly bounded as n grows. Prop. 2.1 thus implies that the WSEPT policy is asymptotically optimal under this scaling.

We prove the result by upper bounding the performance of the WSEPT policy J^{G} in terms of the penalized perfect information bound J_z^{P} and using the fact that the penalized perfect information problem gives a lower bound on the optimal cost, that is, $J^* \geq J_z^{\mathrm{P}}$. Our proof differs from Möhring et al. (1999) in that we use the penalized perfect information bound to derive the result. Möhring et al. (1999) develop an intermediate result that translates valid inequalities for deterministic scheduling problems (Hall et al., 1997) into valid inequalities for the stochastic version of this problem. Crucial in this step in Möhring et al. (1999) is the property, specific to this problem, that the job processing times and job start times are independent for any policy $\pi \in \Pi$. We also use the valid inequalities in Hall et al. (1997), but we apply these directly to the deterministic scheduling problems that arise in the penalized perfect information problem in every sample path.

We now demonstrate that the basic approach here can be useful in analyzing the performance of simple policies for other challenging stochastic dynamic programs.

3 Stochastic knapsack problem

We consider a stochastic knapsack problem, as studied in Dean et al. (2008). There is a set $\mathcal{N} = \{1, \ldots, n\}$ of items available to be inserted to a knapsack of capacity κ. Item $i \in \mathcal{N}$ has a deterministic value denoted by $v_i \geq 0$ and a stochastic size denoted by $s_i \geq 0$. 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 decision maker selects an item i and attempts to insert it into the knapsack. After that, the size of item i is revealed, and a value of v_i is obtained if i is successfully inserted. The decision maker 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.

We let Π denote the set of non-anticipative, adaptive policies. A policy $\pi \in \Pi$ is a mapping $2^{\mathcal{N}} \times [0, \kappa] \to \mathcal{N}$ that determines the next item $\pi(S, c)$ to be inserted to the knapsack given the set of remaining items $S \subseteq \mathcal{N}$ and the remaining knapsack capacity c. We denote the decision epochs by $t = 1, \ldots, n$. For a given $\pi \in \Pi$, we let S_t^π denote the items available for insertion at the beginning of time t, and c_t^π denote the knapsack's remaining capacity. To simplify the notation, we let $\pi_t = \pi(S_t^\pi, c_t^\pi)$ denote the item to be inserted at time t under policy π. All items are initially available for insertion; that is, $S_1^\pi = \mathcal{N}$ and $c_1^\pi = \kappa$. At time t, item $\pi_t = \pi(S_t, c_t)$ is selected for insertion, and the state is updated as $S_{t+1}^\pi = S_t^\pi \setminus \{\pi_t\}$ and $c_{t+1}^\pi = c_t^\pi - s_{\pi_t}$. If the item fits into the knapsack, i.e. if $c_{t+1}^\pi \geq 0$, the value of v_{π_t} is collected. Otherwise, the problem ends.

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

$$V^* \;=\; \max_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{n \wedge (\tau^\pi - 1)} v_{\pi_t}.$$

3.1 A greedy policy

Following Dean et al. (2008), we let $w_i = v_i \mathbb{P}\{s_i \leq \kappa\}$ denote the *effective value* of item i and $\mu_i = \mathbb{E}[\min\{s_i, \kappa\}]$ be the mean truncated size of item i. In the event that $s_i > \kappa$, the actual realization of the size is irrelevant because item i certainly overflows the knapsack, and the decision maker will never collect the item's value in this case. We consider a greedy policy that sorts the items in decreasing order of value per *expected* size, w_i/μ_i, and inserts items in this order until the knapsack overflows or no items remain. Without loss of generality we assume that items are sorted in decreasing order of this ratio, i.e., $w_1/\mu_1 \geq w_2/\mu_2 \geq \ldots \geq w_n/\mu_n$. The expected performance of the greedy policy is given by

$$V^{\mathrm{G}} \;=\; \mathbb{E} \sum_{t=1}^{n \wedge (\tau^{\mathrm{G}} - 1)} v_t \,,$$

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 where the greedy policy can perform arbitrary poorly (see, e.g., a deterministic example of this with $n = 2$ in §4 of Dean et al. (2008)). We might expect the greedy policy to perform poorly, even in deterministic examples, when sizes are large relative to capacity: since 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.

There are, however, positive results for the greedy policy. Derman et al. (1978) show that the greedy policy is optimal in the case of exponentially distributed sizes. Blado et al. (2015) conduct extensive numerical experiments and find the greedy policy often performs well, especially for examples with many items. 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 glean intuition for this from the deterministic case by considering the LP relaxation of the problem that allows the decision maker to insert fractional items. Since the greedy ordering is optimal for the LP relaxation, in the deterministic case we have

$$V^{\mathrm{G}} \;\leq\; V^{*} \;\leq\; V^{\mathrm{LP}} \;\leq\; V^{\mathrm{G}} + \max_{i=1,\ldots,n} v_i \,, \tag{3}$$

where V^{LP} is the optimal objective value of the LP relaxation. In (3), 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, then we conclude from (3) 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 will derive a result analogous to (3) for the stochastic version of the problem where the decision maker optimizes over all possible non-anticipative policies. The result will then allow 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 $\mathbf{s} = \{s_i\}_{i=1}^n$ before selecting any items. Given a sample path $\mathbf{s} \in \mathbb{R}_+^n$, 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, i.e., $V^{*} \leq V^{\mathrm{P}}$. The perfect information problem is equivalent to the deterministic knapsack problem

$$V^{\mathrm{P}}(\mathbf{s}) = \max_{\mathbf{x}\in\{0,1\}^n} \left\{ \sum_{i=1}^{n} v_i x_i \;:\; \sum_{i=1}^{n} s_i x_i \leq \kappa \right\}, \tag{4}$$

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 having access to the realizations

of all sizes in advance, we can avoid inserting potentially large items. Dean et al. (2008) demonstrate this convincingly with the following example: consider the case when all items are symmetric with value one, and the sizes are a Bernoulli random variable with probability 1/2 scaled by $\kappa + \varepsilon$ for some $\varepsilon > 0$, i.e., each item's size is either 0 or $\kappa + \varepsilon$ with equal probability. Since the items are symmetric, the problem is trivial and it is easy to show that $V^{\mathrm{G}} = V^* = 1 - (1/2)^n \leq 1$. On the other hand, in the perfect information problem, it is optimal to select every item with a realized size of zero. Since this occurs with probability 1/2 for each item, and items are independent, this leads us to the very poor upper bound of $V^{\mathrm{P}} = n/2$.

3.3 Penalized perfect information bound

To improve the upper bound, we must impose a penalty that punishes violations of the non-anticipativity constraints. Before proceeding, we first discuss a variation of the problem that will be helpful. Specifically, the greedy policy ranks items using their effective values, so it will be useful for us to work with a variation of the problem in which the values v_i are replaced by the effective value w_i. In this variation, we also need to include the value of the overflowing item. This leads us to the formulation

$$W^* \;=\; \max_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{n \wedge \tau^\pi} w_{\pi_t} \,,$$

where the set of policies Π and the stopping time τ^π are defined as before. We first show that this formulation provides an upper bound.

Proposition 3.1. $V^* \leq W^*$.

We now return to the penalty, which we apply to this effective value formulation. We use a vector of weights $\mathbf{z} = (z_i)_{i=1}^n \in \mathbb{R}^n$ and consider the partial sum $M_j^\pi = \sum_{t=1}^j z_{\pi_t}(\tilde{s}_{\pi_t} - \mu_{\pi_t})$, where $\tilde{s}_i = \min\{s_i, \kappa\}$ and π_t is the item to be inserted by the policy in consideration at time t. We use $M^\pi_{\tau^\pi \wedge n}$ as a penalty; this is dual feasible in the sense of BSS in that it does not penalize, in expectation, any non-anticipative policy. Dual feasibility follows because, for every non-anticipative policy $\pi \in \Pi$, τ^π is a stopping time and M_t^π is a martingale with respect to the natural filtration. Thus the Optional Stopping Theorem implies that $\mathbb{E}\left[M^\pi_{\tau^\pi \wedge n}\right] = M_0^\pi = 0$, since the stopping time $\tau^\pi \wedge n$ is bounded.

Using the effective value formulation, we can write the problem with penalties included as

$$W_z^* = \max_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{n \wedge \tau^\pi} w_{\pi_t} + z_{\pi_t}(\tilde{s}_{\pi_t} - \mu_{\pi_t}) \,. \tag{5}$$

Because the penalty is dual feasible as described above, $W_z^* = W^*$. We let $W_z^{\mathrm{P}}(\mathbf{s})$ denote the optimal

(deterministic) value of the penalized perfect information problem for sample path $\mathbf{s} \in \mathbb{R}^n_+$. Since the set of perfect information policies includes the set Π of non-anticipative policies, and the penalty is dual feasible, we obtain an upper bound $W^* \leq W^{\mathrm{P}}_z$, where $W^{\mathrm{P}}_z = \mathbb{E}_{\mathbf{s}}[W^{\mathrm{P}}_z(\mathbf{s})]$ denotes the penalized perfect information bound. In Section C.1 we discuss how to calculate $W^{\mathrm{P}}_z(\mathbf{s})$ by solving an integer program that includes additional variables representing which item, if any, overflows the knapsack.

With this procedure, any $\mathbf{z} \in \mathbb{R}^n$ provides an upper bound on W^*, and hence by Proposition 3.1, on V^*. How should we choose $\mathbf{z}$ to get a good bound? Recall that the perfect information policy may "cheat" by selecting items with relatively low realized sizes. The penalty we use seeks to align the perfect information policy with the greedy policy by creating incentives for the decision maker with perfect information to resist selecting items with realized sizes that are small relative to their expected sizes. In our analysis, we will use the penalty corresponding to $z_i = w_i/\mu_i$. Notice that in this case, the penalized value of selecting item i then becomes $w_i + z_i(\tilde{s}_i - \mu_i) = (w_i/\mu_i)\tilde{s}_i$. Thus, in the penalized perfect information problem, the decision maker may "cheat" and select items with low realized sizes, but will also receive less value for doing so, as the objective is now proportional to the realized sizes.

It is instructive to see how this works on the example from Dean et al. (2008) as discussed in Section 3.2, with n symmetric items of value one and sizes that are Bernoulli with probability $1/2$, scaled by $\kappa + \varepsilon$. Recall that a greedy policy is (trivially) optimal and the optimal value is $V^* = 1 - (1/2)^n$, but the perfect information bound without penalty provides the poor bound of $V^{\mathrm{P}} = n/2$. In this example, $w_i = 1/2$, $\mu_i = \kappa/2$, and $\tilde{s}_i$ is either 0 or κ, each with probability $1/2$. In the penalized perfect information problem, the value for selecting an item is $(w_i/\mu_i)\tilde{s}_i$, which is 0 if $\tilde{s}_i = 0$ and 1 if $\tilde{s}_i = \kappa$: 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. Thus, $W^{\mathrm{P}}_z(\mathbf{s}) = 1$ if $s_i > 0$ for any i, and $W^{\mathrm{P}}_z(\mathbf{s}) = 0$ otherwise. Because $\mathbb{P}\{s_i = 0 \ \forall \ i\} = (1/2)^n$, the penalized perfect information bound then is

$$W^{\mathrm{P}}_z \ = \ 1 - (1/2)^n \ = \ V^*,$$

i.e., we recover a tight bound for all values of n.

In general, with the choice $z_i = w_i/\mu_i$, 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 due to value collected from an overflowing item, analogous to the analysis of LP relaxations in the deterministic case in (3).

3.4 Performance analysis

We now formalize the above discussion in the general case. We show that the greedy policy incurs a small loss in value compared to 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.

Proposition 3.2. *The performance of the greedy policy satisfies:*

(i) ***Performance guarantee.***

$$V^{\mathrm{G}} \leq V^* \leq W_z^{\mathrm{P}} \leq V^{\mathrm{G}} + \max_i w_i + \mathbb{E}\left[\max_i \frac{w_i \tilde{s}_i}{\mu_i}\right]. \tag{6}$$

(ii) ***Asymptotic optimality.*** *Suppose that* $\lim_{n\to\infty} \frac{1}{\kappa}\mathbb{E}\left[\max_i \frac{w_i \tilde{s}_i}{\mu_i}\right] = 0$, *then*

$$\lim_{n\to\infty} \frac{1}{\kappa}\left(V^* - V^{\mathrm{G}}\right) = 0\,. \tag{7}$$

We prove (6) by relating the optimal value of the penalized perfect information problem W_z^{P} to the performance of the greedy policy V^{G}. Recall that with the penalty $z_i = w_i/\mu_i$, the values for selecting items with low realized sizes are adjusted downwards and thus the decision maker with perfect information has less incentive to "cheat" by selecting items with low realized sizes. The decision maker with perfect information, however, can still "cheat" by choosing a large item to overflow the knapsack since it receives the value of the overflowing item. We handle this issue by decomposing the penalized perfect information problem into (i) a traditional deterministic knapsack problem and (ii) another problem in which the decision maker 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 order is optimal and a loss of at most $\max_i w_i$ 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 (6). In the second problem the decision maker simply chooses the item with largest penalized value $(w_i/\mu_i)\tilde{s}_i$ as a candidate to overflow the knapsack, which leads to the second loss term in (6).

Proposition 3.2 shows that the greedy policy is asymptotically optimal when the expected maximum penalized value grows more slowly than the capacity of the knapsack; this, in turn, limits the value the penalized perfect information policy can obtain by overflowing the knapsack.[1] Note that asymptotic optimality only requires the loss from the second term in (6) to vanish: this follows from the fact that $\max_i w_i \leq \mathbb{E}[\max_i(w_i\tilde{s}_i/\mu_i)]$, as we show in the proof of the result. The condition given in Proposition 3.2(ii),

[1] 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.

while general, may be difficult to verify directly. The next result provides sufficient conditions that are easy to check and are not very restrictive in order for asymptotic optimality to hold. We say a_n is *little omega* of b_n or $a_n = \omega(b_n)$ if $\lim_{n=\infty} a_n/b_n = \infty$, that is, a_n grows asymptotically faster than b_n.

Corollary 3.3. *Suppose that the ratios of value-to-size are uniformly bounded, that is, $w_i/\mu_i \leq \bar{r}$ for some $\bar{r} < \infty$ independent of the number of items n. Then (7) holds if:*

(a) Sizes are uniformly bounded, that is, $s_i \leq \bar{s} < \infty$, and capacity scales as $\kappa = \omega(1)$.

(b) Sizes have uniformly bounded means and variances, that is, $\mathbb{E}[s_i] \leq \bar{m} < \infty$ and $Var(s_i) \leq \bar{\sigma}^2 < \infty$, and capacity scales as $\kappa = \omega(\sqrt{n})$.

(c) Sizes are normally distributed with mean $m_i \leq \bar{m} < \infty$ and standard deviation $\sigma_i \leq \bar{\sigma} < \infty$, and capacity scales as $\omega(\sqrt{\log n})$.

(d) Sizes are Pareto distributed with scale $m_i \leq \bar{m} < \infty$ shape $\alpha_i \geq \underline{\alpha} > 0$, and capacity scales as $\kappa = \omega(n^{1/\underline{\alpha}})$.

Intuitively, the growth of the maximum penalized value $\max_i(w_i\tilde{s}_i)/\mu_i$ is governed to a large extent by the tails of the distributions of sizes. When items are symmetric, roughly speaking, we have that $\mathbb{E}[\max_i s_i] \approx F^{-1}(n/(n+1))$ where F is the cumulative distribution function of sizes, and thus we need capacity to grow at least as $F^{-1}(n/(n+1))$ for (7) to hold. The previous result 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 are normally distributed (an example of a light-tailed distribution), it suffices that capacity grow at a logarithmic rate. Similar conditions can be verified as sufficient for sub-gaussian distributions. When sizes are Pareto distributed (an example of a heavy-tailed distribution), it suffices that capacity grow at a power-law rate.

3.5 Alternative bounds

In the performance analysis above, it was convenient to work with the effective value formulation, which provides an upper bound on V^*, as a starting point. This is helpful because it links to the greedy policy, which uses effective values w_i and mean truncated sizes μ_i. Although this suffices in our theoretical analysis, in specific examples, the upper bound W^* itself may have some slack due to the fact that this formulation receives full value for an item that overflows the knapsack.

We can also apply a penalty directly to the original formulation of the problem (i.e., with values and sizes not adjusted, and no value received for an overflowing item). We use the same form of penalty as above,

except different values for the vector $\mathbf{z}$. Including this penalty, the problem is

$$V_z = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{n \wedge (\tau^\pi - 1)} v_{\pi_t} + \sum_{t=1}^{n \wedge \tau^\pi} z_{\pi_t}(s_{\pi_t} - \mathbb{E}[s_{\pi_t}]) \right].$$

As above, the penalty is mean zero for any feasible policy, so $V_z = V^*$. The only differences between this formulation and (5) are: (a) we use actual values and sizes rather than effective values and truncated sizes; and (b) values can only be collected prior to overflow, as in the original statement of the problem. Note that despite point (b), we must still include the penalty terms until τ^π, i.e., until it is known that an overflow has occurred. This is required to ensure dual feasibility of the penalty: for any feasible policy $\pi \in \Pi$, τ^π is a stopping time, but $\tau^\pi - 1$ is not.

With this approach, we now choose $z_i = v_i/\mathbb{E}[s_i]$, which is analogous to the choice above, except with non-adjusted values and sizes. The optimal value with perfect information $V_z^{\mathrm{P}}(\mathbf{s})$ for a given sample path $\mathbf{s}$ can again be calculated as an integer program, very similar in form to that for $W_z^{\mathrm{P}}(\mathbf{s})$ but with different objective coefficients. This is discussed in Section C.2. We have found in our numerical examples that we often get a stronger bound from $V_z^{\mathrm{P}} = \mathbb{E}_{\mathbf{s}}[V_z^{\mathrm{P}}(\mathbf{s})]$ than W_z^{P}; although this need not always be true, this is often the case due to the fact that $V_z^{\mathrm{P}}(\mathbf{s})$ credits an overflowing item somewhat differently. (We can also use the bound V_z^{P} to derive a result similar to Prop. 3.2, but comparing to a greedy policy that ranks items by $v_i/\mathbb{E}[s_i]$. Such a greedy policy, however, does not appear to be standard in the literature on this problem.)

In our examples, we also compare to upper bounds derived in Dean et al. (2008). Specifically, we also calculate:

$$V^{\mathrm{DGV}} \;=\; \max\left\{ \sum_{i=1}^{n} w_i x_i \;:\; \sum_{i \in S} \mu_i x_i \le 2\kappa \left(1 - \prod_{i \in S}(1-\mu_i)\right) \;\forall\, S \subseteq \mathcal{N},\; 0 \le x_i \le 1,\; \forall\, i \in \mathcal{N} \right\}. \qquad (8)$$

Dean et al. (2008) show that $V^{\mathrm{DGV}} \ge V^*$ and that V^{DGV} involves optimization over a polymatroid and can thus be done easily following the greedy ordering; they also show that V^{DGV} is a strengthening of an upper bound from a deterministic LP relaxation that replaces sizes with their mean truncated sizes and allows the knapsack to have double capacity (i.e., 2κ). The goals of the analysis in Dean et al. (2008) are somewhat different than ours in that they seek constant factor approximation guarantees across all problem instances. Dean et al. (2008) show that $V^{\mathrm{DGV}} \le 4V^*$, a result that is useful in obtaining constant factor guarantees, and also provide examples where this factor of 4 is nearly attained. Thus, we should not necessarily expect V^{DGV} to provide an especially good bound on a specific problem instance, but it does provide a benchmark upper bound that is easy to compute.

Blado et al. (2015) provide sophisticated bounds based on approximate dynamic programming and find

that these bounds perform well in examples. These bounds, however, are somewhat involved to implement (e.g., requiring column generation) and may be computationally challenging (e.g., Blado et al. (2015) report runtimes of several hours on examples with $n = 100$). For these reasons, we did not attempt to calculate these bounds.

3.6 Numerical examples

We demonstrate the upper bounds on a set of randomly generated examples. We consider $n = 50$, 100, 500, and 1000. We generate values as $v_i \sim \mathrm{U}[0, 1]$ and mean sizes as $\mathbb{E}[s_i] \sim \mathrm{U}[0, 1]$, all i.i.d. Thus the mean value of the expected sizes is 1/2, and we let $\kappa = \theta n/2$, where $\theta \in \{1/8, 1/4, 1/2\}$. The case $\theta = 1/8$ corresponds to tight capacity relative to the case $\theta = 1/2$ (or, equivalently, relatively larger sizes in the $\theta = 1/8$ case). For each value of n and each value of θ, we randomly generate 20 sets of values and sizes as described above, and for each set, we consider an instance with:

(i) Exponential sizes, with rate $1/\mathbb{E}[s_i]$ for each item;

(ii) Bernoulli sizes, with probability 1/2, supported on $\{0, 2\mathbb{E}[s_i]\}$ for each item; and

(iii) Uniform sizes, supported on $[0, 2\mathbb{E}[s_i]]$ for each item.

Note that, for consistency across the distributions, we are using the same values and mean sizes for each item for a given instance. For each n, we have 20 instances for 3 different θ values and 3 different distributions, or 180 total instances for each n.

For each instance, we calculate the lower bound V^{G} corresponding to the greedy policy, and the upper bounds corresponding to: (a) V_z^{P} the perfect information bound with penalty as discussed in Section 3.5; (b) W_z^{P} the perfect information bound with penalty, using the effective value formulation; (c) V^{P}, the perfect information bound without penalty; and (d) the upper bound V^{DGV} from Dean et al. (2008). All bounds (other than V^{DGV}, which does not require simulation) are estimated with 100 sample paths; this led to low mean standard errors in the results. All calculations were done using Matlab on a desktop computer; we used the MOSEK Optimization Toolbox for solving the integer programs in the upper bound calculations. We note that for instances with $n = 500$ and $n = 1000$, we solved the LP relaxations for the integer programs corresponding to W_z^{P} and V_z^{P}. This is of course sufficient to obtain upper bounds and had small impact on the results while significantly reduces the runtime for these large instances. These bounds can be computed efficiently in practice: for example, in instances with $n = 1000$ the calculation of V_z^{P}, W_z^{P}, and V^{P} typically took around 20 seconds per sample without much code optimization.

Table 1 summarizes the results for the intermediate value of $\theta = 1/4$; the other results are similar and in Section E of the appendix. Table 1 shows the 25^{th}, 50^{th}, and 75^{th} percentiles of the gaps relative to the

Exponential Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	10.59%	6.89%	2.15%	1.25%
		50%	11.34%	7.26%	2.23%	1.30%
		75%	12.28%	7.90%	2.35%	1.33%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	15.46%	9.35%	2.71%	1.53%
		50%	16.38%	9.89%	2.77%	1.58%
		75%	17.57%	10.72%	2.92%	1.62%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	25.21%	27.11%	27.60%	26.91%
		50%	26.92%	28.91%	29.19%	27.94%
		75%	33.40%	32.71%	29.81%	29.07%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	39.07%	38.99%	39.22%	39.17%
		50%	43.24%	41.80%	40.63%	39.84%
		75%	47.10%	44.57%	41.60%	40.87%

Bernoulli Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	3.88%	2.34%	0.63%	0.33%
		50%	4.18%	2.41%	0.69%	0.34%
		75%	4.54%	2.61%	0.70%	0.35%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	9.71%	5.36%	1.28%	0.67%
		50%	10.72%	5.65%	1.38%	0.69%
		75%	11.13%	5.97%	1.41%	0.69%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	32.14%	35.87%	36.23%	35.01%
		50%	36.77%	37.48%	37.62%	36.03%
		75%	41.15%	42.16%	38.74%	37.81%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	40.18%	38.32%	40.02%	39.41%
		50%	43.12%	42.32%	41.25%	40.44%
		75%	46.97%	44.93%	42.23%	40.96%

Uniform Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	5.22%	3.03%	0.73%	0.37%
		50%	5.70%	3.23%	0.75%	0.39%
		75%	6.25%	3.46%	0.78%	0.40%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	10.75%	5.88%	1.35%	0.69%
		50%	11.76%	6.19%	1.39%	0.71%
		75%	12.59%	6.64%	1.45%	0.73%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	15.41%	17.12%	15.92%	15.78%
		50%	17.63%	18.37%	16.76%	16.35%
		75%	19.57%	19.61%	17.79%	16.80%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	40.11%	40.44%	39.84%	39.84%
		50%	46.11%	42.77%	40.75%	40.43%
		75%	48.25%	45.75%	42.25%	41.35%

Table 1: Stochastic knapsack results for $\theta = \frac{1}{4}$.

greedy policy, across the 20 instances in each case. We report percentiles, rather than expected values, to provide a better sense of the spread of the gaps. The bound V^{DGV} (rows (d)) consistently leads to a gap of around 40%. The perfect information bound without penalty V^{P} (rows (c)) is also poor, although somewhat better (e.g., gaps around $15-20\%$ in the case of uniform sizes). These bounds alone shed no light on the asymptotic optimality of the greedy policy for large n. The penalized perfect information bounds V_z^{P} and W_z^{P} (rows (a) and (b), respectively) perform much better and clearly convey the asymptotic optimality of the greedy policy. Typically, V_z^{P} provides a stronger bound than W_z^{P}, although the difference diminishes as n grows. The gaps are somewhat larger with exponential sizes, as the distributions have heavier tails than in the other cases; note that we know from existing results that greedy is optimal with exponential sizes, so the gaps in examples with exponential sizes are solely due to the upper bounds. Tables 3 and 4 show similar results, with somewhat larger (smaller) gaps with relatively tighter (looser) capacity, i.e., $\theta = 1/8$ ($\theta = 1/2$).

Although the theory tells us the penalized perfect information bounds must converge, we were surprised by the performance of these bounds on these specific examples.

4 Stochastic project completion

We consider a model of a firm working on a project that, upon completion, generates reward for the firm. A motivating example is a firm developing a new product, with the reward being the expected NPV from launching the product. The firm is concurrently working on n different *alternatives*, and finishing *any* of these alternatives completes the project. The firm faces uncertainty in completing these alternatives, and needs to optimally balance development costs with the pressure to quickly complete the project. Although the model we discuss is stylized, at a high level the model captures a problem faced by firms in a variety of industries, where different alternatives are being considered towards some desired end goal, e.g. a pharmaceutical company studying the efficacy of various compounds when developing a new cancer treatment, or a tech company exploring the viability of different designs for a new smart phone.

Associated with each alternative i a state variable x_i, which represents the number of steps remaining to complete alternative i. In each period prior to the completion, the firm incurs a baseline cost of $c_0 \geq 0$, which is the cost of concurrently working on all of the alternatives. Upon the first completion of any of the n alternatives, i.e., when $x_i = 0$ for any i, the firm generates a reward (i.e., an expected net present value of future profits) of $R = r/(1-\delta)$, where $r \geq 0$ and δ is the discount factor, and the problem ends.

Prior to completion, the state of each alternative evolves randomly and independently (over both time and alternatives) according to a discrete-time Bernoulli process. Under baseline development, each alternative transitions from x_i to $x_i - 1$ with probability q_i or remains at x_i with probability $1 - q_i$. If the state of

alternative i changes from x_i to $x_i - 1$, we say that alternative i *improves*. In each period, the firm can also devote additional resources in an effort to accelerate the completion of an alternative. If the firm accelerates alternative i in a given period, the cost of working in that period is $c_i \geq c_0$, and alternative i improves with probability $p_i \geq q_i$; all other alternatives $j \neq i$ improve with probability q_j as in baseline development. We assume the firm can accelerate at most one alternative per period. The goal is to maximize the expected discounted reward over an infinite horizon.

This problem resembles a restless bandit problem in that the firm can either work "actively" (accelerate) or "passively" (not accelerate) on alternatives in each period, and the states of alternatives evolve stochastically whether or not alternatives are accelerated. Distinguishing this problem from a restless bandit problem, however, is a coupled termination condition: the problem ends upon completion of *any* alternative.

We can write this problem as a stochastic dynamic program with state variable $\boldsymbol{x} := (x_1, \ldots, x_n)$. Prior to completion, i.e., when $x_i > 0$ for all i, the value function is

$$V(\boldsymbol{x}) \;=\; \max_{i=0,\ldots,n} \left\{ -c_i + \delta \mathbb{E} V(\boldsymbol{x} - \boldsymbol{d}^i) \right\}, \tag{9}$$

and $V(\boldsymbol{x}) = R$ whenever $x_i = 0$ for some i. In (9), the expectation is over $\boldsymbol{d}^i \in \{0,1\}^n$, which is a vector of independent Bernoulli random variables such that $d^i_i = 1$ with probability p_i, and $d^i_j = 1$ with probability q_j for all $j \neq i$. The random vector $\boldsymbol{d}^i$ captures the improvements across all alternatives when accelerating alternative i. Note that because $p_i \geq q_i$ for all i, we can restrict to three outcomes for each alternative in each period: (i) the alternative improves only if it is accelerated; (ii) the alternative improves regardless of decision (i.e., under baseline development and if accelerated); and (iii) the alternative does not improve regardless of decision. In (9), the index 0 represents the option of accelerating no alternative. Finally, we let $\boldsymbol{x}_0$ denote the initial state, which we assume is known.

It will be convenient to write the expected reward of a given, feasible policy π, which maps from states $\boldsymbol{x}$ to an action in $\{0, \ldots, n\}$. We denote this expected reward by V^π and can write this as

$$V^\pi \;=\; \mathbb{E}\left[\delta^{\tau^\pi} R - \sum_{t=0}^{\tau^\pi - 1} \delta^t c_{\pi(x_t)} \right] \;=\; R - \mathbb{E} \sum_{t=0}^{\tau^\pi - 1} \delta^t \left(r + c_{\pi(x_t)} \right),$$

where $\tau^\pi = \inf\{t : x^\pi_{t,i} = 0 \text{ for some } i = 1, \ldots, n\}$ denotes the stopping (completion) time of the policy, $\boldsymbol{x}^\pi_t$ denotes the state process of the policy, and to simplify the notation we dropped the dependence on the initial state $\boldsymbol{x}_0$. Thus, this problem is equivalent to a stochastic shortest path problem with discounting and costs given by $\alpha_i = r + c_i$ when alternative i is accelerated. We let Π denote the set of feasible, non-anticipative

policies, and we can equivalently define the optimal reward as

$$V^* \;=\; R - \min_{\pi\in\Pi} \mathbb{E} \sum_{t=0}^{\tau^\pi - 1} \delta^t \alpha_{\pi(x_t)}. \tag{10}$$

From the principle of dynamic programming it follows that $V^* = V(\boldsymbol{x}_0)$. Following (10), it will sometimes be convenient to focus simply on the expected discounted costs until completion, and we will use J in place of V accordingly, with the understanding that $J^\pi = R - V^\pi$.

We will assume that the acceleration costs are sufficiently small: specifically, we assume

$$\frac{c_i - c_0}{r + c_0} \leq \delta(p_i - q_i)\,, \tag{11}$$

for $i = 1, \ldots, n$. Although (11) does somewhat restrict our study of this problem, given that the firm views the project as attractive to develop, we may expect acceleration costs would be small relative to rewards. Note that (11) requires $c_i - c_0 = 0$ for any alternatives i such that $p_i = q_i$, which is intuitive: the marginal cost of accelerating an alternative should be zero if accelerating the alternative provides no benefit. It can also be shown that, in the simple case with $n = 1$ alternative in isolation and with one step to completion remaining, (11) implies that acceleration is optimal.

4.1 Static policies

Solving for an optimal policy may be very difficult, as the number of states is exponential in n; if there are many alternatives or several alternatives with many uncompleted steps, many possible states must be considered in the recursion in (9). Intuitively, however, we may expect policies that accelerate alternatives that appear most promising - as in, closest (or cheapest) to completion - to perform well. Motivated by this intuition, we will study the quality of static policies, which commit to accelerating a fixed alternative (or no alternative) in every state prior to completion. We can write the expected reward V^{S_i} of the i^{th} static policy, which accelerates alternative i in every period, as

$$V^{\mathrm{S}_i} \;=\; R - \alpha_i \mathbb{E} \sum_{t=0}^{\tau^{\mathrm{S}_i} - 1} \delta^t \;=\; R - \underbrace{\frac{\alpha_i}{1-\delta} \cdot \mathbb{E}[1 - \delta^{\tau^{\mathrm{S}_i}}]}_{:= J^{\mathrm{S}_i}}.$$

Analogous to above, τ^{S_i} denotes the stopping time of the i^{th} static policy, which can be written as

$$\tau^{\mathrm{S}_i} \;=\; \min\left\{Y_1^i, \ldots, Y_n^i\right\}, \tag{12}$$

where Y_j^i is a negative binomial distribution supported on $\{x_{0,j}, x_{0,j}+1, \ldots\}$ with $x_{0,j}$ successes and success probability q_j if $j \neq i$ (or p_j if $j = i$).[2] Although calculating the exact distribution of τ^i is difficult, we can easily estimate V^{S_i} by evaluating (12) using Monte Carlo simulation. We then select the best static policy, which has expected reward $V^{\mathrm{S}} = \max_{i=0,\ldots,n} V^{\mathrm{S}_i}$.

If the alternatives can never improve without acceleration, i.e., if $q_i = 0$ for all $i = 1, \ldots, n$, then a static policy is optimal. This follows from the observation that in this case, if it is optimal to accelerate alternative i in a given state, then it will also be optimal to accelerate i in the ensuing state: either the state of i improves or it does not, and the states of all other alternatives do not change in this case. If $q_i > 0$ for some alternatives, however, a static policy will not be optimal in general: alternatives that are not accelerated in a given period can nonetheless improve, and in later periods it may be optimal to accelerate such alternatives if they become close enough to completion.

Although static policies are not generally optimal, we would expect the best static policy to perform well in states close to completion. For example, when $x_{0,i} = 1$ for all i, then it is straightforward to argue that a static policy is optimal: in the next period, the firm will either complete some alternative or not. If not, the state remains unchanged ($x_{0,i} = 1$ for all i since no alternative was completed). Thus, there is only one state to consider and a static policy is trivially optimal.

We will show that static policies perform well in the other extreme: namely, we will show that static policies are asymptotically optimal in states *far* from completion. The asymptotic analysis that we consider will focus on the case when the horizon effectively becomes very large, i.e., $\delta \to 1$, and the initial state $x_{0,i}$ grows commensurately as $O\left((1-\delta)^{-1}\right)$. One interpretation of this scaling is that number of steps to complete alternatives grows at the same rate as the length of the horizon $(1-\delta)^{-1}$, which is reminiscent of the fluid scalings used in revenue management (see, e.g., Gallego and van Ryzin (1994)).

4.2 Perfect information bound

As a starting point for our analysis, we first consider the perfect information upper bound, which is given by the expected reward of a firm that knows the future state transitions for all alternatives in advance. We can write the optimal expected reward with perfect information V^{P} as

$$V^{\mathrm{P}} \;=\; R - \mathbb{E}_{\mathbf{d}} \min_{\pi \in \Pi^{\mathrm{P}}} \sum_{t=0}^{\tau^\pi - 1} \delta^t \alpha_\pi, \tag{13}$$

[2] We define a negative binomial distribution supported on $\{x, x+1, \ldots\}$ with $x \in \mathbb{N}$ successes and success probability $p \in (0,1]$ (or $\mathrm{NB}(x,p)$ for short) as the *total* number of trials needed to get x successes when the probability of each individual success is p. The probability generating function is given by $\mathbb{E}\left[z^{\mathrm{NB}(x,p)}\right] = (zp/(1-(1-p)z))^x$ for $|z| \leq 1/(1-p)$.

where the expectation is over the sample path $\mathbf{d} := (\boldsymbol{d}_t^0, \ldots, \boldsymbol{d}_t^n)_{\{t\geq 0\}}$ describing alternative state transitions in each period for each possible action (the sample path $\mathbf{d}$ is independent of the policy), and Π^{P} denotes the set of all policies with perfect information on the particular sample path. It is clear that V^{P} is an upper bound on V^*, since all non-anticipative, feasible policies are contained in Π^{P}.

In Appendix D.1 we show that the perfect information problem can be efficiently solved via Monte Carlo simulation by using two observations. First, for every sample path $\mathbf{d}$, the perfect information problem is separable in the alternatives: that is, it suffices to find the policy that completes each alternative in isolation with least cost and then take the associated alternative with the minimum cost. Second, we show that under assumption (11) the optimal policy to complete alternative i in isolation has a simple structure: namely, accelerate alternative i only in periods when the alternative requires acceleration to improve. (Note that if (11) does not hold, the analysis is more complicated because always accelerating i in such periods need not be optimal. For example, if the cost of acceleration is large, the perfect information policy may not accelerate alternative i when this alternative can be improved shortly thereafter without acceleration. Thus, if (11) does not hold, the optimal action in each period could depend on the remainder of the sample path.) Thus the perfect information policy can "cheat" and save acceleration costs in time periods when an alternative will improve without acceleration, as well as in time periods when the alternative will not improve even if accelerated.

To illustrate the performance of the perfect information bound, we consider a simple example with $n = 1$ alternative, and the values $r = 4$, $c_0 = 0$, $c_1 = 2\delta$, $p_1 = 1/2$, and $q_1 = 0$. These parameters satisfy assumption (11) with equality. Since $q_1 = 0$, the firm can never complete the alternative without accelerating it, and the static policy that always accelerates this alternative is in fact optimal. The expected reward of this policy is

$$V^{\mathrm{S}} \;=\; R - \frac{\alpha_1}{1-\delta} \cdot \mathbb{E}[1-\delta^{\tau}] \;=\; \frac{1}{1-\delta}\left((4+2\delta)\mathbb{E}\delta^{\tau} - 2\delta\right),$$

where τ has a negative binomial distribution with parameters x_1 and $p_1 = 1/2$.

In the perfect information problem in this example, the firm observes $\mathbf{d} = (d_t^0, d_t^1)_{\{t\geq 0\}}$ before making acceleration decisions. Since $q_1 = 0$, we have $d_t^0 = 0$ in every period; that is, the alternative never improves without acceleration. Also, with probability $p_1 = 1/2$, d_t^1 equals 1 in a given period, which corresponds to an improvement of the alternative if the firm accelerates the alternative in period t. Conversely, with probability $1 - p_1 = 1/2$, $d_t^1 = 0$, which corresponds to no improvement of the alternative in period t, regardless of the firm's decision. It is clear that the optimal policy in the perfect information problem is, in all periods up to completion, to accelerate the alternative (at cost $r + c_1$) in periods when $d_1^t = 1$ and to not accelerate the alternative (at cost r) in periods when $d_t^1 = 0$. Using this policy in (13), we can show that the optimal

expected reward with perfect information is

$$V^{\mathrm{P}} \;=\; R - \frac{(1-p_1)r + p_1(r+c_1)}{1-\delta} \cdot \mathbb{E}[1-\delta^{\tau}] \;=\; \frac{1}{1-\delta}\left((4+\delta)\,\mathbb{E}\delta^{\tau} - \delta\right),$$

where the first equality follows from a basic martingale argument (see Lemma B.3). The relative suboptimality gap then scales, when $x_1 = \lceil \bar{x}/(1-\delta) \rceil$ for some $\bar{x} > 0$, as

$$(1-\delta)(V^{\mathrm{P}} - V^{\mathrm{S}}) \;=\; \delta\,(1 - \mathbb{E}\delta^{\tau}) \;\xrightarrow[\delta \to 1]{}\; 1 - e^{-2\bar{x}},$$

where the limiting expression follows from using the probability generating function for a negative binomial random variable and taking limits. (The factor of $1-\delta$ normalizes the reward difference to be in units of average reward per period and facilitates comparisons between different discount factors.)

This analysis on its own may lead one to conclude that, in the limit as the number of states grows to be very large, static policies are suboptimal, perhaps significantly so, sacrificing an average per period reward up to a constant factor of $1 - e^{-2\bar{x}}$. This conclusion is false: as argued above, a static policy is optimal for this problem. The gap in this analysis is arising from slack in the perfect information bound. In this problem, there is substantial value to the information about the future state transitions of the alternatives, and we require a penalty to compensate for this information and obtain a better upper bound.

4.3 Penalized perfect information bound

The perfect information problem does not provide a tight upper bound because, with advance knowledge about state transitions, the firm can complete the project cheaply. In the example above, the firm can save acceleration costs in time periods when $d_t^1 = 0$, as the alternative will not improve in these periods even with acceleration, whereas in the static (and optimal in that example) policy, the firm incurs acceleration costs in every period prior to completion. We will attempt to improve the perfect information bound by incorporating a penalty.

We will consider penalties that depend on the selected action in every period and $\mathbf{d}$. The penalty in a given time period t prior to completion when alternative i is accelerated is denoted $z_i(\mathbf{d}_t)$, where $\mathbf{d}_t$ is the period t component of $\mathbf{d}$, i.e., $\mathbf{d}_t = (\boldsymbol{d}_t^0, \ldots, \boldsymbol{d}_t^n)$ and the total discounted penalty is $\sum_{t=0}^{\tau^\pi - 1} \delta^t z_\pi(\mathbf{d}_t)$ when the policy is π. Following BSS, in order to obtain an upper bound on V^*, we require a dual feasible penalty:

$$\mathbb{E}_{\mathbf{d}}\left[\sum_{t=0}^{\tau^\pi - 1} \delta^t z_\pi(\mathbf{d}_t)\right] \le 0 \qquad \text{for all } \pi \in \Pi. \tag{14}$$

A sufficient condition for (14) is $\mathbb{E}_{\mathbf{d}_t}[z_i(\mathbf{d}_t)] \leq 0$ for each fixed action i since $\tau^\pi - 1$ is a stopping time. For a given penalty z, we denote the optimal reward with perfect information by V_z^{P}, which satisfies

$$V_z^{\mathrm{P}} \;=\; R - \mathbb{E}_{\mathbf{d}} \min_{\pi \in \Pi^{\mathrm{P}}} \sum_{t=0}^{\tau^\pi - 1} \delta^t (\alpha_\pi + z_\pi(\mathbf{d}_t)).$$

Our goal is to find a penalty that (i) is dual feasible according to (14), and (ii) leads to an upper bound that we can relate to the reward of the static policy. We first provide an intermediate result along these lines. In the following, we denote by $J_z^{\mathrm{P}} = R - V_z^{\mathrm{P}}$ the penalized cost until completion with perfect information.

Proposition 4.1. *There exists a dual feasible penalty z such that*

$$J_z^{\mathrm{P}} \;=\; \frac{1}{1-\delta} \mathbb{E}_{\mathbf{d}} \min_{i=1,\ldots,n} \sum_{t=0}^{\tau^i - 1} \delta^t (\alpha_i + z_i(\mathbf{d}_t)), \tag{15}$$

where τ^i is the completion time of alternative i when the firm accelerates i in every time period and, in addition, $\mathbb{E}_{\mathbf{d}_t}[\alpha_i + z_i(\mathbf{d}_t)] \geq (J^{\mathrm{S}}/J^{\mathrm{S}_i})\alpha_i$ for all $i \in \{1, \ldots, n\}$.

Proposition 4.1 states that we can find a dual feasible penalty and associated perfect information lower bound (on costs) that is no smaller than the expected value of the best (penalized) static cost in every sample path. This alone does not show that the static policy performs well, because the expectation in (15) is outside the minimization; however, Proposition 4.1 also adds that the expected per stage penalized costs are sufficiently large relative to the expected completion costs associated with the static policy. This latter fact will be essential in using (15) to provide a bound on the performance of the static policy. The proof of Proposition 4.1 describes explicitly a penalty that satisfies the desired properties. Before proceeding, we provide some intuition for how this penalty works.

Recall that the perfect information policy can potentially "cheat" (incur no acceleration costs) in time periods when alternatives will improve without acceleration (or not improve even with acceleration). The penalty partially aligns the perfect information policy with the best static policy that aims to complete alternative i by creating incentives for the decision maker with perfect information to accelerate i in every period. The penalty accomplishes this by effectively reducing the acceleration costs for i in periods when i will improve (or not) regardless of whether i is accelerated, but increasing the cost of accelerating i in periods when i can only improve if accelerated. The proposed penalty carefully balances these cost changes to ensure that the perfect information policy accelerates alternative i in all periods, and also to ensure that the dual feasibility condition (14) holds, so that the procedure leads to an upper bound on V^*.

4.4 Performance analysis

Proposition 4.1 leads us to our main result for this problem.

Proposition 4.2. *Suppose* $(1-\delta)x_{0,i} \leq \bar{x}$ *for* $i = 1, \ldots, n$ *for some* $\bar{x}$ *independent of* δ*, and all other parameters are held constant. Then the expected reward of the static policy satisfies:*

(i) ***Performance guarantee****. For some* β *independent of* δ*,*

$$V^{\mathrm{S}} \leq V^* \leq V_z^{\mathrm{P}} \leq V^{\mathrm{S}} + \frac{\beta}{\sqrt{1-\delta}}.$$

(ii) ***Asymptotic optimality****.*

$$\lim_{\delta \to 1} (1-\delta)\left(V^* - V^{\mathrm{S}}\right) = 0.$$

Part (i) of Prop. 4.2 provides a uniform bound on the suboptimality of the static policy, and the proof of this result in Appendix B.6 provides the constant β. Part (ii) of Prop. 4.2 is a straightforward consequence of part (i) and states that the expected reward of the static policy approaches that of the optimal policy as the number of states grow and δ approaches one. Note that part (i) characterizes the rate of convergence of this asymptotic optimality.

What is the intuition for this result? We can imagine partitioning the state space into regions where accelerating each alternative is the best action under the optimal policy. If it is ever optimal to accelerate a particular alternative i, then it should be optimal to accelerate i when alternative i is much closer to completion relative to all other alternatives. The static policy commits upfront to the alternative that, in terms of costs, appears "closest" in the initial state. If this initial state is far from a boundary state where the optimal policy would switch between accelerating alternatives, then, it is unlikely the alternative states will ever evolve so as to reach this boundary. If the initial state is, on the other hand, close to one of these boundary states, then it is relatively inconsequential which alternative is accelerated, as both are nearly optimal. Either way, the difference in completion costs relative to the static policy will tend to be small as the number of steps to completion grows.

4.5 Bounds from approximate dynamic programming

As a benchmark for the upper bounds described above, we also consider a bound based on the linear programming approach to approximate dynamic programming (ADP). It is well-known that the optimal value function can be calculated by formulating the DP (9) as a linear program with decision variables corresponding to the value function in every possible state and constraints that impose Bellman optimality

of the value function. The difficulty with this approach is that both number of variables and constraints will scale with the number of states, which may be prohibitively large.

Instead of solving the exact LP formulation for the DP, we can consider an approximation within a parameterized class of functions as is now standard in ADP (de Farias and Roy, 2003). In particular, we approximate the value function as a separable function across alternatives $V^{\text{ADP}}(\boldsymbol{x}) = \sum_{i=1}^{n} V_i(x_i)$ for some alternative-specific value functions V_i. The values $\{V_i(x_i)\}_{0 \le x_i \le x_{0,i}} \in \mathbb{R}^{|x_{0,i}|+1}$ are then variables in the LP:

$$
\begin{aligned}
V^{\text{ADP}} = \underset{V_i(x_i)}{\text{minimize}} \quad & \sum_{i=1}^{n} V_i(x_{0,i}) && && \text{(16a)}\\
\text{subject to} \quad & \sum_{i=1}^{n} V_i(x_i) \ge -c_i + \delta \sum_{i=1}^{n} \mathbb{E} V_i(x_i - d_j^i), && \forall i = 0, \dots, n,\ \boldsymbol{x} > 0, && \text{(16b)}\\
& \sum_{i=1}^{n} V_i(x_i) \ge R, && \forall \boldsymbol{x} : x_i = 0 \text{ for some } i = 1, \dots, n, && \text{(16c)}\\
& V_i(x_i - 1) \ge V_i(x_i), && \forall i = 1, \dots, n,\ x_i = 1, \dots, x_{0,i}. && \text{(16d)}
\end{aligned}
$$

Constraints (16b)-(16c) are relaxations of the Bellman optimality equalities, and (16d) enforce the natural condition that the alternative-specific approximate value functions be nonincreasing in the alternative state. The optimal objective value of this ADP provides an upper bound on the optimal value, i.e., $V^* \le V^{\text{ADP}}$. In Appendix D.2 we show that, by leveraging the separability of the value function, this ADP can be further simplified to an LP with $O(\sum_{i=1}^{n} x_{0,i})$ variables and $O(\sum_{i=1}^{n} x_{0,i})$ constraints.

Although we do not have any formal analysis of the quality of this ADP bound, we will see in our examples that these bounds are somewhat loose. The issue is that the ADP insists that the Bellman inequalities hold in all possible states, and these constraints may bind in states that are very unlikely to occur. In particular, the optimal value function satisfies $V(\boldsymbol{x}) = R$ for all completed states, i.e., for all states $\boldsymbol{x}$ such that $x_i = 0$ for some i. Similarly, the approximate value function V^{ADP} needs to satisfy $V^{\text{ADP}}(\boldsymbol{x}) \ge R$ for all completed states, but because V^{ADP} is restricted to be separable across alternatives, the constraints will tend to bind at "corners" of the sets of completed states, but be slack otherwise. Thus, we may obtain poor approximations of the value function in many of the likely completed states, which can lead to poor approximations in earlier states (in particular, the initial state).

Figure 1 provides an illustration of an example with $n = 2$ alternatives and initial states of $(100, 100)$. This example is small enough that we can solve for the optimal value function. The parameters in this example are $\delta = 0.99$, $(c_0, c_1, c_2) = (0, 0.08, 0.15)$, $(p_1, p_2) = (0.8, 0.9)$, $(q_1, q_2) = (0.08, 0.27)$, and $r = 10$. The upper bound from the ADP has about 19% slack; we will see examples where the ADP upper bound

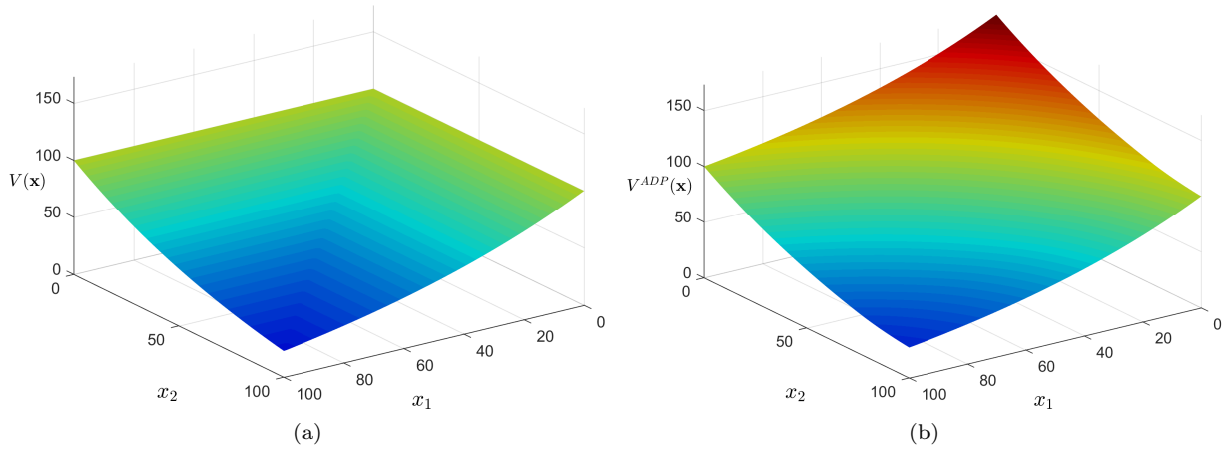

Figure 1: Value functions for an example with $n = 2$ alternatives: (a) optimal vs. (b) ADP.

is worse. In this example, it is evident from Figure 1 that the ADP constraints are binding in the states $(100, 0)$ and $(0, 100)$, which are extremely unlikely to occur under any policy.

4.6 Numerical examples

We investigate the performance of the static policy and the upper bounds on a set of randomly generated examples with $n = 10$ alternatives; these examples have far too many states to solve DP (9). We study four discount factors: 0.9, 0.99, 0.999, and 0.9999. We generate 150 example problems with 10 alternatives and use these same 150 examples across each discount factor. These examples have $r = 10$, and $c_0 \sim \mathrm{U}[0, 1]$. For the alternative advancement probabilities, for each alternative in each example, we generate two $\mathrm{U}[0, 1]$ values and take p_i (q_i) to be the maximum (minimum) of these two values. The initial state $x_{0,i}$ for each alternative is set to $\lceil \bar{x}_i/(1-\delta) \rceil$, where $\bar{x}_i \sim \mathrm{U}[0, 3]$. Finally, we set the acceleration costs of each alternative in each example to be $c_0 + (p_i - q_i)(c_0 + r)\delta$, which ensures that cost assumption (11) holds with equality. These examples are generated to ensure that the optimal static expected reward was positive.

For each example and each discount factor, we evaluate: the lower bound given by the expected reward of the static policy V^{S}, the perfect information upper bound V^{P}, the penalized perfect information upper bound V_z^{P}, and the ADP upper bound V^{ADP}. These first three bounds were calculated using Monte Carlo simulation of $1,000$ sample paths for each example; the resulting mean standard errors of these bounds are small in all examples. Evaluating the static policy is straightforward by calculating the reward for each fixed action until completion in each sample path, then selecting the action with maximum mean reward. This procedure also provides values for the expected costs $J^{\mathrm{S}i}$ for each static policy, which we use in the penalty. To evaluate V_z^{P}, we take advantage of Prop. 4.1, which shows that the penalized perfect information bound can be calculated by taking the expected value of the minimum of the penalized costs for each static

		%ile	δ			
			0.9	0.99	0.999	0.9999
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{S}}}{V^{\mathrm{S}}}$	25%	1.5%	0.061%	-0.0014%	-0.0006%
		50%	4.1%	0.44%	0.06%	0.017%
		75%	18%	4%	1.1%	0.37%
(b)	$\frac{V^{\mathrm{P}} - V^{\mathrm{S}}}{V^{\mathrm{S}}}$	25%	5.32%	2.89%	2.71%	2.73%
		50%	12.1%	9.08%	8.83%	8.78%
		75%	32.3%	26.3%	26.1%	26.1%
(c)	$\frac{V^{\mathrm{ADP}} - V^{\mathrm{S}}}{V^{\mathrm{S}}}$	25%	3.54%	2.19%	2.06%	2.04%
		50%	11.9%	9.19%	8.61%	8.6%
		75%	39.9%	33.8%	33.2%	33.2%

Table 2: Example results for stochastic project completion. Suboptimality gaps relative to static policy for (a) penalized perfect information upper bounds; (b) perfect information upper bounds; and (c) ADP upper bounds.

policy. The calculation of the perfect information bound without penalty is similar, except no penalties are used, and is described in Appendix D.1. Finally, the ADP upper bound is easy to evaluate using the LP formulation discussed in Appendix D.2.

Table 2 reports the gap between the upper bounds and the lower bound V^{S} in relative terms. For each upper bound and each discount factor, we report the 25^{th}, 50^{th}, and 75^{th} percentiles of these relative gaps across all 150 examples. The relative gaps using the penalized perfect information upper bounds are, overall, small, and the asymptotic optimality of the static policy discussed in Prop. 4.1 is apparent from the results. (Note that two of the reported numbers are negative. This is due to sample error; the gaps with the penalized perfect information upper bound are not statistically different from zero in these cases). The perfect information bound with no penalty and the ADP upper bound are similar in terms of performance. Although these upper bounds are within $2-3\%$ in some examples, they can often be quite weak (30% or more). Moreover, these bounds do not convey asymptotic optimality of the static policy: the relative gaps using these bounds appear to converge to nonzero values (the median across all examples is about $8-9\%$). These examples demonstrate that perfect information bounds can be useful both in theoretical analysis as well as in specific numerical examples, but underscore the fact that an effective penalty may be essential for getting good results.

5 Conclusion

In this paper, we consider the use of information relaxations and penalties for analyzing the suboptimality gap of heuristic policies in complex dynamic programs, and we demonstrated this on three challenging problems. As we demonstrate with each problem, the penalty is essential for obtaining a good bound. In

each problem, we provide a simple example where the perfect information bound performs quite poorly, but with the inclusion of a penalty we recover a tight bound.

We are hopeful that this technique can be applied successfully in many other problems, but the key is finding a good penalty in the analysis. Although there is no general recipe to analyze the suboptimality of heuristic policies in stochastic DPs, at a high level the steps involved in our approach are as follows. First, we identify a heuristic policy and design a dual feasible penalty that somehow "aligns" the perfect information policy with the heuristic policy in consideration. Second, we analyze the penalized perfect information problem for each sample path using some problem specific technique. Finally, we take expectations of the penalized perfect information problem over all sample paths and connect this upper bound (or lower bound if costs) with the expected performance of the heuristic policy.

As was the case with the three problems we study here, what penalties will be effective depends on the specifics of the problem. It may be helpful to summarize the intuition for why the penalties worked in each of these three problems. First, in the stochastic scheduling problem we penalize the decision maker with perfect information for scheduling early the jobs with realized processing times that are small relative to their expected values. This tilts incentives in the perfect information problem towards scheduling early jobs with long realized processing times, which, absent the penalty, would be scheduled later. We choose the constants of proportionality on the penalty to make the ratio of weights to processing times in the penalized perfect information problem exactly equal to the ratio used by the WSEPT policy in ranking jobs. Second, in the stochastic knapsack problem, we deduct a penalty that is proportional to the difference between the expected size of the item and the realized size of the item. Again this tilts incentives in the perfect information problem towards selecting items with large realized sizes, which, absent the penalty, would be better to avoid in the perfect information problem. We choose the constants of proportionality on this penalty to make the ratio of value to size in the penalized perfect information bound exactly equal to the ratio used by the greedy policy to select items. Third, in the stochastic project completion problem the penalty increases the cost of acceleration in periods when an alternative can only be improved through acceleration, and lowers costs otherwise. We choose these cost changes induced by the penalty to ensure that it is optimal with perfect information to follow a policy that always accelerates one particular alternative in every period until completion. Although the optimal such alternative varies across each sample path, this is enough to align the perfect information policy with the best static policy that accelerates a fixed alternative.

Although penalized perfect information analysis weaves a common thread in our study of these problems, we must acknowledge that problem specific battles in each problem remained in order to connect the penalized perfect information bound with the expected performance of the heuristic policy. In the stochastic scheduling problem, we use existing valid inequalities from deterministic scheduling and bound the penalized perfect

information problem via polymatroid optimization. In the stochastic knapsack problem, we cast the penalized perfect information problem as a integer program and then apply LP relaxation bounds. In the stochastic project completion problem, we decompose the penalized perfect information problem across the alternatives and provide a bound on that decomposition. We believe that penalized perfect information analysis can be effective in other problems, but we do not expect that problem-specific insights can be avoided.

Moving forward, it would be interesting to consider other problems with more complicated structure and the need for more sophisticated policies and penalties. The policies we study here are all inherently non-adaptive, and in many problems adaptivity may be essential for getting good performance. As for the penalty, in each problem we study, the penalty punishes "local" violations of the non-anticipativity constraint, in that at every point in time the penalty punishes the decision maker for actions viewed in isolation but does not use the "global" system state. In other problems it might be necessary to consider more sophisticated penalties that punish global violations of the non-anticipativity constraint, i.e., penalties that have explicit state dependance.

A Review of Information Relaxation Duality

In this appendix we provide a review of information relaxation duality. The treatment here closely follows BSS. As in BSS, we consider the case where time is discrete and the horizon is finite; similar results can be derived for continuous time problems or infinite horizon problems.

A.1 The Primal Problem

Uncertainty in the DP is described by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the set of possible outcomes or *scenarios* ω, $\mathcal{F}$ is a σ-algebra that describes the set of possible events, and $\mathbb{P}$ is a probability measure describing the likelihood of each event.

Time is discrete and indexed by $t = 0, \ldots, T$. The DM's state of information evolves over time and is described by a filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ where the σ-algebra $\mathcal{F}_t$ describes the DM's state of information at the beginning of period t; we will refer to $\mathbb{F}$ as the *natural filtration*. The filtrations must satisfy $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$ for all $t < T$ so the DM does not forget what she once knew.

The DM must choose an action a_t in period t from a set A_t; we let $A(\omega) \subseteq A_0 \times \cdots \times A_T$ denote the set of all feasible action sequences $\boldsymbol{a} = (a_0, \ldots, a_T)$ given scenario ω. The DM's choice of actions is described by a *policy* π that selects a sequence of actions $\boldsymbol{a}$ in A for each scenario ω in Ω (i.e., $\pi : \Omega \to A$). To ensure the DM knows the feasible set when choosing actions in period t, we assume that the set of actions available in period t depends on the prior actions $(a_0, \ldots, a_{t\text{-}1})$ and is $\mathcal{F}_t$-measurable for each set of prior actions. We let Π denote the set of all *feasible* policies, i.e., those that ensure that $\pi(\omega)$ is in $A(\omega)$.

In the primal problem, we require the DM's choices to be *nonanticipative* in that the choice of action a_t in period t depends only on what is known at the beginning of period t; that is, we require policies to be adapted to the natural filtration $\mathbb{F}$ in that a policy's selection of action a_t in period t must be measurable with respect to $\mathcal{F}_t$. We let $\Pi_{\mathbb{F}}$ be the set of feasible policies that are nonanticipative.

The DM's goal is to select a feasible nonanticipative policy to maximize the expected total reward. The rewards are defined by a $\mathbb{F}$-adapted sequence of reward functions $(r_0, \ldots, r_T)$ where the reward r_t in period t depends on the first t+1 actions $(a_0, \ldots, a_t)$ of the action sequence $\boldsymbol{a}$ and the scenario ω. We let $r(\boldsymbol{a}, \omega) = \sum_{t=0}^{T} r_t(\boldsymbol{a}, \omega)$ denote the total reward. The primal DP is then:

$$\sup_{\pi \in \Pi_{\mathbb{F}}} \mathbb{E}[\, r(\pi)\,]. \tag{17}$$

Here $\mathbb{E}[\, r(\pi)\,]$ could be written more explicitly as $\mathbb{E}[\, r(\pi(\omega), \omega)\,]$ where policy π selects an action sequence that depends on the random scenario ω and the rewards r depend on the action sequence selected by π and the scenario ω. We suppress the dependence on ω and interpret $r(\pi)$ as a random variable representing the total reward generated under policy π.

It will be helpful to rewrite the primal DP (17) as a Bellman-style recursion in terms of the optimal value functions V_t. We let $\boldsymbol{a}_t = (a_0, \ldots, a_t)$ denote the sequence of actions up to and including period t. Since the period-t reward r_t depends only on the first t+1 actions $(a_0, \ldots, a_t)$, we will write $r_t(\boldsymbol{a})$ as $r_t(\boldsymbol{a}_t)$ with the understanding that the actions are selected from the full sequence of actions $\boldsymbol{a}$; we will use a similar convention for V_t. For $t > 0$, let $A_t(\boldsymbol{a}_{t\text{-}1})$ be the subset of period-t actions A_t that are feasible given the prior choice of actions $\boldsymbol{a}_{t\text{-}1}$: r_t and A_t are both implicitly functions of the scenario ω. We take the terminal value function $V_{T+1}(\boldsymbol{a}_T) = 0$ and, for $t = 0, \ldots, T$, we define

$$V_t(\boldsymbol{a}_{t\text{-}1}) = \sup_{a_t \in A_t(\boldsymbol{a}_{t\text{-}1})} \Big\{ r_t(\boldsymbol{a}_{t\text{-}1}, a_t) + \mathbb{E}[\, V_{t+1}(\boldsymbol{a}_{t\text{-}1}, a_t) \,|\, \mathcal{F}_t\,] \Big\}. \tag{18}$$

Here both sides are random variables (and therefore implicitly functions of the scenario ω) and we select an optimal action a_t for each scenario ω.

A.2 Duality Results

In the dual problem, we relax the requirement that the policies be nonanticipative and impose penalties that punish violations of these constraints. We define relaxations of the nonanticipativity constraints by

considering alternative information structures. We say that a filtration $\mathbb{G} = (\mathcal{G}_0, \ldots, \mathcal{G}_T)$ is a *relaxation* of the natural filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ if, for each t, $\mathcal{F}_t \subseteq \mathcal{G}_t$; we abbreviate this by writing $\mathbb{F} \subseteq \mathbb{G}$. $\mathbb{G}$ being a relaxation of $\mathbb{F}$ means that the DM knows more in every period under $\mathbb{G}$ than she knows under $\mathbb{F}$. For example, the perfect information relaxation is given by taking $\mathcal{G}_t = \mathcal{F}$ for all t. We let $\Pi_{\mathbb{G}}$ denote the set of feasible policies that are adapted to $\mathbb{G}$. For any relaxation $\mathbb{G}$ of $\mathbb{F}$, we have $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$; thus, as we relax the filtration, we expand the set of feasible policies. In this paper, we will focus on the *perfect information relaxation*, where the set of relaxed policies is the set of all policies Π and actions are selected with full knowledge of the scenario ω.

The set of penalties Z is the set of functions z that, like the total rewards, depend on actions $\boldsymbol{a}$ and the scenario ω. As with rewards, we will typically write the penalties as an action-dependent random variable $z(\boldsymbol{a})$ $(= \pi(\boldsymbol{a}, \omega))$ or a policy-dependent random variable $z(\pi)$ $(= z(\pi(\omega), \omega))$, suppressing the dependence on the scenario ω. We define the set $Z_{\mathbb{F}}$ of *dual feasible penalties* to be those that do not penalize nonanticipative policies in expectation, that is

$$Z_{\mathbb{F}} \quad = \quad \{z \in Z \ : \ \mathbb{E}[\, z(\pi)\,] \leq 0 \quad \text{for all } \pi \text{ in } Z_{\mathbb{F}}\}\,. \tag{19}$$

Policies that do not satisfy the nonanticipativity constraints (and thus are not feasible to implement) may have positive expected penalties.

We can obtain an upper bound on the expected reward associated with any nonanticipative policy by relaxing the nonanticipativity constraint on policies and imposing a dual feasible penalty, as stated in the following weak duality lemma. BSS show that this result holds for any information relaxation $\mathbb{G}$ of $\mathbb{F}$. We state a simpler result for the perfect information relaxation.

Lemma A.1 (Weak Duality). *If $\pi_{_F}$ and z are primal and dual feasible respectively (i.e., $\pi_{_F} \in \Pi_{\mathbb{F}}$ and $z \in Z_{\mathbb{F}}$), then*

$$\mathbb{E}[\, r(\pi_{_F})\,] \leq \sup_{\pi \in \Pi} \mathbb{E}[\, r(\pi) - z(\pi)\,] = \mathbb{E}\left[\sup_{\boldsymbol{a} \in A(\omega)} \{r(\boldsymbol{a}, \omega) - \pi(\boldsymbol{a}, \omega)\} \right]. \tag{20}$$

Proof. With z and $\pi_{_F}$ as defined in the lemma, we have

$$\mathbb{E}[\, r(\pi_{_F})\,] \leq \mathbb{E}[\, r(\pi_{_F}) - z(\pi_{_F})\,] \leq \sup_{\pi \in \Pi} \mathbb{E}[\, r(\pi) - z(\pi)\,]\,.$$

The first inequality holds because $z \in Z_{\mathbb{F}}$ (thus $\mathbb{E}[\, z(\pi_{_F})\,] \leq 0$) and the second because $\pi_{_F} \in \Pi_{\mathbb{F}}$ and $\Pi_{\mathbb{F}} \subseteq \Pi$. □

Thus any dual feasible penalty will provide an upper bound on the expected reward generated by any primal feasible policy. If we minimize over the dual feasible penalties in (20), we obtain the dual of the primal DP (17):

$$\inf_{z \in Z_{\mathbb{F}}} \left\{ \sup_{\pi \in \Pi} \mathbb{E}[\, r(\pi) - z(\pi)\,] \right\}. \tag{21}$$

The following result shows that in principle, we could determine the maximal expected reward in the primal DP (17) by solving the dual problem (21). This result is analogous to the strong duality theorem of linear programming. We refer the reader to BSS for a proof of this result.

Theorem A.2 (Strong Duality). *We have*

$$\sup_{\pi \in \Pi_{\mathbb{F}}} \mathbb{E}[\, r(\pi)\,] = \inf_{z \in Z_{\mathbb{F}}} \left\{ \sup_{\pi \in \Pi} \mathbb{E}[\, r(\pi) - z(\pi)\,] \right\}.$$

Furthermore, if the primal problem on the left is bounded, the dual problem on the right has an optimal solution $z^\star \in Z_{\mathbb{F}}$ that achieves this bound.

A.3 Penalties

BSS provides a general approach for constructing "good" penalties, based on a set of generating functions. We will show that we can, in principle, generate an optimal penalty using this approach.

Proposition A.3 (Constructing Good Penalties). *Let $(w_0, \ldots, w_T)$ be a sequence of generating functions defined on $A \times \Omega$ where each w_t depends only on the first $t+1$ actions $(a_0, \ldots, a_t)$ of $\boldsymbol{a}$. Define $z_t(\boldsymbol{a}) = w_t(\boldsymbol{a}) - \mathbb{E}[\, w_t(\boldsymbol{a}) \,|\, \mathcal{F}_t \,]$ and $z(\boldsymbol{a}) = \sum_{t=0}^{T} z_t(\boldsymbol{a})$. Then, for all $\pi_{\scriptscriptstyle F}$ in $\Pi_{\mathbb{F}}$, we have $\mathbb{E}[\, z_t(\pi_{\scriptscriptstyle F}) \,|\, \mathcal{F}_t \,] = 0$ for all t, and $\mathbb{E}[\, z(\pi_{\scriptscriptstyle F}) \,] = 0$.*

The result implies that the penalties z generated in this way will be dual feasible (i.e., $\mathbb{E}[\, z(\pi_{\scriptscriptstyle F}) \,] \leq 0$ for $\pi_{\scriptscriptstyle F}$ in $\Pi_{\mathbb{F}}$), but is stronger in that it implies the inequality defining dual feasibility (19) holds with equality: i.e., $\mathbb{E}[\, z(\pi) \,] = 0$ for all π in $\Pi_{\mathbb{F}}$. In this case, we say the penalty *has no slack*. A penalty that has slack can certainly be improved by eliminating the slack. Good penalties are thus, by construction, dual feasible with no slack. We refer the reader to BSS for a proof and further discussion of this result.

Considering a sequence of generating functions $(w_0, \ldots, w_T)$, we can write the dual problem recursively as follows. Take the terminal dual value function to be $\bar{V}_{T+1}(\boldsymbol{a}_T) = 0$. For $t = 0, \ldots, T$, we have

$$\bar{V}_t(\boldsymbol{a}_{t\text{-}1}) = \sup_{a_t \in A_t(\boldsymbol{a}_{t\text{-}1})} \Big\{ r_t(\boldsymbol{a}_{t\text{-}1}, a_t) - w_t(\boldsymbol{a}_{t\text{-}1}, a_t) \;+\; \mathbb{E}[\, w_t(\boldsymbol{a}_{t\text{-}1}, a_t) \,|\, \mathcal{F}_t \,] + \bar{V}_{t+1}(\boldsymbol{a}_{t\text{-}1}, a_t) \Big\}. \tag{22}$$

The expected initial value, $\mathbb{E}\big[\, \bar{V}_0 \,\big]$, provides an upper bound on the primal DP (17).

We can construct an optimal penalty using Proposition A.3 by taking the generating functions to be based on the optimal DP value function given by (18). Specifically, if we take generating functions $w_t(\boldsymbol{a}) = V_{t+1}(\boldsymbol{a}_t)$, we obtain an optimal penalty of the form:

$$z^\star(\boldsymbol{a}) \;\; = \;\; \sum_{t=0}^{T} V_{t+1}(\boldsymbol{a}_t) - \mathbb{E}[\, V_{t+1}(\boldsymbol{a}_t) \,|\, \mathcal{F}_t \,]. \tag{23}$$

It is easy to show by induction that the dual value functions are equal to the corresponding primal value functions, i.e., $\bar{V}_t = V_t$. This is trivially true for the terminal values (both are zero). If we assume inductively that $\bar{V}_{t+1} = V_{t+1}$, terms cancel and (22) reduces to the expression for V_t given in equation (18). Thus, with this choice of generating function, we obtain an optimal penalty that we refer to as the *ideal penalty*: the inner problem is equal to V_0 in every scenario and, moreover, the primal and dual problems will have the same sets of optimal policies.

Of course, in practice, we will not know the true value function and cannot construct this ideal penalty. The key to obtaining a good bound is to find a penalty that approximates the differences in the value function given in (23).

B Proofs

B.1 Proof of Proposition 2.1

We prove the following chain of inequalities

$$J_z^{\mathrm{P}} \leq J^* \leq J^{\mathrm{G}} \leq J_z^{\mathrm{P}} + \frac{m-1}{2m}(\Delta + 1) \sum_{i=1}^{n} w_i \mathbb{E}[p_i],$$

where the penalty is given by $z_i = r_i = w_i / \mathbb{E}[p_i]$. The first two inequalities in the chain follow trivially. In the remainder of the proof we prove the last inequality, that is, we aim to upper bound the performance of the WSEPT policy using the penalized perfect information policy. We prove the result in three steps. First, we upper bound the performance of the WSEPT policy. Second, we lower bound the objective value of the penalized perfect information problem for a fixed realization of processing times. Finally, we combine the previous bounds to bound the performance of the WSEPT policy in terms of the *expected* penalized perfect information bound.

Step 1. We first upper bound the performance of the WSEPT policy. For any list scheduling policy π that sequences jobs according to $\pi_1, \pi_2, \ldots, \pi_n$, it is well known that the completion time of job $C^{\pi}_{\pi_j}$ satisfies that

$$C^{\pi}_{\pi_j} \leq \frac{1}{m} \sum_{\ell=1}^{j-1} p_{\pi_\ell} + p_{\pi_j} \,.$$

See for example, Lemma 3.3 of Hall et al. (1997) for a proof. As a result we obtain that the performance of the WSEPT policy satisfies that

$$J^{\mathrm{G}} = \mathbb{E}\left[\sum_{i=1}^{n} w_i C^{\mathrm{G}}_i\right] \leq \mathbb{E}\left[\sum_{i=1}^{n} w_i \left(p_i + \frac{1}{m}\sum_{j=1}^{i-1} p_j\right)\right] = \sum_{i=1}^{n} w_i \mathbb{E}\left[p_i\right] + \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{i-1} w_i \mathbb{E}\left[p_j\right] \,, \tag{24}$$

where the inequality follows from the fact that WSEPT sequences jobs according to $1, \ldots, n$ because jobs are assumed to be sorted in decreasing order w.r.t the ratio of weight to expected processing time $r_i = w_i/\mathbb{E}[p_i]$.

Step 2. We next lower bound the objective value of the penalized perfect information problem for a fixed realization of processing times. Using that the penalties are given by $z_i = r_i = w_i/\mathbb{E}[p_i]$ we obtain from (2) that the perfect information problem can be lower bounded by

$$J^{\mathrm{P}}_z(\mathbf{p}) \geq \underline{J}^{\mathrm{P}}_z(\mathbf{p}) + \sum_{i=1}^{n} w_i p_i - r_i p_i^2 \,, \tag{25}$$

where $\underline{J}^{\mathrm{P}}_z(\mathbf{p})$ is the objective value of the deterministic scheduling problem $PM//\sum_i w^z_i C_i$ with weights $w^z_i = w_i + z_i(p_i - \mathbb{E}[p_i]) = r_i p_i$. Leveraging known results from deterministic parallel machine scheduling, Lemma B.1 provides the following lower bound on the objective value of the previous scheduling problem

$$\underline{J}^{\mathrm{P}}_z(\mathbf{p}) \geq \sum_{i=1}^{n} r_i \left(\frac{1}{m} p_i \sum_{j=1}^{i-1} p_j + \frac{m+1}{2m} p_i^2\right) \,. \tag{26}$$

Step 3. We conclude by combining our previous results to bound the performance of the WSEPT policy in terms of the *expected* penalized perfect information bound. Taking expectations w.r.t. the sample path $\mathbf{p}$ we obtain that

$$\begin{aligned}
J^{\mathrm{P}}_z &= \mathbb{E}_{\mathbf{p}}\left[J^{\mathrm{P}}_z(\mathbf{p})\right] \geq \mathbb{E}_{\mathbf{p}}[\underline{J}^{\mathrm{P}}_z(\mathbf{p})] + \sum_{i=1}^{n} w_i \mathbb{E}[p_i] - r_i \mathbb{E}[p_i^2] \\
&\geq \sum_{i=1}^{n} w_i \mathbb{E}[p_i] + \frac{1}{m}\sum_{i=1}^{n} r_i \mathbb{E}\left[p_i \sum_{j=1}^{i-1} p_j\right] - \frac{m-1}{2m}\sum_{i=1}^{n} r_i \mathbb{E}[p_i^2] \\
&= \sum_{i=1}^{n} w_i \mathbb{E}[p_i] + \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{i-1} w_i \mathbb{E}\left[p_j\right] - \frac{m-1}{2m}\sum_{i=1}^{n} w_i \frac{\mathbb{E}[p_i^2]}{\mathbb{E}[p_i]} \\
&\geq J^{\mathrm{G}} - \frac{m-1}{2m}\sum_{i=1}^{n} w_i \frac{\mathbb{E}[p_i^2]}{\mathbb{E}[p_i]} \geq J^{\mathrm{G}} - \frac{m-1}{2m}(\Delta+1)\sum_{i=1}^{n} w_i \mathbb{E}[p_i] \,,
\end{aligned}$$

where the first inequality follows from (25); the second inequality from (26); the second equation from the fact that $r_i = w_i/\mathbb{E}[p_i]$ and using that the processing times are independent; the third inequality from the bound on the performance of the WSEPT policy given in (24); and the last inequality because $\mathbb{E}[p_i^2]/\mathbb{E}[p_i]^2 = \mathrm{Var}[p_i]/\mathbb{E}[p_i]^2 + 1 \leq \Delta + 1$.

B.2 Proof of Proposition 3.1

Let $\mathcal{N}^\pi$ be the (stochastic) set of items that policy π attempts to insert into the knapsack and let c_i^π be the capacity remaining before policy π attempts to insert an item $i \in \mathcal{N}^\pi$ into the knapsack. For any $\pi \in \Pi$,

$$V^\pi = \mathbb{E} \sum_{t=1}^{n \wedge (\tau^\pi - 1)} v_{\pi_t} = \sum_{i=1}^{n} v_i \mathbb{P}\{i \in \mathcal{N}^\pi, s_i \le c_i^\pi\}$$
$$\le \sum_{i=1}^{n} v_i \mathbb{P}\{i \in \mathcal{N}^\pi, s_i \le \kappa\} = \sum_{i=1}^{n} w_i \mathbb{P}\{i \in \mathcal{N}^\pi\} = \mathbb{E} \sum_{t=1}^{n \wedge \tau^\pi} w_{\pi_t} = W^\pi \,,$$

where the inequality follows because $s_i \le c_i^\pi$ implies $s_i \le \kappa$ since $c_i^\pi \le \kappa$. The third equality follows because the events $i \in \mathcal{N}^\pi$ and $s_i \le \kappa$ are independent (for non-anticipative policies, recall that the size is not revealed until after an item is selected), and because $w_i = v_i \mathbb{P}\{s_i \le \kappa\}$. Since this holds for all $\pi \in \Pi$, this variation provides an upper bound on the optimal value in original formulation, i.e., $V^* \le W^*$.

B.3 Proof of Proposition 3.2

Item (ii) follows from (i) and using that $\max_i w_i = \max_i \mathbb{E}\left[\frac{w_i \tilde{s}_i}{\mu_i}\right] \le \mathbb{E}\left[\max_i \frac{w_i \tilde{s}_i}{\mu_i}\right]$ from Jensen's inequality together with the fact that $\mu_i = \mathbb{E}\tilde{s}_i$. In the remainder of the proof we prove item (i).

We prove the following chain of inequalities

$$V^{\mathrm{G}} \le V^* \le W^* \le W_z^{\mathrm{P}} \le V^{\mathrm{G}} + \max_i w_i + \mathbb{E}[\max_i w_i \tilde{s}_i / \mu_i] \,.$$

The first inequality in the chain follow trivially, the second inequality was argued previously and the third inequality follows because the penalty $\mathbf{z}$ is dual feasible. In the remainder of the proof we prove the last inequality, that is, we aim to upper bound the penalized perfect information policy in terms of the performance of the greedy policy.

Setting $z_i = w_i/\mu_i$ and relaxing constraints (33b) and (33d) we obtain that problem (33) decouples in terms of the decision variables $\mathbf{x}$ and $\mathbf{y}$. Thus, we obtain the upper bound

$$\bar{W}_z^{\mathrm{P}}(\mathbf{s}) \le \underbrace{\begin{array}{ll} \max\limits_{\mathbf{x}\in\{0,1\}^n} & \sum_{i=1}^n z_i \tilde{s}_i x_i \\ \text{s.t.} & \sum_{i=1}^n \tilde{s}_i x_i \le 1 \,. \end{array}}_{\clubsuit} + \underbrace{\begin{array}{ll} \max\limits_{\mathbf{y}\in\{0,1\}^n} & \sum_{i=1}^n z_i \tilde{s}_i y_i \\ \text{s.t.} & \sum_{i=1}^n y_i \le 1 \,. \end{array}}_{\spadesuit} \,, \tag{27}$$

where we also relaxed the knapsack constraint (33a) to $\sum_{i=1}^n \tilde{s}_i x_i \le 1$ because $\tilde{s}_i \le s_i$. 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 z_i, as in the greedy policy. We obtain by considering the continuous relaxation to $x_i \in [0,1]$ that $x_i = 1$ for $i \le \tau^{\mathrm{G}} \wedge n$ and $x_i \in (0,1]$ for $i = \tau^{\mathrm{G}}$. Rounding up to one the last fractional item, we obtain the upper bound

$$\clubsuit \le \sum_{i=1}^{n \wedge \tau^{\mathrm{G}}} z_i \tilde{s}_i \,.$$

Taking expectations over the sample path $\mathbf{s}$, we obtain

$$\mathbb{E}_{\mathbf{s}}[\clubsuit] \le \mathbb{E}_{\mathbf{s}}\left[\sum_{i=1}^{n \wedge \tau^{\mathrm{G}}} z_i \tilde{s}_i\right] = \mathbb{E}_{\mathbf{s}}\left[\sum_{i=1}^{n \wedge \tau^{\mathrm{G}}} w_i\right] = \mathbb{E}_{\mathbf{s}}\left[\sum_{i=1}^{n \wedge \tau^{\mathrm{G}} - 1} w_i\right] + \mathbb{E}_{\mathbf{s}}\left[w_{\tau^{\mathrm{G}}} \mathbf{1}\{\tau^{\mathrm{G}} \le n\}\right]$$
$$\le \mathbb{E}_{\mathbf{s}}\left[\sum_{i=1}^{n \wedge \tau^{\mathrm{G}} - 1} v_i\right] + \max_{i \in \mathcal{N}} w_i = V^{\mathrm{G}} + \max_{i \in \mathcal{N}} w_i \,, \tag{28}$$

where the first equality follows from considering the martingale $R_t = \sum_{i=1}^{t} z_i \tilde{s}_i - w_i$ and using the Optional Stopping Theorem to obtain that $\mathbb{E}_{\mathbf{s}}\left[R_{n\wedge\tau^{\mathrm{G}}}\right] = 0$; the second equality from excluding the overflowing item; the second inequality follows from the facts that $w_i = v_i \mathbb{P}\{s_i \le \kappa\} \le v_i$ and $w_{\tau^{\mathrm{G}}} \le \max_{i\in\mathcal{N}} w_i$.

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

$$\spadesuit = \max_{i\in\mathcal{N}} z_i \tilde{s}_i \,. \tag{29}$$

Taking expectations in equation (27) and using the bounds in (28) and (29) we obtain that

$$W_z^{\mathrm{P}} \le \mathbb{E}_{\mathbf{s}}\left[\bar{W}_z^{\mathrm{P}}(\mathbf{s})\right] = V^{\mathrm{G}} + \max_{i\in\mathcal{N}} w_i + \mathbb{E}[\max_{i\in\mathcal{N}} z_i \tilde{s}_i]\,,$$

as required because $z_i = w_i/\mu_i$.

B.4 Proof of Corollary 3.3

Because the value-to-size ratios are bounded is suffices to show that $E_n = \frac{1}{\kappa}\mathbb{E}\left[\max_{i=1,\dots,n} \min\{s_i, \kappa\}\right] \to 0$ as $n \to \infty$. We prove each case at a time.

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

Case (b). Using $\min\{s_i, \kappa\} \le s_i$ together with Theorem B.2 we obtain

$$E_n \le \frac{1}{\kappa}\mathbb{E}\left[\max_i s_i\right] \le \frac{1}{\kappa}\max_i \mathbb{E}s_i + \frac{1}{\kappa}\sqrt{\frac{n-1}{n}\sum_i \mathrm{Var}(s_i)} \le \frac{\bar{m}}{\kappa} + \frac{\sqrt{n}}{\kappa}\bar{\sigma}\,,$$

which converges to zero because $\kappa = \omega(\sqrt{n})$.

Case (c). Some definitions are in order. We denote by $x \le_{\mathrm{sd}} y$ if random variable x is first-order stochastically dominated by y, that is, $\mathbb{P}\{x \le z\} \ge \mathbb{P}\{y \le z\}$ for all z. Because $s_i \sim \mathcal{N}(m_i, \sigma_i^2)$ we have $s_i \le_{\mathrm{sd}} t_i$ where $t_i \sim \max\{\mathcal{N}(\bar{m}, \bar{\sigma}^2), \bar{m}\}$. From extreme value theory we have that $(\max_{i=1,\dots,n} t_i - a_n)/b_n$ converges in distribution to a Gumbel random variable where $a_n = \bar{m} + \bar{\sigma}\sqrt{2\log(n)} - \bar{\sigma}\frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}}$ and $b_n = \bar{\sigma}/\sqrt{2\log(n)}$ because the limit distribution is not sensitive to left truncation (Arnold et al., 2008, p.215). Lemma B.5 implies that $E_n \to 0$ as $n \to \infty$ when $\kappa = \omega\big(\max(a_n, b_n)\big) = \omega\big(\sqrt{\log(n)}\big)$.

Case (d). Because s_i is Pareto distributed with scale m_i and shape α_i we have $s_i \le_{\mathrm{sd}} t_i$ where t_i is Pareto distributed with scale $\bar{m}$ and shape $\underline{\alpha}$. From extreme value theory we have that $(\max_{i=1,\dots,n} t_i - a_n)/b_n$ converges in distribution to a Frèchet random variable where $a_n = 0$ and $b_n = n^{1/\underline{\alpha}}$ (Arnold et al., 2008, p.215). Lemma B.5 implies that $E_n \to 0$ as $n \to \infty$ when $\kappa = \omega\big(\max(a_n, b_n)\big) = \omega\big(n^{1/\underline{\alpha}}\big)$.

B.5 Proof of Proposition 4.1

We let Π_i^{P} denote the set of perfect information policies that incur costs until (and do not garner rewards until) alternative i is completed, and we use π_i to denote a policy in Π_i^{P}, with associated completion time τ^i. We first observe that, for any z such that $\alpha_i + z_i \ge 0$ for all i, and any sample path $\mathbf{d}$,

$$\min_{\pi\in\Pi^{\mathrm{P}}} \sum_{t=0}^{\tau^\pi - 1} \delta^t(\alpha_\pi + z_\pi(\mathbf{d}_t)) \quad = \quad \min_{i=1,\dots,n} \min_{\pi_i\in\Pi_i^{\mathrm{P}}} \sum_{t=0}^{\tau^i - 1} \delta^t(\alpha_{\pi_i} + z_{\pi_i}(\mathbf{d}_t)). \tag{30}$$

To see this, let $\pi^\star$ be a minimizer of the left-hand side; by the definition of τ^π, the policy $\pi^\star$ completes some alternative $i^\star$. Then $\pi^\star \in \Pi_{i^\star}^{\mathrm{P}}$, and the costs on the right-hand side of (30) with $\pi_{i^\star} = \pi^\star$ will equal those on the left-hand side. By minimizing over i, we can do no worse on the right-hand side. Conversely, let $i^\star$ and $\pi_{i^\star}$ be a minimizer of the left-hand side. The policy $\pi_{i^\star}$ is feasible for the problem on the left-hand side and will complete alternative $i^\star$ in time $\tau^{i^\star}$, and perhaps some other alternative earlier. Since $\alpha_\pi + z_\pi(\cdot) \ge 0$ by

assumption (i.e., penalized costs are nonnegative), earlier completion can only improve the costs, and the left-hand side can be no larger than the right-hand side.

We now describe a penalty z such that, for each $i \in \{1, \ldots, n\}$, the minimization of $\pi_i \in \Pi_i^{\mathrm{P}}$ on the right-hand side of (30) will be attained by the policy that exclusively accelerates alternative i until completion. It is first helpful to describe the relevant events for alternative i. In each period t, for each alternative i, exactly one of the following occur:

(a) The firm fails to improve i regardless of the decision, i.e., $d_{t,i}^j = 0$ for all $j = 0, \ldots, n$ (w.p. $1 - p_i$).

(b) The firm improves i regardless of the decision, i.e., $d_{t,i}^j = 1$ for all $j = 0, \ldots, n$ (w.p. q_i).

(c) The firm improves i if and only if i is accelerated, i.e., $d_{t,i}^i = 1$ and $d_{t,i}^j = 0$ for all $j \neq i$ (w.p. $p_i - q_i$).

To this end, we take $z_0(\mathbf{d}_t) = 0$ for all possible outcomes $\mathbf{d}_t$ in period t. For $i > 0$, we take $z_i(\mathbf{d}_t)$ such that

$$\alpha_i + z_i(\mathbf{d}_t) \quad = \quad \begin{cases} \alpha_0 + \frac{1}{p_i - q_i} \left(\frac{J^{\mathrm{S}}}{J^{\mathrm{S}_i}} \alpha_i - \alpha_0 \right)^+ & \text{if } d_{t,i}^i = 1 \text{ and } d_{t,i}^j = 0 \text{ for all } j \neq i, \\ \alpha_0 & \text{otherwise.} \end{cases} \tag{31}$$

We first verify that this z satisfies the dual feasibility condition (14). First, $\mathbb{E}_{\mathbf{d}_t}[z_0(\mathbf{d}_t)] = 0$ trivially. For all $i > 0$, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{d}_t}[\alpha_i + z_i(\mathbf{d}_t)] &= \alpha_i + \mathbb{E}_{\mathbf{d}_t}[z_i(\mathbf{d}_t)] \\ &= \alpha_0 + ((J^{\mathrm{S}}/J^{\mathrm{S}_i})\alpha_i - \alpha_0)^+ \\ &= \max\{\alpha_0, (J^{\mathrm{S}}/J^{\mathrm{S}_i})\alpha_i\} \\ &\leq \alpha_i, \end{aligned}$$

where we use the definition of z in (31), the fact that the event $d_{t,i}^i = 1$ and $d_{t,i}^j = 0$ for all $j \neq i$ happens with probability $p_i - q_i$, and the fact that $\alpha_i \geq \alpha_0$ and $J^{\mathrm{S}} \leq J^{\mathrm{S}_i}$, since $J^{\mathrm{S}} = \min_{i=0,\ldots,n} J^{\mathrm{S}_i}$. Subtracting α_i from both sides, we have $\mathbb{E}_{\mathbf{d}_t}[z_i(\mathbf{d}_t)] \leq 0$, which implies the dual feasibility condition (14). This in turn implies $J_z^{\mathrm{P}} \leq J^*$ in (15).

Now we argue that the policy that accelerates alternative i will minimize the costs on the right-hand side of (30) over all $\pi_i \in \Pi_i^{\mathrm{P}}$ for every $\mathbf{d}$. The reasoning is similar to the zero penalty perfect information bound case discussed in Section 4.2. For a fixed alternative i, if either events (a) or (b) occur, the alternative will not improve or will improve regardless of the firm's acceleration decision. Thus, the firm can simply choose the action with the smallest cost in these outcomes; since $\alpha_i + z_i(\mathbf{d}_t) = \alpha_0$, which is the lowest possible cost among all actions, accelerating i is optimal when either (a) or (b) occurs.

If (c) occurs, the firm can only improve alternative i by accelerating i. The firm could choose not to accelerate and instead wait for a cheaper improvement without acceleration in the future. Thus in order for the firm to accelerate alternative i in the current period, we should compare to the lowest possible penalized cost of improving i, which could happen in the ensuing period since all penalized costs are nonnegative. If the penalized cost of accelerating i when (c) occurs is no larger than this lowest possible penalized cost of improving, it will always be optimal to accelerate i for such outcomes. This is equivalent to:

$$\alpha_0 + \frac{1}{p_i - q_i} \left(\frac{J^{\mathrm{S}}}{J^{\mathrm{S}_i}} \alpha_i - \alpha_0 \right)^+ \quad \leq \quad \alpha_0 + \delta\alpha_0,$$

which is equivalent to $((J^{\mathrm{S}}/J^{\mathrm{S}_i})\alpha_i - \alpha_0)^+/\alpha_0 \leq \delta(p_i - q_i)$. Since $J^{\mathrm{S}}/J^{\mathrm{S}_i} \leq 1$, it is sufficient for $(\alpha_i - \alpha_0)/\alpha_0 \leq \delta(p_i - q_i)$, which is equivalent to condition (11). This establishes that an optimal policy in Π_i^{P} is the static policy that always works on i.

We then have

$$J_z^{\mathrm{P}} \quad = \quad \mathbb{E}_{\mathbf{d}} \min_{\pi \in \Pi^{\mathrm{P}}} \sum_{t=0}^{\tau^\pi - 1} \delta^t (\alpha_\pi + z_\pi(\mathbf{d}_t))$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathbf{d}} \min_{i=1,\ldots,n} \min_{\pi_i \in \Pi_i^{\mathrm{P}}} \sum_{t=0}^{\tau^i-1} \delta^t (\alpha_{\pi_i} + z_{\pi_i}(\mathbf{d}_t)) \\
&= \mathbb{E}_{\mathbf{d}} \min_{i=1,\ldots,n} \sum_{t=0}^{\tau^i-1} \delta^t (\alpha_i + z_i(\mathbf{d}_t))
\end{aligned}
$$

where the first equality is the definition of J_z^{P}, the second equality follows from (30) and the fact that $\alpha_i + z_i(\mathbf{d}_t) \geq \alpha_0 \geq 0$, and the third equality uses the fact that always accelerating i minimizes cost over $\pi_i \in \Pi^{\mathrm{P}}$ for every $\mathbf{d}$.

That $\mathbb{E}_{\mathbf{d}_t}[\alpha_i + z_i(\mathbf{d}_t)] \geq (J^{\mathrm{S}}/J^{\mathrm{S}_i})\alpha_i$ follows from the discussion above showing dual feasibility of z, in which we showed $\mathbb{E}_{\mathbf{d}_t}[\alpha_i + z_i(\mathbf{d}_t)] = \max\{\alpha_0, (J^{\mathrm{S}}/J^{\mathrm{S}_i})\alpha_i\}$.

B.6 Proof of Proposition 4.2

The second result follows readily from the performance guarantee in the first result. In the remainder of the proof we show the first result. Because the static policy is primal feasible and the penalty z is dual feasible we have that $V^S \leq V^* \leq V_z^{\mathrm{P}}$. The proof of the result follows from upper bounding the optimal reward with perfect information in terms of the expected reward of the static policy, or alternatively lower bounding the optimal cost with perfect information in terms of the expected cost of the static policy.

Let $J_z^{\mathrm{P},i}(\mathbf{d}) := \sum_{t=0}^{\tau^i-1} \delta^t(\alpha_i + z_i(\mathbf{d}_t))$ denote the cost of the penalized perfect information policy that aims to complete alternative i for a given sample path $\mathbf{d}$. From Proposition 4.1 we obtain that

$$
\begin{aligned}
J_z^{\mathrm{P}} = \mathbb{E}_{\mathbf{d}} \min_{i=1,\ldots,n} \sum_{t=0}^{\tau^i-1} \delta^t(\alpha_i + z_i(\mathbf{d}_t)) = \mathbb{E}_{\mathbf{d}} \min_{i=1,\ldots,n} J_z^{\mathrm{P},i}(\mathbf{d}) = \\
\geq \min_{i=1,\ldots,n} \left\{ \mathbb{E}_{\mathbf{d}} J_z^{\mathrm{P},i}(\mathbf{d}) \right\} - \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \mathrm{Var}(J_z^{\mathrm{P},i}(\mathbf{d}))}\,,
\end{aligned}
$$

where the inequality follows from the lower bound on the expected value of the minimum of random variables in terms of their expected values and variances given in Theorem B.2. We next bound the mean and variance of the perfect information costs $J_z^{\mathrm{P},i}(\mathbf{d})$.

For the mean we have by Lemma B.3

$$
\mathbb{E}_{\mathbf{d}} J_z^{\mathrm{P},i}(\mathbf{d}) = \mathbb{E}_{\mathbf{d}_t}[\alpha_i + z_i(\mathbf{d}_t)] \frac{1 - \mathbb{E}[\delta^{\tau^i}]}{1-\delta} \geq \alpha_i \frac{J^{\mathrm{S}}}{J^{\mathrm{S}_i}} \frac{1 - \mathbb{E}[\delta^{\tau^i}]}{1-\delta} = J^{\mathrm{S}} \frac{1 - \mathbb{E}[\delta^{\tau^i}]}{1 - \mathbb{E}[\delta^{\tau^{\mathrm{S}_i}}]} \geq J^{\mathrm{S}}\,,
$$

where the first inequality follows from Proposition 4.1, the second equality because $J^{\mathrm{S}_i} = \frac{\alpha_i}{1-\delta} \cdot \mathbb{E}[1 - \delta^{\tau^{S_i}}]$ where τ^{S_i} is the stopping time of the i^{th} static policy, and the last inequality because τ^{S_i} is first-order stochastically smaller than τ^i (because, under the static policy, some alternative different that i can be completed while accelerating i) together with the fact that $1 - \delta^t$ is increasing in $t \in \mathbb{N}$ for $\delta \in (0,1)$.

For the variance, first note that the expected reward per period is bounded by $\alpha_0 \leq \mathbb{E}_{\mathbf{d}_t}[\alpha_i + z_i(\mathbf{d}_t)] \leq \alpha_i$. The reward per period is bounded by $\alpha_0 \leq \alpha_i + z_i(\mathbf{d}_t) \leq (1+\delta)\alpha_0$ because $J^{\mathrm{S}} \leq J^{\mathrm{S}_i}$ and from (11), which implies that the variance of the reward per period is bounded by $\mathrm{Var}(\alpha_i + z_i(\mathbf{d}_t)) \leq \alpha_0^2/4$. We have by Lemma B.3 that

$$
\mathrm{Var}(J_z^{\mathrm{P},i}(\mathbf{d})) \leq \frac{2}{1-\delta} \left(\mathrm{Var}(\alpha_i + z_i(\mathbf{d}_t)) + \mathbb{E}[\alpha_i + z_i(\mathbf{d}_t)]^2 \frac{\mathrm{Var}(\delta^{\tau^i})}{1-\delta} \right) \leq \frac{\alpha_0^2}{1-\delta} \left(\frac{1}{2} + 8 \frac{\mathrm{Var}(\delta^{\tau^i})}{1-\delta} \right).
$$

Recall that τ^i corresponds to the completion time of alternative i under a policy that accelerates alternative i exclusively. Thus the completion satisfies $\tau^i \overset{(d)}{=} Y_i^i$ where Y_i^i is distributed as a negative binomial distribution

supported on $\{x_{0,i}, x_{0,j}+1, \ldots\}$ with $x_{0,j}$ successes and success probability p_i. Because $(1-\delta)x_{0,i} \leq \bar{x}$, Lemma B.4 implies that

$$\mathrm{Var}(\delta^{\tau^i}) \;\leq\; \underbrace{2\bar{x}a_i\frac{1-p_i}{p_i}\exp\left(2\bar{x}(a_i-1)\frac{1-p_i}{p_i}\right)}_{:=\gamma_i}(1-\delta),$$

where $a_i = (4+p_i)/(2p_i)$. Putting everything together, we obtain

$$J_z^{\mathrm{P}} \;\geq\; J^{\mathrm{S}} - \frac{\beta}{\sqrt{1-\delta}},$$

where

$$\beta \;:=\; \alpha_0\sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}+8\gamma_i\right)},$$

and the first result follows.

B.7 Auxiliary results

The following result is a strengthened version of Lemma 3.2 from Hall et al. (1997) and provides a lower bound on the objective value of the deterministic scheduling problem $PM//\sum_i r_i p_i C_i$. We reproduce the result for the sake of completeness.

Lemma B.1. *The objective value of the deterministic scheduling problem $PM//\sum_i r_i p_i C_i$, denoted by $\underline{J}_z^{\mathrm{P}}(\mathbf{p})$, is lower bounded by*

$$\underline{J}_z^{\mathrm{P}}(\mathbf{p}) \geq \sum_{i=1}^{n} r_i\left(\frac{1}{m}p_i\sum_{j=1}^{i-1}p_j + \frac{m+1}{2m}p_i^2\right).$$

Proof. A lower bound on the objective value can be obtained from the observation that for any feasible schedule on m machines the completion times should satisfy the following inequalities

$$\sum_{i\in A} p_i C_i \geq \frac{1}{2m}\left(\sum_{i\in A} p_i\right)^2 + \frac{1}{2}\sum_{i\in A} p_i^2\,,$$

for every subset $A \in \mathcal{N}$ (see, e.g., Hall et al. (1997)). Lemma 3.2 from Hall et al. (1997) proves a similar result under a weaker class of valid inequalities.

Optimizing over the completion times we obtain that the latter deterministic scheduling problem can be lower bounded by the following linear program

$$\begin{aligned}\underline{J}_z^{\mathrm{P}}(\mathbf{p}) \geq \min_{\mathbf{C}\in\mathbb{R}^n} & \sum_{i=1}^{n} r_i p_i C_i \\ \text{s.t.} & \sum_{i\in A} p_i C_i \geq f(A)\,, \quad \forall A \subseteq \mathcal{N}\,, \\ & C_i \geq 0\,,\end{aligned} \tag{32}$$

where the set function $f : 2^{\mathcal{N}} \to R$ is given by $f(A) = \frac{1}{2m}p(A)^2 + \frac{1}{2}p^2(A)$, where we denote by $p(A) = \sum_{i\in A} p_i$ and $p^2(A) = \sum_{i\in A} p_i^2$. It is not hard to see that the set function is super-modular, that is, for every $j, k \notin A$ we have that

$$f(A+j) + f(A+k) = f(A) + f(A+j+k) - 2p_j p_k \leq f(A) + f(A+j+k)\,.$$

Setting $y_i = p_i C_i$ we obtain that the feasible set of problem (32) is a polymatroid. Because the objective is linear in y_i, we can apply the greedy algorithm of Edmonds to characterize its optimal solution. Since the objective's coefficients satisfy $r_1 \geq r_2 \geq \ldots \geq r_n$, we obtain that the optimal solution is given by $y_i^* = f(\{1,\ldots,i\}) - f(\{1,\ldots,i-1\})$, or equivalently that

$$\begin{aligned} y_i^* &= f(\{1,\ldots,i\}) - f(\{1,\ldots,i-1\}) \\ &= \frac{1}{2m}\left(p(\{1,\ldots,i-1\}) + p_i\right)^2 - \frac{1}{2m} p(\{1,\ldots,i-1\})^2 + \frac{1}{2} p_i^2 \\ &= \frac{1}{m} p_i p(\{1,\ldots,i-1\}) + \frac{m+1}{2m} p_i^2 \,. \end{aligned}$$

As a result the objective value is given by $\sum_{i=1}^n r_i y_i^*$, implying that

$$\underline{J}_z^{\mathrm{p}}(\mathbf{p}) \geq \sum_{i=1}^n r_i y_i^* = \sum_{i=1}^n r_i \left(\frac{1}{m} p_i \sum_{j=1}^{i-1} p_j + \frac{m+1}{2m} p_i^2 \right) . \qquad \square$$

The following result provides an lower bound on the expected value of the minimum of random variables in terms of their expected values and variances.

Theorem B.2 (Aven (1985)). *For any sequence of random variables $\{X_i\}_{i=1}^n$ we have that $\mathbb{E}[\min_i X_i] \geq \min_i \mathbb{E}X_i - \sqrt{\frac{n-1}{n}\sum_i Var(X_i)}$ and $\mathbb{E}[\max_i X_i] \leq \min_i \mathbb{E}X_i + \sqrt{\frac{n-1}{n}\sum_i Var(X_i)}$.*

We next provide bounds for moments of stopped discounted partial sums. Some definitions are in order. Let $(y_i)_{i\geq 1} \in \mathbb{N}^\infty$ be i.i.d. random variables and consider the partial sum $S_n = \sum_{i=1}^n y_i$. Let τ be the stopping time corresponding to the first time that the partial sum hits a target $X \in \mathbb{N}$, that is, $\tau = \inf\{n \geq 1 : S_n = X\}$. Let $(z_i)_{i\geq 1} \in \mathbb{N}^\infty$ be i.i.d. random variables with bounded support. We aim to bound the mean and variance of the stopped discounted partial sum

$$C = \sum_{i=1}^{\tau} \delta^{i-1} z_i \,.$$

Here the process $(z_i)_{i\geq 1}$ *need not be independent* of the process $(y_i)_{i\geq 1}$.

Lemma B.3. *Suppose that the stopping time τ is integrable and the increments z_i have bounded support.*

(i) ***Expected value****. The expected value of the stopped discounted partial sum is given by:*

$$\mathbb{E}[C] = \mathbb{E}[z_1] \frac{1 - \mathbb{E}\left[\delta^\tau\right]}{1-\delta} \,.$$

(ii) ***Variance****. The variance of the stopped discounted partial sum is bounded by:*

$$Var\,(C) \leq \frac{2}{1-\delta} \left(Var(z_1) + \mathbb{E}[z_1]^2 \frac{Var(\delta^\tau)}{1-\delta} \right) .$$

Proof. We first prove the result for the expected value and then prove the result for the variance.

Expected value. Consider the martingale $M_n = \sum_{i=1}^n \delta^{i-1}(z_i - \mathbb{E}[z_1])$ with $M_0 = 0$. Because the stopping time τ is integrable and the increments z_i have bounded support, we conclude by the Optional Stopping

Theorem that $\mathbb{E}[M_\tau] = 0$. By construction we have that $M_\tau = C - \mathbb{E}[z_1]\sum_{i=1}^{\tau}\delta^{i-1}$. As a result the expected value of the stopped partial sum is given by:

$$\mathbb{E}[C] = \mathbb{E}[z_1]\mathbb{E}\left[\sum_{i=1}^{\tau}\delta^{i-1}\right] = \mathbb{E}[z_1]\frac{1-\mathbb{E}\left[\delta^\tau\right]}{1-\delta}\,.$$

Variance. Consider the quadratic martingale $R_n = M_n^2 - \mathrm{Var}(z_1)\sum_{i=1}^{n}(\delta^2)^{i-1}$. Another application of the Optional Stopping Theorem yields that $\mathbb{E}\left[R_\tau\right] = 0$, which implies that $\mathbb{E}\left[M_\tau^2\right] = \mathrm{Var}(z_1)\mathbb{E}\left[\sum_{i=1}^{\tau}(\delta^2)^{i-1}\right] = \mathrm{Var}(z_1)(1-\mathbb{E}\left[\delta^{2\tau}\right])/(1-\delta^2)$ where $M_\tau = C - \mathbb{E}[z_1]\sum_{i=1}^{\tau}\delta^{i-1}$. We can bound the variance as follows

$$\begin{aligned}
\sqrt{\mathrm{Var}\left(C\right)} = \left\|C - \mathbb{E}[C]\right\|_2 &\le \left\|C - \mathbb{E}[z_1]\sum_{i=1}^{\tau}\delta^{i-1}\right\|_2 + \left\|\mathbb{E}[z_1]\sum_{i=1}^{\tau}\delta^{i-1} - \mathbb{E}[C]\right\|_2 \\
&= \left\|M_\tau\right\|_2 + \frac{\mathbb{E}[z_1]}{1-\delta}\left\|\delta^\tau - \mathbb{E}\left[\delta^\tau\right]\right\|_2 \\
&= \sqrt{\mathrm{Var}(z_1)\frac{1-\mathbb{E}\left[\delta^{2\tau}\right]}{1-\delta^2}} + \frac{\mathbb{E}[z_1]}{1-\delta}\sqrt{\mathrm{Var}(\delta^\tau)}
\end{aligned}$$

where the first equation follows the definition of variance and denoting the L_2 norm of X as $\|X\|_2 = \sqrt{\mathbb{E}[X^2]}$, and the first inequality follows from Minkowski's inequality. From the AM-GM inequality we have that $\left(\sqrt{a}+\sqrt{b}\right)^2 \le 2(a+b)$ for $a, b > 0$. Therefore

$$\begin{aligned}
\mathrm{Var}\left(C\right) &\le 2\mathrm{Var}(z_1)\frac{1-\mathbb{E}\left[\delta^{2\tau}\right]}{1-\delta^2} + 2\frac{\mathbb{E}[z_1]^2}{(1-\delta)^2}\mathrm{Var}(\delta^\tau) \\
&\le \frac{2}{1-\delta}\left(\mathrm{Var}(z_1) + \mathbb{E}[z_1]^2\frac{\mathrm{Var}(\delta^\tau)}{1-\delta}\right)
\end{aligned}$$

where the second inequality follows from $1-\delta^2 \ge 1-\delta$ and $0 \le \mathbb{E}\left[\delta^{2\tau}\right]$ because $\tau \ge 1$ and $\delta \in (0,1)$. □

The next result bounds the variance of the random variable δ^τ, where τ captures the time to complete an alternative when working exclusively on it.

Lemma B.4. *Suppose τ is distributed as a negative binomial random variable supported on $\{x, x+1, \ldots\}$ with $x \in \mathbb{N}$ successes and success probability $p \in (0,1)$. Then for $\delta \in (0,1)$ we have:*

$$\mathit{Var}(\delta^\tau) \le zae^{z(a-1)}(1-\delta)\,,$$

where $z = 2x(1-\delta)\frac{1-p}{p}$ and $a = \frac{(4+p)}{2p}$.

Proof. Using the probability generating function of the negative binomial random variable (see footnote 2) we obtain that the variance can be written as

$$\mathrm{Var}(\delta^\tau) = \mathbb{E}\left[\left(\delta^2\right)^\tau\right] - \mathbb{E}\left[\delta^\tau\right]^2 = \delta^{2x}\left[\underbrace{\left(\frac{p}{1-(1-p)\delta^2}\right)^x}_{\clubsuit} - \underbrace{\left(\frac{p}{1-(1-p)\delta}\right)^{2x}}_{\spadesuit}\right].$$

The factor δ^{2x} is trivially bounded as $\delta^{2x} \le 1$ because $\delta \in (0,1)$ and $x \ge 0$. In the remainder of the proof we bound the terms in the parenthesis.

For the second term we have

$$\spadesuit = \exp\left(2x\log\left(\frac{p}{1-(1-p)\delta}\right)\right) \ge \exp\left(2x\left(1 - \frac{1-(1-p)\delta}{p}\right)\right)$$

$$= \exp\left(-2x(1-\delta)\frac{1-p}{p}\right) = \exp(-z)\,,$$

where the first inequality follows because $\log x \geq 1 - \frac{1}{x}$ for $x \geq 0$ and the last from setting $z = 2x(1-\delta)\frac{1-p}{p}$. For the first term we have

$$\begin{aligned}\clubsuit &= \exp\left(x \log\left(\frac{p}{1-(1-p)\delta^2}\right)\right) \leq \exp\left(x\left(\frac{p}{1-(1-p)\delta^2} - 1\right)\right) \\ &= \exp\left(-x\frac{(1-p)(1-\delta^2)}{1-(1-p)\delta^2}\right) = \exp\left(-z\frac{p(1+\delta)}{2\left(p\delta^2+1-\delta^2\right)}\right)\,,\end{aligned}$$

where the first inequality follows because $\log x \leq x - 1$ for $x \geq 0$ and the last equality from the definition of z and using that $1 - \delta^2 = (1+\delta)(1-\delta)$. The second factor in the exponential can be lower bounded as

$$\frac{p(1+\delta)}{2\left(p\delta^2+1-\delta^2\right)} \geq \frac{p(1+\delta)}{2\left(p+2(1-\delta)\right)} = 1 - \frac{(4+p)(1-\delta)}{2\left(p+2(1-\delta)\right)} \geq 1 - \frac{(4+p)}{2p}(1-\delta) = 1 - a(1-\delta)\,,$$

where the first inequality follows because $\delta \leq 1$ and $(1-\delta) = (1+\delta)(1-\delta) \leq 2(1-\delta)$, the second inequality from using that $1 - \delta \geq 0$ in the denominator, and the last equality from setting $a = \frac{(4+p)}{2p}$.

Putting everything together, we obtain

$$\mathrm{Var}(\delta^\tau) \leq \clubsuit - \spadesuit \leq e^{-z}\left(e^{za(1-\delta)} - 1\right) \leq zae^{z(a-1)}(1-\delta)\,,$$

where the inequality follows because $e^{c(1-\delta)} - 1 \leq ce^c(1-\delta)$ for $c \geq 0$ and $\delta \in (0,1)$. □

The following result bounds the growth of the expected value of the maximum of a sequence of random variables when the appropriately normalized maximum of this same sequence converges in distribution.

Lemma B.5. *Let $(s_i)_{i=1}^\infty$ be a sequence of non-negative i.i.d. random variables such that $(\max_{i=1,\dots,n} s_i - a_n)/b_n$ converges in distribution to some random variable Z where $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ are deterministic sequences with $b_n \geq 0$. Let*

$$E_n = \frac{1}{\kappa_n}\mathbb{E}\left[\max_{i=1,\dots,n} \min\{s_i, \kappa_n\}\right]\,.$$

Suppose that the sequence $(\kappa_n)_{n=1}^\infty$ satisfies $\kappa_n = \omega\big(\max(a_n, b_n)\big)$, then

$$\lim_{n=\infty} E_n = 0\,.$$

Proof. Let $M_n = \max_{i=1,\dots,n} s_i$ be the maximum among the first n random variables. The expression in consideration can be written as $E_n = \mathbb{E}\left[\min\left\{\frac{M_n}{\kappa_n}, 1\right\}\right]$, where we used the fact that maximum and minimum are commutative. Note that

$$\frac{M_n}{\kappa_n} = \frac{b_n}{\kappa_n}\frac{M_n - a_n}{b_n} + \frac{a_n}{\kappa_n}\,.$$

By assumption we have that $b_n/\kappa_n \to 0$ and $a_n/\kappa_n \to 0$ as $n \to \infty$, while $(M_n - a_n)/b_n \Rightarrow Z$ as $n \to \infty$. By Slutsky's Theorem we conclude that $M_n/\kappa_n \Rightarrow 0$, which implies that $M_n/\kappa_n \xrightarrow{p} 0$ in probability because the limit is constant. Because the function $\min\{x, 1\}$ is continuous in x we obtain by the continuous mapping theorem that $\min\{M_n/\kappa_n, 1\} \xrightarrow{p} 0$. Because $0 \leq \min\{M_n/\kappa_n, 1\} \leq 1$ the sequence is uniformly integrable and we conclude that $E_n \to 0$ as required. □

C Computing bounds for stochastic knapsack

In this section we explain how to solve the penalized perfect information problem for the stochastic knapsack problem.

C.1 Integer programming formulation for $W_z^{\mathrm{P}}(\mathbf{s})$

To calculate $W_z^{\mathrm{P}}(\mathbf{s})$, because the item that overflows the knapsack now counts towards 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

$$\begin{aligned}
\bar{W}_z^{\mathrm{P}}(\mathbf{s}) := \max_{\mathbf{x},\mathbf{y}\in\{0,1\}^n} & \sum_{i=1}^n (w_i + z_i(\tilde{s}_i - \mu_i))(x_i + y_i) \\
\text{s.t.} & \sum_{i=1}^n s_i x_i \le \kappa\,, && \text{(33a)} \\
& x_i + y_i \le 1\,, \quad \forall i \in \mathcal{N}\,, && \text{(33b)} \\
& \sum_{i=1}^n y_i \le 1\,, && \text{(33c)} \\
& \sum_{i=1}^n s_i(x_i + y_i) \ge \kappa(1 - x_i)\,, \quad \forall i \in \mathcal{N}\,. && \text{(33d)}
\end{aligned}$$

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 (33b) imposes that an item either fits the knapsack or overflows it. Constraint (33c) guarantees that there is at most one overflowing item. Constraint (33d) requires the overflowing item, if one exists, causes the selected capacity to exceed the capacity of knapsack. This constraint is vacuous when all items fit the knapsack, i.e., if there is no overflow. Note that we can only be sure that $W_z^{\mathrm{P}}(\mathbf{s}) \le \bar{W}_z^{\mathrm{P}}(\mathbf{s})$, because the "overflowing" item y_i can be chosen to exactly match the capacity of the knapsack. In order for item y_i to actually overflow the knapsack we need to make inequality (33d) 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 $i \in \mathcal{N}$), replacing constraint (33d) by $\sum_{i=1}^n s_i(x_i + y_i) \ge (\kappa + \epsilon)(1 - x_i)$ for some $\epsilon > 0$ in problem (33) gives that $\bar{W}_z^{\mathrm{P}}(\mathbf{s}) = W_z^{\mathrm{P}}(\mathbf{s})$. The bound given by $\bar{W}_z^{\mathrm{P}}(\mathbf{s})$, however, suffices for our analysis.

C.2 Integer programming formulation for $V_z^{\mathrm{P}}(\mathbf{s})$

The calculation of $V_z^{\mathrm{P}}(\mathbf{s})$ is similar to that of $W_z^{\mathrm{P}}(\mathbf{s})$. We can obtain an upper bound on $V_z^{\mathrm{P}}(\mathbf{s})$ for a fixed sample path $\mathbf{s}$ by solving the integer programming problem

$$\begin{aligned}
\bar{V}_z^{\mathrm{P}}(\mathbf{s}) := \max_{\mathbf{x},\mathbf{y}\in\{0,1\}^n} & \sum_{i=1}^n (v_i + z_i(s_i - \mathbb{E}[s_i]))x_i + \sum_{i=1}^n z_i(s_i - \mathbb{E}[s_i])y_i \\
\text{s.t.} & \sum_{i=1}^n s_i x_i \le \kappa\,, && \text{(34a)} \\
& x_i + y_i \le 1\,, \quad \forall i \in \mathcal{N}\,, && \text{(34b)} \\
& \sum_{i=1}^n y_i \le 1\,, && \text{(34c)} \\
& \sum_{i=1}^n s_i(x_i + y_i) \ge \kappa(1 - x_i)\,, \quad \forall i \in \mathcal{N}\,. && \text{(34d)}
\end{aligned}$$

Formulation (34) is essentially identical to (33), with the difference that we use values instead of effective values, sizes instead of truncated sizes, and we do not receive value for the overflowing item (though the overflowing item does affect the objective through the penalty in the terms involving y_i). Again, as with $W_z^{\mathrm{P}}(\mathbf{s})$, if the distribution of sizes is continuous or lattice, we could exactly obtain $V_z^{\mathrm{P}}(\mathbf{s})$ by replacing the constraints (34d) with $\sum_{i=1}^{n} s_i(x_i + y_i) \geq (\kappa + \epsilon)(1 - x_i)$ for some $\epsilon > 0$, but we nonetheless obtain an upper bound with the integer program (34).

D Computing bounds for stochastic project completion

In this section we explain how to efficiently solve the perfect information problem and the approximate dynamic programming bound for the stochastic project completion problem.

D.1 Perfect information problem

In the perfect information problem, the firm observes the sample path $\mathbf{d} = (\boldsymbol{d}_t^0, \ldots, \boldsymbol{d}_t^n)_{\{t \geq 0\}}$ before making acceleration decisions, where $d_{t,j}^i \in \{0, 1\}$ indicates whether or not the alternative j improves in period t when the firm accelerates alternative i. We let Π_i^{P} denote the set of perfect information policies that incur costs until (and do not garner rewards until) alternative i is completed, and we use π_i to denote a policy in Π_i^{P}, with associated completion time τ^i. Following the steps of the proof of Proposition 4.1 we first observe that for every sample path $\mathbf{d}$,

$$\min_{\pi \in \Pi^{\mathrm{P}}} \sum_{t=0}^{\tau^\pi - 1} \delta^t \alpha_\pi \quad = \quad \min_{i=1,\ldots,n} \min_{\pi_i \in \Pi_i^{\mathrm{P}}} \sum_{t=0}^{\tau^i - 1} \delta^t \alpha_{\pi_i}. \tag{35}$$

This implies that the perfect information problem is separable in the alternatives; that is, it suffices to find the best policy that aims to complete each alternative in isolation and then take the alternative with the minimum cost.

In order to solve each inner minimization problem in the right-hand side of (35) it is first helpful to describe the relevant events for alternative i. In each period t, for each alternative i, exactly one of the following occur:

(a) The firm fails to improve i regardless of the decision, i.e., $d_{t,i}^j = 0$ for all $j = 0, \ldots, n$ (w.p. $1 - p_i$).

(b) The firm improves i regardless of the decision, i.e., $d_{t,i}^j = 1$ for all $j = 0, \ldots, n$ (w.p. q_i).

(c) The firm improves i if and only if i is accelerated, i.e., $d_{t,i}^i = 1$ and $d_{t,i}^j = 0$ for all $j \neq i$ (w.p. $p_i - q_i$).

Because policy π_i is only concerned about the previous three outcomes we can consider the alternative sample path $(\hat{d}_{t,i}^0, \hat{d}_{t,i}^1)_{\{t \geq 0\}}$ with three outcomes generated independently over time with the following distribution:

$$(\hat{d}_{t,i}^0, \hat{d}_{t,i}^1) \quad = \quad \begin{cases} (0,0) & \text{(fail regardless), w.p. } 1 - p_i, \\ (1,1) & \text{(improve regardless), w.p. } q_i, \\ (0,1) & \text{(improve only if accelerated), w.p. } p_i - q_i. \end{cases}$$

For a given sample path $(\hat{d}_{t,i}^0, \hat{d}_{t,i}^1)_{\{t \geq 0\}}$, the objective is to minimize the discounted costs until completion of alternative i. If the outcome in a given period is improve (not improve) regardless, the optimal action in the perfect information problem is to not accelerate, since the alternative will improve (not improve) whether or not the alternative is accelerated in that period.

Thus, the only case to consider is what the optimal action should be when $(\hat{d}_{t,i}^0, \hat{d}_{t,i}^1) = (0, 1)$, i.e., the alternative will improve in a given period if and only if accelerated. For such outcomes, we claim it is always optimal to accelerate the alternative. If the firm were to not accelerate the alternative for such outcomes, at best the firm could improve the alternative in the ensuing period at discounted cost $\delta\alpha_0 = \delta(r + c)$, which

occurs when an "improve regardless" outcome of $(1,1)$ occurs at $t+1$. Thus, for it to be optimal to always accelerate when $(\hat{d}^0_{t,i}, \hat{d}^1_{t,i}) = (0,1)$, it is sufficient that

$$\underbrace{\alpha_i}_{\text{cost of advancing at } t} \leq \underbrace{\alpha_0 + \delta\alpha_0}_{\text{lowest possible cost of advancing at } t+1 \text{ but not } t}$$

or, equivalently, $(c_i - c)/(r + c) \leq \delta$, which is implied by (11).

Thus, the optimal perfect information policy for alternative i is to accelerate in each period t prior to completion if and only if $(\hat{d}^0_{t,i}, \hat{d}^1_{t,i}) = (0,1)$. Since the alternative is improved in each period with probability $q_i + p_i - q_i = p_i$, the completion time τ^i of the perfect information policy satisfies is distributed as negative binomial random variable supported on $\{x_0, x_0 + 1, \ldots\}$ with x_0 successes and success probability p_i. Thus the optimal cost of completing alternative i in isolation is

$$\min_{\pi_i \in \Pi_i^{\mathrm{P}}} \sum_{t=0}^{\tau^i - 1} \delta^t \alpha_{\pi_i} = \sum_{t=0}^{\tau^i - 1} \delta^t \left(\alpha_0 + \mathbb{1}_{\{(\hat{d}^0_{t,i}, \hat{d}^1_{t,i}) = (0,1)\}} \cdot (\alpha_i - \alpha_0) \right)$$

and the optimal perfect information expected reward is given by

$$V^{\mathrm{P}} = R - \mathbb{E}_{\mathbf{d}} \min_{i=1,\ldots,n} \sum_{t=0}^{\tau^i - 1} \delta^t \left(\alpha_0 + \mathbb{1}_{\{(\hat{d}^0_{t,i}, \hat{d}^1_{t,i}) = (0,1)\}} \cdot (\alpha_i - \alpha_0) \right) .$$

The latter bound can be efficiently computed via Monte Carlo simulation.

D.2 Approximate dynamic programming

Here we present a linear programming formulation for the ADP given in Section 4.5 with $O(\sum_{i=1}^n x_{0,i})$ variables and $O(\sum_{i=1}^n x_{0,i})$ constraints. Using the separability of the value function under the alternative-specific approximation we obtain that for $i \neq 0$ the constraints (16b) can be written as

$$c_i \geq [\delta p_i \left(V_i(x_i - 1) - V_i(x_i)\right) - (1 - \delta) V_i(x_i)] + \sum_{j \neq i} \left[\delta q_j \left(V_j(x_j - 1) - V_j(x_j)\right) - (1 - \delta) V_j(x_i)\right] .$$

This constraint can be handled by introducing auxiliary variables z_i and y_i, and imposing instead that $c_i \geq z_i + \sum_{j \neq i} y_i$ together with

$$\begin{aligned} z_i &\geq \delta p_i \left(V_i(x_i - 1) - V_i(x_i)\right) - (1 - \delta) V_i(x_i) , \\ y_i &\geq \delta q_i \left(V_i(x_i - 1) - V_i(x_i)\right) - (1 - \delta) V_i(x_i) , \end{aligned}$$

for all $i = 1 \ldots n$ and $x_i = 1 \ldots x_{0,i}$. Similarly for $i = 0$ we can write this constraint as $c_0 \geq \sum_{j=1}^n y_j$. Because the value functions are non-increasing in the state, we can write the boundary constraints (16c) involving R for each i as

$$V_i(0) + \sum_{j \neq i}^{n} V_j(x_{0,j}) \geq R ,$$

since $V_j(x_j) \geq V_j(x_{0,j})$. Putting everything together we obtain the following LP:

$$\begin{aligned} V^{\mathrm{ADP}} = \min_{V_i(x_i), z_i, y_i} \quad & \sum_{i=1}^{n} V_i(x_{0,i}) \\ \text{s.t. } & z_i \geq \delta p_i \left(V_i(x_i - 1) - V_i(x_i)\right) - (1 - \delta) V_i(x_i) , \quad \forall i = 1 \ldots n,\ x_i = 1 \ldots x_{0,i} , \\ & y_i \geq \delta q_i \left(V_i(x_i - 1) - V_i(x_i)\right) - (1 - \delta) V_i(x_i) , \quad \forall i = 1 \ldots n,\ x_i = 1 \ldots x_{0,i} , \\ & c_i \geq z_i + \sum_{j \neq i} y_j , \quad \forall i = 1 \ldots n , \end{aligned}$$

$$
\begin{aligned}
& c_0 \geq \sum_{j=1}^{n} y_j \,, \\
& V_i(0) + \sum_{j \neq i}^{n} V_j(x_{0,j}) \geq R \,, \quad \forall i = 1 \dots n \,, \\
& V_i(x_i - 1) \geq V_i(x_i) \,, \quad \forall i = 1 \dots n, \, x_i = 1 \dots x_{0,i} \,.
\end{aligned}
$$

E Additional numerical results for stochastic knapsack

In this section we present additional numerical results for the stochastic knapsack problem for the cases when the load factor is $\theta = 1/8$ (tight capacity) and when the load factor is $\theta = 1/2$ (moderate capacity).

Exponential Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	4.85%	3.32%	1.20%	0.68%
		50%	5.96%	3.52%	1.26%	0.71%
		75%	6.67%	3.72%	1.29%	0.75%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	7.62%	4.84%	1.54%	0.86%
		50%	9.07%	5.05%	1.62%	0.89%
		75%	10.01%	5.33%	1.67%	0.93%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	8.38%	9.82%	10.38%	9.85%
		50%	11.93%	11.01%	11.30%	10.64%
		75%	14.24%	11.55%	12.45%	11.28%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	18.93%	21.11%	22.05%	21.62%
		50%	26.01%	23.43%	23.84%	23.11%
		75%	31.24%	24.80%	25.52%	24.20%

Bernoulli Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	1.53%	1.11%	0.32%	0.18%
		50%	2.04%	1.23%	0.34%	0.18%
		75%	2.37%	1.36%	0.37%	0.18%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	4.49%	2.76%	0.70%	0.36%
		50%	5.44%	2.93%	0.71%	0.37%
		75%	5.81%	3.12%	0.77%	0.38%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	10.47%	10.63%	11.60%	10.84%
		50%	13.10%	12.11%	12.28%	11.57%
		75%	15.98%	13.00%	13.37%	12.20%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	20.51%	21.36%	22.80%	21.93%
		50%	26.44%	23.88%	24.27%	23.38%
		75%	32.64%	25.46%	25.91%	24.54%

Uniform Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	2.34%	1.44%	0.38%	0.19%
		50%	2.89%	1.55%	0.40%	0.20%
		75%	3.23%	1.60%	0.41%	0.21%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	5.32%	3.06%	0.75%	0.37%
		50%	6.30%	3.22%	0.76%	0.39%
		75%	6.91%	3.36%	0.81%	0.41%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	4.74%	5.46%	5.86%	5.66%
		50%	7.73%	6.47%	6.41%	6.20%
		75%	9.17%	7.30%	6.58%	6.46%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	19.89%	21.04%	22.47%	22.14%
		50%	27.55%	24.34%	24.51%	23.76%
		75%	33.14%	25.80%	25.32%	24.91%

Table 3: Stochastic knapsack results for $\theta = \frac{1}{2}$.

Exponential Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	16.86%	10.69%	3.41%	1.94%
		50%	18.08%	11.53%	3.50%	2.04%
		75%	19.11%	11.96%	3.73%	2.09%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	21.88%	14.33%	4.21%	2.33%
		50%	23.75%	15.19%	4.31%	2.46%
		75%	25.92%	16.12%	4.58%	2.51%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	40.08%	38.98%	42.63%	42.76%
		50%	45.00%	43.81%	44.03%	44.83%
		75%	47.93%	47.05%	47.96%	45.66%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	40.06%	40.94%	39.98%	40.52%
		50%	43.40%	43.99%	40.80%	41.81%
		75%	47.86%	46.51%	43.74%	42.83%

Bernoulli Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	5.22%	3.27%	0.92%	0.46%
		50%	6.09%	3.59%	0.96%	0.48%
		75%	6.41%	3.93%	1.01%	0.51%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	13.55%	7.79%	1.88%	0.94%
		50%	15.18%	8.46%	1.93%	0.96%
		75%	16.72%	8.88%	2.03%	1.00%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	65.41%	64.89%	71.00%	70.34%
		50%	74.58%	72.73%	73.58%	73.34%
		75%	80.39%	79.93%	78.74%	74.65%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	42.47%	40.34%	40.58%	40.91%
		50%	43.84%	44.14%	41.29%	41.97%
		75%	48.17%	46.68%	43.68%	42.83%

Uniform Sizes

		%ile	n			
			50	100	500	1000
(a)	$\frac{V_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	7.68%	4.41%	1.07%	0.56%
		50%	8.29%	4.84%	1.11%	0.58%
		75%	9.75%	5.25%	1.17%	0.60%
(b)	$\frac{W_z^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	15.44%	8.64%	1.97%	1.03%
		50%	17.30%	9.42%	2.04%	1.05%
		75%	19.71%	10.04%	2.14%	1.08%
(c)	$\frac{V^{\mathrm{P}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	24.89%	24.39%	24.94%	25.02%
		50%	27.77%	27.41%	26.63%	25.88%
		75%	31.54%	29.50%	27.30%	27.06%
(d)	$\frac{V^{\mathrm{DGV}} - V^{\mathrm{G}}}{V^{\mathrm{G}}}$	25%	42.78%	43.61%	40.59%	41.08%
		50%	45.06%	45.43%	41.65%	42.02%
		75%	51.45%	48.87%	43.67%	43.15%

Table 4: Stochastic knapsack results for $\theta = \frac{1}{8}$.

References

Andersen, L. M. and Broadie, M. (2004), 'Primal-dual simulation algorithm for pricing multidimensional american options', *Management Science* **50**, 1222–1234.

Arnold, B., Balakrishnan, N. and Nagaraja, H. (2008), *A First Course in Order Statistics*, Society for Industrial and Applied Mathematics.

Aven, T. (1985), 'Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables', *Journal of Applied Probability* **22**(3), 723–728.

Blado, D., Hu, W. and Toriello, A. (2015), Semi-infinite relaxations for the dynamic knapsack problem with stochastic item sizes. Working Paper.

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

Brown, D. B. and Haugh, M. B. (2014), Information relaxation bounds for infinite horizon markov decision processes. Working Paper.

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

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

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

Childs, P. D. and Triantis, A. J. (1999), 'Dynamic r&d investment policies', *Management Science* **45**(10), 1359–1377.

de Farias, D. P. and Roy, B. V. (2003), 'The linear programming approach to approximate dynamic programming', *Operations Research* **51**(6), 850–865.

Dean, B. C., Goemans, M. X. and Vondrák, J. (2008), 'Approximating the stochastic knapsack problem: The benefit of adaptivity', *Mathematics of Operations Research* **33**(4), 945–964.

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

Desai, V., Farias, V. F. and Moallemi, C. C. (2012), 'Pathwise optimization for optimal stopping problems', *Management Science* **58**, 2292–2308.

Devalkar, S., Anupindi, R. and Sinha, A. (2011), 'Integrated optimization of procurement, processing, and trade of commodities', *Operations Research* **59**, 1369–1381.

Ding, M. and Eliashberg, J. (2002), 'Structuring the new product development pipeline', *Management Science* **48**(3), 343–363.

Feldman, J., Henzinger, M., Korula, N., Mirrokni, V. S. and Stein, C. (2010), Online stochastic packing applied to display ad allocation, *in* 'Proceedings of the 18th annual European conference on Algorithms: Part I', ESA'10, Springer-Verlag, pp. 182–194.

Gallego, G. and van Ryzin, G. (1994), 'Optimal dynamic pricing of inventories with stochastic demand over finite horizons', *Management Science* **40**(8), 999–1020.

Garg, N., Gupta, A., Leonardi, S. and Sankowski, P. (2008), Stochastic analyses for online combinatorial optimization problems, *in* 'Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms', SODA '08, Society for Industrial and Applied Mathematics, pp. 942–951.

Grandoni, F., Gupta, A., Leonardi, S., Miettinen, P., Sankowski, P. and Singh, M. (2008), Set covering with our eyes closed, *in* 'Foundations of Computer Science, 2008. FOCS '08. IEEE 49th Annual IEEE Symposium on', pp. 347–356.

Hall, L. A., Schulz, A. S., Shmoys, D. B. and Wein, J. (1997), 'Scheduling to minimize average completion time: Off-line and on-line approximation algorithms', *Mathematics of Operations Research* **22**(3), 513–544.

Haugh, M. B., Iyengar, G. and Wang, C. (2014), Tax-aware dynamic asset allocation. Working Paper.

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

Haugh, M. B. and Lim, A. E. (2012), 'Linear-quadratic control and information relaxations', *Operations Research Letters* **40**(6), 521–528.

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

Manshadi, V. H., Gharan, S. O. and Saberi, A. (2012), 'Online stochastic matching: Online actions based on offline statistics', *Mathematics of Operations Research* **37**(4), 559–573.

Möhring, R. H., Schulz, A. S. and Uetz, M. (1999), 'Approximation in stochastic scheduling: The power of lp-based priority policies', *J. ACM* **46**(6), 924–942.

Nadarajah, S., Margot, F. and Secomandi, N. (2015), 'Relaxations of approximate linear programs for the real option management of commodity storage', *Management Science (Articles in Advance)* .

Papstavrou, J. D., Rajagopalan, S. and Kleywegt, A. J. (1996), 'The dynamic and stochastic knapsack problem with deadlines', *Management Science* **42**(12), 1706–1718.

Pinedo, M. L. (2012), *Scheduling*, Springer US.

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

Rogers, L. (2007), 'Pathwise stochastic optimal control', *SIAM Journal on Control and Optimization* **46**, 1116–1132.

Rothkopf, M. (1966), 'Scheduling with random service times', *Management Science* **12**, 703–713.

Santiago, L. and Vakili, P. (2005), 'On the value of flexibility in r&d projects', *Management Science* **51**(8), 1206–1218.

Smith, W. (1956), 'Various optimizers for single-stage production', *Nav. Res. Log. Quarterly* **3**, 59–66.

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

Weiss, G. (1990), 'Approximation results in parallel machines stochastic scheduling', *Annals of Operations Research* **26**(1), 195–242.

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