1-1-2012

Essays in Problems of Optimal Sequential Decisions

Alessandro Arlotto
University of Pennsylvania, ale.arlotto@gmail.com

Essays in Problems of Optimal Sequential Decisions

Abstract
In this dissertation, we study several Markovian problems of optimal sequential decisions by focusing on research questions that are driven by probabilistic and operations-management considerations. Our probabilistic interest is in understanding the distribution of the total reward that one obtains when implementing a policy that maximizes its expected value. With this respect, we study the sequential selection of unimodal and alternating subsequences from a random sample, and we prove accurate bounds for the expected values and exact asymptotics. In the unimodal problem, we also note that the variance of the optimal total reward can be bounded in terms of its expected value. This fact then motivates a much broader analysis that characterizes a class of Markov decision problems that share this important property. In the alternating subsequence problem, we also outline how one could be able to prove a Central Limit Theorem for the number of alternating selections in a finite random sample, as the size of the sample grows to infinity. Our operations-management interest is in studying the interaction of on-the-job learning and learning-by-doing in a workforce-related problem. Specifically, we study the sequential hiring and retention of heterogeneous workers who learn over time. We model the hiring and retention problem as a Bayesian infinite-armed bandit, and we characterize the optimal policy in detail. Through an extensive set of numerical examples, we gain insights into the managerial nature of the problem, and we demonstrate that the value of active monitoring and screening of employees can be substantial.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Operations & Information Management

First Advisor
Noah Gans

Second Advisor
J. Michael Steele

Keywords
Gittins index, learning, Markov decision problem, optimal sequential decision

Subject Categories
Operational Research | Other Education

ESSAYS IN PROBLEMS OF OPTIMAL SEQUENTIAL DECISIONS

Alessandro Arlotto

A DISSERTATION

in

Operations and Information Management

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2012

Supervisor of Dissertation

Noah Gans, Professor of Operations and Information Management

Co-Supervisor of Dissertation

J. Michael Steele, Professor of Statistics, Operations and Information Management

Graduate Group Chairperson

Eric Bradlow, Professor of Marketing, Statistics, and Education

Dissertation Committee

Noah Gans, Professor of Operations and Information Management
J. Michael Steele, Professor of Statistics, Operations and Information Management
Sergei Savin, Associate Professor of Operations and Information Management
Stephen E. Chick, Professor of Technology and Operations Management

ESSAYS IN PROBLEMS OF OPTIMAL SEQUENTIAL DECISIONS

Alessandro Arlotto

Ai miei genitori,

a nonna Maria,

e alla memoria di Zio Aldo.

ACKNOWLEDGEMENTS

Completing a Ph.D. is an arduous process that is not possible without the support of dedicated advisors, talented collaborators, and caring friends.

During my time at Wharton, I had the privilege of working with the best advisors one could imagine. Noah Gans and Mike Steele have been wonderful, and it is a great honor to be one of their students.

Throughout the years, Noah was kind enough to bring me back to the real world any time I was dangerously departing. Noah has been the *Riccardo Muti* of my Ph.D. work, and I could not have been more fortunate. His full-time availability, his comments, and his experience convinced me (at last!) to find a way to put together something of broader interest to the academic community, and for this I am very thankful.

Besides being an incredible mathematician, Mike has inspired me with his passion for being a real scholar, his genuine interest in problems that span all of mathematics, and his continued guidance, patience, and wisdom. His web-page *Advice For Graduate Students* has been my best friend in the worst moments, always pointing me toward the right direction.

Noah's and Mike's contributions in my training as a scholar and, more importantly, as a person have been inestimable, and I will always be profoundly grateful to them.

My gratitude also goes to Steve Chick, Sergei Savin and Larry Shepp, whom I had the privilege to work with and learn from at different levels. My sincere thanks also go to Omar Besbes, Gérard Cachon, Robert Chen, Robin Pemantle, and Maria Rieders for sharing their knowledge and many important moments throughout my studies. Moreover, I am particularly indebted to Marco Scarsini, without whom none of this would have ever been possible, and to Igor Prünster for his enduring support.

The Wharton staff in the Operations and Information Management department and in the Doctoral Office has been incredibly supportive, and my thanks go to Andrea, Beth, JP, Jamie, Kim, Mallory, Rochelle, and Stan.

During my time in Philadelphia, I was fortunate to encounter many warm-hearted friends who walked along this journey. Thanks to Jose, Jun, Necati, Tom, Toni, and Vibs. All together we made it, and your friendship is most valuable. Thanks also to "coach" Santiago, Bob, Brent, Fazil, Jaelynn, and Pnina.

Thanks to the Italian Community in Philadelphia. You guys have been the best gift when most was needed. Alberto, Francesco, Gloria, Lorenzo, and Marco, there are no words for describing my gratitude. Thanks for your trustworthiness and friendship. Thanks also to the always-welcomed "foreigners" Chiara and Fabrizio. I wish you guys all the best.

Infine, un grazie particolare va ai miei genitori e a mio fratello Davide che hanno reso questi anni di distanza solo fisica, e mai affettiva. Vi voglio bene.

Alessandro Arlotto

Philadelphia, PA

August 9, 2012

ABSTRACT

ESSAYS IN PROBLEMS OF OPTIMAL SEQUENTIAL DECISIONS

Alessandro Arlotto

Noah Gans

J. Michael Steele

In this dissertation, we study several Markovian problems of optimal sequential decisions by focusing on research questions that are driven by probabilistic and operations-management considerations. Our probabilistic interest is in understanding the distribution of the total reward that one obtains when implementing a policy that maximizes its expected value. With this respect, we study the sequential selection of unimodal and alternating subsequences from a random sample, and we prove accurate bounds for the expected values and exact asymptotics. In the unimodal problem, we also note that the variance of the optimal total reward can be bounded in terms of its expected value. This fact then motivates a much broader analysis that characterizes a class of Markov decision problems that share this important property. In the alternating subsequence problem, we also outline how one could be able to prove a Central Limit Theorem for the number of alternating selections in a finite random sample, as the size of the sample grows to infinity. Our operations-management interest is in studying the interaction of on-the-job learning and learning-by-doing in a workforce-related problem. Specifically, we study the sequential hiring and retention of heterogeneous workers who learn over time. We model the hiring and retention problem as a Bayesian infinite-armed bandit, and we characterize the optimal policy in detail. Through an extensive set of numerical examples, we gain insights into the managerial nature of the problem, and we demonstrate that the value of active monitoring and screening of employees can be substantial.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

This dissertation studies several Markovian problems of optimal sequential decisions (a.k.a. Markov Decision Problems) in which a decision-maker faces uncertain outcomes and needs to make decisions throughout a discrete time horizon with finitely- or infinitely-many decision times. Each period's decision has both immediate and long-term consequences, and the decision-maker needs to balance them in some way. In fact, by not accounting for these intertemporal relationships, the decision-maker may not achieve a good overall performance.

To see how these decisions are interconnected, let's consider a classical operation research problem, the stochastic knapsack problem (c.f. Martello and Toth, 1990). A decision-maker is given a knapsack with finite capacity, and he is sequentially presented with arriving items of random size. Any time he sees an item arriving, the size of the item is revealed, and the decision-maker chooses whether to include the item in the knapsack or not. If he chooses to include an item, then he loses some of the capacity he has available, and this constrains his future decisions. If he chooses not to include an item, then capacity is preserved, but the decision-maker, who does not know the sizes of the items arriving in the future, might have lost one opportunity for inclusion.

The way in which the decision maker should balance the immediate and the long-term effects of his decisions depends on his objective. For instance, in the knapsack problem discussed earlier, decisions will be very different if the decision-maker's goal is to maximize the expected number of items included in the knapsack, or to minimize the expected time it takes to fill the knapsack up. Throughout this dissertation, we fix the decision-maker's objective to multiperiod expected reward maximization (or, equivalently, expected cost minimization), possibly with discounting.

With this objective in mind, scholars have been usually interested in

(a) establishing the existence of an optimal policy, i.e. a policy that maximizes the total expected reward over the time horizon in consideration;

(b) obtaining properties of the optimal policy to gain intuition into the nature of the sequential decision process;

(c) evaluating the performance of the optimal policy, and comparing it with heuristics that are usually easy to implement.

These questions are at the basis of our dissertation work, which is also motivated by probabilistic and operation-management considerations, as we discuss below.

A Probabilistic Approach to Markov Decision Problems

In this dissertation, we go beyond the classical questions, (a)-(c), described above, as we are also interested in gaining a probabilistic understanding of the optimal total reward that one earns when implementing a policy that maximizes its expected value. This probabilistic reasoning is what has motivated the work in Chapters 2, 3, and 4, in which we study specific Markov decision problems, and we consider the following research questions:

(i) Given the existence of a Markov deterministic policy (often unique) that maximizes the total expected multiperiod reward, what can we say about the actual size of such expected reward? Can we prove accurate bounds for any given horizon length, n, so that, as n grows to infinity, we obtain exact asymptotics?

(ii) For the same policy, what can we say about the variability of the total reward?

(iii) And, what about limit theorems? Can we characterize the limiting distribution of the total reward obtained when one implements the policy that maximizes its expected value when the horizon length, n, grows to infinity?

Questions (i)-(iii) are quite natural to probabilists, but have received remarkably little attention in the Markov-decision-problem literature. Ultimately, the total reward that one

earns when implementing an optimal policy is a random variable, and it can be seen as a sum of *dependent* one-period rewards, where the dependency is induced by the decision-maker's actions.

In studying the distributional aspects of the optimal total reward, we make extensive use of the following techniques:

- **Prophet inequalities:** *a priori* bounds on the performance of any feasible selection policy that depends only on the sequential nature of the problem and not on the decision-maker's actions.

- **Martingale methods:** Markov decision problems have an immediate connection with martingale methods. The value function that describes the expected reward to-go from any decision time $1 \leq t \leq n$ to the end of the horizon, n, can be used to construct a martingale that becomes useful in characterizing the variance and the limiting distribution of the optimal total reward.

- **Time-inhomogeneous Markov chain theory:** A Markov decision problem is mathematically described through a time-inhomogeneous Markov chain, and the one period rewards are functions of such a Markov chain. Time-inhomogeneity plays a crucial role in finite-horizon problems, as the decision-maker's actions vary with the decision time.

- **Relationships between finite and infinite horizon problems:** infinite-horizon discounted problems are usually better behaved and, for this reason, people often prefer to pass from finite to infinite horizon set-ups. In the specific *alternating-subsequence* problem (Chapter 4), we pass back from infinite to finite horizon, and we quantify the size of the expected total-reward for the finite horizon formulation with accuracy.

An Operations-Management Application of Markov Decision Problems

Markov decision problems are an important framework for modeling real-life decision-making. With this respect, we study a work-force management application in which an employer needs to decide on the sequential employment of heterogeneous workers who learn over time. Heterogeneity reflects the fact that workers' abilities may differ, while the learning over time (also known as on-the-job learning) takes into account the fact that workers' experiences might affect their performance. On-the-job learning is, in fact, an important element of the operations of call-centers, manufacturing and other activities, especially when there may be high turnover of employees.

In this set-up, our analysis focuses on the standard dynamic programming questions, (a)-(c), described above, and our main contribution is on the managerial implications of monitoring and screening employees. After formulating the hiring and retention problem as a Bayesian infinite-armed bandit, we characterize the optimal policy in detail, and, through an extensive set of numerical examples, we demonstrate that the value of active monitoring and screening of employees can be substantial.

Overview of the Subsequent Chapters

In Chapter 2, we consider the problem of selecting sequentially a longest *unimodal subsequence* from a sequence of independent, identically distributed, continuous random variables, and we find that a person doing optimal sequential selection does within a factor of the square root of two as well as a prophet who knows all of the random observations in advance of any selections. Our analysis applies, in fact, to the selection of subsequences that have $d+1$ monotone blocks, and, by including the case $d = 0$, our analysis also covers monotone subsequences. We also show that the variance of the optimal number of unimodal selections can be expressed in terms of its expected value. This phenomenon turns out to be typical in a much larger class of Markov decision problems, which we study in Chapter 3.

In Chapter 4, we consider the sequential selection of a longest *alternating subsequence* from a sequence of independent, identically distributed, continuous random variables, and we determine the exact asymptotic behavior of the expected optimal number of alternating selections. Moreover, we find that a person who is constrained to make sequential selections does only about 12% worse than a person who can make selections with full knowledge of the random sequence. We also consider the limiting distribution of the optimal number of alternating selections, and we outline the steps that are needed for proving a Central Limit Theorem.

In Chapter 5, we consider the hiring and retention of heterogeneous workers who learn over time, as described earlier in this introduction.

CHAPTER 2 : Optimal Sequential Selection of Unimodal Subsequences

A classic result of Erdős and Szekeres (1935) tells us that in any sequence $x_1, x_2, \ldots, x_n$ of n real numbers there is a subsequence of length $k = \lceil n^{1/2} \rceil$ that is either monotone increasing or monotone decreasing. More precisely, given $x_1, x_2, \ldots, x_n$ one can always find a subsequence $1 \leq n_1 < n_2 < \cdots < n_k \leq n$ for which we either have

$$x_{n_1} \leq x_{n_2} \leq \cdots \leq x_{n_k}, \quad \text{or} \quad x_{n_1} \geq x_{n_2} \geq \cdots \geq x_{n_k}.$$

Many years later, Fan Chung (1980) considered the analogous problem for unimodal sequences. Specifically, she sought to determine the maximum value ℓ_n such that in any sequence of n real values $x_1, x_2, \ldots, x_n$ one can find a subsequence $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ of length $k = \ell_n$ and a "turning place" $1 \leq t \leq k$ for which one either has

$$x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_t} \geq x_{i_{t+1}} \geq \cdots \geq x_{i_k}, \quad \text{or}$$

$$x_{i_1} \geq x_{i_2} \geq \cdots \geq x_{i_t} \leq x_{i_{t+1}} \leq \cdots \leq x_{i_k}.$$

Through a sustained and instructive analysis, she surprisingly obtained an exact formula:

$$\ell_n = \left\lceil (3n - 3/4)^{1/2} - 1/2 \right\rceil.$$

Shortly afterwards, Steele (1981) considered unimodal subsequences of permutations, or equivalently, unimodal subsequences of a sequence of n independent, uniformly distributed

This chapter is written under the supervision of Prof. J. Michael Steele. The results presented here are also in the joint paper Arlotto and Steele (2011), published in *Combinatorics, Probability and Computing*.

random variables $X_1, X_2, \ldots, X_n$. For the random variables

$$U_n = \max\{k : X_{i_1} \le X_{i_2} \le \cdots \le X_{i_t} \ge X_{i_{t+1}} \ge \cdots \ge X_{i_k}, \text{ where } 1 \le i_1 < i_2 < \cdots < i_k \le n\},$$

and

$$D_n = \max\{k : X_{i_1} \ge X_{i_2} \ge \cdots \ge X_{i_t} \le X_{i_{t+1}} \le \cdots \le X_{i_k}, \text{ where } 1 \le i_1 < i_2 < \cdots < i_k \le n\},$$

it was established that

$$\mathbb{E}\left[\max\{U_n,\, D_n\}\right] \sim \mathbb{E}[U_n] \sim \mathbb{E}[D_n] \sim 2(2n)^{1/2} \quad \text{as } n \to \infty. \tag{2.1}$$

Here, we consider analogs of the random variables U_n, D_n and $L_n = \max\{U_n, D_n\}$ but instead of seeing the whole sequence all at once, one observes the variables sequentially. Thus, for each $1 \le i \le n$, the chooser must decide at time i when X_i is first presented whether to accept or reject X_i as an element of the unimodal subsequence. The sequential (or on-line) selection for the simpler problem of a monotone subsequence — the analog of the original Erdős and Szekeres (1935) problem — was considered long ago in Samuels and Steele (1981).

Main Results

We denote by $\Pi(n)$ the set of all feasible policies for the unimodal sequential selection problem for $\{X_1, X_2, \ldots, X_n\}$ where these random variables are independent with a common continuous distribution function F. Given any feasible sequential selection policy $\pi_n \in \Pi(n)$, if we let τ_k denote the index of the k'th selected element, then for each k the value τ_k is a stopping time with respect to the increasing sequence of σ-fields $\mathcal{F}_i = \sigma\{X_1, X_2, ..., X_i\}$,

$1 \leq i \leq n$. In terms of these stopping times, the random variable

$$U_n^o(\pi_n) = \max\{k : X_{\tau_1} \leq X_{\tau_2} \leq \cdots \leq X_{\tau_t} \geq X_{\tau_{t+1}} \geq \cdots \geq X_{\tau_k}, \text{ where}$$
$$1 \leq \tau_1 < \tau_2 < \cdots < \tau_k \leq n\},$$

is the length of the unimodal subsequence that is selected by the policy π_n. For the moment, we just consider unimodal subsequences that begin with an increasing piece and end with a decreasing piece; either of these pieces is permitted to have size one.

For each n there is a policy $\pi_n^* \in \Pi(n)$ that maximizes the expected length of the selected subsequence, and the main issue is to determine the asymptotic behavior of this expected value. The answer turns out to have an informative relationship to the off-line selection problem. A prophet with knowledge of the whole sequence before making his choices will do better than an optimal on-line chooser, but he will only do better by a factor of $\sqrt{2}$.

Theorem 2.1 (Expected Length of Optimal Unimodal Subsequences)**.** *For each $n \geq 1$, there is a $\pi_n^* \in \Pi(n)$, such that*

$$\mathbb{E}[U_n^o(\pi_n^*)] = \sup_{\pi_n \in \Pi(n)} \mathbb{E}[U_n^o(\pi_n)],$$

and for such an optimal policy one has the upper bound

$$\mathbb{E}[U_n^o(\pi_n^*)] < 2n^{1/2}$$

and the lower bound

$$2n^{1/2} - 4(\pi/6)^{1/2}n^{1/4} - O(1) < \mathbb{E}[U_n^o(\pi_n^*)]$$

which combine to give the asymptotic formula

$$\mathbb{E}[U_n^o(\pi_n^*)] \sim 2n^{1/2} \quad \textit{as } n \to \infty.$$

In a natural sense that we will shortly make precise, the optimal policy π_n^* is unique. Consequently, one can ask about the *distribution* of the length $U_n^o(\pi_n^*)$ of the subsequence that is selected by the optimal policy, and there is a pleasingly general argument that gives an upper bound for the variance. Moreover, that bound is good enough to provide a weak law for $U_n^o(\pi_n^*)$.

Theorem 2.2 (Variance Bound)**.** *For the unique optimal policy $\pi_n^* \in \Pi(n)$, one has the bounds*

$$\mathrm{Var}[U_n^o(\pi_n^*)] \leq \mathbb{E}[U_n^o(\pi_n^*)] < 2n^{1/2}. \tag{2.2}$$

Corollary 2.3 (Weak Law for Unimodal Sequential Selections)**.** *For the sequence of optimal policies $\pi_n^* \in \Pi(n)$, one has the limit*

$$U_n^o(\pi_n^*)/\sqrt{n} \xrightarrow{p} 2 \quad \textit{as } n \to \infty.$$

The variance bound in Theorem 2.2 and its corollary hold for a much larger class of Markov decision problems that go beyond the unimodal subsequence problem, which we study in Chapter 3.

Organization of the Proofs

The proof of Theorem 2.1 comes in two halves. First, we show by an elaboration of an argument of Gnedin (1999) that there is an *a priori* upper bound for $\mathbb{E}[U_n^o(\pi_n)]$ for all n and all $\pi_n \in \Pi(n)$. This argument uses almost nothing about the structure of the selection policy beyond the fact, from Section 2.3, that it suffices to consider policies that are specified by acceptance *intervals*. For the lower bound, we simply construct a good (but suboptimal) policy. Here, there is an obvious candidate, but the proof of its efficacy seems to be more delicate than one might have expected.

The proof of Theorem 2.2 in Section 2.2 exploits a martingale that comes naturally from the Bellman equation. The summands of the quadratic variation of this martingale are then

found to have a fortunate relationship to the probability that an observation is selected. It is this "self-bounding" feature that leads one to the bound (2.2) of the variance by the mean.

In Section 2.4, we outline analogs of Theorems 2.1 and 2.2 for subsequences that can be decomposed into $d+1$ alternating monotone blocks (rather than just two). If one takes $d = 0$, this reduces to the monotone subsequence problem, and in this case only the variance bound is new. Finally, in Section 2.5 we comment briefly on two conjectures. These deal with a more refined understanding of $\mathrm{Var}[U_n^o(\pi_n^*)]$ and with the naturally associated central limit theorem.

2.1. Mean Bounds and Exact Asymptotics: Proof of Theorem 2.1

Since the distribution F is assumed to be continuous and since the problem is unchanged by replacing X_i by its monotone transformation $F^{-1}(X_i)$, we can assume without loss of generality that the X_i are uniformly distributed on $[0, 1]$. Next, we introduce two tracking variables. First, we let S_i denote the value of the last element that has been selected up to and including time i. We then let R_i denote an indicator variable that tracks the monotonicity of the selected subsequence; specifically we set $R_i = 0$ if the selections made up to and including time i are increasing; otherwise we set $R_i = 1$.

The sequence of real values $\{S_i : R_i = 0,\ 1 \le i \le n\}$ is thus a monotone increasing sequence, though of course not in the strict sense because there will typically be long patches where the successive values of S_i do not change. Similarly, $\{S_i : R_i = 1,\ 1 \le i \le n\}$ is a monotone decreasing sequence, and the full sequence $\{S_i :\ 1 \le i \le n\}$ is a unimodal sequence — in the non-strict sense that permits "flat spots." As a convenience for later formulas, we also set $S_0 = 0$ and $R_0 = 0$.

The Class of Feasible Interval Policies

Here, we will consider feasible policies that have acceptance sets that are given by intervals. It is reasonably obvious that any optimal policy must have this structure, but for completeness we give a formal proof of this fact in Section 2.3.

Now, if the value X_i is under consideration for selection, two possible scenarios can occur: if $R_{i-1} = 0$ (so one is in the "increasing part" of the selected subsequence), then a selectable X_i can be *above or below* S_{i-1}. On the other hand, if $R_{i-1} = 1$ (and one is in the "decreasing part" of the selected subsequence), then any selectable X_i has to be smaller than S_{i-1}. Thus, to specify a feasible interval policy, we just need to specify for each i an interval $[a, b] \subset [0, 1]$ where we accept X_i if $X_i \in [a, b]$ and we reject it otherwise. Here, the values of the end-points of the interval are functions of i, S_{i-1}, and R_{i-1}. In longhand, we write the acceptance interval as

$$\Delta_i(S_{i-1}, R_{i-1}) \equiv [a(i, S_{i-1}, R_{i-1}),\, b(i, S_{i-1}, R_{i-1})].$$

There are some restrictions on the functions $a(i, S_{i-1}, R_{i-1})$ and $b(i, S_{i-1}, R_{i-1})$. To make these explicit, we consider two sets of functions, $\mathcal{A}$ and $\mathcal{B}$. We say $a \in \mathcal{A}$ provided that $a : \{1, 2, ..., n\} \times [0, 1] \times \{0, 1\} \to [0, 1]$ and

$$0 \leq a(i, s, r) \leq s \quad \text{for all } s \in [0, 1],\ r \in \{0, 1\} \text{ and } 1 \leq i \leq n.$$

Similarly, we say $b \in \mathcal{B}$ provided that $b : \{1, 2, ..., n\} \times [0, 1] \times \{0, 1\} \to [0, 1]$ and

$$s \leq b(i, s, 0) \leq 1 \quad \text{for all } s \in [0, 1] \text{ and } 1 \leq i \leq n;$$

$$0 \leq b(i, s, 1) = s \quad \text{for all } s \in [0, 1] \text{ and } 1 \leq i \leq n.$$

Together a pair $(a, b) \in \mathcal{A} \times \mathcal{B}$ defines an *interval policy* $\pi_n \in \Pi(n)$ where we accept X_i at

time i if and only if $X_i \in \Delta_i(S_{i-1}, R_{i-1})$. We let $\Pi'(n)$ denote the set of feasible interval policies.

Three Representations

First, we note that for S_i we have a simple update rule driven by whether X_i is rejected or accepted:

$$S_i = \begin{cases} S_{i-1} & \text{if } X_i \notin \Delta_i(S_{i-1}, R_{i-1}) \\ X_i & \text{if } X_i \in \Delta_i(S_{i-1}, R_{i-1}). \end{cases}$$

For the sequence $\{R_i\}$, the update rule is initialized by setting $R_0 = 0$; one should then note that only one change takes place in the values of the sequence $\{R_i\}$. Specifically, we change to $R_i = 1$ at the first i such that $S_i < S_{i-1}$, i.e. the first instance where we have a decrease in our sequence of selected values. For specificity, we can rewrite this rule as

$$R_i = \begin{cases} 1 & \text{if } X_i \in \Delta_i(S_{i-1}, R_{i-1}) \\ & \text{and } S_{i-1} = \max\{S_k : 1 \leq k \leq i\} \\ R_{i-1} & \text{otherwise.} \end{cases} \tag{2.3}$$

Finally, using $\mathbb{1}(E)$ to denote the indicator function of the event E, we see by counting the occurrences of the "selection events" $X_i \in \Delta_i(S_{i-1}, R_{i-1})$, that for each $1 \leq k \leq n$ the number of selections made up to and including time k is given by the sum of the indicators

$$U_k^o(\pi_n) = \sum_{i=1}^{k} \mathbb{1}\left(X_i \in \Delta_i(S_{i-1}, R_{i-1})\right). \tag{2.4}$$

Proof of the Upper Bound (An a priori *Prophet Inequality)*

The immediate task is to show that for all $n \geq 1$ and all $\pi_n \in \Pi'(n)$, one has the inequality

$$\mathbb{E}[U_n^o(\pi_n)] < 2n^{1/2}. \tag{2.5}$$

It will then follow from Proposition 2.8 in Section 2.3 that the bound (2.5) holds for all $\pi_n \in \Pi(n)$. We start with the representation (2.4) and then after two applications of the Cauchy-Schwarz inequality we have

$$\begin{aligned}
\mathbb{E}[U_n^o(\pi_n)] &= \sum_{i=1}^{n} \mathbb{E}\left[b(i, S_{i-1}, R_{i-1}) - a(i, S_{i-1}, R_{i-1})\right] \\
&\leq n^{1/2} \left\{ \sum_{i=1}^{n} \left(\mathbb{E}\left[b(i, S_{i-1}, R_{i-1}) - a(i, S_{i-1}, R_{i-1})\right]\right)^2 \right\}^{1/2} \\
&\leq n^{1/2} \left\{ \sum_{i=1}^{n} \mathbb{E}\left[\left(b(i, S_{i-1}, R_{i-1}) - a(i, S_{i-1}, R_{i-1})\right)^2\right] \right\}^{1/2}.
\end{aligned}$$

The target bound (2.5) is therefore an immediate consequence of the following — curiously general — lemma.

Lemma 2.4 (Telescoping Bound)**.** *For each $n \geq 1$ and for any strategy $\pi_n \in \Pi'(n)$, one has the inequality*

$$\sum_{i=1}^{n} \mathbb{E}\left[\left(b(i, S_{i-1}, R_{i-1}) - a(i, S_{i-1}, R_{i-1})\right)^2\right] < 4. \tag{2.6}$$

Proof. We first introduce a bookkeeping function $g : [0,1] \times \{0,1\} \to [0,2]$ by setting

$$g(s,r) = \begin{cases} s, & \text{if } r = 0 \\ 2 - s, & \text{if } r = 1. \end{cases}$$

Trivially g is bounded by 2, and we will argue by conditioning and telescoping that the left side of inequality (2.6) is bounded above by $2\,\mathbb{E}\left[g(S_n, R_n)\right] < 4$. Specifically, if we condition on $\mathcal{F}_{i-1}$, then the independence and uniform distribution of X_i gives us, after a few lines of

straightforward calculation, that

$$\begin{aligned}
&\mathbb{E}[g(S_i,R_i) - g(S_{i-1},0) \mid \mathcal{F}_{i-1}] \\
&\qquad = \int_{a(i,S_{i-1},0)}^{S_{i-1}} (g(x,1) - S_{i-1})\,dx + \int_{S_{i-1}}^{b(i,S_{i-1},0)} (g(x,0) - S_{i-1})\,dx \\
&\qquad = \frac{1}{2}\left(b(i,S_{i-1},0) - a(i,S_{i-1},0)\right)^2 \\
&\qquad\quad + \left(S_{i-1} - a(i,S_{i-1},0)\right)\left(2 - S_{i-1} - b(i,S_{i-1},0)\right).
\end{aligned}$$

Since last summand is non-negative we have the tidier bound

$$\left(b(i,S_{i-1},0) - a(i,S_{i-1},0)\right)^2 \leq 2\,\mathbb{E}[g(S_i,R_i) - g(S_{i-1},0) \mid \mathcal{F}_{i-1}]. \tag{2.7}$$

By an analogous direct calculation one also has the identity

$$\begin{aligned}
\mathbb{E}[g(S_i,1) - g(S_{i-1},1) \mid \mathcal{F}_{i-1}] &= \int_{a(i,S_{i-1},1)}^{S_{i-1}} (g(x,1) - g(S_{i-1},1))\,dx \\
&= \frac{1}{2}\left(b(i,S_{i-1},1) - a(i,S_{i-1},1)\right)^2.
\end{aligned} \tag{2.8}$$

Since $R_{i-1} = 1$ implies $R_i = 1$, we can write $g(S_i,R_i) - g(S_{i-1},R_{i-1})$ as the sum

$$\{g(S_i,R_i) - g(S_{i-1},0)\}\mathbb{1}(R_{i-1}=0) + \{g(S_i,1) - g(S_{i-1},1)\}\mathbb{1}(R_{i-1}=1),$$

so the two bounds (2.7) and (2.8) give us the key estimate

$$\left(b(i,S_{i-1},R_{i-1}) - a(i,S_{i-1},R_{i-1})\right)^2 \leq 2\,\mathbb{E}[g(S_i,R_i) - g(S_{i-1},R_{i-1}) \mid \mathcal{F}_{i-1}].$$

Finally, when we take the total expectation and sum, one sees that telescoping gives

$$\sum_{i=1}^{n} \mathbb{E}\left[\left(b(i,S_{i-1},R_{i-1}) - a(i,S_{i-1},R_{i-1})\right)^2\right] \leq 2\,\mathbb{E}\left[g(S_n,R_n)\right] < 4,$$

just as needed. □

Proof of the Lower Bound (Exploitation of Suboptimality)

We construct an explicit policy $\widetilde{\pi}_n \in \Pi(n)$ that is close enough to optimal to give us the bound

$$2n^{1/2} - 4(\pi/6)^{1/2}n^{1/4} - O(1) < \mathbb{E}[U_n^o(\pi_n^*)]. \tag{2.9}$$

The basic idea is to make an approximately optimal choice of an increasing subsequence from the sample $\{X_i : 1 \le i \le n/2\}$ and an approximately optimal choice of a decreasing subsequence from the sample $\{X_i : n/2+1 \le i \le n\}$. The cost of giving up a flexible choice of the "turn-around time" is substantial, but this class of policies is still close enough to optimal to give the required bound (2.9).

For the moment, we assume that n is even. We then select observations according to the following process:

- For $1 \le i \le n/2$, we select the observation X_i if and only if X_i falls in the interval between S_{i-1} and $\min\{1,\, S_{i-1} + 2n^{-1/2}\}$.
- We set $S_{n/2} = 1$ and for $n/2 + 1 \le i \le n$, we select the observation X_i if and only if X_i falls in the interval between $\max\{0,\, S_{i-1} - 2n^{-1/2}\}$ and S_{i-1}.

Here, of course, the selections for $1 \le i \le n/2$ are increasing and the selections for $n/2+1 \le i \le n$ are decreasing, so the selected subsequence is indeed unimodal.

We then consider the stopping time

$$\nu = \min\{i : S_i > 1 - 2n^{-1/2} \text{ or } i \ge n/2\},$$

and we note that the representation (2.4), the suboptimality of the policy $\widetilde{\pi}_n$, and the symmetry between our policy on $1 \le i \le n/2$ and on $n/2+1 \le i \le n$ will give us the lower bound

$$2\,\mathbb{E}\left[\sum_{i=1}^{\nu} \mathbb{1}\left(X_i \in [S_{i-1}, S_{i-1} + 2n^{-1/2}]\right)\right] \le \mathbb{E}[U_n^o(\widetilde{\pi}_n)] \le \mathbb{E}[U_n^o(\pi_n^*)]. \tag{2.10}$$

Wald's Lemma now tells us that

$$\mathbb{E}\left[\sum_{i=1}^{\nu} \mathbb{1}\left(X_i \in [S_{i-1}, S_{i-1} + 2n^{-1/2}]\right)\right] = 2\, n^{-1/2}\mathbb{E}[\nu],$$

so we have

$$4\, n^{-1/2}\, \mathbb{E}[\nu] \leq \mathbb{E}[U_n^o(\pi_n^*)].$$

The main task is to estimate $\mathbb{E}[\nu]$. It is a small but bothersome point that the summands $\mathbb{1}\left(X_i \in [S_{i-1}, S_{i-1} + 2n^{-1/2}]\right)$ are not i.i.d. over the entirety of the range $i \in [1, n/2]$; the distribution of the last terms differ from that of the predecessors. To deal with this nuisance, we take Z_j, $1 \leq j < \infty$, to be a sequence of random variables defined by setting

$$Z_j = \begin{cases} 0 & \text{w.p. } 1 - 2n^{-1/2} \\ U_j & \text{w.p. } 2n^{-1/2}, \end{cases}$$

where the U_j's are independent and uniformly distributed on $[0, 2n^{-1/2}]$. Easy calculations now give us for all $1 \leq j < \infty$ that

$$\mathbb{E}Z_j = \frac{2}{n}, \quad \mathrm{Var}[Z_j] = \frac{8n^{1/2} - 12}{3n^2} < \frac{8}{3n^{3/2}}, \quad \text{and } |Z_j - \mathbb{E}Z_j| < \frac{2}{n^{1/2}}. \tag{2.11}$$

Next, if we set $\widetilde{S}_0 \equiv 0$ and put

$$\widetilde{S}_i = \sum_{j=1}^{i} Z_j, \quad \text{for } 1 \leq i \leq n,$$

for $1 \leq i \leq \nu$, we have $S_i \stackrel{d}{=} \widetilde{S}_i$. Setting $\widetilde{\nu} = \min\{i : \widetilde{S}_i > 1 - 2n^{-1/2} \text{ or } i \geq n/2\}$ we also have $\nu \stackrel{d}{=} \widetilde{\nu}$, so to estimate $\mathbb{E}[\nu]$ it then suffices to estimate

$$\mathbb{E}[\widetilde{\nu}] = \sum_{i=0}^{n/2-1} \mathbb{P}\left(\widetilde{\nu} > i\right) = \sum_{i=0}^{n/2-1} \mathbb{P}\left(\widetilde{S}_i \leq 1 - 2n^{-1/2}\right) = \frac{n}{2} - \sum_{i=0}^{n/2-1} \mathbb{P}\left(\widetilde{S}_i > 1 - 2n^{-1/2}\right).$$

The proof of the lower bound (2.9) will then be complete once we check that

$$\sum_{i=0}^{n/2-1} \mathbb{P}\left(\widetilde{S}_i > 1 - 2n^{-1/2}\right) < (\pi/6)^{1/2} n^{3/4} + \lceil n^{1/2} \rceil. \tag{2.12}$$

This bound turns out to be a reasonably easy consequence of Bernstein's inequality (c.f., Lugosi, 2009, Theorem 6) which asserts that for *any* i.i.d sequence $\{Z_j\}$ with the almost sure bound $|Z_j - \mathbb{E}Z_j| \leq M$ one has for all $t > 0$ that

$$\mathbb{P}\left(\sum_{j=1}^{i} \{Z_j - \mathbb{E}Z_j\} > t\right) \leq \exp\left\{-\frac{t^2}{2i\,\mathrm{Var}[Z_1] + 2Mt/3}\right\}.$$

If we set $n^* = \lfloor n/2 - n^{1/2} - 1 \rfloor$, then Bernstein's inequality together with the bounds (2.11) and some simplification will give us

$$\begin{aligned}\sum_{i=0}^{n/2-1} \mathbb{P}\left(\widetilde{S}_i > 1 - 2n^{-1/2}\right) &\leq \lceil n^{1/2} \rceil + \sum_{i=0}^{n^*} \mathbb{P}\left(\widetilde{S}_i > 1 - 2n^{-1/2}\right) \\ &\leq \lceil n^{1/2} \rceil + \sum_{i=0}^{n^*} \exp\left\{-\frac{3\left(-2i - 2n^{1/2} + n\right)^2}{8n\left(n^{1/2} - 1\right)}\right\}.\end{aligned}$$

The summands are increasing, so the sum is bounded by

$$\int_0^{n/2-n^{1/2}} \exp\left\{-\frac{3\left(-2u - 2n^{1/2} + n\right)^2}{8n\left(n^{1/2} - 1\right)}\right\} du = (2/3)^{1/2}(n^{3/2} - n)^{1/2} \int_0^{\alpha(n)} e^{-u^2}\, du,$$

where $\alpha(n) = (3/8)^{1/2}\left(n^{1/2} - 2\right)\left(n^{1/2} - 1\right)^{-1/2}$. Upon bounding the last integral by $\pi^{1/2}/2$, one then completes the proof of the target bound (2.12). Finally, we note that if n is odd, one can simply ignore the last observation at the cost of decreasing our lower bound by at most one.

Remark 2.5. A benefit of Bernstein's inequality (and the slightly sharper Bennett inequality) is that one gets to take advantage of the good bound on $\mathrm{Var}[Z_j]$. The workhorse Hoeffding inequality would be blind to this useful information.

2.2. Variance Bound: Proof of Theorem 2.2

To prove the variance bound in Theorem 2.2, we need some of the machinery of the Bellman equation and dynamic programming. To introduce the classical backward induction, we first set $v_i(s,r)$ equal to the expected length of the longest unimodal subsequence of $\{X_i, X_{i+1}, \ldots, X_n\}$ that is obtained by sequential selection when $S_{i-1} = s$ and $R_{i-1} = r$. We then have the "terminal conditions"

$$v_n(s,0) = 1, \quad v_n(s,1) = s, \quad \text{for all } s \in [0,1]$$

and we set

$$v_{n+1}(s,r) \equiv 0 \quad \text{for all } s \in [0,1] \text{ and } r \in \{0,1\}.$$

For $1 \le i \le n-1$, we have the *Bellman equation*:

$$v_i(s,r) = \begin{cases} \int_0^s \max\{v_{i+1}(s,0), 1 + v_{i+1}(x,1)\}\, dx & \text{if } r = 0 \\ \quad + \int_s^1 \max\{v_{i+1}(s,0), 1 + v_{i+1}(x,0)\}\, dx & \\ & \\ (1-s) v_{i+1}(s,1) & \text{if } r = 1 \\ \quad + \int_0^s \max\{v_{i+1}(s,1), 1 + v_{i+1}(x,1)\}\, dx. & \end{cases} \tag{2.13}$$

One should note that the map $s \mapsto v_i(s,0)$ is continuous and strictly decreasing on $[0,1]$ for $1 \le i \le n-1$ with $v_n(s,0) = 1$ for all $s \in [0,1]$. In addition, the map $s \mapsto v_i(s,1)$ is continuous and strictly increasing on $[0,1]$ for all $1 \le i \le n$.

If we now define $a^* : \{1,2,\ldots,n\} \times [0,1] \times \{0,1\} \to [0,1]$ by setting

$$a^*(i,s,r) = \inf\{x \in [0,s] : v_{i+1}(s,r) \le 1 + v_{i+1}(x,1)\}, \tag{2.14}$$

then we have $a^* \in \mathcal{A}$. Similarly, if we define $b^* : \{1,2,\ldots,n\} \times [0,1] \times \{0,1\} \to [0,1]$ by

setting

$$b^*(i,s,r) = \begin{cases} \sup\{x \in [s,1] : v_{i+1}(s,0) \leq 1 + v_{i+1}(x,0)\} & \text{if } r = 0. \\ s & \text{if } r = 1. \end{cases} \tag{2.15}$$

then we have $b^* \in \mathcal{B}$. Here, $a^*(i,s,r)$ and $b^*(i,s,r)$ are state-dependent thresholds for which one is indifferent between (i) selecting the current observation x, adjusting r to r' as in (2.3), and continuing to act optimally with new state pair (x,r'), or (ii) rejecting the current observation, x, and continuing to act optimally with unchanged state pair, (s,r).

By the Bellman equation (2.13) and the continuity and monotonicity properties of the value function, the values a^* and b^* provide us with a unique acceptance interval for all $1 \leq i \leq n$ and all pairs (s,r). The policy π_n^* associated with a^* and b^* then accepts X_i at time $1 \leq i \leq n$ if and only if

$$X_i \in \Delta_i^*(S_{i-1}, R_{i-1}) \equiv [a^*(i, S_{i-1}, R_{i-1}),\, b^*(i, S_{i-1}, R_{i-1})],$$

where, as in Section 2.1, S_{i-1} is the value of the last observation selected up to and including time $i-1$, and R_{i-1} tracks the direction of the monotonicity of the subsequence selected up to and including time $i-1$. In Section 2.3, we will prove that this policy is indeed the unique optimal policy for the sequential selection of a unimodal subsequence.

We do not need a detailed analysis of a^* and b^*, but it is useful to collect some facts. In particular, one should note that $a^*(i,s,r) = 0$ whenever $v_{i+1}(s,r) \leq 1$ and $b^*(i,s,0) = 1$ whenever $v_{i+1}(s,0) \leq 1$. In addition, the difference $b^*(i,s,r) - a^*(i,s,r)$ provides us with an explicit bound on the increments of the value function $v_i(s,r)$, as the following lemma suggests.

Lemma 2.6. *For all $s \in [0,1]$, $r \in \{0,1\}$ and $1 \leq i \leq n$, we have*

$$0 \leq v_i(s,r) - v_{i+1}(s,r) \leq b^*(i,s,r) - a^*(i,s,r) \leq 1. \tag{2.16}$$

Proof. The lower bound is trivial and it follows by the fact that $v_i(s,r)$ is strictly decreasing in i for each $(s,r) \in [0,1] \times \{0,1\}$.

For the upper bound, we first assume that $r = 0$. Then, subtracting $v_{i+1}(s,0)$ on both sides of equation (2.13) when $r = 0$ and using the definition of a^* and b^*, we obtain

$$\begin{aligned} v_i(s,0) - v_{i+1}(s,0) = &-(b^*(i,s,r) - a^*(i,s,r))v_{i+1}(s,0) \\ &+ \int_{a^*(i,s,r)}^{s} (1 + v_{i+1}(x,1))\,dx + \int_{s}^{b^*(i,s,r)} (1 + v_{i+1}(x,0))\,dx. \end{aligned}$$

Recalling the monotonicity property of $s \mapsto v_{i+1}(s,r)$, we then have

$$\begin{aligned} v_i(s,0) - v_{i+1}(s,0) \leq &-(b^*(i,s,r) - a^*(i,s,r))v_{i+1}(s,0) \\ &+ (s - a^*(i,s,r))(1 + v_{i+1}(s,1)) + (b^*(i,s,r) - s)(1 + v_{i+1}(s,0)), \end{aligned}$$

and since $v_{i+1}(s,1) \leq v_{i+1}(s,0)$, we finally obtain

$$v_i(s,0) - v_{i+1}(s,0) \leq b^*(i,s,r) - a^*(i,s,r) \leq 1,$$

as (2.16) requires. The proof for $r = 1$ is very similar and it is therefore omitted. □

We now come to the main lemma of this section.

Lemma 2.7. *The process defined by*

$$Y_i = U_i^o(\pi_n^*) + v_{i+1}(S_i, R_i) \quad \textit{for all } 0 \leq i \leq n,$$

is a martingale with respect to the natural filtration $\{\mathcal{F}_i\}_{0 \leq i \leq n}$. *Moreover, for the martingale difference sequence* $d_i = Y_i - Y_{i-1}$ *one has that*

$$|d_i| = |\ Y_i - Y_{i-1}\ | \leq 1 \quad \textit{for all } 1 \leq i \leq n.$$

Proof. We first note that Y_i is $\mathcal{F}_i$-measurable and bounded. Then, from the definition of

$v_i(s,r)$ we have that $v_i(S_{i-1}, R_{i-1}) = \mathbb{E}\left[U_n^o(\pi_n^*) - U_{i-1}^o(\pi_n^*) \mid \mathcal{F}_{i-1}\right]$. Thus,

$$Y_i = U_i^o(\pi_n^*) + \mathbb{E}\left[U_n^o(\pi_n^*) - U_i^o(\pi_n^*) \mid \mathcal{F}_i\right] = \mathbb{E}\left[U_n^o(\pi_n^*) \mid \mathcal{F}_i\right],$$

which is clearly a martingale.

To see that the martingale differences are bounded let

$$W_i = v_{i+1}(S_{i-1}, R_{i-1}) - v_i(S_{i-1}, R_{i-1})$$

represents the change in Y_i if we do not select X_i, and let

$$Z_i = (1 + v_{i+1}(X_i, \mathbb{1}(X_i < S_{i-1})) - v_{i+1}(S_{i-1}, R_{i-1}))\mathbb{1}(X_i \in \Delta_i^*(S_{i-1}, R_{i-1}))$$

represents the change when we do select X_i. We then have that

$$d_i = W_i + Z_i,$$

and by our Lemma 2.6 we know that $-1 \leq W_i \leq 0$. Moreover, the definition of the threshold functions a^* and b^* and the monotonicity property of $s \mapsto v_{i+1}(s,r)$ give us that $0 \leq Z_i \leq 1$, so that $|d_i| \leq 1$, as desired. □

Final Argument for the Variance Bound

For the martingale differences $d_i = Y_i - Y_{i-1}$ we have

$$Y_n - Y_0 = \sum_{i=1}^n d_i, \quad \text{and} \quad \text{Var}[Y_n] = \mathbb{E}\left[\sum_{i=1}^n d_i^2\right],$$

and we also have the initial representation

$$Y_0 = U_0^o(\pi_n^*) + v_1(S_0, R_0) = v_1(0,0) = \mathbb{E}[U_n^o(\pi_n^*)]$$

and the terminal identity

$$Y_n = U_n^o(\pi_n^*) + v_{n+1}(S_n, R_n) = U_n^o(\pi_n^*).$$

We now recall the decomposition $d_i = W_i + Z_i$ introduced in the proof of Lemma 2.7, where

$$W_i = v_{i+1}(S_{i-1}, R_{i-1}) - v_i(S_{i-1}, R_{i-1})$$

and

$$Z_i = (1 + v_{i+1}(X_i, \mathbb{1}(X_i < S_{i-1})) - v_{i+1}(S_{i-1}, R_{i-1}))\mathbb{1}(X_i \in \Delta_i^*(S_{i-1}, R_{i-1})).$$

Since W_i is $\mathcal{F}_{i-1}$ measurable, we have

$$\mathbb{E}\left[d_i^2 \mid \mathcal{F}_{i-1}\right] = \mathbb{E}\left[Z_i^2 \mid \mathcal{F}_{i-1}\right] + 2\, W_i\, \mathbb{E}\left[Z_i \mid \mathcal{F}_{i-1}\right] + W_i^2.$$

We also have $0 = \mathbb{E}\left[d_i \mid \mathcal{F}_{i-1}\right] = W_i + \mathbb{E}\left[Z_i \mid \mathcal{F}_{i-1}\right]$ so

$$\mathbb{E}\left[d_i^2 \mid \mathcal{F}_{i-1}\right] = \mathbb{E}\left[Z_i^2 \mid \mathcal{F}_{i-1}\right] - W_i^2. \tag{2.17}$$

Finally, from the definition of Z_i, a^* and b^* we obtain

$$\begin{aligned}\mathbb{E}\left[Z_i^2 \mid \mathcal{F}_{i-1}\right] &= \int_{a^*(i,S_{i-1},R_{i-1})}^{b^*(i,S_{i-1},R_{i-1})} (1 + v_{i+1}(x, \mathbb{1}(x < S_{i-1})) - v_{i+1}(S_{i-1}, R_{i-1}))^2 \, dx \\ &\leq b^*(i, S_{i-1}, R_{i-1}) - a^*(i, S_{i-1}, R_{i-1}),\end{aligned}$$

since the integrand is bounded by 1. Summing (2.17), applying the last bound, and taking expectations gives us

$$\mathrm{Var}[U_n^o(\pi_n^*)] \leq \sum_{i=1}^{n} \mathbb{E}\left[b^*(i, S_{i-1}, R_{i-1}) - a^*(i, S_{i-1}, R_{i-1})\right] = \mathbb{E}[U_n^o(\pi_n^*)],$$

where the last equality follows from our basic representation (2.4).

2.3. Intermezzo: Optimality and Uniqueness of Interval Policies

The unimodal sequential selection problem is a finite horizon Markov decision problem with bounded rewards and finite action space, and for such a problem, it is known that there exists a non-randomized Markov policy π_n^* that is optimal (c.f. Bertsekas and Shreve, 1978, Corollary 8.5.1) . This amounts to saying that there exists an optimal strategy π_n^* such that for each i, S_{i-1} and R_{i-1}, there is a Borel set $D_i^*(S_{i-1}, R_{i-1}) \subseteq [0,1]$ such that X_i is accepted if and only if $X_i \in D_i^*(S_{i-1}, R_{i-1})$. Here, we just want to show that the Borel sets $D_i^*(S_{i-1}, R_{i-1})$ are actually intervals (up to null sets).

Given the optimal acceptance sets $D_i^*(S_{i-1}, R_{i-1})$, $1 \leq i \leq n$, we now set

$$v_i(S_{i-1}, R_{i-1}) = \mathbb{E}\left[\sum_{k=i}^{n} \mathbb{1}(X_k \in D_k^*(S_{k-1}, R_{k-1})) \,\middle|\, \mathcal{F}_{i-1}\right],$$

so we have the recursion

$$v_i(S_{i-1}, R_{i-1}) = \mathbb{E}\left[\mathbb{1}(X_i \in D_i^*(S_{i-1}, R_{i-1})) + v_{i+1}(S_i, R_i) \,\middle|\, \mathcal{F}_{i-1}\right], \tag{2.18}$$

and $v_i(s,r)$ is just the optimal expected number of selections made from the subsample $\{X_i, X_{i+1}, \ldots, X_n\}$ given that $S_{i-1} = s$ and $R_{i-1} = r$. We then note that $v_n(s,0) = 1$ for all $s \in [0,1]$, and one can check by induction on i that the map $s \mapsto v_i(s,0)$ is continuous and strictly decreasing in s for $1 \leq i \leq n-1$. A similar argument also gives that the map $s \mapsto v_i(s,1)$ is continuous and strictly increasing in s for all $1 \leq i \leq n$.

If we now set

$$a(i, S_{i-1}, R_{i-1}) = \operatorname{ess\,inf} D_i^*(S_{i-1}, R_{i-1}) \quad \text{and}$$
$$b(i, S_{i-1}, R_{i-1}) = \operatorname{ess\,sup} D_i^*(S_{i-1}, R_{i-1}),$$

then we want to show for all $1 \leq i \leq n$ and all (S_{i-1}, R_{i-1}) that we have

$$\mathbb{P}(\{D_i(S_{i-1}, R_{i-1})^c \cap [a(i, S_{i-1}, R_{i-1}), b(i, S_{i-1}, R_{i-1})]\}) = 0.$$

To argue by contradiction, we suppose that there is an $1 \leq i \leq n$ and an acceptance set $D_i^* \equiv D_i^*(S_{i-1}, R_{i-1})$ that is not equivalent to an interval; i.e. we suppose

$$\mathbb{P}(\{D_i^{*c} \cap [a^*(i, S_{i-1}, R_{i-1}), b^*(i, S_{i-1}, R_{i-1})]\}) > 0. \tag{2.19}$$

We then consider the sets

$$L_i = [0, S_{i-1}] \cap D_i^* \quad \text{and} \quad U_i = [S_{i-1}, 1] \cap D_i^*,$$

and we introduce the intervals

$$\widetilde{L}_i = [S_{i-1} - |L_i|,\, S_{i-1}] \quad \text{and} \quad \widetilde{U}_i = [S_{i-1},\, S_{i-1} + |U_i|],$$

where $|A|$ denotes the Lebesgue measure of a set A. The set $\widetilde{D}_i = \widetilde{L}_i \cup \widetilde{U}_i$ is also an interval and $|\widetilde{D}_i| = |D_i^*|$, so, if we can show that

$$\mathbb{E}[\mathbb{1}(X_i \in D_i^*) + v_{i+1}(S_i, R_i)] < \mathbb{E}[\mathbb{1}(X_i \in \widetilde{D}_i) + v_{i+1}(S_i, R_i)], \tag{2.20}$$

then the representation (2.18) tells us that policy π_n^* is not optimal, a contradiction.

To prove the bound (2.20), we note that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(X_i \in \widetilde{D}_i) + v_{i+1}(S_i, R_i) \,\middle|\mathcal{F}_{i-1}\right] - \mathbb{E}\left[\mathbb{1}(X_i \in D_i^*) + v_{i+1}(S_i, R_i)\middle|\mathcal{F}_{i-1}\right] \\
&\quad = \mathbb{E}\left[v_{i+1}(X_i, R_i)\mathbb{1}(X_i \in \widetilde{D}_i) \,\middle|\mathcal{F}_{i-1}\right] - \mathbb{E}\left[v_{i+1}(X_i, R_i)\mathbb{1}(X_i \in D_i^*)\middle|\mathcal{F}_{i-1}\right]
\end{aligned}$$

since $\widetilde{D}_i$ and D_i^* are $\mathcal{F}_{i-1}$-measurable and $\mathbb{E}[\mathbb{1}(X_i \in \widetilde{D}_i)|\mathcal{F}_{i-1}] = \mathbb{E}[\mathbb{1}(X_i \in D_i^*)|\mathcal{F}_{i-1}]$. By

our construction, we also have the identities

$$\mathbb{E}\left[v_{i+1}(X_i,R_i)\mathbb{1}(X_i\in\widetilde{D}_i)\,\middle|\mathcal{F}_{i-1}\right]=\int_{\widetilde{L}_i}v_{i+1}(x,1)\,dx+\int_{\widetilde{U}_i}v_{i+1}(x,0)\,dx, \tag{2.21}$$

and

$$\mathbb{E}\left[v_{i+1}(X_i,R_i)\mathbb{1}(X_i\in D_i^*)\,\middle|\mathcal{F}_{i-1}\right]=\int_{L_i}v_{i+1}(x,1)\,dx+\int_{U_i}v_{i+1}(x,0)\,dx. \tag{2.22}$$

Now since $|L_i|=|\widetilde{L}_i|$ implies that $|\widetilde{L}_i\cap L_i^c|=|L_i\cap\widetilde{L}_i^c|$, we can write

$$\begin{aligned}\int_{\widetilde{L}_i}v_{i+1}(x,1)\,dx-\int_{L_i}v_{i+1}(x,1)\,dx&=\int_{\widetilde{L}_i\cap L_i^c}v_{i+1}(x,1)\,dx-\int_{L_i\cap\widetilde{L}_i^c}v_{i+1}(x,1)\,dx\\&=(\beta_i-\alpha_i)|\widetilde{L}_i\cap L_i^c|,\end{aligned} \tag{2.23}$$

where $\alpha_i=\alpha_i(S_{i-1},R_{i-1})$, and $\beta_i=\beta_i(S_{i-1},R_{i-1})$ are chosen according to the mean value theorem for integrals. The sets $\widetilde{L}_i\cap L_i^c$ and $L_i\cap\widetilde{L}_i^c$ are almost surely disjoint since $\widetilde{L}_i\cap L_i^c\subset[S_{i-1}-|L_i|,S_{i-1}]$ and $L_i\cap\widetilde{L}_i^c\subset[0,S_{i-1}-|L_i|]$. So, we find that $\alpha_i<\beta_i$ since $v_{i+1}(x,1)$ is strictly decreasing in x.

An analogous argument tells us that we can write

$$\int_{\widetilde{U}_i}v_{i+1}(x,1)\,dx-\int_{U_i}v_{i+1}(x,1)\,dx=(\delta_i-\gamma_i)|\widetilde{U}_i\cap U_i^c|, \tag{2.24}$$

where $\gamma_i<\delta_i$ and γ_i and δ_i depend on (S_{i-1},R_{i-1}). If we now set

$$c_i(S_{i-1},R_{i-1})=\min\{\beta_i-\alpha_i,\delta_i-\gamma_i\},$$

then the identities (2.21) and (2.22) and the differences (2.23) and (2.24) give us the bound

$$c_i(S_{i-1},R_{i-1})|\widetilde{D}_i\cap D_i^{*c}|\leq\mathbb{E}\Big[v_{i+1}(X_i,R_i)\mathbb{1}(X_i\in\widetilde{D}_i)-v_{i+1}(X_i,R_i)\mathbb{1}(X_i\in D_i^*)\Big|\mathcal{F}_{i-1}\Big].$$

Since $c_i(S_{i-1},R_{i-1})>0$, the assumption (2.19) implies that the left hand-side above is

strictly positive. When we take total expectation we get

$$0 < \mathbb{E}\left[v_{i+1}(X_i, R_i)\mathbb{1}(X_i \in \widetilde{D}_i) - v_{i+1}(X_i, R_i)\mathbb{1}(X_i \in D_i^*)\right].$$

In view of the recursion (2.18), this contradicts the optimality of π^*. This completes the proof of (2.20), and, in summary we have the following proposition.

Proposition 2.8. *If π_n^* is an optimal non-randomized Markov policy for the unimodal sequential selection problem, then, up to sets of measure zero, π^* is an interval policy.*

Corollary 2.9. *There is a unique policy $\pi_n^* \in \Pi(n)$ that is optimal.*

To prove the corollary, one combines the optimality of the interval policy given by Proposition 2.8 with the monotonicity properties of the Bellman equation (2.13). Specifically, the map $s \mapsto v_i(s, 0)$ is strictly decreasing in s for all $1 \leq i \leq n-1$ and the map $s \mapsto v_i(s, 1)$ is strictly increasing in s for all $1 \leq i \leq n$, so the equations (2.14) and (2.15) determine the values $a^*(\cdot)$ and $b^*(\cdot)$ uniquely.

2.4. Generalizations and Specializations: d-Modal Subsequences

There are natural analogs of Theorems 2.1 and 2.2 for "d-modal subsequences," by which we mean subsequences that are allowed to make "d-turns" rather than just one. Equivalently, these are subsequences that are the concatenation of (at most) $d+1$ monotone subsequences. If we let $U_n^{o,d}(\pi_n^*)$ denote the analog of $U_n^o(\pi_n^*)$ when the selected subsequence is d-modal, then the arguments of the preceding sections may be adapted to provide information on the expected value of $U_n^{o,d}(\pi_n^*)$ and its variance. Here, one should keep in mind that the case $d = 0$ is *not* excepted; the arguments of the preceding sections do indeed apply to the selection of monotone subsequences.

Theorem 2.10 (Expected Length of Optimal d-Modal Subsequences)**.** *If $\Pi(n)$ denotes the class of feasible policies for the d-modal subsequence selection problem, then there is a*

unique $\pi_n^* \in \Pi(n)$ *such that*

$$\mathbb{E}[U_n^{o,d}(\pi_n^*)] = \sup_{\pi_n \in \Pi(n)} \mathbb{E}[U_n^{o,d}(\pi_n)].$$

Moreover, for all $n \geq 1$ *and* $d \geq 0$ *one has*

$$c(d)^{1/2} n^{1/2} - c(d)^{3/4} (\pi/3)^{1/2} n^{1/4} - O(1) < \mathbb{E}[U_n^{o,d}(\pi_n^*)] < c(d)^{1/2} n^{1/2}, \tag{2.25}$$

where $c(d) = 2(d+1)$. *In particular, one has*

$$\mathbb{E}[U_n^{o,d}(\pi_n^*)] \sim \{2(d+1)\}^{1/2} n^{1/2} \quad \text{as } n \to \infty.$$

One should note that the case $d = 0$ corresponds to the *monotone subsequence selection problem* studied by Samuels and Steele (1981) and more recently by Gnedin (1999). The monotone selection problem is also equivalent to certain bin packing problems studied by Bruss and Robertson (1991) and Rhee and Talagrand (1991).

In the special case of $d = 0$, our upper bound (2.25) agrees with that of Bruss and Robertson (1991) as well as with the result of Gnedin (1999). Our lower bound (2.25) on the mean for $d = 0$ turns out to be slightly worse than that of Rhee and Talagrand's (1991) since our constant for the $n^{1/4}$ term is $2^{3/4}(\pi/3)^{1/2} \sim 1.72$, while theirs is $8^{1/4} \sim 1.68$.

For the d-modal problem, one can also prove the a variance bound that generalizes Theorem 2.2 in a natural way.

Theorem 2.11 (Variance Bound for d-Modal Subsequences)**.** *For the unique optimal policy* $\pi_n^* \in \Pi(n)$, *one has the bound*

$$\mathrm{Var}[U_n^{o,d}(\pi_n^*)] \leq \mathbb{E}[U_n^{o,d}(\pi_n^*)].$$

Chebyshev's inequality and Theorem 2.11 now combine as usual to provide a weak law for

$U_n^{o,d}(\pi_n^*)$. Even for $d = 0$ this variance bound is new.

2.5. Two Conjectures

Numerical studies for small d and moderate n, support the conjecture that one has the asymptotic relation

$$\mathrm{Var}[U_n^{o,d}(\pi_n^*)] \sim \frac{1}{3}\mathbb{E}[U_n^{o,d}(\pi_n^*)] \quad \text{as } n \to \infty. \tag{2.26}$$

As observed by an anonymous reader, the methods of Section 2.2 and the concavity of the value function established in Samuels and Steele (1981) are in fact enough to prove an appropriate lower bound

$$\frac{1}{3}\,\mathbb{E}[U_n^{o,d}(\pi_n^*)] - 2 < \mathrm{Var}[U_n^{o,d}(\pi_n^*)] \quad \text{where } d = 0. \tag{2.27}$$

Here, one should now be able to prove an upper bound on $\mathrm{Var}[U_n^{o,d}(\pi_n^*)]$ that is strong enough to establish the case $d = 0$ of the conjecture (2.26), but confirmation of this has eluded us.

Also, by numerical calculations of the optimal policy π_n^* and by subsequent simulations of $U_n^{o,d}(\pi_n^*)$ for $d = 0$, $d = 1$, and modest values of n, it seems likely that the random variable $U_n^{o,d}(\pi_n^*)$ obeys a central limit theorem. Specifically, the natural conjecture is that for all $d \geq 0$ one has

$$\frac{\sqrt{3}\left(U_n^{o,d}(\pi_n^*) - \sqrt{2(d+1)n}\right)}{(2(d+1)n)^{1/4}} \Longrightarrow N(0,1) \quad \text{as } n \to \infty. \tag{2.28}$$

Implicit in this conjecture is the belief that the lower bound (2.25) can be improved to $\{2(d+1)n\}^{1/2} - o(n^{1/4})$, or better.

So far, the only central limit theorem available for a sequential selection problem is that obtained by Bruss and Delbaen (2001, 2004) for a Poissonized version of the monotone subsequence problem. Given the sequential nature of the problem, it appears to be difficult

to de-Poissonize the results of Bruss and Delbaen (2004) to obtain conclusions about the distribution of $U_n^{o,d}(\pi_n^*)$ even for $d = 0$.

For completeness, we should note that even for the *off-line* unimodal subsequence problem, not much more is known about the random variable U_n than its asymptotic expected value (2.1). Here one might hope to gain some information about the distribution of U_n by the methods of Bollobás and Brightwell (1992) and Bollobás and Janson (1997), and it is even feasible — but only remotely so — that one could extend the famous distributional results of Baik et al. (1999) to unimodal subsequences. More modestly, one certainly should be able to prove that the distribution of U_n is *not* asymptotically normal. One motivation for going after such a result would be to underline how the restriction to sequential strategies can bring one back to the domain of the central limit theorem.

CHAPTER 3 : Markov Decision Problems where Means Bound Variances

The reward $R_n(\pi_n^*)$ that one receives by following an optimal policy π_n^* for a Markov decision problem (MDP) with $n < \infty$ decision periods is a random variable, and its expected value is often well understood. Still, just knowing the mean of a random variable leaves much that is unknown. Given the extensive literature on MDPs, it is striking that we typically fail to have much understanding of the distribution of $R_n(\pi_n^*)$ that goes beyond what can be said about its expected value.

Moreover, in many MDPs, the reward $R_n(\pi_n^*)$ has a direct economic interpretation, and ultimately one is expected to make a well-founded judgment about the utility of an optimal policy π_n^*. Here, it is clear that one needs to take into account the riskiness of the reward. The simplest measures of riskiness are the standard deviation and the variance of $R_n(\pi_n^*)$, yet even these are often unstudied for MDPs.

Our main goal here is to identify a substantial class of MDPs for which we can say something general and useful about the variance of $R_n(\pi_n^*)$. Specifically, we identify an example-rich class of MDPs for which the variance of $R_n(\pi_n^*)$ can be bounded by a small constant multiple of its expectation. Several useful consequences follow from this bound, including practical constraints on the riskiness of the realized reward and a straightforward weak law of large numbers.

A Typical Example

To fix ideas and to gain some intuition, we first consider a simple version of the sequential knapsack problem. The capacity $c \in (0, \infty)$ of the knapsack is given, and we are sequentially presented with non-negative values $Y_1, Y_2, \ldots, Y_n$ that we view as item sizes. We assume that the item sizes are independent random variables with common distribution F, and for

This chapter is written under the supervision of Prof. Noah Gans and J. Michael Steele. The results presented here are also in a joint research paper with Noah Gans and J. Michael Steele.

specificity, we assume there are constants $A > 0$ and $\alpha > 0$ such that $F(x) \sim Ax^\alpha$ as $x \to 0$. In the simplest — but most important — case, the random variables Y_i, $1 \leq i \leq n$, are uniformly distributed on $[0, 1]$. In this case we have $F(x) = x$ for $x \in [0, 1]$, so we have $A = 1$ and $\alpha = 1$.

Now, at time t when Y_t is first presented, the decision maker must determine whether to include or to exclude Y_t from the knapsack. The goal of the decision maker is to maximize the expected number of items that can be included without exceeding the capacity constraint.

We let $\Pi(n)$ denote the set of all non-anticipating knapsack policies, and for any policy $\pi_n \in \Pi(n)$ we let τ_i denote the index of the ith item that is chosen for inclusion in the knapsack. Non-anticipation of the policy π_n is equivalent to saying that each τ_i is a stopping time with respect to the increasing sequence of σ-fields $\mathcal{F}_t = \sigma\{Y_1, Y_2, \ldots, Y_t\}$, $1 \leq t \leq n$. Moreover, we have a concrete representation for the number of items included in the knapsack when one follows the policy π_n; the reward is simply the number of inclusions

$$R_n(\pi_n) = \max\left\{k : \sum_{i=1}^{k} Y_{\tau_i} \leq c\right\}.$$

Given this setup, classical results from dynamic programming — and common sense — now assure us that for each n there is a Markov deterministic policy $\pi_n^* \in \Pi(n)$ that maximizes the expected number of items in the knapsack.

A great deal is known about the expected value $\mathbb{E}\left[R_n(\pi_n^*)\right]$ of the optimal policy under this model. In particular, the analysis of Coffman et al. (1987) tells us that

$$\mathbb{E}\left[R_n(\pi_n^*)\right] \sim \left[A\alpha^{-\alpha}(\alpha+1)^\alpha cn\right]^{1/(1+\alpha)} \qquad \text{as } n \to \infty.$$

This relation was subsequently refined by an upper bound in Bruss and Robertson (1991) and by a lower bound in Rhee and Talagrand (1991), so this is a problem where the mean is genuinely well-understood.

The Variance Bound

The main result obtained here now tells us that one can supplement this characterization of the expected value with a bound on the variance. Specifically, in this problem we have

$$\mathrm{Var}\left[R_n(\pi_n^*)\right] \leq \mathbb{E}\left[R_n(\pi_n^*)\right] \quad \text{for each } 1 \leq n < \infty. \tag{3.1}$$

As an immediate corollary of this bound, we obtain a weak law of large numbers for $R_n(\pi_n^*)$. Specifically, from the variance bound (3.1) and Chebyshev's inequality, one finds that as $n \to \infty$ we have that

$$n^{-1/(1+\alpha)} R_n(\pi_n^*) \quad \xrightarrow{p} \quad \left[A\alpha^{-\alpha}(\alpha+1)^{\alpha}c\right]^{1/(1+\alpha)}.$$

Here, one should notice that the distribution of the realized reward is a consequence of the underlying model and of the optimality criterion, which focuses exclusively on the expected reward $\mathbb{E}[R_n(\pi_n^*)]$. Since the optimality of strategy π_n^* only takes the expected reward into consideration, it is noteworthy that there is any relation at all between the mean of the reward $R_n(\pi_n^*)$ and its variance. One cannot rule out the *a priori* possibility that the optimal strategy π_n^* might perversely inflate the variance $\mathrm{Var}[R_n(\pi_n^*)]$ just to eke out a modest increment to the mean $\mathbb{E}[R_n(\pi_n^*)]$, and there might be MDPs for which such unsightly behavior occurs. Still, for a substantial class of natural problems, the optimal policy is not so short sighted.

Tools, Proofs, and Further Examples

The behavior exhibited by this example turns out to be typical of a large class of MDPs. The identifying feature of these problems is that they satisfy a natural *paid-to-play* property that we describe in detail in Section 3.1. In Section 3.2 we state our main result and discuss some immediate implications.

The proof of the main result follows in Section 3.3, and in Section 3.4 we provide some easy-to-check sufficient conditions that assure an MDP is paid-to-play. Finally, in Sections 3.5 and 3.6 we offer examples of paid-to-play MDPs from operations research, operations management, financial engineering and combinatorial optimization. We also review some connections with related literature, and we underscore some open problems.

3.1. Markov Decision Problems and the Paid-to-Play Property

Here, a discrete-time Markov decision problem is specified by a 5-tuple $(\mathcal{X}, \mathcal{A}, f, r, n)$ where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, f is a deterministic state transition function, r is the one-period reward function, and n is the time horizon. We will further suppose that $\mathcal{X}$ has the form $\mathcal{X}_1 \times \mathcal{X}_2$, so the state of the system at time t can be written as (x_t, y_t) with $x_t \in \mathcal{X}_1$ and $y_t \in \mathcal{X}_2$. This extra structure on $\mathcal{X}$ allows us to accommodate states (x_t, y_t) for which x_t provides some appropriate summary of the "past" and y_t reflects the current state of an exogenous random process. Specifically, we take $y_t = Y_t$, where the random variables $\{Y_t : 1 \leq t \leq n\}$ are independent with known distributions $\{F_t : 1 \leq t \leq n\}$.

Now, given a state $(x, y) \in \mathcal{X}$ and a time $1 \leq t \leq n$ we let $\mathcal{A}_t(x, y) \subseteq \mathcal{A}$ denote the set of feasible actions, and we write

$$\Gamma_t = \{(x, y, a) : (x, y) \in \mathcal{X},\, a \in \mathcal{A}_t(x, y)\}$$

for the set of the time-t admissible state-action pairs.

When the system is in state (x_t, y_t) and the action $a_t \in \mathcal{A}_t(x_t, y_t)$ is chosen, we receive a reward $r_t(x_t, y_t, a_t)$. After we choose action a_t, the system moves from (x_t, y_t) to (x_{t+1}, y_{t+1}) where

$$x_{t+1} = f(t, x_t, y_t, a_t) \quad \text{and} \quad y_{t+1} = Y_{t+1}. \tag{3.2}$$

In general, a *Markov policy* is a sequence $\pi_n = (\mu_1, \mu_2, \ldots, \mu_n)$ of time-indexed stochastic kernels μ_t that associate with each state $(x_t, y_t) \in \mathcal{X}$ a probability measure on the set of

feasible actions $\mathcal{A}_t(x_t, y_t)$. Here we are only concerned with *Markov deterministic policies* where for each t and (x_t, y_t), the selection kernel μ_t assigns mass one to some action $a_t \in \mathcal{A}_t(x_t, y_t)$.

To complete the description of the MDP $(\mathcal{X}, \mathcal{A}, f, r, n)$, we should add that all spaces are assumed to be Polish and all maps are assumed to be measurable. These properties are present in almost any problem of practical interest.

Policies and Optimality

In this general formulation, again $\Pi(n)$ denotes the set of all non-anticipative policies, and for any $\pi_n \in \Pi(n)$ we let

$$R_k(\pi_n) = \sum_{t=1}^{k} r_t(X_t, Y_t, A_t), \qquad 1 \leq k \leq n$$

be the reward accrued up to and including time k.

The optimality criterion of interest in this paper is the so-called *expected total reward*. Hence, we are interested in finding the policy $\pi_n^* \in \Pi(n)$ such that

$$\mathbb{E}[R_n(\pi_n^*)] = \sup_{\pi \in \Pi(n)} \mathbb{E}[R_n(\pi)]. \tag{3.3}$$

Standard dynamic programming techniques allow us to express the value of the optimization problem (3.3) in recursive form. Specifically, we have the *Bellman equation* which tells us that, at each time $1 \leq t \leq n$ and for each state $(X_t, Y_t) = (x, y)$, the expected value of the optimal policy over periods t to n is

$$v_t(x, y) = \sup_{a \in \mathcal{A}_t(x,y)} \left\{ r_t(x, y, a) + \mathbb{E}\left[v_{t+1}(f(t, x, y, a), Y_{t+1}) \right] \right\}, \tag{3.4}$$

and this backwards recursion is initialized by setting $v_{n+1}(x, y) = 0$ for all state values $(x, y) \in \mathcal{X}$. We let $a_t^* \in \mathcal{A}_t(x, y)$ denote the optimal action for period t when in state (x, y),

and we recall that

$$\mathbb{E}[v_1(X_1, Y_1)] = \mathbb{E}[R_n(\pi_n^*)];$$

that is, the expected value of the value function at time one is equal to the expected value of the optimal reward over the full set of n time periods.

Under mild integrability conditions on the reward function, it is known that there is a Markov policy that is optimal (c.f. Bertsekas and Shreve, 1978, Corollary 8.1.1) and, if the supremum in (3.4) is achieved for each $(x, y) \in \mathcal{X}$ and each $1 \leq t \leq n$, the optimal Markov policy can be taken to be deterministic (c.f. Bertsekas and Shreve, 1978, Proposition 8.5.) Moreover, this deterministic Markov policy is generated by the Bellman recursion (3.4). Here, we restrict our analysis to problems in which there is an optimal Markov deterministic policy, and we let π_n^* denote such a policy.

Paid-to-Play Markov Decision Problems

Intuitively, in a *paid-to-play* MDP the decision maker needs to be rewarded (*paid*) to take an action that changes its current state (*to-play*). Thus, without an appropriate reward, the decision maker will always prefer to stay in the current state rather than make a transition to some other feasible state.

More formally, a paid-to-play MDP has three properties that are expressed most easily with help from the Bellman equation (3.4). To specify these, we first let $x_{t+1}^* = f(t, x_t, y_t, a_t^*)$ denote the state that one reaches by taking the optimal action, a_t^*, at time t when in state (x_t, y_t). We can now lay out the full definition.

Definition 3.1 (**Paid-to-Play MDPs).** A Markov decision problem $(\mathcal{X}, \mathcal{A}, f, r, n)$ is said to have the *paid-to-play* property if

(a) the reward function $r_t : \Gamma_t \to \mathbb{R}$ is non-negative and uniformly bounded; i.e. there is a $K < \infty$ such that $0 \leq r_t(x, y, a) \leq K$ for all $(x, y, a) \in \Gamma_t$ and all $1 \leq t \leq n$;

(b) for each state (x_t, y_t) and each decision time $1 \leq t \leq n$, the set of actions $\mathcal{A}_t(x_t, y_t)$ includes a *do-nothing* action a^0 for which one has $r_t(x_t, y_t, a^0) = 0$ and $f(t, x_t, y_t, a^0) = x_t$;

(c) for each time $1 \leq t \leq n$ and for each state (x_t, y_t) one has that

$$\mathbb{E}\left[v_{t+1}(x^*_{t+1}, Y_{t+1})\right] \leq \mathbb{E}\left[v_{t+1}(x_t, Y_{t+1})\right]. \tag{3.5}$$

Condition (a) is self-explanatory. It is automatically satisfied in many (but not all) problems of interest. At a later point, we will consider how this condition may be relaxed.

Condition (b) assures us that we always have the possibility of remaining in state x_t at time $t + 1$, and it tells us that if we take this do-nothing action then we receive no reward. One consequence of this assumption is that it assures us that the quantity $\mathbb{E}\left[v_{t+1}(x_t, Y_{t+1})\right]$ that appears in the equation (3.5) of Condition (c) is always well defined.

Condition (c) gets to the essence of the paid-to-play property. It says that if the decision maker does not receive a reward for moving to a new state, then the decision maker always stays in the current state. Thus, unless there is a strictly positive reward for moving, we have $a^*_t = a^0_t$, $x^*_{t+1} = x_t$, and the inequality (3.5) becomes an equality.

3.2. Paid-to-Play MDPs: Bounding the Variance by the Mean

We can now state our main theorem and explore some of its immediate consequences.

Theorem 3.2 (Variance Bound)**.** *Let $(\mathcal{X}, \mathcal{A}, f, r, n)$ be a Markov decision problem and let*

$\pi_n^* \in \Pi(n)$ *be a Markov deterministic policy such that*

$$\mathbb{E}[R_n(\pi_n^*)] = \sup_{\pi \in \Pi(n)} \mathbb{E}[R_n(\pi)].$$

If the Markov decision problem has the paid-to-play property, then

$$\mathrm{Var}[R_n(\pi_n^*)] \leq K\, \mathbb{E}[R_n(\pi_n^*)], \tag{3.6}$$

where K is the uniform bound on the one-period reward function.

Theorem 3.2 provides us with an immediate measure of the dispersion of the optimal total reward, $R_n(\pi_n^*)$. Specifically, it tells us that we have a bound on coefficient of variation of the optimal total reward:

$$\mathrm{CoeffVar}[R_n(\pi_n^*)] = \frac{(\mathrm{Var}[R_n(\pi_n^*)])^{1/2}}{\mathbb{E}[R_n(\pi_n^*)]} \leq \left(\frac{K}{\mathbb{E}[R_n(\pi_n^*)]} \right)^{1/2}.$$

Here, K bounds the *one-period reward* and $\mathbb{E}[R_n(\pi_n^*)]$ is the *multi-period optimal expected reward* which typically goes to infinity as $n \to \infty$. Consequently, for the typical paid-to-play MDP, the coefficient of variation goes to zero as $n \to \infty$.

The variance bound (3.6) and Chebyshev's inequality also provide estimates of concentration for the distribution of the optimal total reward. Specifically, for any $\epsilon > 0$, Chebyshev's inequality tells us that

$$\mathbb{P}\left(|R_n(\pi_n^*) - \mathbb{E}[R_n(\pi_n^*)]| > \epsilon \right) \leq \epsilon^{-2}\, K\, \mathbb{E}[R_n(\pi_n^*)],$$

so if we take where $\alpha > 1$ and set $\epsilon = \alpha \left\{ K\, \mathbb{E}[R_n(\pi_n^*)] \right\}^{1/2}$, then we have

$$\mathbb{P}\left(|R_n(\pi_n^*) - \mathbb{E}[R_n(\pi_n^*)]| > \alpha \left\{ K\, \mathbb{E}[R_n(\pi_n^*)] \right\}^{1/2} \right) \leq \alpha^{-2}.$$

In the typical case when $\mathbb{E}[R_n(\pi_n^*)] \to \infty$ as $n \to \infty$, the Chebyshev bound gives us a weak

law of large numbers, which is worth setting out as a corollary.

Corollary 3.3 (Weak Law for Optimal Total Rewards with Large Horizon)**.** *In any paid-to-play Markov decision problem where* $\mathbb{E}[R_n(\pi_n^*)] \to \infty$ *as* $n \to \infty$*, one has*

$$\frac{R_n(\pi_n^*)}{\mathbb{E}[R_n(\pi_n^*)]} \xrightarrow{p} 1 \quad \text{as } n \to \infty.$$

This corollary is good news for the variability-adverse decision maker. It tells us that in the typical case, where $\mathbb{E}[R_n(\pi_n^*)] \to \infty$ as $n \to \infty$, the reward that is realized by the optimal strategy will (with increasingly high probability) behave like its mean. In particular, the corollary provides the decision maker with a *ex-ante* justification for viewing the expected reward as a credible MDP objective function. In a paid-to-play MDP, what one gets is probably close to what one expects.

3.3. A Martingale Proof

The proof of Theorem 3.2 begins with the easy observation that the Bellman equation (3.4) leads to a general martingale. We then find that this martingale carries all the information that is needed to bound the variance of the optimal total reward, once we check that the paid-to-play property gives us useful control of the martingale differences.

Lemma 3.4 (Bellman Martingale)**.** *For* $0 \le t \le n$*, the process defined by*

$$M_t = R_t(\pi_n^*) + \mathbb{E}\left[v_{t+1}(X_{t+1}, Y_{t+1}) \,|\, \mathcal{F}_t\right]$$

is a martingale with respect to the natural filtration $\mathcal{F}_t = \sigma\{X_1, Y_1, Y_2, \ldots, Y_t\}$*.*

Proof of Lemma 3.4. We first note that M_t is $\mathcal{F}_t$ measurable and bounded. We then observe that

$$v_{t+1}(X_{t+1}, Y_{t+1}) = \mathbb{E}[R_n(\pi_n^*) - R_t(\pi_n^*) \,|\, \mathcal{F}_{t+1}].$$

Since $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, an application of the tower property gives us

$$\mathbb{E}\left[v_{t+1}(X_{t+1}, Y_{t+1}) \,|\, \mathcal{F}_t\right] = \mathbb{E}[R_n(\pi_n^*) - R_t(\pi_n^*) \,|\, \mathcal{F}_t],$$

and since $R_t(\pi_n^*)$ is $\mathcal{F}_t$-measurable, we then obtain

$$M_t = \mathbb{E}[R_n(\pi_n^*) \,|\, \mathcal{F}_t],$$

which is clearly a martingale. □

For $t = 0$ and $t = n$, we have the initial and terminal values of M_t:

$$M_0 = \mathbb{E}[v_1(X_1, Y_1) \,|\, \mathcal{F}_0] = \mathbb{E}[R_n(\pi_n^*)] \qquad \text{and} \qquad M_n = R_n(\pi_n^*).$$

Also, for each $1 \leq t \leq n$, we have the martingale difference sequence

$$\begin{aligned} d_t &= M_t - M_{t-1} \\ &= r_t(X_t, Y_t, A_t^*) + \mathbb{E}\left[v_{t+1}(X_{t+1}, Y_{t+1}) \,|\, \mathcal{F}_t\right] - \mathbb{E}\left[v_t(X_t, Y_t) \,|\, \mathcal{F}_{t-1}\right]. \end{aligned} \tag{3.7}$$

By telescoping sums and orthogonality of the martingale differences, we also have

$$M_n - M_0 = \sum_{t=1}^{n} d_t \qquad \text{and} \qquad \text{Var}\left[M_n\right] = \mathbb{E}\left[\sum_{t=1}^{n} d_t^2\right],$$

where $M_n = R_n(\pi_n^*)$ and $M_0 = \mathbb{E}[R_n(\pi_n^*)]$.

Now we use Condition (b) of the paid-to-play property to rewrite the martingale difference d_t more conveniently. By adding and subtracting $\mathbb{E}\left[v_{t+1}(X_t, Y_{t+1}) \,|\, \mathcal{F}_{t-1}\right]$ on the right-hand side of (3.7), we obtain

$$d_t = A_t + B_t$$

where

$$A_t = \mathbb{E}\left[v_{t+1}(X_t, Y_{t+1}) \,|\, \mathcal{F}_{t-1}\right] - \mathbb{E}\left[v_t(X_t, Y_t) \,|\, \mathcal{F}_{t-1}\right]$$

and

$$B_t = r_t(X_t, Y_t, A_t^*) + \mathbb{E}\left[v_{t+1}(X_{t+1}^*, Y_{t+1}) \,|\, \mathcal{F}_t\right] - \mathbb{E}\left[v_{t+1}(X_t, Y_{t+1}) \,|\, \mathcal{F}_{t-1}\right].$$

The deterministic state transition function (3.2) implies that X_t is $\mathcal{F}_{t-1}$-measurable, so, by the independence of the random variables $\{Y_t : 1 \leq t \leq n\}$, we have

$$\mathbb{E}\left[v_{t+1}(X_t, Y_{t+1}) \,|\, \mathcal{F}_{t-1}\right] = \mathbb{E}\left[v_{t+1}(X_t, Y_{t+1}) \,|\, \mathcal{F}_t\right],$$

and we rewrite B_t more nicely as

$$B_t = r_t(X_t, Y_t, A_t^*) + \mathbb{E}\left[v_{t+1}(X_{t+1}^*, Y_{t+1}) \,|\, \mathcal{F}_t\right] - \mathbb{E}\left[v_{t+1}(X_t, Y_{t+1}) \,|\, \mathcal{F}_t\right]. \tag{3.8}$$

We also see that A_t is $\mathcal{F}_{t-1}$-measurable, so

$$\mathbb{E}[d_t^2 \,|\, \mathcal{F}_{t-1}] = \mathbb{E}[B_t^2 \,|\, \mathcal{F}_{t-1}] + 2A_t\mathbb{E}[B_t \,|\, \mathcal{F}_{t-1}] + A_t^2 \tag{3.9}$$

and, since $0 = \mathbb{E}[d_t \,|\, \mathcal{F}_{t-1}] = A_t + \mathbb{E}[B_t \,|\, \mathcal{F}_{t-1}]$, we obtain $\mathbb{E}[B_t \,|\, \mathcal{F}_{t-1}] = -A_t$. Thus, from (3.9) we have

$$\mathbb{E}[d_t^2 \,|\, \mathcal{F}_{t-1}] = \mathbb{E}[B_t^2 \,|\, \mathcal{F}_{t-1}] - A_t^2.$$

For each state realization $(X_t, Y_t) = (x_t, y_t)$, the optimality of action a_t^* and Condition (c) of the paid-to-play property together imply that

$$0 \leq r_t(x_t, y_t, a_t^*) + \mathbb{E}\left[v_{t+1}(x_{t+1}^*, Y_{t+1})\right] - \mathbb{E}\left[v_{t+1}(x_t, Y_{t+1})\right] \leq r_t(x_t, y_t, a_t^*),$$

so, by recalling the representation (3.8), we have

$$0 \leq B_t \leq r_t(X_t, Y_t, A_t^*).$$

The uniform boundedness of the rewards then implies that

$$B_t^2 \leq r_t(X_t, Y_t, A_t^*)^2 \leq K\, r_t(X_t, Y_t, A_t^*),$$

and taking conditional expectations, we obtain

$$\mathbb{E}[B_t^2 \,|\, \mathcal{F}_{t-1}] \leq K\, \mathbb{E}[r_t(X_t, Y_t, A_t^*) \,|\, \mathcal{F}_{t-1}].$$

Finally, by taking total expectations and summing, we have

$$\mathrm{Var}[R_n(\pi_n^*)] \leq K\, \mathbb{E}\left[\sum_{t=1}^{n} r_t(X_t, Y_t, A_t^*)\right] = K\, \mathbb{E}[R_n(\pi_n^*)], \tag{3.10}$$

as desired.

Remark 3.5. The paid-to-play property and the decomposition $d_t = A_t + B_t$ combine nicely to imply that the martingale M_t, $0 \leq t \leq n$, has bounded differences. In particular, we have $|d_t| \leq K$. To see this, first notice that $0 \leq B_t \leq K$, where the lower bound follows from the optimality of action A_t^* and the upper bound follows from conditions (a)–(c) of the paid-to-play property. At the same time, the representation $\mathbb{E}[B_t \,|\, \mathcal{F}_{t-1}] = -A_t$ gives us $-K \leq A_t \leq 0$, so indeed we have the uniform bound $|d_t| = |A_t + B_t| \leq K$.

Remark 3.6. The uniform bound on the reward function r_t, $1 \leq t \leq n$, can be relaxed with a much milder requirement on the second moment of r_t. In fact, when taking total expectations in (3.10), we see that Theorem 3.2 still holds if there is a constant $K < \infty$ such that

$$\mathbb{E}[r_t^2(X_t, Y_t, A_t^*)] \leq K\mathbb{E}[r_t(X_t, Y_t, A_t^*)] \quad \text{uniformly in } t.$$

This condition holds rather widely; in particular it holds for rewards with exponential tails. Naturally, the bounded difference property of the martingale M_t in Remark 3.5, would no longer hold in this case.

3.4. A Sufficient Condition for Paid-to-Play

If the value function $v_t(x, y)$ defined by the Bellman equation is monotone non-decreasing in x for each given y and for all $1 \leq t \leq n$, and if the state variable x_t is a monotone non-increasing function of time, $1 \leq t \leq n$, then the key Condition (c) of the paid-to-play property is satisfied. This gives us a simple criterion for a MDP to have the paid-to-play property, and the situation is common enough to justify summarization as a proposition.

Proposition 3.7 (Sufficient Conditions)**.** *A Markov decision problem* $(\mathcal{X}, \mathcal{A}, f, r, n)$ *satisfies Condition* (c) *of the paid-to-play property if:*

(i) *the fist component* $\mathcal{X}_1$ *of the state space* $(\mathcal{X}_1, \mathcal{X}_2)$ *is a partially ordered set;*

(ii) *for any* x_t*, the optimal transition* $x^*_{t+1} = f(t, x_t, y_t, a^*_t)$ *satisfies* $x^*_{t+1} \leq x_t$*;*

(iii) *the value function* $x \mapsto v_{t+1}(x, y)$ *is non-decreasing in* x *for each* $y \in \mathcal{X}_2$*.*

Remark 3.8. As a small variation, one should also note that the key Condition (c) is also satisfied whenever we have $x_t \leq x^*_{t+1}$, and the map $x \mapsto v_{t+1}(x, y)$ is non-increasing in x for each $y \in \mathcal{X}_2$.

3.5. Three Examples

Markov decision problems with the paid-to-play property are remarkably common. Examples occur in operations research, operations management, financial engineering, and combinatorial optimization. Here, we note that the unimodal subsequence selection problem in Chapter 2 can be cast as a paid-to-play Markov decision problem, and we focus on three additional examples. These should be sufficiently general to suggest many further examples.

3.5.1. Dynamic and Stochastic Knapsack Problem

The knapsack problem is one of the most studied problems in operations research (c.f. Martello and Toth, 1990; Kellerer et al., 2004). Its theory is rich, and its persistent real-

world appearances support both deterministic and stochastic interpretations.

Here, we will focus just on the paper of Papastavrou et al. (1996), which considers a knapsack of capacity $0 < c < \infty$ and items that arrive over a discrete time horizon with n periods. For each time period $1 \leq t \leq n$, the probability of an arrival in period t is assumed to be a constant $p > 0$, and associated with each arriving item there is a pair (W, Z) of random variables, where W is viewed as the size of the arriving item "to be packed" and Z is viewed as the reward that one earns if the currently presented item is selected for placement in the knapsack.

The sequence of size-reward pairs (W_t, Z_t), $1 \leq t \leq n$ is assumed to be independent with a common distribution $F(w, z) = \mathbb{P}(W \leq w, Z \leq z)$ with support in $\mathbb{R} \times [0, K] \subset \mathbb{R}^2$, where $K < \infty$. An arriving item can be accepted only if its size is smaller than or equal to the remaining capacity of the knapsack, and the goal is to determine a strategy that maximizes the expected reward that is accumulated by the end of the time horizon.

To derive the Bellman equation for this problem, we first suppose that at time t we have remaining capacity equal to x. With probability $1 - p$, no new arrival occurs, and the remaining level of capacity x does not change. In this case, one is left with the expected reward over the remaining time that is equal to $v_{t+1}(x)$. On the other hand, with probability p, an arrival occurs and the size-reward pair (w, z) becomes known to the decision maker. With probability $1 - F(x, K)$, the size w exceeds the remaining capacity, in which case the arriving item cannot be accepted, and one is again left with the expected reward to-go, $v_{t+1}(x)$. Finally, if $w \leq x$, then it is feasible to accept the arriving item, and one chooses the action that yields the largest expected reward-to-go. If we do not accept the new item we have $v_{t+1}(x)$, but if we accept the new item then we have $z + v_{t+1}(x - w)$.

Assembling these observations, we see that for each time $1 \leq t \leq n$ and each level of

remaining capacity $x \in [0, c]$ the Bellman equation is given by

$$v_t(x) = (1-p)v_{t+1}(x) + p(1 - F(x, K))v_{t+1}(x) \\ + p \int_{[0,x]\times[0,K]} \max\{v_{t+1}(x), z + v_{t+1}(x-w)\} dF(w, z), \tag{3.11}$$

together with the boundary conditions

$$v_t(0) = 0 \quad \text{for} \quad 1 \le t \le n \quad \text{and } v_{n+1}(x) = 0 \quad \text{for } x \in [0, c].$$

For this MDP we have non-negative, uniformly bounded rewards, and the remaining capacity X_t is a non-increasing function of t under any feasible policy. Also, at any time, the decision maker can refuse to accept the offered item, and this leaves the capacity unchanged and yields zero reward. Finally, according to Lemma 1 of Papastavrou et al. (1996), the value function $v_t(x)$ is non-decreasing in x, so all of the conditions of Proposition 3.7 are met. Hence the knapsack problem of Papastavrou et al. (1996) is indeed a paid-to-play MDP.

In just the same way, one can verify that knapsack problems studied by Coffman et al. (1987), Bruss and Robertson (1991) and Rhee and Talagrand (1991) are all paid-to-play MDPs. In any knapsack problem one always has the option of a do-nothing action, and the paid-to-play property is then easily checked from the monotonicity of the value function $x \mapsto v_t(x)$, $1 \le t \le n$.

One should also note that the MDPs of capacity-control revenue management (c.f. Talluri and van Ryzin, 2004, Section 2.5.1) share much of the structure of the classical knapsack problem. In capacity control problems the initial capacity is discrete and "item arrivals" are now replaced with customer arrivals. Each newly arriving customer offers a price $Z = z$ for one unit of capacity. The decision maker needs to decide whether to sell at price z, or to reject the offer and wait for the next arriving customer.

Here, one can derive a Bellman equation that is quite close to (3.11). Rejection of the offer corresponds to the required do-nothing action, and the monotonicity of the value function $x \mapsto v_t(x)$ is also immediate from the problem definition. So, just as before, one checks that the capacity-control revenue problem is a paid-to-play MDP.

3.5.2. Investment Problems with Stochastic Opportunities

Derman et al. (1975) and Prastacos (1983) study a sequential investment problem with initial capital of c. At each time $1 \leq t \leq n$, an investment opportunity arises independently with probability $0 < p \leq 1$, and the investor gets to see its quality $Y_t = y$. The investor then decides the amount, a, that is to be invested in the opportunity, and this generates a return, $r(y, a)$, that is a deterministic, non-negative, non-decreasing and bounded function of the pair (y, a), such that $r(y, 0) = 0$ for all y.

To derive the Bellman equation of this problem, suppose that at time t the investor capital x on hand. With probability $1 - p$ no investment opportunity arises, no capital gets invested, and the investor is left with the expected return over periods $t + 1$ to n, $v_{t+1}(x)$. With probability p, however, an investment opportunity arises and the investor sees its quality $Y_t = y$. He then chooses the investment amount $a \leq x$ that maximizes the return function $g(a) = r(y, a) + v_{t+1}(x - a)$. Thus, for each $1 \leq t \leq n$ the investor's Bellman equation is given by

$$v_t(x) = (1 - p)v_{t+1}(x) + p \int \max_{0 \leq a \leq x} \{r(y, a) + v_{t+1}(x - a)\} \, dF(y), \tag{3.12}$$

together with the boundary conditions

$$v_t(0) = 0 \text{ for all } 1 \leq t \leq n \qquad \text{and} \qquad v_{n+1}(x) = 0 \text{ for all } x \in [0, c].$$

We now note that Condition (a) of paid-to-play MDPs is satisfied since the return function is non-negative, time independent, and bounded. The investor always has the possibility of investing zero capital in new opportunities. This yields zero return and does not change the level of remaining capital. Thus, Condition (b) is also met. Finally, one can check that the

map $x \mapsto v_t(x)$ is non-decreasing in x for all $1 \leq t \leq n$ (see also Prastacos, 1983, Theorem 2.1) and that $x^*_{t+1} \leq x_t$ for each t. Hence, by appealing to Proposition 3.7, we find that Condition (c) is also met, and therefore the investment problem is a paid-to-play MDP.

Extensions that retain the paid-to-play property include known and time-dependent probabilities $\{p_t, 1 \leq t \leq n\}$, as well as known and time-dependent quality distributions $\{F_t, 1 \leq t \leq n\}$.

3.5.3. Network Capacity Control and Stochastic Depletion Problems with Deterministic Transitions

In the basic version of the problem (c.f. Talluri and van Ryzin, 2004, Section 3.2) a network has ℓ resources and a firm sells m products. Each product is a bundle of the ℓ resources sold at a given price. For each resource $1 \leq i \leq \ell$ and each product $1 \leq j \leq m$, we let $c_{ij} = 1$ if product j uses resource i, and $c_{ij} = 0$ otherwise. This gives us an $\ell \times m$ incidence matrix $\boldsymbol{C} = [c_{ij}]$, where $\boldsymbol{c}_j$ is the jth column vector of $\boldsymbol{C}$, and it includes all of the resources used by product j.

At each time $1 \leq t \leq n$, a decision maker is sequentially presented with an arriving customer who offers nonzero prices for subsets of the m products. More formally, we let $\boldsymbol{Y}_t = (Y_{1,t}, Y_{2,t}, \ldots, Y_{m,t})$ be the demand vector for period t in which $Y_{j,t} = y_j > 0$ indicates a request for product j at price y_j. The sequence $\{\boldsymbol{Y}_t : 1 \leq t \leq n\}$ is assumed to be independent across time, with known joint probability distribution F_t.

To derive the Bellman equation for this problem suppose that, at time t, the state of the network is described by a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_\ell)^T$ of resource capacities. A decision maker then sees a vector of offered prices $\boldsymbol{Y}_t = \boldsymbol{y} = (y_1, y_2, \ldots, y_m)$ and, for each $y_j > 0$, he needs to decide whether to sell product j at price y_j. Thus, the decision maker chooses an allocation vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)^T$ that maximizes the sum of the one-period revenues, $\boldsymbol{ya}$, plus the expected revenues to-go, $v_{t+1}(\boldsymbol{x} - \boldsymbol{Ca})$.

We let $\mathcal{A}_t(\boldsymbol{x}, \boldsymbol{y}) = \{\boldsymbol{a} \in \{0,1\}^m : \boldsymbol{C}\boldsymbol{a} \leq \boldsymbol{x}\}$ be the set of product bundles that are available for sale at time t, and we write the Bellman equation as

$$v_t(\boldsymbol{x}) = \mathbb{E}\left[\max_{\boldsymbol{a} \in \mathcal{A}_t(\boldsymbol{x},\boldsymbol{y})} \{\boldsymbol{y}\boldsymbol{a} + v_{t+1}(\boldsymbol{x} - \boldsymbol{C}\boldsymbol{a})\}\right]. \tag{3.13}$$

As usual, the backwards induction in (3.13) begins by setting $v_{n+1}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$.

Given the real-word motivation of the problem we can assume that prices are non-negative and bounded so that Condition (a) of the paid-to-play property is met. The allocation vector $\boldsymbol{a} = (0, 0, \ldots, 0)$ is a feasible choice for every time $1 \leq t \leq n$ and for any state $(\boldsymbol{x}, \boldsymbol{y})$, and it yields zero revenues, so Condition (b) is also met. Finally, the use of resources over time implies that the state-of-the-network process, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, is non-increasing, and one can check that the value function (3.13) is non-decreasing in $\boldsymbol{x}$. Proposition 3.7 then gives us that Condition (c) is also met, and we can conclude that the network capacity control problem is a paid-to-play MDP.

An alternative way of considering this setup is to see it as a special case of a stochastic depletion problem with deterministic transitions (c.f. Chan and Farias, 2009). In fact, any problem for which the choice of an optimal action generates a deterministic depletion of some system resources is a paid-to-play MDP. Another example that falls into this more general framework is the sequential selection of random vectors under a sum constraint, studied by Stanke (2004).

3.6. Connections with Related Literature and Open Problems

Related Literature

The theory of MDPs has considered variance criteria that aim to incorporate the attitude of the decision maker toward risk or variability. This stream of research has evolved by studying non-standard objectives (different from the total expected reward in (3.3)) that change the nature of the optimization problem. In particular, scholars have been interested

in variance-penalized MDPs as well as in expected reward maximization, subject to variance constraints. See White (1988) and references therein as well as Filar et al. (1989), Baykal-Gürsoy and Ross (1992), Sobel (1994), and Puterman (1994).

Our focus in this paper is different. We take the total expected reward criterion as given, and we study the variance generated under the optimal Markov deterministic policy for a particular class of finite-horizon MDPs. The work of Sobel (1982) and Feinberg and Fei (2009) similarly considers the variance of the optimal total reward in infinite-horizon, discounted MDPs.

The literature on variance bounds for functions of independent random variables also has bearing on our work. In particular, the inequality established by Efron and Stein (1981) and improved by Steele (1986) (see also Lugosi, 2009, Section 4) provides an *a priori* upper bound for the variance of $g(Y_1, \ldots, Y_n)$, where $g : \mathbb{R}^n \to \mathbb{R}$ is a measurable function, and $Y_1, \ldots, Y_n$ are arbitrary independent random variables. However, this bound is difficult to use in a sequential setting, as it requires that we estimate what happens when each random variable is replaced (one at the time) by an independent copy of itself, or is left out. In the context of sequential problems, this estimation becomes difficult as each replacement may change subsequent decisions and subsequent one-period rewards in a way that is difficult to control. Our approach, based on the Bellman martingale in Lemma 3.4, is much more natural in a sequential setting, and it has the additional benefit of leading to quantities that are easier to estimate.

Open Problems

The fact that optimal Markov deterministic policies for finite-horizon MDPs are usually non-stationary makes the distributional analysis of $R_n(\pi_n^*)$ challenging. In this paper, we have isolated a substantial class of MDPs that might be suitable for further research in at least two directions.

The first set of work would complement the variance upper bound in Theorem 3.2 with

some useful lower bound, possibly of the same order. A general result seems difficult, but one can resort to specific analysis of each decision problem that might provide the desired lower bound.

The second direction seeks to understand the relation between finite-horizon and infinite-horizon discounted, or poissonized, MDPs. The question that arises in this context is whether one can obtain useful information on the former by studying the latter two. In fact, infinite-horizon discounted and poissonized problems are often characterized by stationary optimal policies that are easier to analyze. One usually can go back and forth between different formulations to obtain useful information about the first moment of $R_n(\pi_n^*)$, but it is unclear at the moment how this technique can be used to extract useful information about moments other than the first, or for distributional results.

CHAPTER 4 : Optimal Sequential Selection of Alternating Subsequences

Given a finite (or infinite) sequence $\boldsymbol{x} = \{x_1, x_2, ..., x_n, \ldots\}$ of real numbers, we say that a subsequence $x_{i_1}, x_{i_2}, \ldots, x_{i_k}, \ldots$ with $1 \leq i_1 < i_2 < \ldots < i_k < \cdots$ is *alternating* if we have $x_{i_1} < x_{i_2} > x_{i_3} < x_{i_4} \cdots$. When $\boldsymbol{x}$ is an element of the set of permutations $\mathcal{S}_n$ of the integers $\{1, 2, \ldots, n\}$, the study of the set of alternating permutations goes back to Euler (c.f Stanley, 2010).

Here, we are mainly concerned with the length $a(\boldsymbol{x})$ of the longest alternating subsequence of $\boldsymbol{x}$. This function has been more recently studied by Widom (2006), Pemantle (c.f. Stanley, 2007, p. 568) and Stanley (2008). In particular, they consider the situation in which $\boldsymbol{x}$ is chosen at random from $\mathcal{S}_n$. By exploiting explicit formulas for generating functions and delicate applications of the saddle point method, they were able to obtain exact formulas for the first two moments and to prove a central limit theorem. Specifically, if $\boldsymbol{x}$ is chosen according to the uniform distribution on the set of permutations $\mathcal{S}_n$ and if $A_n := a(\boldsymbol{x})$ denotes the length of the longest alternating subsequence of $\boldsymbol{x}$, then for $n \geq 4$ one has

$$\mathbb{E}[A_n] = \frac{2n}{3} + \frac{1}{6} \quad \text{and} \quad \mathrm{Var}[A_n] = \frac{8n}{45} - \frac{13}{180}.$$

More recently, Houdré and Restrepo (2010) used purely probabilistic means to obtain a simpler proof of this result and the corresponding central limit theorem. Moreover, the methods of Houdré and Restrepo also apply to models of random words that are more refined than simple random selection from a set of permutations.

Here, we study the problem of making *on-line selection* of an alternating subsequence. That is, we now regard the sequence $x_1, x_2, ...$ as being presented to us sequentially, and, at the

This chapter is written under the supervision of Prof. J. Michael Steele. The results presented here concerning the behavior of the expected number of optimal alternating selections (finite and infinite horizon) are also in the joint paper Arlotto, Chen, Shepp, and Steele (2011), published in the *Journal of Applied Probability*.

time i when x_i is presented, we must choose to include x_i as a term of our subsequence — or we must reject x_i as a member of the subsequence.

We will consider the sequence to be given by independent random variables $X_1, X_2, \ldots$ that have a common continuous distribution F, and, since we are only concerned with order properties, one can without loss of generality take the distribution to be uniform on $[0, 1]$. We now need to be more explicit about the set Π of feasible strategies for on-line selection. At time i, when presented with X_i we must decide to select X_i based on its value, the value of earlier members of the sequence, and the actions we have taken in the past. All of this information can be captured by saying that τ_k, the index of the k'th selection, must be a stopping time with respect to the increasing sequence of σ-fields, $\mathcal{F}_i = \sigma\{X_1, X_2, \ldots, X_i\}$, $i = 1, 2, \ldots$. Given any feasible policy $\pi \in \Pi$, the random variable of most interest here is $A_n^o(\pi)$, the number of selections made by the policy π up to and including time n. In other words, $A_n^o(\pi)$ is equal to the largest k for which there are stopping times $1 \leq \tau_1 < \tau_2 < \cdots < \tau_k \leq n$ such that $\{X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k}\}$ is an alternating sequence.

Theorem 4.1 (Asymptotic Selection Rate for Large Samples)**.** *For each $n = 1, 2, ...$, there is a policy $\pi_n^* \in \Pi$ such that*

$$\mathbb{E}[A_n^o(\pi_n^*)] = \sup_{\pi \in \Pi} \mathbb{E}[A_n^o(\pi)],$$

and for such an optimal policy one has for all $n \geq 1$ that

$$(2 - \sqrt{2})n \leq \mathbb{E}[A_n^o(\pi_n^*)] \leq (2 - \sqrt{2})n + C,$$

where C is a constant with $C < 11 - 4\sqrt{2} \sim 5.343$. In particular, one has

$$\mathbb{E}[A_n^o(\pi_n^*)] \sim (2 - \sqrt{2})n \quad \text{as } n \to \infty.$$

The proof of this result exploits the analysis of a closely related selection problem in which

one considers a sample of size N where N is geometrically distributed with parameter $0 < \rho < 1$ (so one has $\mathbb{P}(N = k) = \rho^{k-1}(1-\rho)$, $k = 1, 2, 3, \ldots$.) Here, we also assume that N is independent of the sequence $X_1, X_2, \ldots$.

Parallel to our first theorem, we consider the number $A_N^o(\pi)$ of selections made by a feasible policy π up to and including the random time N. The geometric smoothing provided by N gives us a useful "shift symmetry" that is missing in the fixed n problem, and the analysis of a geometric sample turns out to be far more tractable. In particular, one can determine the *exact* expected length of the sequence selected by an optimal policy.

Theorem 4.2 (Expected Selection Size in Geometric Samples)**.** *For each $0 < \rho < 1$, there is a $\pi^* \in \Pi$, such that*

$$\mathbb{E}[A_N^o(\pi^*)] = \sup_{\pi \in \Pi} \mathbb{E}[A_N^o(\pi)],$$

and for such an optimal policy one has

$$\mathbb{E}[A_N^o(\pi^*)] = \frac{3 - 2\sqrt{2} - \rho + \rho\sqrt{2}}{\rho(1-\rho)} \sim (2 - \sqrt{2})(1-\rho)^{-1} \quad \text{as } \rho \to 1.$$

These theorems respectively tell us that optimal on-line selection yields subsequences that grow at a linear rate $(2 - \sqrt{2})n \sim 0.585\, n$ or $(2 - \sqrt{2})\mathbb{E}N \sim 0.585\mathbb{E}N$. This is about a 12% discount off the rate $(2/3)n \sim 0.667\, n$ that one would obtain with *a priori* knowledge of the full finite sample $\{X_1, X_2, ..., X_n\}$, and this discount seems quite modest given the great difference in the knowledge that one has.

To build some intuition about these rates, one should also consider the "maximally timid strategy" where one chooses the first observation that falls in [0, 0.5], then one chooses the next observation that falls in [0.5, 1], and then the next that falls in [0, 0.5], and so on. This strategy obviously leads to an asymptotic selection rate of $0.5\, n$. Finally, one should also consider the "purely greedy strategy" where one accepts any new arrival that is feasible given the previous selections. Curiously enough, by a reasonably quick Markov chain calculation one can show that the greedy strategy leads to the same selection rate

$0.5\,n$ that one finds for the "maximally timid strategy."

We begin by proving Theorem 4.2 which will give us an exact formula for the expected number of selections made under the optimal policy for geometric samples. This result will then be used to prove the upper and lower bounds of Theorem 4.1.

4.1. Infinite Horizon Formulation: Mean

We now let S_i denote the value of the last member of the subsequence selected up to and including time i. To keep track of the up-down nature of our selections, we then set $R_i = 0$ if S_i is a local minimum of $\{S_0, S_1, \ldots, S_i\}$ and set $R_i = 1$ if S_i is a local maximum. To initialize our process, we set $S_0 = 1$ and $R_0 = 1$.

Next, we make the class Π of feasible policies more explicit. For each $1 \leq i < \infty$ and for each pair (S_{i-1}, R_{i-1}) a feasible strategy π specifies a set $\Delta_i(S_{i-1}, R_{i-1})$ such that

$$\Delta_i(S_{i-1}, 0) \subseteq [S_{i-1}, 1] \quad \text{and} \quad \Delta_i(S_{i-1}, 1) \subseteq [0, S_{i-1}],$$

and X_i is selected for our subsequence if and only if $X_i \in \Delta_i(S_{i-1}, R_{i-1})$. For each $\pi \in \Pi$, we have the basic relation

$$A_N^o(\pi) = \sum_{i=1}^{N} \mathbb{1}(X_i \in \Delta_i(S_{i-1}, R_{i-1})) = \sum_{i=1}^{\infty} \mathbb{1}(X_i \in \Delta_i(S_{i-1}, R_{i-1}))\mathbb{1}(i \leq N),$$

and by taking expectations on both sides we have

$$\mathbb{E}[A_N^o(\pi)] = \mathbb{E}\left[\sum_{i=1}^{\infty} \rho^{i-1}\mathbb{1}(X_i \in \Delta_i(S_{i-1}, R_{i-1}))\right].$$

We come to this relation by considering random sample sizes with the geometric distribution, but the right side of this identity can also be interpreted as the infinite-horizon discounted expected length of the alternating subsequence selected by π. We are interested in the

policy $\pi^* \in \Pi$ such that

$$\mathbb{E}[A_N^o(\pi^*)] = \sup_{\pi \in \Pi} \mathbb{E}\left[\sum_{i=1}^{\infty} \rho^{i-1} \mathbb{1}(X_i \in \Delta_i(S_{i-1}, R_{i-1}))\right],$$

and from the general theory of Markov decision problems, we know that an optimal policy can be characterized as the solution of an associated Bellman equation.

First Bellman Equation

For any i such that $S_{i-1} = s$ and $R_{i-1} = r$, we let $v(s, r)$ denote the expected number of selections made after time i by an optimal policy. By the lack of memory property of the geometric distribution and by the usual considerations of dynamic programming, one can now check that $v(s, r)$ satisfies Bellman equation:

$$v(s,r) = \begin{cases} \rho s v(s,0) + \int_s^1 \max\{\rho v(s,0), 1 + \rho v(x,1)\}\, dx & \text{if } r = 0 \\ \rho(1-s) v(s,1) + \int_0^s \max\{\rho v(s,1), 1 + \rho v(x,0)\}\, dx & \text{if } r = 1. \end{cases} \tag{4.1}$$

To see why this equation holds, first consider the case when $r = 0$ (so the next selection is to be a local maximum). With probability ρ we get to see another observation X_{i+1} and, with probability s, the value we observe is less than the previously selected value. In this case, we do not have the opportunity to make a selection, and this observation contributes the term $\rho s v(s, 0)$ to our equation.

Next, consider the case when $s < X_{i+1} \leq 1$. Now one can choose to select $X_{i+1} = x$ or not. If we do not select $X_{i+1} = x$, the expected number of subsequent selections is $\rho v(s, 0)$ and, if we do select $X_{i+1} = x$, we increment sequence by 1 and the expected number of subsequence selections that are made by an optimal policy in the future given by $\rho v(x, 1)$. Since X_{i+1} is uniformly distributed in $[s, 1]$ the expected optimal contribution is given by the second term of our Bellman equation (top line). The proof of the second line of the Bellman equation is completely analogous.

Finally, given a solution $v(s,r)$ to the Bellman equation (4.1), we have

$$v(1,1) = \mathbb{E}[A_N^o(\pi^*)],$$

so, now our goal is to determine $v(1,1)$. To do this it will be useful to reorganize the Bellman equation (4.1) in a tidier form. This is possible since the solution $v(s,r)$ of the Bellman equation has a useful symmetry property.

Lemma 4.3 (Reflection Identity)**.** *For all $s \in [0,1]$, the solution $v(s,r)$ of the Bellman equation* (4.1) *satisfies*

$$v(s,0) = v(1-s,1). \tag{4.2}$$

Proof. The Bellman equation (4.1) is a fixed point equation, and by the classical theory of dynamic programming, it can be solved by value iteration (c.f. Bertsekas and Shreve, 1978, Sec. 9.5) . We will prove the identity (4.2) by showing that it holds for the sequence of approximations, so it also holds for the limit.

We first set $v^0(s,r) = 0$ for all $(s,r) \in [0,1] \times \{0,1\}$ and we note that v^0 trivially satisfies the Reflection Identity (4.2). Next, for our induction hypothesis, we assume that we have $v^{n-1}(s,0) = v^{n-1}(1-s,1)$ for all $s \in [0,1]$. The next iterate in the sequence is then given by

$$v^n(s,0) = \rho s v^{n-1}(s,0) + \int_s^1 \max\left\{\rho v^{n-1}(s,0), 1 + \rho v^{n-1}(x,1)\right\}\, dx.$$

By applying our induction hypothesis on v^{n-1}, we then obtain

$$v^n(s,0) = \rho s v^{n-1}(1-s,1) + \int_s^1 \max\left\{\rho v^{n-1}(1-s,1), 1 + \rho v^{n-1}(1-x,0)\right\}\, dx.$$

Now, after changing variables in the integral on the right-hand side, we find

$$\begin{aligned} v^n(s,0) &= \rho s v^{n-1}(1-s,1) + \int_0^{1-s} \max\left\{\rho v^{n-1}(1-s,1), 1 + \rho v^{n-1}(x,0)\right\}\, dx \\ &= v^n(1-s,1), \end{aligned}$$

and this completes the induction step. Now, for all $(s,r) \in [0,1] \times \{0,1\}$, we have $v^n(s,r) \to v(s,r)$ as $n \to \infty$ so taking limits in the last identity completes the proof of the reflection identity. □

A Simpler Equation

Using the reflection identity (4.2) we can put the Bellman equation (4.1) into a more graceful form. Specifically, if we introduce a single variable function $v(y)$ defined by setting

$$v(y) \equiv v(y,0) = v(1-y,1),$$

then substitution into our original equation (4.1) gives us

$$v(y) = \rho y v(y) + \int_y^1 \max\{\rho v(y), 1 + \rho v(1-x)\}\, dx. \tag{4.3}$$

Here we should note that by the definition of $v(y) = v(y,0)$ we have that $v(\cdot)$ is continuous, $v(1) = 0$, and v is non-increasing on $[0,1]$. We will show shortly that v is actually piecewise linear and it is constant on an initial segment of $[0,1]$.

An Alternative Interpretation

The symmetrized equation (4.3) can be used to obtain a new probabilistic interpretation of $v(y)$. To set this up, we first put

$$g(y) = \inf\{x \in [y,1] :\ \rho v(y) \le 1 + \rho v(1-x)\}. \tag{4.4}$$

With this definition, we can rewrite (4.3) a bit more nicely as

$$v(y) = \rho g(y) v(y) + \int_{g(y)}^1 \{1 + \rho v(1-x)\}\, dx. \tag{4.5}$$

Thus, one removes the maximum from the integrand (4.3) at the price of introducing a threshold function g that depends on v.

We now recursively define random variables $\{Y_i : i = 1, 2, \ldots\}$ by setting $Y_0 = 0$ and taking

$$Y_i = \begin{cases} Y_{i-1} & \text{if } X_i < g(Y_{i-1}) \\ 1 - X_i & \text{if } X_i \geq g(Y_{i-1}), \end{cases}$$

and finally introduce a new value function

$$v_0(y) \equiv \mathbb{E}\left[\sum_{i=1}^{\infty} \rho^{i-1} \mathbb{1}(X_i \geq g(Y_{i-1})) \;\middle|\; Y_0 = y \right]. \tag{4.6}$$

The next proposition shows that $v_0(y)$ is actually equal to $v(y)$. As part of the bargain, we obtain a concrete characterization of the threshold function g.

Proposition 4.4 (Structure of the Solution of the Bellman Equation)**.** *We have the following characterizations of g and v_0:*

(i) *There is a unique $\xi_0 \in [0, 1]$ such that*

$$g(y) = \max\{\xi_0, y\} \quad \textit{for all } 0 \leq y \leq 1,$$

and moreover $0 \leq \xi_0 < 1/2$.

(ii) *The function $v_0(\cdot)$ is a solution of the Bellman equation (4.3), so, by uniqueness, we have $v_0(y) = v(y)$ for all $0 \leq y \leq 1$.*

Proof. From the definition of g we see that

$$\rho v(y) \leq 1 + \rho v(1 - y) \quad \Rightarrow \quad g(y) = y. \tag{4.7}$$

Now, for $1/2 \leq y$ we have $1 - y \leq y$, so the monotonicity of v gives us the bound $\rho v(y) \leq 1 + \rho v(1 - y)$; consequently, we have $g(y) = y$ for $y \in [1/2, 1]$.

If the condition (4.7) holds for all $y \in [0, 1/2)$, then $g(y) = y$ for all $y \in [0, 1]$ and we can

take $\xi_0 = 0$. Otherwise there is a $y_0 \in [0, 1/2)$ for which we have

$$1 + \rho v(1 - y_0) < \rho v(y_0).$$

For $\Delta(y) = 1 + \rho v(1 - y) - \rho v(y)$, we then have $\Delta(y_0) < 0$ and $\Delta(1) = 1 + \rho v(0) > 0$, so by continuity we have $S = \{y : \Delta(y) = 0\} \neq \emptyset$. If we now take ξ_0 to be the infimum of S, then $\xi_0 \in [y_0, 1/2) \subset [0, 1/2)$ and $\rho v(\xi_0) = 1 + \rho v(1 - \xi_0)$. The definition of g now tells us that $g(y) = \xi_0$ for $y \leq \xi_0$ and $g(y) = y$ for $\xi_0 \leq y$. This completes the proof of the first part of the proposition.

Finally, to check that v_0 solves the equation (4.6), we just condition on the value of X_1 and calculate the expectation of the sum. When we take the total expectation, we get the right side of (4.5). □

Characterization of the Critical Value

Now that we know that the threshold function g for the solution of Bellman equation (4.3) has the form $g(y) = \max\{\xi_0, y\}$ for some $\xi_0 \in [0, 1/2)$, the main problem is to find ξ_0. The natural plan is to fix $\xi \in [0, 1/2]$ and to consider a general selection function of the form $f(y) = \max\{\xi, y\} \equiv (\xi \vee y)$. We then want to calculate the associated value function and to optimize over ξ.

The associated value function is given by

$$V(y, \xi, \rho) = \mathbb{E}\left[\sum_{i=1}^{\infty} \rho^{i-1} \mathbb{1}(X_i \geq \max\{\xi, Y_{i-1}\}) \;\middle|\; Y_0 = y\right], \tag{4.8}$$

and Proposition 4.4 then tells us that

$$\max_{\xi \in [0,1/2]} V(y, \xi, \rho) = v(y) \quad \text{for all } y \in [0, 1].$$

If we abbreviate $V(y, \xi, \rho)$ by setting $V(y) \equiv V(y, \xi, \rho)$, then by conditioning on X_1 in

equation (4.8) we see that $V(y)$ satisfies the integral equation

$$\begin{aligned} V(y) &= (\xi \vee y)\rho V(y) + \int_{\xi \vee y}^{1} \{1 + \rho V(1-x)\}\, dx \\ &= (\xi \vee y)\rho V(y) + \int_{0}^{1-(\xi \vee y)} \{1 + \rho V(x)\}\, dx. \end{aligned} \tag{4.9}$$

This equation has several attractive features. In particular, if we set $y = 1$ then from $0 < \rho < 1$ we see $V(1) = 0$. Also, by writing

$$V(y) = \frac{1}{1 - \rho(\xi \vee y)} \int_{0}^{1-(\xi \vee y)} \{1 + \rho V(x)\}\, dx,$$

we see that the right side does not change when $y \in [0, \xi]$, so we have

$$V(y) = V(y') \quad \text{for all } 0 \le y,\, y' \le \xi. \tag{4.10}$$

From now on, we will let $V'(\xi)$ denote the right derivative of the integral equation (4.9) evaluated at ξ, and let $V'(1-\xi)$ denote the left derivative of (4.9) evaluated at $1-\xi$. Elsewhere $V'(y)$ simply denotes the derivative of (4.9) evaluated at y.

Lemma 4.5. *The solution of equation* (4.9) *satisfies the following four conditions:*

(i) $V(1-\xi)(1-\rho+\rho\xi) = \xi + \rho\xi V(\xi)$;

(ii) $V'(\xi)(1-\rho\xi) = \rho[V(\xi) - V(1-\xi)] - 1$;

(iii) $V'(1-\xi)(1-\rho+\rho\xi) = \rho[V(1-\xi) - V(\xi)] - 1$;

(iv) $V'(1-\xi)(1-\rho+\rho\xi)^2(1-\rho\xi) = V'(\xi)(1-\rho\xi)^2(1-\rho+\rho\xi) + (1-\rho+\rho\xi)^2 - (1-\rho\xi)^2$.

Proof. Conditions (i)–(iii) are easy to check. Condition (i) is just (4.9) evaluated at $1-\xi$ together with (4.10). Conditions (ii) and (iii) simply follow by evaluating (4.9) at ξ and $1-\xi$ respectively and by differentiating both sides with respect to ξ.

The proof of Condition (iv) requires more work. Consider $y \in (\xi, 1-\xi)$ so that the integral equation (4.9) becomes

$$V(y) = y\rho V(y) + \int_0^{1-y} \{1 + \rho V(x)\}\, dx.$$

Differentiating once we have

$$V'(y)(1-\rho y) = \rho[V(y) - V(1-y)] - 1, \tag{4.11}$$

and differentiating again gives us

$$V''(y)(1-\rho y) - \rho V'(y) = \rho V'(y) + \rho V'(1-y). \tag{4.12}$$

To estimate the value of $V'(1-y)$ we note that $1-y \in (\xi, 1-\xi)$, and we evaluate the integral equation (4.9) at $1-y$. We then differentiate with respect to y to obtain

$$V'(1-y)(1-\rho+\rho y) = \rho[V(1-y) - V(y)] - 1. \tag{4.13}$$

By combining equations (4.11) and (4.13), we then have

$$V'(1-y) = (1-\rho+\rho y)^{-1}(-V'(y)(1-\rho y) - 2),$$

which we can plug into the last addend of (4.12) to obtain

$$V''(y)(1-\rho y)(1-\rho+\rho y) = V'(y)\rho(1-2\rho+3\rho y) - 2\rho. \tag{4.14}$$

By multiplying both sides of (4.14) by $(1-\rho y)$, we obtain the critical identity

$$V''(y)(1-\rho y)^2(1-\rho+\rho y) = V'(y)\rho(1-\rho y)(1-2\rho+3\rho y) - 2\rho(1-\rho y). \tag{4.15}$$

For $h(y) = (1-\rho y)^2(1-\rho+\rho y)$ notice that $h'(y) = -\rho(1-\rho y)(1-2\rho+3\rho y)$, so that we

can rewrite the identity (4.15) as

$$V''(y)h(y) + V'(y)h'(y) - \left[(1-\rho y)^2\right]' = 0.$$

An immediate integration then gives us

$$V'(y)h(y) - (1-\rho y)^2 = C,$$

where C is a constant, and if we take $C = V'(\xi)h(\xi) - (1-\rho\xi)^2$ we find

$$V'(y) = V'(\xi)\frac{h(\xi)}{h(y)} + \frac{(1-\rho y)^2 - (1-\rho\xi)^2}{h(y)} \quad \text{for all } \xi < y < 1-\xi. \tag{4.16}$$

Finally, on setting $y = 1-\xi$ we recover the desired condition (iv). □

Calculation of the Critical Value.

Conditions (i)–(iv) in Lemma 4.5 generate a system of four equations in four unknowns, $V(\xi), V(1-\xi), V'(\xi)$, and $V'(1-\xi)$. By solving this system one finds

$$\begin{aligned} V(\xi) &= \frac{2-2\xi-\rho+2\rho\xi-2\rho\xi^2}{2(1-\rho)(1-\rho\xi)} && (4.17)\\ V(1-\xi) &= \frac{\rho\left(2-4\rho\xi-\rho^2+4\rho^2\xi-2\rho^2\xi^2\right)}{2(1-\rho)(1-\rho\xi)^2(1-\rho+\rho\xi)} \\ V'(\xi) &= \frac{-2+4\rho-4\rho\xi-\rho^2+2\rho^2\xi^2}{2(1-\rho\xi)^2(1-\rho+\rho\xi)} && (4.18)\\ V'(1-\xi) &= \frac{-2+4\rho\xi+\rho^2-4\rho^2\xi+2\rho^2\xi^2}{2(1-\rho\xi)(1-\rho+\rho\xi)^2}. \end{aligned}$$

Finally, by substituting (4.18) into (4.16) we get

$$V'(y) = \frac{-(2-\rho)^2 + 2(1-\rho y)^2}{2(1-\rho+\rho y)(1-\rho y)^2} \quad \text{for all } \xi < y < 1-\xi.$$

Now, given any ξ, we want to compute $V(0,\xi,\rho)$. We first recall that we have $V(1,\xi,\rho) = 0$

and $V(y,\xi,\rho) = V(\xi,\xi,\rho)$ for all $0 \le y \le \xi$. We therefore find that $\frac{\partial}{\partial y}V(y,\xi,\rho) = 0$ on $0 \le y \le \xi$, so on integrating we have

$$V(1,\xi,\rho) - V(0,\xi,\rho) = \int_0^1 V'(y)\,dy = \int_\xi^1 V'(y)\,dy$$

and hence

$$V(0,\xi,\rho) = -\int_\xi^1 V'(y)\,dy.$$

We now optimize this last quantity with respect to ξ. By differentiating both sides with respect to ξ we get

$$\frac{\partial}{\partial \xi}V(0,\xi,\rho) = V'(\xi)$$

and we are interested in the value ξ_0 such that

$$V'(\xi_0) = 0.$$

Our formula (4.18) for $V'(\xi_0)$ tells us that $V'(\xi_0) = 0$ if and only if

$$2(1-\rho\xi_0)^2 = (2-\rho)^2.$$

We therefore find that the unique choice for ξ_0 is given by

$$\xi_0 = \frac{1}{\sqrt{2}} + \frac{1-\sqrt{2}}{\rho}. \tag{4.19}$$

A routine calculation verifies that $V''(\xi_0) < 0$, so we have found our maximum.

When we evaluate $V(\xi_0,\xi_0,\rho)$ using equation (4.17), we find

$$V(\xi_0,\xi_0,\rho) = \frac{3-2\sqrt{2}-\rho+\rho\sqrt{2}}{\rho(1-\rho)},$$

and this gives us the main formula of Theorem 4.2. From this formula it is immediate that

$$\lim_{\rho\uparrow 1}(1-\rho)V(\xi_0,\xi_0,\rho) = 2-\sqrt{2},$$

so the proof of Theorem 4.2 is complete.

4.2. Finite Horizon Formulation: Mean Bounds and Exact Asymptotics

We will use our results for geometric sample sizes to get both lower and upper bounds for the finite sample size selection problem. The lower bound is the easiest. For fixed n, one can use the (now suboptimal) policy from an appropriately chosen geometric sample size problem. The proof of the upper bound is considerably harder, and the method will be described later in this section. Before making these arguments, we need to organize a few structural observations.

Selection Policies and a Bellman Equation for Finite Samples

When the sample size n is deterministic and known, the feasible policies need to take this information into account. In particular, the selection thresholds will no longer be stationary; they will depend on the number of sample elements that remain to be seen.

Just as in Section 4.1, we consider the pairs (S_{i-1}, R_{i-1}), $1 \le i \le n$, where S_{i-1} is the size of the last selection made before time i and R_{i-1} is 0 or 1 accordingly as the last selection was a local minimum or a local maximum. A feasible policy $\pi \in \Pi$ again specifies a set $\Delta_{n-i+1}(S_{i-1}, R_{i-1})$ that depends only on past actions, but now we have dependence on the number of remaining periods, $n-i+1$. For any policy $\pi \in \Pi$, the expected size of the selected sample can then be written as

$$\mathbb{E}[A_n^o(\pi)] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}(X_i \in \Delta_{n-i+1}(S_{i-1}, R_{i-1}))\right]$$

and there is an optimal policy π_n^* for which we have

$$\mathbb{E}[A_n^o(\pi_n^*)] = \sup_{\pi \in \Pi} \mathbb{E}[A_n^o(\pi)].$$

In this case, an optimal policy can be characterized as the solution to a finite sample Bellman equation. Specifically, we set $v_0(s,r) \equiv 0$ for all (s,r) in $[0,1] \times \{0,1\}$, and for $k \geq 1$ we let

$$v_k(s,r) = \begin{cases} sv_{k-1}(s,0) + \int_s^1 \max\{v_{k-1}(s,0), 1 + v_{k-1}(x,1)\}\, dx & \text{if } r = 0 \\ (1-s)v_{k-1}(s,1) + \int_0^s \max\{v_{k-1}(s,1), 1 + v_{k-1}(x,0)\}\, dx & \text{if } r = 1. \end{cases}$$

This equation is justified by the same considerations that were used in the derivation of equation (4.1), and we note that, here, the subscript k denotes the number of periods left to the end of the time horizon.

Symmetry and Simplification

For the finite sample size problem, one loses much of the nice symmetry of the geometric sample size problem. Nevertheless, the solution of the finite sample Bellman equation still has a reflection identity analogous to that given by Lemma 4.3.

Lemma 4.6. *The solution of the finite sample Bellman equation satisfies*

$$v_k(s,0) = v_k(1-s,1) \quad \text{for all } k \geq 1 \text{ and all } s \in [0,1]. \tag{4.20}$$

Proof. Again we use an induction argument, but this time we do not need to take limits of an infinite sequence of approximate solutions. Instead we simply use backward induction and always work with exact solutions.

Since we have $v_1(s,0) = 1 - s$ and $v_1(1-s,1) = 1 - s$, we see that equation (4.20) holds for $k = 1$, so we suppose by induction that $v_{k-1}(s,0) = v_{k-1}(1-s,1)$. One then has

$$v_k(s,0) = sv_{k-1}(s,0) + \int_s^1 \max\{v_{k-1}(s,0), 1 + v_{k-1}(x,1)\}\, dx,$$

so by applying the induction hypothesis on the right-hand side one obtains

$$v_k(s,0) = sv_{k-1}(1-s,1) + \int_s^1 \max\{v_{k-1}(1-s,1), 1 + v_{k-1}(1-x,0)\}\, dx.$$

If we now change variable in this last integral, we get

$$\begin{aligned} v_k(s,0) &= sv_{k-1}(1-s,1) + \int_0^{1-s} \max\{v_{k-1}(1-s,1), 1 + v_{k-1}(x,0)\}\, dx \\ &= v_k(1-s,1), \end{aligned}$$

and this completes the induction step. □

We can now define a new single variable function $v_k(y)$ by setting

$$v_k(y) = v_k(y,0) = v_k(1-y,1) \tag{4.21}$$

and, by substitution into the original finite sample Bellman equation we have

$$v_k(y) = yv_{k-1}(y) + \int_y^1 \max\{v_{k-1}(y), 1 + v_{k-1}(1-x)\}\, dx. \tag{4.22}$$

Here we should also note that $v_k(\cdot)$ is continuous and non-increasing on $[0,1]$ for all $k \geq 1$.

The Threshold Functions

We now define the finite-sample equivalent of the threshold function (4.4) by setting

$$g_k(y) = \inf\{x \in [y,1] :\ v_{k-1}(y) \leq 1 + v_{k-1}(1-x)\}. \tag{4.23}$$

If we then set $Y_0 = 0$ and define Y_i recursively by setting

$$Y_i = \begin{cases} Y_{i-1} & \text{if } X_i < g_{n-i+1}(Y_{i-1}) \\ 1 - X_i & \text{if } X_i \geq g_{n-i+1}(Y_{i-1}), \end{cases} \tag{4.24}$$

then, in complete parallel to the geometric case, we see that the solution of the finite sample Bellman equation (4.22) can be written more probabilistically as

$$v_1(y) = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}(X_i \geq g_{n-i+1}(Y_{i-1})) \;\middle|\; Y_0 = y\right]. \tag{4.25}$$

Finally, from equation (4.21) we have

$$v_1(0) = v_1(0,0) = v_1(1,1) = \mathbb{E}[A_n^o(\pi_n^*)],$$

and this gives us the last piece of structural information that we need.

Proof of the Lower Bound

To prove that

$$(2-\sqrt{2})n \leq \mathbb{E}[A_n^o(\pi_n^*)] \quad \text{for all } n \geq 1$$

we only need to choose a good suboptimal policy. We now fix $\xi \in [0, 1/2]$ and we consider the policy in which X_i is selected if and only if $X_i \geq \max\{\xi, Y_{i-1}\}$. Here, $Y_0 = y$ is in the interval $[0, 1-\xi]$ and the Y_i's are defined recursively by setting

$$Y_i = \begin{cases} Y_{i-1} & \text{if } X_i < \max\{\xi, Y_{i-1}\} \\ 1 - X_i & \text{if } X_i \geq \max\{\xi, Y_{i-1}\}. \end{cases}$$

The sequence $\{Y_i : i = 0, 1, \ldots\}$ is a discrete-time Markov Chain on the state space $[0, 1-\xi]$. For a measurable $C \subseteq [0, 1-\xi]$ we let $|C|$ denote the Lebesgue measure of C, and we write the transition kernel of the process $\{Y_i : i = 0, 1, \ldots\}$ as

$$K(y, C) = \mathbb{1}(y \in C)(\xi \vee y) + |C \cap [0, 1 - (\xi \vee y)]|.$$

It is now easy to check that the process $\{Y_i\}$ has a unique stationary distribution γ, and in fact γ is just the uniform distribution on $[0, 1-\xi]$, (i.e., $\gamma(C) = (1-\xi)^{-1}|C|$ for all

measurable $C \subseteq [0, 1-\xi]$).

For any starting value $Y_0 = y \in [0, 1-\xi]$, the suboptimality of the selection functions $\max\{\xi, Y_{i-1}\}$ gives that

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}(X_i \geq \max\{\xi, Y_{i-1}\}) \;\middle|\; Y_0 = y\right] \leq v_1(y).$$

Since $v_1(y)$ is non-increasing in y, we see that for any starting distribution μ supported on $[0, 1-\xi]$ one has

$$\mathbb{E}_\mu\left[\sum_{i=1}^{n} \mathbb{1}(X_i \geq \max\{\xi, Y_{i-1}\})\right] \leq \mathbb{E}_\mu[v_1(Y_0)] \leq v_1(0) = \mathbb{E}[A_n^o(\pi_n^*)].$$

If one chooses the starting distribution μ to be the stationary distribution γ, then

$$\mathbb{E}_\gamma\left[\sum_{i=1}^{n} \mathbb{1}(X_i \geq \max\{\xi, Y_{i-1}\})\right] = n\,\mathbb{E}_\gamma\left[1 - \max\{\xi, Y_0\})\right] \leq \mathbb{E}[A_n^o(\pi_n^*)], \tag{4.26}$$

and we can compute the first expression explicitly. So, we have

$$\mathbb{E}_\gamma\left[1 - \max\{\xi, Y_0\}\right] = \frac{1}{1-\xi}\int_0^{1-\xi} 1 - \max\{\xi, y\}\, dy = \frac{1-2\xi^2}{2(1-\xi)}.$$

We can maximize this by taking $\xi = 1 - 2^{-1/2}$ (as in (4.19) when $\rho = 1$), and we then obtain

$$\mathbb{E}_\gamma\left[1 - \max\{\xi, Y_0\})\right] = 2 - \sqrt{2}.$$

Together with the inequality (4.26), this completes the proof of our lower bound.

Proof of the Upper Bound

The proof of the upper bound in Theorem 4.1 requires a more sustained argument. Unlike the problem for geometric samples, the value function $v_k(\cdot)$ is no longer constant on an initial segment of $[0, 1]$. Nevertheless, the next proposition tells us that the value function does have a useful uniform boundedness on an initial segment. This is the first of several

structural observations that we will need to obtain our upper bound for $\mathbb{E}[A_n^o(\pi_n^*)]$.

Proposition 4.7 (Value Function Initial Segment Bounds)**.** *For all $0 \leq u < 1/6$, the functions $v_k(\cdot)$ defined by the Bellman recursion* (4.22) *satisfy*

(i) $1 < v_k(u) - v_k(5/6)$*, for all $k \geq 2$;*

(ii) $v_k(u) - v_k(1/6) < 1$*, for all $k \geq 1$.*

Moreover, the threshold functions $g_k(y)$ defined by equation (4.23) *are guaranteed to satisfy $1/6 \leq g_k(y)$ for all $y \in [0,1]$ and all $k \geq 3$.*

Naturally enough, the proof of this proposition depends on inductive arguments that exploit the defining Bellman equation. The first of these arguments gives us some control over the changes of $v_k(u)$ when we change both k and u.

Lemma 4.8 (Restricted Supermodularity)**.** *For $y \in [0, 1/2]$ and $u \in [y, 1-y]$, the functions $\{v_k(\cdot)\}$ defined by the Bellman recursion* (4.22) *satisfy*

$$v_{k-1}(u) - v_{k-1}(1-y) \leq v_k(u) - v_k(1-y) \quad \textit{for all } k \geq 1.$$

Proof. We proceed by induction on k. For $k = 1$, we have $v_0(u) = 0$ for all $u \in [0,1]$. Moreover, $v_1(u) = 1 - u$ and $v_1(1-y) = y$, so we have

$$v_0(u) - v_0(1-y) \leq v_1(u) - v_1(1-y) \quad \text{for all } u \in [y, 1-y].$$

Now, for our backward induction, we can assume more generally that

$$v_{k-1}(u) - v_{k-1}(1-y) \leq v_k(u) - v_k(1-y) \quad \text{for all } u \in [y, 1-y].$$

The Bellman equation (4.22) then gives us

$$v_{k+1}(u) - v_{k+1}(1-y) = uv_k(u) + \int_u^1 \max\{v_k(u), 1 + v_k(1-x)\}\, dx \\ - (1-y)v_k(1-y) - \int_{1-y}^1 \max\{v_k(1-y), 1 + v_k(1-x)\}\, dx,$$

and, since $u \leq 1-y$, we can break up the first integral to obtain

$$v_{k+1}(u) - v_{k+1}(1-y) = uv_k(u) - (1-y)v_k(1-y) + \int_u^{1-y} \max\{v_k(u), 1 + v_k(1-x)\}\, dx \\ + \int_{1-y}^1 \max\{v_k(u), 1 + v_k(1-x)\} - \max\{v_k(1-y), 1 + v_k(1-x)\}\, dx. \quad (4.27)$$

For $x \in [1-y, 1]$, we have $v_k(y) \leq v_k(1-x)$ since $v_k(\cdot)$ is non-increasing on $[0,1]$. Therefore, since $y \leq u \leq 1-y$ we have $v_k(1-y) \leq v_k(u) \leq v_k(y)$ so that for $x \in [1-y, 1]$ we have

$$\max\{v_k(u), 1 + v_k(1-x)\} = \max\{v_k(1-y), 1 + v_k(1-x)\} = 1 + v_k(1-x),$$

and we see that the integral (4.27) equals 0. We now have just the identity

$$v_{k+1}(u) - v_{k+1}(1-y) = uv_k(u) - (1-y)v_k(1-y) + \int_u^{1-y} \max\{v_k(u), 1 + v_k(1-x)\}\, dx$$

or, equivalently,

$$v_{k+1}(u) - v_{k+1}(1-y) = u\left(v_k(u) - v_k(1-y)\right) \\ + \int_u^{1-y} \max\{v_k(u) - v_k(1-y), 1 + v_k(1-x) - v_k(1-y)\}\, dx.$$

Changing variables in this last integral then gives us the convenient identity

$$v_{k+1}(u) - v_{k+1}(1-y) = u\left(v_k(u) - v_k(1-y)\right) \\ + \int_y^{1-u} \max\{v_k(u) - v_k(1-y), 1 + v_k(x) - v_k(1-y)\}\, dx. \quad (4.28)$$

Since $y \leq u$ and $1-u \leq 1-y$, we can now use our induction assumption to obtain

$$\begin{aligned} v_{k+1}(u) - v_{k+1}(1-y) &\geq u\,(v_{k-1}(u) - v_{k-1}(1-y)) \\ &\quad + \int_y^{1-u} \max\left\{v_{k-1}(u) - v_{k-1}(1-y), 1 + v_{k-1}(x) - v_{k-1}(1-y)\right\} dx \\ &= v_k(u) - v_k(1-y), \end{aligned}$$

where the last equality follows from the recursion (4.28). □

We can now complete the proof of the Value Function Bounds in Proposition 4.7.

Proof of Proposition 4.7. We begin by proving (i) by induction on k. For $k = 2$, one iteration of the recursive definition of the Bellman equation (4.22) gives us that $v_2(x) = (3/2)(1-x^2)$, so $v_2(u) - v_2(5/6) = (3/2)(25/36 - u^2) > 1$ since by hypothesis we have $u < 1/6$. We now make the induction assumption

$$1 < v_{k-1}(u) - v_{k-1}(5/6) \quad \text{for } 0 \leq u < 1/6,$$

and observe from the Bellman equation (4.22) that

$$\begin{aligned} v_k(u) - v_k(5/6) &= u v_{k-1}(u) + \int_u^1 \max\{v_{k-1}(u), 1 + v_{k-1}(1-x)\}\, dx \\ &\quad - 5/6\, v_{k-1}(5/6) - \int_{5/6}^1 \max\{v_{k-1}(5/6), 1 + v_{k-1}(1-x)\}\, dx. \end{aligned}$$

Since $u < 5/6$, the monotonicity of $v_{k-1}(\cdot)$ implies $v_{k-1}(5/6) \leq v_{k-1}(u)$. So, for $x \in [5/6, 1]$, we have $\max\{v_{k-1}(5/6), 1 + v_{k-1}(1-x)\} \leq \max\{v_{k-1}(u), 1 + v_{k-1}(1-x)\}$. This gives us the lower bound

$$\begin{aligned} u\,(v_{k-1}(u) - v_{k-1}(5/6)) + \int_u^{5/6} \max\{v_{k-1}(u) - v_{k-1}(5/6), 1 + v_{k-1}(1-x) - v_{k-1}(5/6)\}\, dx \\ \leq v_k(u) - v_k(5/6). \end{aligned}$$

To get a lower bound for the integral of the maximum, we replace the integrand by $v_{k-1}(u) - v_{k-1}(5/6)$ on $[u, 1/6)$ and replace it by $1 + v_{k-1}(1-x) - v_{k-1}(5/6)$ on $[1/6, 5/6]$. Changing variables then gives us

$$\frac{1}{6}\left(v_{k-1}(u) - v_{k-1}(5/6)\right) + \int_{1/6}^{5/6} \{1 + v_{k-1}(x) - v_{k-1}(5/6)\}\, dx \leq v_k(u) - v_k(5/6). \tag{4.29}$$

By our induction hypothesis, the first addend satisfies the bound

$$\frac{1}{6} < \frac{1}{6}\left(v_{k-1}(u) - v_{k-1}(5/6)\right), \tag{4.30}$$

and by Lemma 4.8, the second integral satisfies the bound

$$\int_{1/6}^{5/6} \{1 + v_1(x) - v_1(5/6)\}\, dx \leq \int_{1/6}^{5/6} \{1 + v_{k-1}(x) - v_{k-1}(5/6)\}\, dx.$$

If we now recall that $v_1(x) = 1 - x$ and compute the integral on the left-hand side, we then obtain

$$\frac{32}{36} \leq \int_{1/6}^{5/6} \{1 + v_{k-1}(x) - v_{k-1}(5/6)\}\, dx. \tag{4.31}$$

Finally, adding (4.30) and (4.31) and recalling (4.29) gives us our target bound

$$1 < \frac{38}{36} \leq v_k(u) - v_k(5/6).$$

To prove condition (ii) we again use induction. For $k = 1$, we have $v_1(u) = 1 - u$, so $v_1(u) - v_1(1/6) = 1/6 - u < 1$. Suppose now that

$$v_{k-1}(u) - v_{k-1}(1/6) < 1 \quad \text{for } 0 \leq u < 1/6.$$

The Bellman recursion (4.22) then gives us

$$\begin{aligned} v_k(u) - v_k(1/6) \leq & \int_0^{1/6} \max\{v_{k-1}(u) - v_{k-1}(1/6), 1 + v_{k-1}(1-x) - v_{k-1}(1/6)\}\, dx \\ & + \int_{1/6}^{5/6} \max\{v_{k-1}(u), 1 + v_{k-1}(x)\} - \max\{v_{k-1}(1/6), 1 + v_{k-1}(x)\}\, dx \\ & + \int_{5/6}^{1} \max\{v_{k-1}(u), 1 + v_{k-1}(1-x)\} - \max\{v_{k-1}(1/6), 1 + v_{k-1}(1-x)\}\, dx. \end{aligned}$$

For $x \in [0, 1/6]$, we can check that first integrand is bounded by 1. To see this, we first note that left maximand is bounded by 1 by the induction assumption. Next, we note that $v_{k-1}(1-x) \leq v_{k-1}(5/6)$ so, for the second maximand one has the bound $1 + v_{k-1}(1-x) - v_{k-1}(1/6) \leq 1 + v_{k-1}(5/6) - v_{k-1}(1/6)$ and this last term is non-positive by the inequality (i).

For $x \in [1/6, 5/6]$, the second integrand is bounded by

$$\max\{v_{k-1}(u) - v_{k-1}(1/6), 1 + v_{k-1}(x) - v_{k-1}(1/6)\} \leq 1,$$

since both maximands are bounded by 1; the first one because of the induction assumption and the second one because it is non-increasing in x and attains its maximum for $x = 1/6$.

Finally, for $x \in [5/6, 1]$ the third integrand is bounded by

$$\max\{v_{k-1}(u) - 1 - v_{k-1}(1-x), 0\}, \leq 0$$

since $-v_{k-1}(1-x) \leq -v_{k-1}(1/6)$, and by the induction assumption, we see that the left maximand $v_{k-1}(u) - 1 - v_{k-1}(1/6)$ is also non-positive. So, at last we have

$$v_k(u) - v_k(1/6) \leq 5/6 < 1,$$

and this completes the proof of condition (ii).

The last claim of Proposition 4.7 is that $1/6 \leq g_k(y)$ for all $y \in [0,1]$ and all $k \geq 3$. If $y \in [1/6, 1]$ this bound is trivial since $y \leq g_k(y)$ for all $1 \leq i \leq n$. If $y \in [0, 1/6)$, then the inequality (i) gives us that $1 < v_{k-1}(y) - v_{k-1}(5/6)$ for all $k \geq 3$, so that the definition of $g_k(y)$ in (4.23) gives the required lower bound. This completes the proof of Proposition 4.7. □

Proof of the Upper Bound — The Last Step

We now have all the tools that we need to prove that there is a constant $C < 11 - 4\sqrt{2} \sim 5.343$ such that

$$\mathbb{E}[A_n^o(\pi_n^*)] \leq (2 - \sqrt{2})n + C \quad \text{for all } n \geq 1.$$

We first note that the bound is trivial for $n = 1$ and $n = 2$. For $n \geq 3$, we let $\{g_n, \ldots, g_1\}$ denote the optimal threshold functions determined by recursive solution of the Bellman equation (4.22) for the finite horizon problem with sample size n. We will use the first $n-2$ of these functions to construct a suboptimal selection policy for the geometric sample size problem. From the suboptimality of this policy we will obtain an inequality that will lead to our upper bound.

Construction of a Suboptimal Policy for the Infinite Horizon Problem

We now consider the infinite horizon problem, and, as before, we let $\{X_1, X_2, \ldots\}$ denote the sequence of observations. Here is our selection process:

- We let T_0 denote the index of the first observation in the sequence that falls in the interval $[5/6, 1]$. We select that observation as first element of our subsequence and we set $Y_{T_0} = 1 - X_{T_0}$. We note that Y_{T_0} has the uniform distribution in $[0, 1/6]$.

- Next we use the functions $\{g_n, \ldots, g_3\}$ to decide which of the next $n-2$ observations are to be selected. Specifically, we make our i'th selection in the series if $X_{T_0+i} \geq$

$g_{n-i+1}(Y_{T_0+i-1})$, where as usual the Y_{T_0+i} are defined by the recursion

$$Y_{T_0+i} = \begin{cases} Y_{T_0+i-1} & \text{if } X_{T_0+i} < g_{n-i+1}(Y_{T_0+i-1}) \\ 1 - X_{T_0+i} & \text{if } X_{T_0+i} \geq g_{n-i+1}(Y_{T_0+i-1}). \end{cases}$$

Here one should recall that, by Proposition 4.7, we have $1/6 \leq g_k(Y_{T_0+i-1})$ for all $k \geq 3$, so we have $0 \leq Y_{T_0+i} \leq 5/6$ for $1 \leq i \leq n-2$.

- We will now show how our selection process can be repeated in a stationary way. For $j = 0, 1, 2, \ldots$ we proceed as follows:

  1. If $Y_{T_j+n-2} \in (1/6,\, 5/6]$, then we let

  $$\tau_j = \inf\{i \geq 1 : X_{T_j+n-2+i} \geq 5/6\},$$

  and we select the observation $X_{T_j+n-2+\tau_j}$. We note that the random variable $Y_{T_j+n-2+\tau_j} = 1 - X_{T_j+n-2+\tau_j}$ is uniformly distributed on $[0, 1/6]$.

  2. If $Y_{T_j+n-2} \leq 1/6$, then we simply let $\tau_j = 0$, and we again note that $Y_{T_j+n-2+\tau_j}$ is uniformly distributed on $[0, 1/6]$.

  3. We set $T_{j+1} = T_j + n - 2 + \tau_j$ and set $j = j + 1$.

  4. Just as in the second bullet, we use the functions $\{g_n, \ldots, g_3\}$ to decide which observations to select from $\{X_{T_j+1}, X_{T_j+2}, ..., X_{T_j+n-2}\}$. At time $T_j + n - 2$ we are left with some value Y_{T_j+n-2}, and we return to Step 1 of this bullet.

Analysis of the Policy

The suboptimal policy we constructed provides us with an increasing sequence of stopping times $0 < T_0 < T_1 < T_2 < \cdots$ such that the times $\{T_j : j \geq 1\}$ are regeneration times for the process $\{Y_i : i \geq T_0\}$. Moreover, we also have an i.i.d. sequence of stopping times

$\{\tau_j : j \geq 1\}$ with distribution

$$\tau_j \stackrel{d}{=} \begin{cases} 0 & \text{if } Y_{T_0+n-2} \leq 1/6 \\ \inf\{i \geq 1 : X_i > 5/6\} & \text{if } Y_{T_0+n-2} > 1/6. \end{cases}$$

These regeneration times $\{T_j : j \geq 1\}$ can be written as function of the stopping times $\{\tau_j : j \geq 1\}$; specifically, we have

$$T_j = T_0 + (n-2)j + \sum_{\ell=1}^{j} \tau_\ell. \tag{4.32}$$

For any pair (T_j, Y_{T_j}), $1 \leq j < \infty$, the number $r(T_j, Y_{T_j})$ of selections made from $\{X_{T_j+1}, \ldots, X_{T_j+n-2}\}$ is then given by the sum

$$r(T_j, Y_{T_j}) \stackrel{\text{def}}{=} \sum_{i=1}^{n-2} \mathbb{1}(X_{T_j+i} \geq g_j(Y_{T_j+i-1})).$$

For each $0 < \rho < 1$, the selection process described gives us a feasible policy that provides a lower bound on the expected length – $\mathbb{E}[A_N^o(\pi^*)]$ – of the alternating subsequence selected by an optimal policy form a sample of geometric size.

Moreover, if for discounting purposes we view the number of selections $r(T_j, Y_{T_j})$ as being counted all at time $T_j + n - 2$, then we obtain a lower bound for the expected value achieved by our suboptimal policy. We therefore have the bound

$$\mathbb{E}\left[\sum_{j=0}^{\infty} \rho^{T_j+n-2} r(T_j, Y_{T_j})\right] \leq \mathbb{E}[A_N^o(\pi^*)]. \tag{4.33}$$

We now note that T_0 and Y_{T_0} are independent, and we also note that, for each $j \geq 1$, the

post-T_j process $\{Y_{T_j+i} : i \geq 0\}$ is independent of T_j. Consequently, we have the factorization

$$\mathbb{E}\left[\rho^{T_j+n-2} r(T_j, Y_{T_j})\right] = \mathbb{E}\left[\rho^{T_j+n-2}\right] \mathbb{E}\left[r(T_j, Y_{T_j})\right] \quad \text{for all } j \geq 0, \tag{4.34}$$

and since T_j is a regeneration epoch, we also have

$$\mathbb{E}\left[r(T_j, Y_{T_j})\right] = \mathbb{E}\left[r(T_0, Y_{T_0})\right] \quad \text{for all } j \geq 0.$$

For $Y_{T_0} = y \in [0, 1/6]$, we recall the identity (4.25) and we observe that

$$v_n(y) - 2 \leq \mathbb{E}[r(T_0, Y_{T_0}) | Y_{T_0} = y],$$

since the policy of the right-hand side agrees with the policy of the left-hand side for the first $n-2$ observations, and the policy of the right-hand side never selects the last two.

The monotonicity of $v_n(\cdot)$ and the inequality (ii) of Proposition 4.7 then give us the lower bound

$$\mathbb{E}[A_n^o(\pi_n^*)] - 3 = v_n(0) - 3 \leq \mathbb{E}[r(T_0, Y_{T_0}) | Y_{T_0} = y] \quad \text{for all } 0 \leq y \leq 1/6,$$

so by recalling that $0 \leq Y_{T_0} \leq 1/6$ and taking total expectations we see that

$$\mathbb{E}[A_n^o(\pi_n^*)] - 3 \leq \mathbb{E}[r(T_0, Y_{T_0})].$$

The factorization (4.34) then gives us the bound

$$\mathbb{E}\left[\rho^{T_j+n-2}\right] (\mathbb{E}[A_n^o(\pi_n^*)] - 3) \leq \mathbb{E}\left[\rho^{T_j+n-2} r(T_j, Y_{T_j})\right] \quad \text{for all } j \geq 0.$$

If we now sum over j, use the representation (4.32) and use the suboptimality condition

(4.33), then we have

$$(\mathbb{E}[A_n^o(\pi_n^*)] - 3)\,\mathbb{E}\left[\sum_{j=0}^{\infty} \rho^{T_0+(n-2)(j+1)+\sum_{\ell=1}^{j}\tau_\ell}\right] \leq \mathbb{E}[A_N^o(\pi^*)]. \tag{4.35}$$

We now note that T_0 is also independent from the random variables $\{\tau_j : j \geq 1\}$, and we recall that the τ_j's are i.i.d., so

$$\mathbb{E}\left[\sum_{j=0}^{\infty} \rho^{T_0+(n-2)(j+1)+\sum_{\ell=1}^{j}\tau_\ell}\right] = \mathbb{E}\left[\rho^{T_0}\right] \sum_{j=0}^{\infty} \rho^{(n-2)(j+1)}\,\mathbb{E}\left[\rho^{\tau_1}\right]^j.$$

Since $x \mapsto \rho^x$ is convex, Jensen's inequality tells us that $\rho^{\mathbb{E}T_0} \leq \mathbb{E}[\rho^{T_0}]$ and that $\rho^{\mathbb{E}\tau_1} \leq \mathbb{E}[\rho^{\tau_1}]$, so we have

$$\rho^{\mathbb{E}T_0+n-2} \sum_{j=0}^{\infty} \left(\rho^{n-2+\mathbb{E}\tau_1}\right)^j \leq \mathbb{E}[\rho^{T_0}] \sum_{j=0}^{\infty} \rho^{(n-2)(j+1)}\,\mathbb{E}[\rho^{\tau_1}]^j.$$

The left-hand side is an easy geometric series, and by substitution in equation (4.35), we obtain the crucial bound

$$\mathbb{E}[A_n^o(\pi_n^*)] \leq 3 + \frac{1-\rho^{n-2+\mathbb{E}\tau_1}}{\rho^{\mathbb{E}T_0+n-2}}\,\mathbb{E}[A_N^o(\pi^*)].$$

From the explicit formula for $\mathbb{E}[A_N^o(\pi^*)]$ in Theorem 4.2, we then have

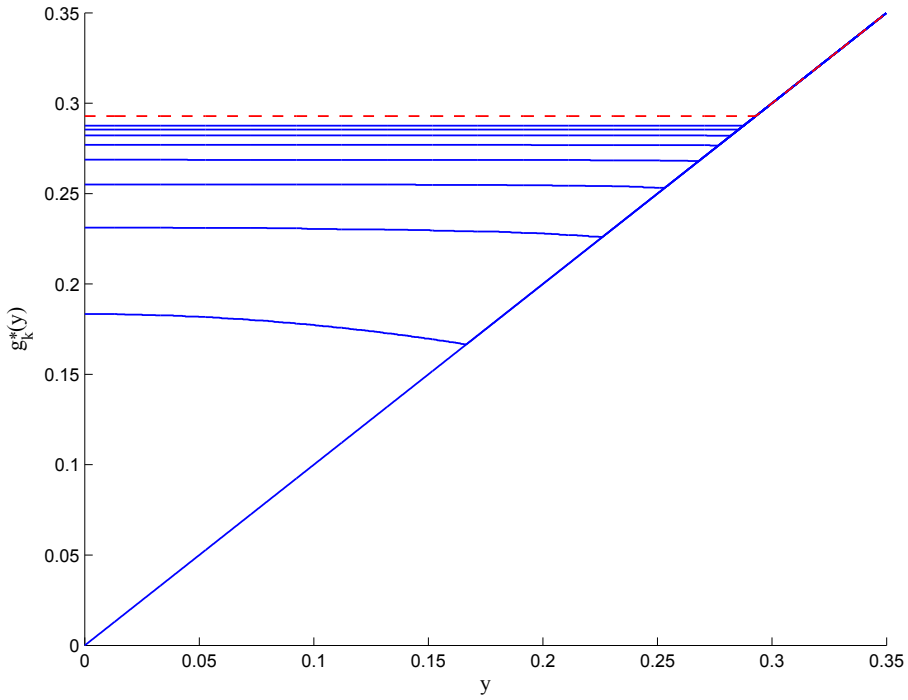
$$\mathbb{E}[A_n^o(\pi_n^*)] \leq 3 + \frac{(1-\rho^{n-2+\mathbb{E}\tau_1})(3-2\sqrt{2}-\rho+\rho\sqrt{2})}{\rho^{\mathbb{E}T_0+n-1}(1-\rho)}.$$

The bound above holds for all $0 < \rho < 1$, so by letting $\rho \uparrow 1$, we obtain

$$\mathbb{E}[A_n^o(\pi_n^*)] \leq 3 + (2-\sqrt{2})(n-2+\mathbb{E}\tau_1) < (2-\sqrt{2})n + 11 - 4\sqrt{2}$$

since $\mathbb{E}[\tau_1] < 6$. This completes the proof of the upper bound.

Figure 1: **Threshold functions** $g_k^*(\cdot),\ k = 1, 2, \ldots, 10$ **and** $g_\infty^*(y)$ **for** $y \in [0, 35/100]$**.**



4.3. A Path to the Central Limit Theorem for the Finite-Horizon Formulation

In a recent working paper, Arlotto and Steele (2012) study a Central Limit Theorem for the optimal number of alternating selections

$$A_n^o(\pi_n^*) = \sum_{i=1}^{n} \mathbb{1}\left(X_i \geq g_{n-i+1}^*(Y_{i-1})\right).$$

Their argument is based on a more refined analysis of the threshold function $g_k^*(\cdot)$, $k \geq 1$ in (4.23). For the reader's benefit, in Figure 1 we plot the threshold functions $g_1^*(y), g_2^*(y), \ldots, g_{10}^*(y)$ (solid) and the limiting function $g_\infty^*(y)$ (dashed) for $y \in [0, 35/100]$, and we note that Arlotto and Steele (2012) actually prove that

- the sequence $\{g_k^*(\cdot) : k \geq 1\}$ is monotonically increasing and bounded, i.e.

$$g_k^*(y) \leq g_{k+1}^*(y) \leq 1 \qquad \text{for all } y \in [0, 1],$$

and therefore it converges uniformly to a limiting threshold function $g^*_\infty(y)$.

- The limiting threshold function $g^*_\infty : [0,1] \to [0,1]$ satisfies

$$g^*_\infty = \max\{\xi_0, y\},$$

where $\xi_0 = 1 - 1/\sqrt{2}$ is the value of (4.19) when $\rho = 1$.

Then, they construct the selection policy π_∞ that uses the threshold function g^*_∞ at all decision times. For $Y'_0 = 0$, one has the Markov chain

$$Y'_i = \begin{cases} Y'_{i-1} & \text{if } X_i < g^*_\infty(Y'_{i-1}) \\ 1 - X_i & \text{if } X_i \geq g^*_\infty(Y'_{i-1}), \end{cases}$$

and the auxiliary random variable

$$A^o_n(\pi_\infty) = \sum_{i=1}^{n} \mathbb{1}\left(X_i \geq g^*_\infty(Y'_{i-1})\right).$$

A functional martingale Central Limit Theorem (c.f. Jones, 2004) for $A^o_n(\pi_\infty)$ then gives that

$$\frac{A^o_n(\pi_\infty) - (2-\sqrt{2})n}{\sqrt{n}} \Longrightarrow N(0, \sigma^2) \qquad \text{as } n \to \infty.$$

This limit theorem can then be used to obtain distributional information on $A^o_n(\pi^*_n)$ and, in particular, one can prove that the variance of $A^o_n(\pi^*_n)$ and $A^o_n(\pi_\infty)$ are asymptotically equivalent, i.e.

$$\text{Var}\left[A^o_n(\pi^*_n)\right] \sim \text{Var}\left[A^o_n(\pi_\infty)\right] \sim \sigma^2 n \qquad \text{as } n \to \infty,$$

and that

$$\frac{A^o_n(\pi^*_n) - (2-\sqrt{2})n}{\sqrt{n}} \underset{\sim}{d} \frac{A^o_n(\pi_\infty) - (2-\sqrt{2})n}{\sqrt{n}} \qquad \text{as } n \to \infty.$$

Thus, one concludes that

$$\frac{A_n^o(\pi_n^*) - (2-\sqrt{2})n}{\sqrt{n}} \Longrightarrow N(0,\sigma^2) \qquad \text{as } n \to \infty,$$

as desired.

4.4. Observations on Methods and Connections

Our principal goal has been to provide a reasonably definitive solution to a concrete problem of sequential optimization. Still, the natural expectation is that the solution of such a problem should also offer some novel methodological perspective. Here, we began by exploiting the well-known idea of passing to the infinite horizon problem, but less commonly (and somewhat doggedly) we made the trek back from the infinite horizon problem to the finite horizon problem. In retrospect, that trek had elements of inevitability to it, but it also had surprises.

In a natural and easy way, the policy for the infinite horizon problem gave us a lower bound for the finite horizon problem, but our first surprise was the discovery (at first numerically) that the lower bound was so close to optimal. There was also something natural about the upper bound for the finite horizon problem, though at first we argued it by contradiction. The idea was that if we had a policy for finite horizon that was "too good", then one should be able to concatenate that policy to give a policy for the infinite horizon problem that would do better than our known optimal policy. The resulting contradiction would then provide an upper bound.

This three-step process would seem to be applicable to many problems of optimal selection, though, from the details of our proof, it is clear that special features must be exploited. For example, without obtaining four relations in Lemma 4.5, we would not have been able to solve the infinite horizon problem. Three of these relations were straightforward, but the critical fourth relation still seems "lucky." We are also fortunate that symmetry relations simplified our Bellman equations. These simplifications have an intuitive basis from the

alternating nature of the problem, but it seems fortuitous that these relations could be made rigorous by inductions (of several kinds) on the Bellman equation.

There are many problems where one would like to go from the infinite horizon problem to the finite horizon problem, but one especially attractive is that of the optimal on-line selection of a monotone subsequence from a sample of independent observations. Here one knows the asymptotic behavior of the means for both finite samples Samuels and Steele (1981) and random samples — including geometric sized samples — (Gnedin 1999; 2000). Most notably, in the infinite horizon case one has a precise understanding of the variance and even a central limit theorem (Bruss and Delbaen 2001; 2004). It would be quite interesting to know if such an analogous CLT can be obtained under the finite horizon formulation.

CHAPTER 5 : Optimal Hiring and Retention Policies for Heterogeneous Workers who Learn

Workers are heterogeneous, and they evolve over time. Evolution often takes the form of on-the-job learning, with attendant decreases in the time required to complete tasks or improvements in quality. When employees turn over (quit) or are terminated they may be replaced by new hires who differ in ability and experience.

Often there may be uncertainty regarding employee attributes. Significant random variations in task times or quality – driven by task-by-task variability – can make it difficult for an employer to infer a given employee's efficiency or quality, particularly for new employees who have little or no previous track record.

Uncertainty, together with these many sources of variation – across employees, across tasks, and over time – makes decisions regarding the retention of workers complex. The longer a worker is retained, the better an inference an employer can make regarding his or her attributes. On-the-job learning, which can lead to quality improvements in incumbent employees, also favors employee retention. Yet the opportunity cost of retaining a poor performer can be great, particularly if there is wide variation in quality across the population of potential hires.

In this chapter, we develop and analyze a model that integrates all of these factors. In our model, an employer (referred to as "she") seeks to hire and retain a fixed number of employees from an infinite, heterogeneous population of potential hires. Each employee (referred to as "he") repeatedly performs the same task, whose cost the employer wishes to minimize or, equivalently, whose quality is to be maximized. Each hire moves down a learning curve, but elements of the curve's parameters are unknown to the employer. The

This chapter is written under the supervision of Prof. Noah Gans. The results presented here are also in a joint research paper with Stephen E. Chick and Noah Gans.

employer takes a Bayesian view of employees' types. By repeatedly observing the task performance of a given worker, she can make increasingly better judgments concerning his quality. After each such task, the employee decides whether he wants to continue working or not. Given that the worker decides to stay, the employer can decide whether to retain him or to replace him with a new hire. Each of these decisions has a cost for the employer. A quitting cost is incurred when a worker quits, a switching cost is incurred when a worker is terminated, and a training cost is incurred for each newly hired employee.

We formulate this problem as an infinite-horizon, discounted problem in which, at any time, the employer uses a single worker. We show that this problem is, essentially, a multi-armed bandit problem with switching costs and an infinite number of arms. (See, e.g. Gittins 1989; Banks and Sundaram 1992; Bergemann and Välimäki 2001; Sundaram 2005.) In our Bayesian setting, we prove that several classical bandit results hold in our case as well.

- The employer can use a worker's prior distribution and tenure to calculate a so-called Gittins index, and at any time it is optimal for the employer to use a Gittins-index minimal employee.

- It is optimal to retain current employees as long as their Gittins indices compare favorably to those of potential hires.

- If a current employee's Gittins index is not minimal, however, then it is optimal to hire a new worker and to never return to the current employee.

This last property is known as "no-recall" and is particularly interesting from an application perspective. Farias and Madan (2011) study bandits that do not recall, or equivalently that are irrevocable.

We also indicate how these Gittins-index results extend to more complex settings: those in which the employer retains multiple employees as well as those in which she hires from multiple, heterogeneous pools of potential hires. In both cases our original results regarding

"no-recall" properties generalize directly.

Given the availability of a Gittins index, the above policy is both intuitive and straightforward to execute. Unfortunately, the Gittins index is difficult to calculate. Nevertheless, for specific common forms of the learning-curve function, and when performance can be appropriately transformed into normally-distributed data with known sampling variances and unknown means (with a conjugate prior distribution), we:

- show that, for a fixed level of experience, the Gittins index is monotone in the posterior mean of the unknown parameter, which allows us to delineate a simple stopping boundary, below which a current worker's employment should continue and above which it should stop;
- develop approximations to the Gittins index that are straightforward to calculate and implement.

These approximations are the basis for numerical examples in that provide insights into the economic nature of the hiring and retention problem. In particular, we:

- demonstrate that the stopping boundary reflects a tradeoff between two types of learning: the performance improvement that is linked to an employee's on-the-job experience, and the statistical learning that allows the employer to make better judgments concerning a worker's ability;
- show that the value of active monitoring and screening of employees can be substantial;
- observe that the early stages of workers' tenures are the most important for the effectiveness of the optimal Gittins-index policy;
- suggest that simple hiring policies with a trial period followed by a one-shot hiring and retention decision have the potential to perform well, within a few percent of the optimal Gittins-index policy.

Sensitivity analysis with respect to model parameters provides further insights. In addition to direct gains that accrue from steeper learning curves, investments in employee learning can provide an important secondary benefit: the optimality of lower termination rates. Reductions in the variability of task performance can improve the sensitivity of screening procedures and similarly reduce optimal termination rates. The ability to terminate employees should motivate managers to consider a broader spectrum of potential hires.

5.1. Literature review

There is a vast empirical literature on learning-curve phenomena (Yelle, 1979), as well as papers devoted to effective managerial control of factors that affect or depend on learning (Dada and Srikanth, 1990; Wiersma, 2007). Much of it is segmented into the individual (e.g., Nembhard and Uzumeri, 2000a; Nembhard, 2001) and organizational levels (e.g., Bailey, 1989; Lapré et al., 2000; Pisano et al., 2001). Nembhard and Uzumeri (2000b) provide a unified study that considers both of them. Our analysis focuses on the individual level.

There also exists a rich literature that addresses labor quality and selection. The literature on secretary problems develops a normative approach to the initial screening and hiring of employees who come from a heterogeneous pool (Freeman, 1983). Similarly, there is work on multi-armed bandit problems that addresses matching problems in labor-markets: typically, problems in which employees choose firms (Jovanovic, 1979; Banks and Sundaram, 1992; Sundaram, 2005). In our context this work can be reinterpreted as addressing firms choosing employees.

The literature that explicitly addresses both worker heterogeneity and learning is much smaller. Most closely related to our work is Nagypál (2007), which models both learning-about-match-quality (between workers and a firm) and learning-by-doing. That paper's aims and results differ significantly from ours. Its model and analysis enable the use of statistical methods to discriminate between the two forms of learning in empirical employment records. We focus on model-based, and normative insights into the nature of effective

retention/termination decisions.

A few recent papers in operations-related fields also address dimensions of heterogeneity in learning and employee retention. Shafer et al. (2001) provide empirical evidence of the heterogeneity of learning curves across individuals who assemble car radios. Pisano et al. (2001) document heterogeneity across hospital units that perform cardiac surgery. Gans et al. (2010) show that the service times of call-center agents reflect on-the-job learning, as well as agent heterogeneity. Mazzola and McCardle (1996, 1997) develop models to estimate uncertain learning curves and to control production run lengths, given that a firm faces this uncertainty. None of these papers considers uncertainty regarding learning curves across individuals or groups, however. Neither do they address employee turnover or employee retention decisions.

Shafer et al. (2001) consider individual learning curves and show that, by not considering learning-parameter variations across workers, one may significantly underestimate overall productivity, given workers who operate independently. Nembhard and Osothsilp (2002) show how task complexity affects the distribution of individual learning and forgetting parameters.

The managerial implications of learning have received less attention. Nembhard (2001) is the first to propose a method that assigns workers to tasks based on learning rates of individuals, considers forgetting as well as learning, and offers heuristics for managers. Our work differs in that we derive optimal policies and our numerical experiments use somewhat different learning curves.

Pinker and Shumsky (2000), Gans and Zhou (2002) and Whitt (2006) study learning with respect to the operations management/human resource management (OM/HRM) interface. Their work does not take into account worker heterogeneity. Gans et al. (2003) and Aksin et al. (2007) are recent surveys that include discussion of learning and HRM in the call-center industry. Gaimon (1997) and Carillo and Gaimon (2000) study the importance of learning

when new technologies are introduced. Gaimon et al. (2011) use mathematical models and empirical data to assess learning-before-doing, which can be modeled as training costs in our analysis, and learning-by-doing, which is modeled by learning curves. Goldberg and Touw (2003) consider statistical inference of learning curve parameters in a managerial context.

5.2. The Hiring and Retention Problem with One Employee

In this section, we define the problem of an employer who requires the services of a single worker and who, at each discrete period of time, decides whether to retain the current employee or to terminate him and hire someone else from an infinite pool of workers. The assumption that there exists an infinite pool of potential hires is appropriate in so-called "employers' markets," in which the potential workforce is sufficiently large that workers who quit need not be considered again. Section 5.4 explores the employment of multiple hires, as well as the presence of several, heterogeneous pools of workers.

At each time $t = 0, 1, 2, \ldots$ the employer requires the service of a single employee, i, drawn from an infinite pool of potential workers, $\mathcal{S}_t$; $\mathcal{S}_0$ represents the initial pool from which the employer can draw. If employee i quits at time t then he is removed from the pool of potential hires and $\mathcal{S}_{t+1} = \mathcal{S}_t \backslash \{i\}$. We let $\pi(t) = i \in \mathcal{S}_t$ denote the employer's choice of employee i at time t and define $\pi = \{\pi(0), \pi(1), \ldots\}$ to be a *hiring and retention policy* that specifies which workers the employer engages over time.

The performance of potential workers is uncertain and evolving over time. If worker $i \in \mathcal{S}_t$ is employed at time t, then his performance is defined by the relation

$$Z_{i,t} = g(\boldsymbol{\theta}_i, n_{i,t}, \epsilon_{i,t}), \tag{5.1}$$

where $\boldsymbol{\theta}_i \in \Omega$ is a vector of parameters that reflects worker i's ability, $n_{i,t} = 0, 1, 2, \ldots$ reflects his experience to date, $\epsilon_{i,t}$ is a noise term with support $\mathcal{E}$, and $g(\cdot)$ is a deterministic function of its arguments. We denote the realization of $Z_{i,t}$ by $z_{i,t}$. For $\boldsymbol{\theta}_i = (a_i, b_i)$, Yelle

(1979) describes the following commonly-used form:

$$Z_{i,t} = \exp\left(a_i + b_i \ln(n_{i,t}+1) + \epsilon_{i,t}\right), \quad n_{i,t} = 0, 1, 2, \ldots. \tag{5.2}$$

Here, a_i is a parameter that determines a base-level of performance and $b_i < 0$ describes the rate of learning. If $Z_{i,t}$ were task time, then a_i and b_i would be scaled in the logarithm of the time unit.

The structural results concerning optimal policies, in Section 5.3, require only the general functional form (5.1), together with some technical assumptions. Furthermore, the function $g(\cdot)$ is quite general and, in addition to learning, might reflect the effect of other factors such as fatigue. While our analysis does hinge on a single measure of performance, the representation of an outcome, $Z_{i,t}$, can be generalized to explicitly represent multiple dimensions (such as revenue, cost, quality) that are aggregated into a single score by using a functional. Section 5.5, in which we develop methods for explicitly calculating the stopping boundaries necessary to implement optimal policies, assumes a more specific form of $Z_{i,t}$, such as that given by (5.2).

At the end of a given period, after his performance, the current employee notifies the employer of his intention to continue working or to leave. So, we associate with each worker a sequence of Bernoulli leaving decisions, $\mathbf{L}_i = (L_{i,0}, L_{i,1}, L_{i,2}, \ldots)$, indexed only by experience, such that worker i leaves or quits at the end of period t, after his $(n_{i,t}+1)$st performance if and only if $L_{i,0} = L_{i,1} = \cdots = L_{i,n_{i,t}-1} = 0$ and $L_{i,n_{i,t}} = 1$. We denote the realization of $\mathbf{L}_i$ and $L_{i,n_{i,t}}$ by $\boldsymbol{\ell}_i$ and $\ell_{i,n_{i,t}}$ respectively. For any hiring policy π and for each worker $i \in \mathcal{S}_0$, we let

$$\Lambda_i(\pi) = \sum_{t=0}^{\infty} \mathbb{1}(\pi(t) = i) \tag{5.3}$$

be i's working lifetime: the number of periods he is employed. In turn, we define worker i's

quitting probability, $q_{i,n}$, to be

$$q_{i,n} = \mathbb{P}\left(L_{i,n} = 1 | \Lambda_i(\pi) \geq n+1\right), \tag{5.4}$$

and call $1 - q_{i,n}$ worker i's *continuation* probability.

For $t \geq 0$, let $\mathcal{H}_{i,t} = \{(z_{\pi(s),s}, \ell_{\pi(s),n_{\pi(s),s}}) : \ \pi(s) = i, s \leq t\}$ ($\mathcal{H}_{i,0} = \emptyset$) denote worker i's *employment history* up to time t. The quitting probability of an employee with experience $n_{i,t}$, $q_{i,n_{i,t}}$, may depend on $\mathcal{H}_{i,t}$ and on his ability $\boldsymbol{\theta}_i$, but it is assumed to be independent of the employer's hiring policy, π:

$$\mathbb{P}\left(L_{i,n} = 1 | \Lambda_i(\pi) \geq n+1\right) = \mathbb{P}\left(L_{i,n} = 1 | \Lambda_i(\pi') \geq n+1\right) \quad \text{for all } \pi \neq \pi' \text{ and all } i, n.$$

This independence assumption is restrictive, and it is not difficult to imagine how employee turnover decisions may be influenced by the employer's retention (and compensation) policies. For example, by paying better performers more, the employer could provide an incentive for employee turnover patterns to change in a manner that is favorable to her. The inclusion of these types of incentives and responses extends the analysis of the employer's hiring and retention problem from the realm of single-decision-maker optimization problems to that of stochastic games and is beyond the focus of our current work. Nevertheless, the strategic interaction of employer and employees is both interesting and important, and we will briefly return to this issue in the numerical results of Section 5.6.

The employer does not know each employee's $\boldsymbol{\theta}_i$ or $\boldsymbol{\ell}_i$ in advance. Rather, she believes that there exists a random vector, $\boldsymbol{\Theta}$, that reflects the distribution of abilities in the population of potential workers, and a random set of leaving decisions, $\mathbf{L}$. The distributions for $\boldsymbol{\Theta}$ and $\mathbf{L}$ can be estimated using historical data and statistical techniques.

Each time the employer hires a new worker, she views that worker's $\boldsymbol{\Theta}_i$ and $\mathbf{L}_i$ as *iid* samples from the population distributions. At time $t = 0$ all potential workers, $i \in \mathcal{S}_0$, have the

same history, $\mathcal{H}_{i,0} = \emptyset$, the same prior distribution for $\boldsymbol{\Theta}_i$, $\nu_{i,0} \equiv \widehat{\nu}$, and no prior experience so that $n_{i,t} \equiv 0$. Thus, at time $t = 0$, the employer is indifferent among her choices.

At any time $t > 0$, each worker, i, has cumulative experience $n_{i,t}$, and the employer uses i's employment history, $\mathcal{H}_{i,t}$, to update her beliefs concerning the distribution of the parameter $\boldsymbol{\Theta}_i$. We denote the posterior distribution that describes the employer's uncertainty concerning $\boldsymbol{\Theta}_i$ at time t as $\nu_{i,t}(X) = \mathbb{P}(\boldsymbol{\Theta}_i \in X | \mathcal{H}_{i,t})$, where $X \subseteq \Omega$ is any Borel set. For $\boldsymbol{\Theta}_i \sim \nu_{i,t}$ we let $Z_{i,t} \equiv Z(\nu_{i,t}, n_{i,t})$, and for $\{\boldsymbol{\Theta}_i = \boldsymbol{\theta}_i\}$, we assume that worker i's performance $\{Z(\nu_{i,t}, n_{i,t}) \,|\, \boldsymbol{\theta}_i\}$ has density $\xi_{n_{i,t}}(z \,|\, \boldsymbol{\theta}_i)$. If worker i is employed at time t, then his experience, $n_{i,t}$, increases deterministically by one, and $n_{i,t+1} = n_{i,t} + 1$. Moreover, the employer updates her belief concerning i' ability distribution according to Bayes' rule. If $\mathcal{P}(\Omega)$ is the set of all probability measures, ν, on Ω, then the Bayes operator $\beta : \mathcal{P}(\Omega) \times \mathbb{R} \to \mathcal{P}(\Omega)$ is defined as

$$\beta(\nu_{i,t}, z)(X) = \frac{\int_X \xi_{n_{i,t}}(z|\boldsymbol{\theta})\, d\nu_{i,t}}{\int_\Omega \xi_{n_{i,t}}(z|\boldsymbol{\theta})\, d\nu_{i,t}} = \nu_{i,t+1}(X), \tag{5.5}$$

for each Borel subset $X \subseteq \Omega$. Thus for any given observation, z, the Bayes operator maps the prior distribution, $\nu_{i,t}$, to its posterior distribution, $\nu_{i,t+1}$.

Within each period, t, the employer incurs a task-related cost that is driven by the selected employee's performance, $c(z_{i,t})$. We assume that $c(z)$ is continuous and nondecreasing in z, which reflects an efficiency-based measure of employee performance. Because the employer does not know employees' true abilities, in each period she uses her belief concerning the distribution of the current employee's ability, $\nu_{i,t}$, to estimate his expected task-related cost:

$$\mathbb{E}[c\,(Z(\nu_{i,t}, n_{i,t}))] = \int_\Omega \left(\int_{\mathcal{E}} c\,(g(\boldsymbol{\theta}, n_{i,t}, x))\, \xi_{n_{i,t}}(g(\boldsymbol{\theta}, n_{i,t}, x) \,|\, \boldsymbol{\theta})\, dx \right) d\nu_{i,t}. \tag{5.6}$$

The employer also incurs costs that are specific to the hiring and retention policy she is implementing. If, at the start of a period, the employer hires a new employee, she incurs an *initial hiring* (or training) cost, c_h. If, at the end of a period, the employee quits, the employer bears a *quitting* cost, c_q, that includes potential separation costs and the cost of

recruiting a replacement. If the employee does not quit, then the employer may decide to terminate him and switch to a different worker, in which case she bears a *switching* cost, c_s. Training, switching and quitting costs are assumed to be nonnegative. To properly account for switching and quitting costs, we introduce for each worker i and each time t a switching indicator, $u_{i,t}$, such that if policy π employs worker i over several, disjoint, time periods, then the index $u_{i,t}$ switches between 0 and 1, and it equals one at every time t such that worker i was not employed at $t-1$. Formally, we set $u_{i,0} = 1$ for all $i \in \mathcal{S}_0$ and for $t \geq 1$ we let

$$u_{i,t} = \begin{cases} 0 & \text{if } \pi(t-1) = i \\ 1 & \text{if } \pi(t-1) \neq i. \end{cases}$$

When $\{u_{i,t-1} : i \in \mathcal{S}_0\} \neq \{u_{i,t} : i \in \mathcal{S}_0\}$, the workers employed at time $t-1$ and at time t differ, and the employer needs to incur the switching or quitting cost for the worker that was employed at time $t-1$.

For any time $\tau \geq 0$ and any set of prior distributions, experiences and switching indices, $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) = \{(\nu_{i,\tau}, n_{i,\tau}, u_{i,\tau}) : i \in \mathcal{S}_0\}$, the infinite-horizon total expected discounted cost of any hiring and retention policy, π, from time τ onwards is

$$\begin{aligned} C_\pi^\tau(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) \;=\; \mathbb{E}\Bigg[\sum_{t=\tau}^{\infty} \gamma^t \;\Bigg\{ & \; c_h \mathbb{1}(n_{\pi(t),t} = 0) + c\left(Z(\nu_{\pi(t),t}, n_{\pi(t),t})\right) && (5.7) \\ & + c_s u_{\pi(t),t} \mathbb{1}(\pi(t-1) \in \mathcal{S}_t \cap t > 0) \\ & + c_q u_{\pi(t),t} \mathbb{1}(\pi(t-1) \notin \mathcal{S}_t \cap t > 0) \Bigg\} \Bigg], && (5.8) \end{aligned}$$

where the discount factor is $\gamma \in [0, 1)$. We note that in each period, t, the employer bears four possible sources of cost. The first, $c_h \mathbb{1}(n_{\pi(t),t} = 0)$, is the hiring and training cost for a new worker, and it is incurred only once, at the beginning of employee $\pi(t)$'s tenure. The second, $c\left(Z(\nu_{\pi(t),t}, n_{\pi(t),t})\right)$, reflects employee $\pi(t)$'s task-related costs. The third, $c_s u_{\pi(t),t} \mathbb{1}(\pi(t-1) \in \mathcal{S}_t \cap t > 0)$, is the cost of switching to a different worker at time t, should the previous employee be terminated. The fourth source of cost, $c_q u_{\pi(t),t} \mathbb{1}(\pi(t-1) \notin$

$\mathcal{S}_t \cap t > 0)$, reflects the cost of switching to a different worker at time t, should the previous employee quit. By observing that $\mathbb{1}(\pi(t-1) \in \mathcal{S}_t \cap t > 0) + \mathbb{1}(\pi(t-1) \notin \mathcal{S}_t \cap t > 0) = \mathbb{1}(t > 0)$, we rewrite (5.7) as

$$C_\pi^\tau(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) = -c_s \mathbb{1}(\tau = 0) + \mathbb{E}\Bigg[\sum_{t=\tau}^{\infty} \gamma^t \quad \Big\{ \quad c_h \mathbb{1}(n_{\pi(t),t} = 0) \tag{5.9}$$
$$+ c\left(Z(\nu_{\pi(t),t}, n_{\pi(t),t})\right) + c_s u_{\pi(t),t}$$
$$+ (c_q - c_s) u_{\pi(t),t} \mathbb{1}(\pi(t-1) \notin \mathcal{S}_t \cap t > 0) \Big\}\Bigg].$$

In this new formulation, the switching cost, c_s, is incurred any time the worker employed at time t is different from that employed at time $t-1$. The difference, $c_q - c_s$, then adjusts the value of the switching cost if the worker employed at $t-1$ has quit. The quantity, $-c_s \mathbb{1}(\tau = 0)$, outside the expectation compensates for the switching cost incurred for the first worker ever employed because $u_{i,0} = 1$ for all $i \in \mathcal{S}_0$.

We let Π denote the set of *non-anticipating* hiring policies, and we assume that the employer seeks a policy $\pi^* \in \Pi$ that minimizes the expected discounted value of future employment costs

$$\pi^* = \text{argmin}_{\pi \in \Pi} C_\pi^0(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}). \tag{5.10}$$

For the problem to be analytically tractable, we assume that the parameter space Ω is a Borel subset of $\mathbb{R}^d$, and we require that the single-period, task-related costs are uniformly bounded, i.e. $c\left(g(\boldsymbol{\theta}, n, x)\right) \in [K_{\text{inf}}, K_{\text{sup}}]$ for each triple $(\boldsymbol{\theta}, n, x) \in \Omega \times \mathbb{N} \times \mathcal{E}$. (See, e.g. Sundaram 2005.)

5.3. Structure of the Optimal Policy

The hiring and retention problem can be formulated as a Bayesian bandit problem with an infinite number of arms. Two elements of the problem complicate the analysis, however. First, when an employee quits, the arm associated with him becomes unavailable. Second, when the employer switches from one employee to another, she incurs the switching costs,

c_s, that cannot be attributed to a single employee. In characterizing the optimal hiring and retention policy, we must address both of these difficulties.

5.3.1. Transformation to Problem with No Quitting

The fact that employees quit can be compensated for by transforming the problem with quitting into one in which workers are always available. Rather than quitting, they become *unproductive*, and their cost exceeds that of any productive worker. To do so we assume that each employee, $i \in \mathcal{S}_0$, becomes unproductive at time t, after his $(n_{i,t}+1)$st performance with probability equal to $q_{i,n_{i,t}}$ in (5.4). When employee i becomes unproductive at time t, his ability distribution changes from $\nu_{i,t}$ to $\nu_{i,t+1} = \mathbb{1}_K$ where $K \in (K_{\sup} + c_h + \max\{c_q, c_s\}, \infty)$ and $c(Z(\mathbb{1}_K, n)) = K$ for every n. Once employee i has become unproductive, he will never be able to go back to the productive state. The choice $K_{\sup} + c_h + \max\{c_q, c_s\} < K$ implies that the cost of an unproductive worker exceeds the cost of any possible realization of any productive worker, plus the largest cost of hiring a new worker. We then define the stopping time

$$\Lambda_i = \inf\{n_{i,t} \geq 1 : \ c\left(Z(\nu_{i,t}, n_{i,t})\right) = K\} \tag{5.11}$$

as the time at which employee i becomes unproductive. Because unproductive workers cannot go back to the productive state, we set $q_{i,k} = 0$ for all $k > n$ when $\Lambda_i = n$, and we modify the Bayes operator (5.5) as follows:

$$\beta(\nu_{i,t}, z)(X) \ = \ \begin{cases} \mathbb{1}_K & \text{if } \nu_{i,t} = \mathbb{1}_K \\ \frac{\int_X \xi_{n_{i,t}}(z|\boldsymbol{\theta})\, d\nu_{i,t}}{\int_\Omega \xi_{n_{i,t}}(z|\boldsymbol{\theta})\, d\nu_{i,t}} & \text{otherwise.} \end{cases} \tag{5.12}$$

Call the original problem in (5.10), in which employees quit, Problem 1, and call the modified problem, in which they become unproductive, Problem 2. The following lemma confirms the fact that the problem with workers who become unproductive is analogous to that of those who quit.

Lemma 5.1. (i) *In Problem 2, any policy that employs unproductive workers is never*

optimal.

(ii) *A policy is optimal for Problem 1 if and only if is optimal for Problem 2.*

Proofs of these claims and of others below are found in Section 5.8.

Lemma 5.1 tells us that for for each policy $\pi \in \Pi$, employee i's working lifetime $\Lambda_i(\pi)$ in (5.3) and the time at which employee i becomes unproductive (5.11) are closely related. In fact, if employee i quits before he is terminated, then $1 + \Lambda_i(\pi) = \Lambda_i$. Otherwise, $1 + \Lambda_i(\pi) < \Lambda_i$.

Lemma 5.2. *If $\mathbb{E}[\Lambda_i] < \infty$ then any policy for Problem 1 uses an infinite number of workers, a.s..*

Thus, if each employee's expected lifetime is finite, then the employer will end up hiring an infinite stream of employees in Problem 1. Similarly, an employer who avoids using employees who have become unproductive in Problem 2 will also use an infinite number of employees if $\mathbb{E}[\Lambda_i] < \infty$.

5.3.2. Transformation to Problem with Retirement Option

We derive the optimal policy for Problem 2 by solving a family of stopping problems in which, at each period, n, the employer chooses between employing a single worker, $i \in \mathcal{S}_0$, or terminating all employment and paying a so-called "retirement" cost, m. Given that we are considering an optimal stopping problem for a single employee, we drop the subscripts for the employee index, i, and the time index, t.

This approach, called the *retirement-option problem*, was introduced by Whittle (1980) for bandit problems with a finite number of arms and extended by Banks and Sundaram (1992) and Sundaram (2005) to study infinite-armed bandit models. In our context, the employer's problem is an infinite-horizon, discounted Markov Decision Process with uniformly bounded costs, a fact that implies that there exists an optimal hiring and retention policy that is stationary and deterministic (Bertsekas and Shreve 1978, Prop. 9.8). The optimal value

A policy is *stationary* if, at any time t, the action it prescribes in a given state is independent of t. A policy is *deterministic* if the action it prescribes is never randomized.

function for the retirement-option approach satisfies the following Bellman equation:

$$V(\nu, n, u, m) = \min\{m, HV(\nu, n, u, m)\} \tag{5.13}$$

where

$$\begin{aligned} HV(\nu, n, u, m) &= c_s u + c_h \mathbb{1}(n = 0) + \mathbb{E}[c(Z(\nu, n))] \\ &\quad + \gamma(1 - q_n)\mathbb{E}[V(\beta(\nu, Z(\nu, n)), n + 1, 0, m)] \\ &\quad + \gamma q_n \left[c_q - c_s + V(\mathbb{1}_K, n + 1, 0, m)\right]. \end{aligned} \tag{5.14}$$

In words, at any decision time, the employer has the choice of retiring at cost m, or continuing the employment of the worker currently on trial. The expected discounted cost of continuing, $HV(\nu, n, u, m)$, can be interpreted by looking at whether the employee is productive ($\nu \neq \mathbb{1}_K$) or not ($\nu = \mathbb{1}_K$). If the employee is productive, then with probability $1 - q_n$, he remains productive and $\beta(\nu, Z(\nu, n)) = \frac{\int_X \xi_n(z|\boldsymbol{\theta})\, d\nu}{\int_\Omega \xi_n(z|\boldsymbol{\theta})\, d\nu}$. With probability q_n, he becomes unproductive and his ability distribution changes to $\mathbb{1}_K$. If the employee is already unproductive at n, then $q_n = 0$, and the modified definition of the Bayes operator (5.12) gives us $\beta(\mathbb{1}_K, Z(\mathbb{1}_K, n)) = \mathbb{1}_K$.

Here, we restrict our attention to values of m such that $m \leq K/(1 - \gamma)$, so that retiring is attractive when $\nu = \mathbb{1}_K$. Then, (5.14) becomes

$$\begin{aligned} HV(\nu, n, u, m) &= c_s u + c_h \mathbb{1}(n = 0) + \mathbb{E}[c(Z(\nu, n))] \\ &\quad + \gamma(1 - q_n)\mathbb{E}[V(\beta(\nu, Z(\nu, n)), n + 1, 0, m)] \\ &\quad + \gamma q_n \left[c_q - c_s + m\right]. \end{aligned} \tag{5.15}$$

If $\nu \neq \mathbb{1}_K$ and the employee is productive at n, the last addend represents the cost difference paid for an employee who has quit, $c_q - c_s$, plus the retirement cost for the employer, m. The quantity $HV(\nu, n, u, m)$ hence represents the cost of employing a worker with ability

distribution, ν, experience, n, and switching indicator, u, for at least one period, followed by an optimal termination decision that depends on the retirement payment, m.

The stopping time

$$\widetilde{\Lambda}(\nu, n, u, m) = \inf\{r \geq 1 : HV(\nu_r, n + r, u_r, m) > m\} \tag{5.16}$$

is the time at which the employer chooses to retire, and $\{\nu_r\}_{r\geq 1}$ and $\{u_r\}_{r\geq 1}$ represent the evolution of the ability distribution and the switching indicator after period n. For $r = 0$, we set $\nu_0 \equiv \nu$ and $u_0 \equiv u$.

Let $Q_n = \{\omega : \Lambda > n,\ \widetilde{\Lambda}(\nu, n, u, m) = \Lambda - n\}$ be the set of sample paths for which a productive worker with ability distribution, ν, experience, n, and switching indicator, u, quits before he is terminated. Notice that, if a worker is already unproductive at n and $\nu = \mathbb{1}_K$, then $\Lambda \leq n$ and therefore $Q_n = \emptyset$. Then, we can write the expected discounted cost of continuing (5.15) as

$$\begin{aligned} HV(\nu, n, u, m) = \mathbb{E}\Bigg[& c_s u + c_h \mathbb{1}(n = 0) \\ & + \sum_{r=0}^{\widetilde{\Lambda}(\nu,n,u,m)-1} \gamma^r c(Z(\nu_r, n + r)) + \gamma^{\widetilde{\Lambda}(\nu,n,u,m)} \{(c_q - c_s)\mathbb{1}_{Q_n} + m\} \Bigg]. \end{aligned} \tag{5.17}$$

This last representation and its properties will be crucial in the proofs of many of our results.

Given the availability of the value function (5.13), we are interested in the value of m for which the employer is indifferent between continuing to employ the current hire or retiring, at cost m. We denote that value by the index

$$M(\nu, n, u) = \sup\{m \in \mathbb{R} : V(\nu, n, u, m) = m\}. \tag{5.18}$$

This index is well-defined because the value function (5.13) is concave and non-decreasing

in m, a fact that is stated and proved in Section 5.8.

5.3.3. Optimal Policy

When the employer switches from one employee to another she incurs a switching cost, c_s, that is not arm specific, a fact that makes the analysis delicate. In particular, if the employer switches away from a given employee, i, and then returns to i at a later period, she pays a switching cost that she would not have incurred had she continued to employee i over contiguous periods. For this reason, the presence of switching costs can make so-called index policies sub-optimal and make the optimal policy extremely difficult to characterize (Banks and Sundaram, 1994; Jun, 2004).

To determine the optimal policy, we therefore proceed in two stages. First, we demonstrate that index policies are optimal for the simpler case without switching costs: $c_s = 0$. Then, we use details of the optimal policy to show that, in fact, index policies continue to be optimal when switching costs $c_s > 0$ are introduced.

When $c_s = 0$, the hiring and retention problem 2 is a simple variant of the infinite-arm bandit problems analyzed by Banks and Sundaram (1992) and Sundaram (2005). In fact, when $c_s = 0$, the optimality equation (5.13) and the index (5.18) are constant with respect to u and workers' states can be reinterpreted as evolving independently of each other. Theorem 5.3 below shows that, in this case, a hiring and retention policy is optimal if and only if it always selects an employee with a minimal index (5.18). Its proof, in Section 5.8, follows the arguments of Gittins and Jones (1974) and of Sundaram (2005).

Theorem 5.3 (Optimality of an Index Policy without Switching Costs)**.** *Assume that $c_s = 0$, and that $\nu_{i,0} = \widehat{\nu}$, $n_{i,0} = 0$, and $u_{i,0} = 1$ for all $i \in \mathcal{S}_0$. A policy π^* is optimal if and only if*

$$\pi^*(t) \in \left\{ i \in \mathcal{S}_0 : \ M_i(\nu_{i,t}, n_{i,t}, u_{i,t}) = \inf_{j \in \mathcal{S}_0} M_j(\nu_{j,t}, n_{j,t}, u_{j,t}) \right\}, \ \textit{a.s. for all } t = 0, 1, 2, \ldots.$$

Remark 5.4. When $c_s = 0$, the assumption that *all* workers $i \in \mathcal{S}_0$ have ability distribution

$\nu_{i,0} \equiv \widehat{\nu}$, experience $n_{i,0} = 0$, and switching indicator $u_{i,0} = 1$ can be relaxed. In fact, Theorem 5.3 also holds for any initial state $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ such that there are infinitely many workers $i \in \mathcal{S}_0$ with ability distribution $\nu_{i,0}$, experience $n_{i,0}$, and switching indicator $u_{i,0}$ such that $M_i(\nu_{i,0}, n_{i,0}, u_{i,0}) = M(\widehat{\nu}, 0, 1)$ (Sundaram, 2005, Theorem 4.1).

Theorem 5.3 generalizes to the case with $c_s > 0$. The intuition behind the generalization is that, if it is optimal to terminate a worker with switching indicator equal to 0, then it is optimal never to employ the same worker again, even if his switching indicator changes to 1. In our problem, this last claim holds because, at any time, there are infinitely many identical workers who have never been employed. Without such availability, one could construct counterexamples in which the index policy is not optimal (Banks and Sundaram, 1994).

Corollary 5.5 (Optimality of an Index Policy with Switching Costs)**.** *Assume that $c_s > 0$ and that $\nu_{i,0} = \widehat{\nu}$, $n_{i,0} = 0$, and $u_{i,0} = 1$ for all $i \in \mathcal{S}_0$. A policy π^* is optimal if and only if*

$$\pi^*(t) \in \left\{ i \in \mathcal{S}_0 : \; M_i(\nu_{i,t}, n_{i,t}, u_{i,t}) = \inf_{j \in \mathcal{S}_0} M_j(\nu_{j,t}, n_{j,t}, u_{j,t}) \right\}, \; \textit{a.s. for all } t = 0, 1, 2, \ldots.$$

Remark 5.6. For Corollary 5.5, we can also relax the assumption that all workers are identical. The corollary still holds for any initial state $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ such that there are infinitely many workers $i \in \mathcal{S}_0$ with ability distribution $\nu_{i,0}$, experience $n_{i,0}$, and switching indicator $u_{i,0}$ such that $M_i(\nu_{i,0}, n_{i,0}, u_{i,0}) = M(\widehat{\nu}, 0, 1)$, and $\inf_{i \in \mathcal{S}_0} M_i(\nu_{i,t}, n_{i,t}, u_{i,t}) = M(\widehat{\nu}, 0, 1)$.

Given the structure of the optimal policy in Theorem 5.3 and Corollary 5.5, we can justifiably call (5.18) a *Gittins index*. Moreover, when the optimal policy is implemented, Corollary 5.5 implies that there is often just one Gittins-index-minimal employee.

Corollary 5.7. *Assume that $\nu_{i,0} = \widehat{\nu}$, $n_{i,0} = 0$, and $u_{i,0} = 1$ for all $i \in \mathcal{S}_0$, and let $\widehat{m} = M(\widehat{\nu}, 0, 1)$ be the Gittins index of a worker who has not yet been tried. Then, at any time, t, at most one worker, i, has Gittins index $M_i(\nu_{i,t}, n_{i,t}, u_{i,t}) < \widehat{m}$.*

Together Lemma 5.1 and Theorem 5.3 also imply the following useful "no-recall" property.

Corollary 5.8 ("No-Recall" Property)**.** *Assume that* $\nu_{i,0} = \widehat{\nu}$*,* $n_{i,0} = 0$*, and* $u_{i,0} = 1$ *for all* $i \in \mathcal{S}_0$*, and let* $t_i = \inf\{t : \ \pi^*(t) = i\}$ *be the first time worker* i *is employed. Then, under an optimal Gittins-index policy* π^**:*

(i) *Worker* i *is employed continuously for* $\Lambda_i(\pi^*)$ *periods; that is* $\pi^*(t) = i$ *for all* $t_i \leq t < t_i + \Lambda_i(\pi^*)$*.*

(ii) *It is never optimal to employ worker* i *from time* $t_i + \Lambda_i(\pi^*)$ *on; that is* $\pi^*(t) \neq i$ *for all* $t \geq t_i + \Lambda_i(\pi^*)$*.*

Therefore, it is never optimal to employ a worker who was previously replaced, and we can reinterpret the switching costs as costs that are due upon firing.

Under the optimal policy, the employer calculates the index for untried workers, $\widehat{m}$. Then she chooses an employee, i, at random from the pool of untried employees, and after i's tth performance, she recalculates i's Gittins index based on the posterior distribution $\nu_{i,t}$. If the new Gittins index has a value of $\widehat{m}$ or less, then it is optimal to retain the current employee. If the updated Gittins index rises above $\widehat{m}$ then it is optimal to terminate him and hire a new employee, at random, from the pool.

For an employer seeking to retain a single employee, the hiring and retention problem decomposes into a sequence of *iid* optimal stopping problems: hire an employee from the pool and retain him until he turns over or his Gittins index rises above $\widehat{m}$, whichever comes first. Given the *iid* nature of the stopping problems, we can show that the Gittins index of the untried workers is closely related to the total expected discounted cost under the optimal policy.

Theorem 5.9. *Assume that* $\nu_{i,0} = \widehat{\nu}$*,* $n_{i,0} = 0$*, and* $u_{i,0} = 1$ *for all* $i \in \mathcal{S}_0$*, and let* $\widehat{m} = M(\widehat{\nu}, 0, 1)$ *be the Gittins index of a worker who has not yet been tried. If* $\mathbb{E}[\Lambda_1] < \infty$ *then* $\widehat{m} - c_s = \inf_{\pi \in \Pi} C_\pi^0(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$.

Theorem 5.9 is appealing because it links the expected total discounted cost under the

optimal policy to the Gittins index. This type of result does not usually hold in a general bandit setting. Here, it relies on the presence of infinitely many identical, untried workers and on the "no-recall" property of the optimal policy described in Corollary 5.8. This allows us to interpret our hiring process as a discounted renewal reward process in which the tenure of every worker is the length of the renewal interval, and the cost of each worker throughout his tenure is the reward. The renewal intervals as well as the rewards are *iid*. In Section 5.5, we use Theorem 5.9 to estimate the expected discounted value of a Gittins-index policy.

5.4. Extensions: Multiple Parallel Workers and Different Pools

Sections 5.2 and 5.3 considered the problem of employing a single worker. We now consider two extensions. Section 5.4.1 considers the problem in which distinct (infinite) pools of heterogeneous workers are available. Section 5.4.2 considers an employer who wishes to retain multiple employees who work in parallel. In both cases, the optimality of an index rule is retained.

5.4.1. Heterogeneous Populations

When the employer faces a finite number of heterogeneous populations, her optimal hiring and retention policy is the same as the one proposed in Corollary 5.5. (See also Remark 5.6.) For example, consider two infinite pools $\mathcal{S}_0^{\nu}$ and $\mathcal{S}_0^{\eta}$, for which the untried workers have common prior distributions $\widehat{\nu}$ and $\widehat{\eta}$, with $\widehat{\nu} \neq \widehat{\eta}$. Let $M(\widehat{\nu}, 0, 1)$ and $M(\widehat{\eta}, 0, 1)$ be the indices of the untried workers in each pool. If $M(\widehat{\nu}, 0, 1) \neq M(\widehat{\eta}, 0, 1)$, then workers belonging to the pool with larger index are never employed by an optimal policy. Otherwise, if $M(\widehat{\nu}, 0, 1) = M(\widehat{\eta}, 0, 1)$, then the employer is indifferent between the two populations.

5.4.2. Hiring and Retention of Multiple Workers

Assume now that $\nu_{i,0} = \widehat{\nu}$, $n_{i,0} = 0$, and $u_{i,0} = 1$ for all $i \in \mathcal{S}_0$, and consider the hiring and retention problem in which the employer wishes to retain a fixed number, D, of people working in parallel.

One can partition the infinite pool of potential employees, $\mathcal{S}_0$, into D separate, countably infinite pools, $\mathcal{S}_{1,0}, \ldots, \mathcal{S}_{D,0}$, of identical workers with common prior distribution, $\widehat{\nu}$, no experience, and common switching indicator equal to 1. When employee i in pool d quits at time t, he is removed from that pool so that $\mathcal{S}_{d,t+1} = \mathcal{S}_{d,t} \backslash \{i\}$. Then, the infinite-horizon total expected discounted cost is

$$C_{\pi}^{0,D}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) = \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t \sum_{d=1}^{D} \Bigg(c_h \mathbb{1}(n_{\pi_d(t),t} = 0) + c(Z(\nu_{\pi_d(t),t}, n_{\pi_d(t),t},)) \\ + u_{\pi_d(t),t} \mathbb{1}(t > 0) c_s \mathbb{1}(\pi_d(t-1) \in \mathcal{S}_{d,t}) \\ + u_{\pi_d(t),t} \mathbb{1}(t > 0) c_q \mathbb{1}(\pi_d(t-1) \notin \mathcal{S}_{d,t}) \Bigg) \Bigg], \tag{5.19}$$

where $\pi_d(t) \in \mathcal{S}_{d,t}$ identifies the index of the worker who is employed from pool d at time t, $\nu_{\pi_d(t),t}$ his ability distribution, $n_{\pi_d(t),t}$ his experience, and $u_{\pi_d(t),t}$ the value of his switching indicator. By interchanging the sums in (5.19) one obtains $C_{\pi}^{0,D}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) = \sum_{d=1}^{D} C_{\pi}^{0,d}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$, where $C_{\pi}^{0,d}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ is the dth position's expected discounted cost, as defined in (5.7). Thus, the D positions' costs are separable so that the total expected discounted cost is minimized when a Gittins-index minimal worker is employed in each pool.

At any time, t, at which the employer seeks to hire a new worker for any of the D positions, she can employ any untried worker who belongs to the pool of potential employees, $\mathcal{S}_t$. This result, due to Bergemann and Välimäki (2001), crucially depends on the assumption that all workers have the same experience and ability distribution at time $t = 0$, so that the artificial splitting of potential hires into D pools is possible.

We note that our analysis of multiple employees also hinges on the independence of the outcomes of various employees' tasks. In many settings, task outcomes may be correlated across workers, however, and the optimality of an allocation index is no longer valid, as is the case for other bandit problems with correlated arms. One potentially promising avenue for addressing such correlations in future work is the knowledge gradient approach (Frazier

et al., 2009).

5.5. Implementing the Optimal Policy

This section shows how analytic properties of the hiring and retention problem can be combined with dynamic programming to enable the computation of the relevant Gittins indices when performance has certain structural properties. As shown in Section 5.8, for any given ν, n, u the value function, $V(\nu, n, u, m)$, is concave and nondecreasing in m. Therefore, given ν, n, u a simple search scheme, such as bisection, can be used to find the largest fixed point, $M(\nu, n, u)$, that defines the Gittins index.

Because our set of *iid* stopping problems allows us to focus on a single employee, we drop the indices i and t as subscripts and let $Z_n = g(\boldsymbol{\theta}, n, \epsilon_n)$. To calculate solution values, we explicitly define the functional form of the $(n+1)$st performance for a worker, Z_n. We assume that $g(\cdot)$ is invertible and that

$$g^{-1}(Z_n) = A + h(n) + \epsilon_n, \quad n = 0, 1, 2, \ldots, \tag{5.20}$$

is a linear model where A determines an unknown base-level that may vary across workers, $h(n)$ is a known learning function, and ϵ_n is normally distributed noise with mean 0 and known variance σ^2.

Because A is unknown, the mean of the noise can be assumed to be zero without loss of generality. We assume that the potential hire's base level of performance, A, has initial prior distribution, $\widehat{\nu}$, that is normally distributed with mean $\widehat{\mu}$ and variance $\widehat{\sigma}^2$, $N(\widehat{\mu}, \widehat{\sigma}^2)$. The form in (5.20) implies another structural property that will be useful for computing the Gittins indices of workers. The random variables $g^{-1}(Z_n) - h(n)$ are normally distributed with unknown mean A and variance $\sigma^2 + \widehat{\sigma}^2$. By standard Bayesian analysis, ν, the posterior distribution of A after observing n tasks, $\mathbf{z}_n = (z_0, z_2, \ldots, z_{n-1})$, is normal with

$$\mathbb{E}[A \mid \mathbf{z}_n] = \frac{\widehat{\mu}(\sigma^2/\widehat{\sigma}^2) + \sum_{k=0}^{n-1}(g^{-1}(z_k) - h(k))}{n + \sigma^2/\widehat{\sigma}^2} \quad \text{and} \quad \mathrm{Var}[A \mid \mathbf{z}_n] = \frac{\widehat{\sigma}^2\sigma^2}{\sigma^2 + n\widehat{\sigma}^2}.$$

Define $\widehat{p} = \sigma^2/\widehat{\sigma}^2$, and let $p = \widehat{p} + n$, where n is the number of samples observed for the single-worker problem. Set $y_p = \widehat{\mu}\,\widehat{p} + \sum_{k=0}^{n-1}(g^{-1}(z_k) - h(k))$ and $w_p = y_p/p$. The posterior distribution, ν, of A given $\mathbf{z}_n$ is thus $N(w_p, \sigma^2/p)$. We can therefore describe (ν, n) by (w_p, p).

These assumptions are sufficient to guarantee that both the Bellman equation (5.13) and the Gittins index (5.18) are monotone in the posterior mean of A, w_p.

Proposition 5.10. *For any given p, u, and m the value function $V(w_p, p, u, m)$ is nondecreasing in w_p. For any given p, and u the Gittins index $M(w_p, p, u)$ is nondecreasing in w_p.*

The monotonicity of the Gittins index with respect to w_p allows us to concisely describe the optimal policy. For each $p = \widehat{p} + n$, there is a simple "stopping" boundary, $\mathfrak{b}(p)$, such that it is optimal to retain the employee (continue) if $w_p < \mathfrak{b}(p)$ and to terminate the employee (stop) if $w_p > \mathfrak{b}(p)$.

Arlotto et al. (2010) provides more detail for how to use the above results to approximate V and the stopping boundary, $\mathfrak{b}$, when (5.20) applies, the functions g and h are known and finite for finite values of their arguments, the noise, ϵ_n, has zero mean and known sampling variance, σ^2, and the prior distribution for A is $N(w_{\widehat{p}}, \sigma^2/\widehat{p})$, so that Proposition 5.10 applies. In summary, we use the common technique of approximating the evolution of the posterior distribution as samples are observed, a Gaussian process, with the evolution of the posterior distribution of a related trinomial process on a grid. We construct the necessary grid of points in the (w, p) coordinate system, estimate the terminal conditions (the period at which the dynamic programming backwards recursion starts, typically a large number of periods in the future) using Monte Carlo simulation, perform a backward recursion using a trinomial tree approximation on the grid of points to approximate both V and the optimal stopping boundary for a given value of m, and then search for the value of m that identifies the Gittins index. This process also identifies the optimal stopping boundary that determines the optimal solution to the hiring and retention problem.

The numerical results in Section 5.6 correspond to a learning function that sets $g(z) = e^z$ and $h(n) = b\ln(n+1)$. This corresponds to (5.2) with a common learning parameter $b_i = b$ and

$$\ln(Z_n) = A + b\ln(n+1) + \epsilon_n, \quad n = 0, 1, 2, \ldots, \tag{5.21}$$

where $\epsilon_n \sim N(0, \sigma^2)$. Here, (5.21) is consistent with empirical studies of various industries. For example, Brown et al. (2005), Shen (2003), and Shen and Brown (2006) provide evidence that handle times for call-centers are frequently lognormally distributed.

The above approach can be used to numerically evaluate other forms of $h(\cdot)$, and we have also tested $h(n) = b\ln(1 + n/(n+\zeta_1))$ and $h(n) = b\ln(1 + \min\{n, \zeta_2\})$. While the details of the stopping boundaries can change with the functional form, the qualitative conclusions we reach from numerical tests with these functions are analogous to what we describe below in Section 5.6. Similarly, we can define a as a common, known parameter and $g^{-1}(Z_n) = a + Bh(n) + \epsilon_n$ to model pools of workers with a common base level of quality and heterogeneous rate of learning. While the theoretical results described in Section 5.3 hold for even more complex settings, such as those with heterogeneous and unknown A and B, the numerical approach here becomes more difficult. In particular, stopping boundaries become multidimensional and monotonicity results, such as those described in Proposition 5.10, may not hold.

5.6. Numerical Examples and the Value of Screening

In this section, we use the methods described in Section 5.5 to calculate Gittins indices, as well as associated optimal stopping boundaries, for several examples. We also use discrete event simulation to estimate rates of termination and voluntary turnover. We compare the performance of the optimal Gittins-index policy with that of other easily implementable policies and demonstrate that an active hiring and retention policy reduces costs and improves the pool of workers who are employed. We perform a sensitivity analysis with respect to the key parameters of our model, and we conclude that increases in employee learning

rates reduce costs, improve the pool of employed workers and lower termination rates. Moreover, we observe that managers favor pools of potential workers with a broader set of abilities.

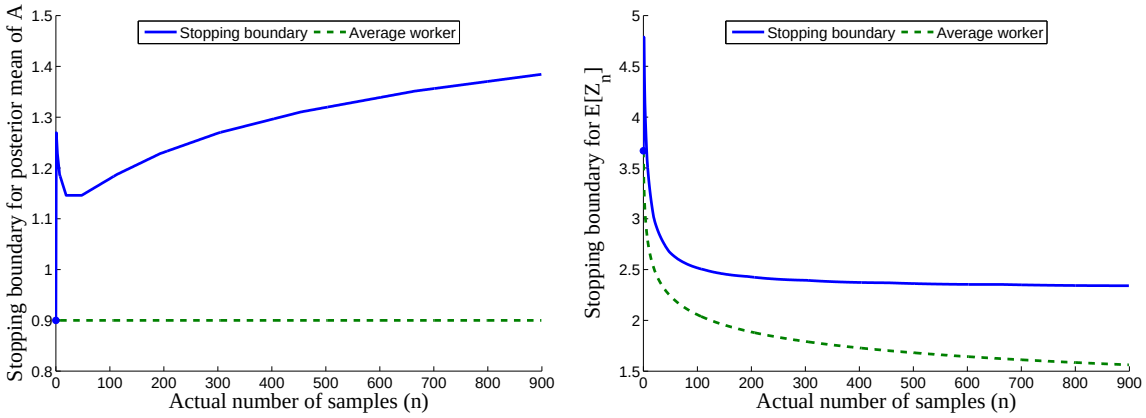
5.6.1. Balancing Uncertainty and Learning Effects

The first example is loosely motivated by a call center. Each Z_n represents the average duration (in minutes) of the calls that an agent handles after n days of experience. We use the log-linear learning curve model (5.21). The distribution of the base-level performance parameter, A, has mean $\widehat{\mu} = 0.90$ and standard deviation $\widehat{\sigma} = 0.40$, and the sampling standard deviation in the daily average of the service times is $\sigma = 0.80$. This implies an expected service time of untried agents of $\mathbb{E}[Z_0] = 3.67$. The annual discount rate is 10%, so the one-period discount rate is $\gamma = 0.9996$ (based on a year of 250 days), and the cost function is linear, $c(z) = cz$, with unit cost $c = 1$. The training cost is $c_h = 30$ which corresponds to the expected cost of employing untried workers for approximately 10 days (2 weeks). Termination and quitting costs are set equal to 0. (See Theorem 5.12, below, in Section 5.6.3.) Learning is deterministic with rate $b = \ln(\alpha)/\ln(250)$, where $\alpha \in [0, 1]$ represents the amount of learning accrued in the first year of tenure so that $\mathbb{E}[Z_{249}] = \alpha\mathbb{E}[Z_0]$. Choosing $\alpha = 0.50$, we obtain $b = -0.1255$.

For lack of real-world data concerning turnover behavior, and to focus our numerical results on the effects of learning, we assume that the quitting probability q_n is constant over time. We let $q_n = 0.01$ for all n, so (in the absence of termination) workers turn over, on average, every 100 days.

Figure 2 displays the stopping boundary associated with the Gittins index for untried employees who, in this example, have $\widehat{m} = 5,491.7$. The left panel plots the stopping boundary with respect to the posterior mean of A, and the solid line in the right panel plots the analogous stopping boundary with respect to the posterior mean of Z_n. From Proposition 5.10, we know that an employee whose posterior mean falls below these stopping boundaries has

Figure 2: **Stopping boundaries for posterior mean of A (left) and for $\mathbb{E}[Z_n]$ (right).**



Parameters: $b = -0.1255, \widehat{\mu} = 0.90, \widehat{\sigma} = 0.40, \sigma = 0.80, \widehat{s} = 4, c_h = 30, c_s = c_q = 0$.

a Gittins index below $\widehat{m}$ and should be retained, and one whose posterior mean falls above the stopping boundary should be replaced by a new hire.

In the left panel, we see that the stopping boundary with respect to the posterior mean of A has an interesting shape. The initial jump from the prior mean, $\widehat{\mu} = 0.90$, up to 1.27 is attributed to the elimination of the training cost, c_h, which is incurred only on day zero. Afterwards, the stopping boundary has a "cupped" shape for the first few periods of an employee's tenure. The dip reflects the effect of statistical learning on the part of the employer. As more samples are collected, uncertainty about the "true" quality of the worker decreases, and the employer can screen workers on the basis of a more informative prior distribution. The subsequent climb reflects the gains the employee enjoys as on-the-job experience makes even relatively poor-quality workers attractive candidates for retention. In its right most reaches, the curve appears to increase to an asymptote involving a constant minus $h(n)$ (a phenomenon that was observed for other learning functions we tested).

The right panel shows the stopping boundary with respect to $\mathbb{E}[Z_n]$. Here, the stopping boundary is unimodal, with a peak on day 1 due to the elimination of the day-zero training cost, followed by a monotone decrease that is initially steep and that later flattens out. Unlike the left panel, the right panel does not explicitly display a "dip" that reflects the

Table 1: **Optimal policy and employee retention.**
(standard errors for the mean in parenthesis)

	Day 1	Days 2 – 10	Days 11 – 20	Total
Terminated workers	0.0196 *(.0006)*	0.2830 *(.0020)*	0.0557 *(.0010)*	0.3982 *(.0022)*
Workers who quit	0.0102 *(.0005)*	0.0692 *(.0011)*	0.0539 *(.0010)*	0.6018 *(.0022)*

problem's two conflicting forces, between the employer's statistical learning and the employees' learning by doing. Instead, after day 1, we find a monotonically decreasing stopping boundary that requires a worker's expected performance to keep improving over time. The dashed line in both panels plots the prior mean, $\widehat{\mu}$, (left) and the expected call times, $\mathbb{E}[Z_n]$, (right) for an "average" employee with base-level service time $A = \widehat{\mu}$. The vertical distance between the two curves is a measure of how much better or worse a "marginally retained" employee is in comparison to an "average" employee. The presence of training costs induces managers to retain workers who are worse than average.

The simulation results in Table 1 describe how the optimal policy affects employee retention. The results are based on 50,000 trials of the single-worker optimal stopping problem, and they show the fraction of workers who are terminated or quit within various time windows.

The policy terminates 39.82% of the employees: 1.96% of workers are terminated on day 1, 28.30% are terminated during periods 2 through 10, and 9.57% thereafter. Hence, much of the termination occurs early on. Of course, termination rates vary significantly with training costs. In Section 5.6.3, we present a sensitivity analysis that addresses this relationship.

5.6.2. How the Optimal Policy compares with Simpler Policies

This section compares the optimal policy with four families of alternative hiring policies. In the first family, workers are never terminated, and they serve until they naturally turn over. In the second, workers are monitored for a limited screening period, during which they can be terminated after each day of performance. If retained at the end of the screening period, they are never terminated. In Table 2, we report results for this type of policy when the screening period is 5, 10 or 20 days long. The third family considers Gittins-index

policies in which workers are screened and termination can occur every 5, 10 or 20 days of performance. (Note that the optimal policy described in this paper is a Gittins-index policy in which screening takes place each day.) Finally, the fourth family considers policies with a trial period of a given length (1, 5, 10 or 20 days) within which workers are not terminated. At the end of the trial period the employer decides whether to retain or terminate the worker, and, if he is retained, he is not terminated until he turns over. In all cases, we use optimal retain/terminate thresholds, given the details of the particular policy.

Table 2 reports infinite-horizon total expected discounted costs, termination rates, long-run average service rates and the expected discounted number of monitored periods for each policy. The results reported are obtained by simulating 1,000 trials with enough workers to cover 50,000 time periods within each trial. We also report analogous simulation results for the optimal policy and note that, because it is estimated via simulation, rather than backward recursion, the Gittins index for this example varies slightly (within one standard error) from that reported in Section 5.6.1.

The results in the second column of Table 2 show that the optimal policy we examined leads to a substantial reduction in cost. For instance, the policy that does not screen employees has a total expected discounted cost that is 10.41% higher than that of the optimal Gittins-index policy. We already know from Table 1 that most termination in the optimal policy occurs relatively early in employees' tenure. It is not surprising then, that the policy that screens workers in each of the first 20 days performs nearly as well as the optimal one. Interestingly, the Gittins-index policy that screens workers every 5 days also performs close to optimally. Thus, screening needs not to occur every period for a policy to be effective. The results for "one-shot" at 5 and 10 periods also suggest that simple, one-shot retention decisions have the potential to perform well, with average discounted costs within a few percent of the optimal Gittins-index policy.

Table 2: **Comparison with other hiring policies.**
(standard errors for the mean in parenthesis)

Policy	Total expected discounted cost		Fraction of terminated workers	Long-run average service rate		Exp. disc. number of monitored periods
Optimal policy	5,494.1 *(12.3)*	-	.3948 *(.0005)*	.6417 *(.0138)*	-	2,333 *(6.6)*
Never screen	6,066.3 *(15.4)*	10.41%	.0000 *(.0000)*	.5364 *(.0149)*	-16.41%	0 *(0.0)*
Screen 1-5	5,618.3 *(12.4)*	2.26%	.3540 *(.0006)*	.6179 *(.0140)*	-3.71%	149 *(21.3)*
Screen 1-10	5,539.8 *(11.9)*	0.83%	.3764 *(.0006)*	.6315 *(.0140)*	-1.59%	288 *(39.9)*
Screen 1-20	5,505.3 *(11.7)*	0.20%	.3960 *(.0005)*	.6405 *(.0137)*	-0.19%	525 *(67.5)*
Gittins every 5	5,528.9 *(11.9)*	0.63%	.3690 *(.0005)*	.6341 *(.0134)*	-1.18%	445 *(4.9)*
Gittins every 10	5,569.3 *(11.5)*	1.37%	.3282 *(.0005)*	.6218 *(.0136)*	-3.10%	212 *(4.2)*
Gittins every 20	5,672.2 *(12.4)*	3.24%	.2623 *(.0005)*	.6015 *(.0130)*	-6.27%	97 *(3.5)*
One-shot at 1	5,896.0 *(14.4)*	7.32%	.2432 *(.0005)*	.5739 *(.0153)*	-10.57%	24 *(3.4)*
One-shot at 5	5,639.6 *(12.7)*	2.65%	.3228 *(.0006)*	.6108 *(.0139)*	-4.81%	23 *(3.1)*
One-shot at 10	5,644.4 *(12.2)*	2.73%	.3234 *(.0006)*	.6146 *(.0137)*	-4.23%	21 *(2.7)*
One-shot at 20	5,696.1 *(12.3)*	3.68%	.2669 *(.0005)*	.6004 *(.0131)*	-6.43%	18 *(2.2)*

For any hiring policy, π, its long-run average service rate is

$$\mu(\pi)^{-1} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\mathbb{E}[Z_{\pi(t),t}]},$$

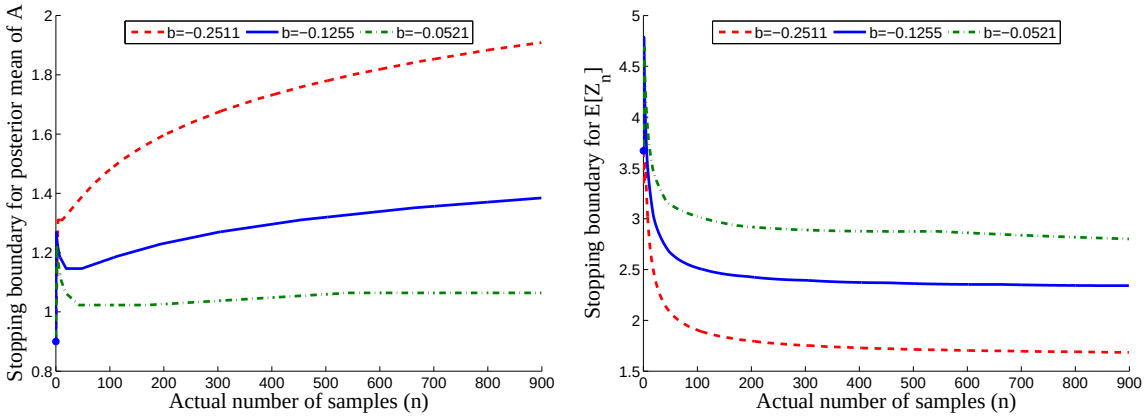
the long-run average number of calls that an agent handles per minute each day. Its numerical values are reported in column four of Table 2, and they suggest that the optimal Gittins-index policy leads to an overall improvement of employee performance. Moreover, the quantity $\mu(\pi)^{-1}$ can then be used to obtain a rough estimate of the number of agents needed for a given call volume. For instance, if we compare the optimal Gittins-index policy with the "never screen" policy, we see that the former requires, on average, 16.41% fewer workers to maintain the same level of capacity. To more clearly understand this, consider the hypothetical scenario in which a call center has an average load of 53.64 calls per minute. With the optimal policy, this requires employing 53.64 / 0.6417 = 83.59 workers – long-run average – to have a "fully-loaded" system. With the policy "never screen", the same "fully-loaded" system requires 53.64 / 0.5364 = 100 workers, and the optimal policy employs 16.41% fewer workers.

The rightmost column of Table 2 counts the expected discounted number of periods in which the employer monitors the performance of its employees. Naturally, the optimal Gittins-index policy in which monitoring occurs every day is the most expensive along this dimension. Interestingly, the policies "Screen 1-20" and "Gittins every 5" perform well with respect to costs and require approximately one fourth of the monitoring effort on the part of the employer. Thus, to the extent that monitoring is an expensive activity, the nature of effective policies may change. While the explicit representation and optimization of monitoring is beyond the scope of the current paper, it certainly merits future work.

5.6.3. Sensitivity analysis

This section examines how the optimal policy depends on four key parameters: employees' learning rates; employer uncertainty regarding employee performance; task-by-task variabil-

Figure 3: **Stopping boundaries for different learning rates.**



Other parameters: $\widehat{\mu} = 0.90, \widehat{\sigma} = 0.40, \sigma = 0.80, \widehat{s} = 4, c_h = 30, c_s = c_q = 0$.

ity; and training costs. The Gittins indices, turnover and termination rates reported in this section are computed as in Section 5.6.1.

Learning rates.

Section 5.6.1 studied a pool of workers whose performance improves by 50% over the first 250-day year ($b = -0.1255$). Here, we compare this performance with that of fast-learning workers who improve by 75% in the first one year ($b = -0.2511$), as well as that of slow-learning workers who improve only by 25% in the same amount of time ($b = -0.0521$). All other parameters are as in Section 5.6.1.

Figure 3 plots the stopping boundary with respect to the posterior mean of A (left) and with respect to $\mathbb{E}[Z_n]$ (right) in these new settings. In the left panel, we notice that the "cupped" shape of the stopping boundary in the early stages of employment is more prominent for the slow learners, and the set of their allowable posterior means is smaller. On the other hand, the fast-learning workers immediately benefit from a tangible performance improvement in their first few days so that the "cupped" part of the stopping boundary disappears. The contribution of this experience-based learning is so high that the screening policy retains workers with a broader set of posterior means. With a faster learning rate, every employee

is faster for each level of experience, and one expects the stopping boundary with respect to $\mathbb{E}[Z_n]$ to decline. This is indeed the case and, in the right panel of Figure 3, we see that the stopping boundary for fast-learning workers is the bottom one. A similar argument explains why the stopping boundary for slow learners is the top one in the right panel.

To more clearly understand the effect of changes in employees' learning, we also look at the values of the Gittins index, at the fraction of terminated workers, and at the long-run average service rate for these three b's. Table 3 shows that the optimal retention policy for pools of fast learners generates the smallest infinite-horizon expected-discounted cost, the lowest fraction of terminated workers and the largest service rate. Conversely, slow learners are the most expensive, have the highest termination rates and the lowest long-run average service rates.

Table 3's results suggest a potentially important, positive sequence of managerial implications. Improvements in on-the-job learning rates make employees with relatively poor initial abilities quickly become attractive relative to untried employees, and it is optimal for the employer to retain them. As a consequence, optimal termination rates decline. Thus, improvements in on-the-job learning rates may allow the employer to enjoy a secondary benefit of being able to retain a wider array of employees. Moreover, there is evidence from the management literature that lower rates of termination may make a company a more desirable place to work and improve its pool of potential hires (Huselid, 1995). Such an employee response to changes in the employment policy is of potential interest. As noted in the introduction, explicit treatment of the phenomenon would extend our analysis in to the realm of stochastic games, however.

Remark 5.11. Empirical evidence in the learning literature shows that slower learners can produce higher value in the long run (see, e.g., March, 1991; Uzumeri and Nembhard, 1998). In our model, this effect could be investigated by segmenting slow learners and fast learners in two different populations. If the prior ability distribution in each population were known, then the optimal policy would be as in Section 5.4.1 and only workers belonging to

the population with better index would be employed. If the prior ability distributions were unknown, however, one would need to construct a hierarchical model that goes beyond the scopes of the current paper.

Variance of base-level performance in prior distribution.

We parameterize the employer's uncertainty concerning the ability of untested workers using the prior standard deviation of A, $\widehat{\sigma}^2$. By varying $\widehat{\sigma}$ while holding σ constant, we can see how the optimal screening policy changes with worker heterogeneity. Here, we analyze three values of the prior standard deviation, 0.20, 0.40, and 0.80 (i.e., $\widehat{\sigma}^2 = 0.04, 0.36, 0.64$ respectively), and we discuss how they affect our results. All other parameters remain constant, as in Section 5.6.1.

Table 4 shows how the Gittins index, the fraction of terminated workers, and the long-run average service rate change with $\widehat{\sigma}^2$. The values obtained in the numerical example agree with the general idea that the Gittins index reflects an option value inherent in the ability to change arms, and it favors arms with more diffuse prior distributions. In our context this implies that, for a given $\widehat{\mu}$, an increase in the variation of ability across workers allows the employer to screen more strictly, thereby increasing termination rates, retaining relatively more capable employees, and lowering total costs.

Sampling variance.

We then perform a sensitivity analysis with respect to the sampling variance, σ^2. The analysis is similar to that for the prior variance, but here we keep $\widehat{\sigma}$ constant as we let σ vary. The values of σ we consider are 0.60, 0.80, 1.00. The other parameters are fixed as in Section 5.6.1.

Table 5 displays the increase in the Gittins index and the decrease in the long-run average service rate as σ increases. It also indicates that, for lower σ, the fractions of employees who are terminated are lower. Thus, reductions in within-period variability improve the selectiv-

Table 3: **Simulation results with different learning rates.**
(standard errors for the mean in parenthesis)

b	Gittins index	Fraction of terminated workers				Long-run average service rate
		Day 1	Days 2–10	Days 11 –20	Total	
-0.2511	3,905.6	0.0102 *(.0004)*	0.1834 *(.0017)*	0.0275 *(.0007)*	0.2366 *(.0019)*	1.0253 *(.0286)*
-0.1255	5,491.7	0.0196 *(.0006)*	0.2830 *(.0020)*	0.0557 *(.0010)*	0.3982 *(.0022)*	0.6417 *(.0138)*
-0.0521	6,762.1	0.0334 *(.0008)*	0.3167 *(.0021)*	0.0822 *(.0012)*	0.4885 *(.0022)*	0.4972 *(.0089)*

Table 4: **Simulation results with different prior variances.**
(standard errors for the mean in parenthesis)

$\widehat{\sigma}$	Gittins index	Fraction of terminated workers				Long-run average service rate
		Day 1	Days 2–10	Days 11 –20	Total	
0.2000	5,715.1	0.0000 *(.0000)*	0.0330 *(.0008)*	0.0351 *(.0008)*	0.1060 *(.0014)*	0.5204 *(.0064)*
0.4000	5,491.7	0.0196 *(.0006)*	0.2830 *(.0020)*	0.0557 *(.0010)*	0.3982 *(.0022)*	0.6417 *(.0138)*
0.8000	4,752.8	0.2406 *(.0019)*	0.2840 *(.0020)*	0.0439 *(.0009)*	0.5842 *(.0022)*	1.1357 *(.0488)*

ity and effectiveness of screening procedures, allowing the employer to reduce termination rates using the optimal policy.

Training costs.

Section 5.6.1 studied a setting in which every time a new worker is employed, the employer incurs a training cost, $c_h = 30$. Here we perform a sensitivity analysis that studies how termination rates and total expected discounted costs vary with training costs. When different values of training costs are considered, the stopping boundaries in Figure 2 change as one would expect. The stopping boundary with respect to the posterior mean of A jumps up as training costs increase, and it retains its peculiar "cupped" shape. Thus, the same observations about two competing forces made in Section 5.6.1 hold here as well. Similarly, an increase in training costs also produces an upward shift of the stopping boundary with respect to $\mathbb{E}[Z_n]$. Naturally, when there are no training costs and $c_h = 0$, the initial jump disappears in both boundaries.

Table 6 shows how the Gittins indices, the fractions of terminated workers, and the long-run average service rates change when $c_h = \{0, 15, 30, 60\}$. It is interesting to note that when training costs are absent, the screening process is very selective and terminates 58.49% of employees on day 1 and 87.36% overall. As training costs enter into the problem, the termination rates quickly drop, and the values of the Gittins indices and of the service rates follow, naturally, the opposite trend.

Switching and quitting costs.

One would expect that changes in switching and quitting costs would similarly affect the optimal policy. However, the theorem below shows that, when the quitting probabilities are constant – so that $q_{i,n} = q$ for all n and for all $i \in \mathcal{S}_0$ – this is not the case.

To state the theorem we need to keep track of how the training, quitting and switching costs affect the Gittins index. To that end, we modify our notation to account for these

Table 5: **Simulation results with different sampling variances.**
(standard errors for the mean in parenthesis)

σ	Gittins index	Fraction of terminated workers				Long-run average service rate
		Day 1	Days 2–10	Days 11 –20	Total	
0.6000	4,993.3	0.0057 *(.0003)*	0.1831 *(.0017)*	0.0595 *(.0011)*	0.2755 *(.0020)*	0.6626 *(.0111)*
0.8000	5,491.7	0.0196 *(.0006)*	0.2830 *(.0020)*	0.0557 *(.0010)*	0.3982 *(.0022)*	0.6417 *(.0138)*
1.0000	6,190.4	0.0534 *(.0010)*	0.3426 *(.0021)*	0.0517 *(.0010)*	0.4961 *(.0022)*	0.6104 *(.0160)*

Table 6: **Simulation results with different training costs.**
(standard errors for the mean in parenthesis)

c_h	Gittins index	Fraction of terminated workers				Long-run average service rate
		Day 1	Days 2–10	Days 11 –20	Total	
0	3,645.2	0.5849 *(.0022)*	0.2600 *(.0020)*	0.0153 *(.0005)*	0.8736 *(.0015)*	0.8316 *(.0153)*
15	4,833.7	0.0833 *(.0012)*	0.3844 *(.0022)*	0.0548 *(.0010)*	0.5546 *(.0022)*	0.6889 *(.0138)*
30	5,491.7	0.0196 *(.0006)*	0.2830 *(.0020)*	0.0557 *(.0010)*	0.3982 *(.0022)*	0.6417 *(.0138)*
45	6,028.9	0.0057 *(.0003)*	0.1949 *(.0018)*	0.0525 *(.0010)*	0.2880 *(.0020)*	0.6122 *(.0136)*
60	6,509.1	0.0017 *(.0002)*	0.1404 *(.0016)*	0.0415 *(.0009)*	0.2201 *(.0019)*	0.5949 *(.0134)*

differences, letting $M(\nu, n, u, c_h, c_s, c_q)$ be the Gittins index (5.18), and

$$\widehat{m}(c_h, c_s, c_q) = M(\widehat{\nu}, 0, 1, c_h, c_s, c_q).$$

Theorem 5.12. *Assume that $\nu_{i,0} = \widehat{\nu}$, $n_{i,0} = 0$, and $u_{i,0} = 1$ for all $i \in \mathcal{S}_0$. Then, if the quitting probabilities are constant, i.e. $q_{i,n} = q$ for all $i \in \mathcal{S}_0$ and all n,*

$$M_i(\nu_{i,t}, n_{i,t}, u_{i,t}, c_h, c_s, c_q) < \widehat{m}(c_h, c_s, c_q)$$

if and only if

$$M_i(\nu_{i,t}, n_{i,t}, u_{i,t}, c_h + c_s, 0, 0) < \widehat{m}(c_h + c_s, 0, 0),$$

for all $t \geq 0$.

Thus, if the hazard rate for quitting is constant for all employees at all times, then changes in switching and quitting costs do not affect the relative ordering of workers' Gittins indices. Of course, the values of the Gittins indices change, as do the (analogous) expected discounted costs of the problem. But because the relative orderings do not change, changes in the switching and quitting costs do not affect the optimal policy, and we therefore do not report a sensitivity analysis with respect to c_s or c_q.

When the quitting probabilities are *not* constant, the specifics of the optimal policy can change with c_s and c_q. Nevertheless, the overall structure of the optimal policy does not change. Theorem 5.3 and Corollary 5.5 hold for any quitting behavior $q_{i,n}$ as in (5.4).

5.7. Conclusions

This paper studies how statistical and on-the-job learning together determine the nature of optimal hiring and retention decisions. Statistical learning arises when workers are heterogeneous and the employer does not know their true quality. On-the-job learning occurs as experience affects workers' performance.

The literature related to this problem comes from various areas, such as labor economics, statistical decision theory, learning-curve theory, and service operations, among others. Our analysis of the hiring and retention problem integrates aspects from all of these streams and adapts the classical Bayesian bandit setup to incorporate training, switching and quitting dynamics. In addition to proving the optimality of an index policy, we show that a "no-recall" property (Corollary 5.8) ensures that worker lifetimes and costs follow the *iid* pattern of a discounted renewal reward process. The *iid* nature of such a sequence allows us to express the optimal infinite-horizon total expected discounted cost as a function of the Gittins index (Theorem 5.9).

Our numerical results show that active screening of employees can significantly improve expected costs and long-run average employee performance. Because most termination takes place early in employees' tenures, relatively simple finite-horizon and one-shot policies also have the potential to perform well. Our sensitivity analysis shows that, as is common in bandit problems, the ability to terminate employees should motivate managers to consider a broader spectrum of potential hires. Moreover, both reductions in within-task variability and improvements in employee learning provide the additional benefit of lowering termination rates.

5.8. Proofs of Mathematical Results

Proofs of mathematical claims are presented in the order of their appearance in the main paper. (The statement "Proof of ..." is presented in bold face). When other technical results are needed, they are stated with a full proof or suitable reference, in the location that they are needed (the result is presented in standard typeface).

To simplify the exposition, we introduce the following shorthand. For any given initial state, $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$, let $M_i \equiv M(\nu_{i,0}, n_{i,0}, u_{i,0})$ denote the initial value of worker i's index, $\widetilde{\Lambda}_i(m) \equiv \widetilde{\Lambda}(\nu_{i,0}, n_{i,0}, u_{i,0}, m)$ be the stopping time (5.16), $HV_i(m) \equiv HV(\nu_{i,0}, n_{i,0}, u_{i,0}, m)$ be i's expected continuation cost (5.17), and $C_{i,t} \equiv c_s u_{i,t} + c_h \mathbb{1}(n_{i,t} = 0) + c\left(Z(\nu_{i,t}, n_{i,t})\right) +$

$(c_q - c_s)\mathbb{1}(\nu_{\pi(t-1),t} = \mathbb{1}_K \cap t > 0)$ be worker i's one-period cost for being employed at time t.

Proof of Lemma 5.1.

For (i), let π be a hiring and retention policy for Problem 2 that employs unproductive workers. Then, let $T = \inf\{t : \ n_{\pi(t),t} \geq \Lambda_{\pi(t)}\}$ to be the first time that such a worker is employed. The one-period cost at time T for employing the unproductive worker $\pi(T)$ is $\gamma^T K$. Construct a new policy π^T such that $\pi^T(t) = \pi(t)$ for $t < T$, and $\pi^T(t) = \pi(t+1)$ for $t \geq T$. For any initial state $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ we have that

$$\begin{aligned} C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) &= -c_s + \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t C_{\pi(t),t} + \gamma^T K + \sum_{t=T+1}^{\infty} \gamma^t C_{\pi(t),t}\right] \\ C^0_{\pi^T}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) &= -c_s + \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t C_{\pi(t),t} + \sum_{t=T+1}^{\infty} \gamma^{t-1} C_{\pi(t),t}\right], \end{aligned}$$

and

$$\begin{aligned} C^0_{\pi^T}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) - C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) &= \mathbb{E}\left[(1-\gamma)\left(\sum_{t=T+1}^{\infty} \gamma^{t-1} C_{\pi(t),t}\right) - \gamma^T K\right] \\ &< \mathbb{E}\left[(1-\gamma)\left(\sum_{t=T+1}^{\infty} \gamma^{t-1} K\right) - \gamma^T K\right] \\ &= \mathbb{E}\left[(1-\gamma)\frac{\gamma^T K}{1-\gamma} - \gamma^T K\right] = 0. \end{aligned}$$

Thus, the infinite horizon total expected discounted cost of π^T is strictly smaller than that of π and π cannot be optimal.

For (ii) we begin with the *if* part. *If:* Let π^* be optimal for Problem 2. Then by part (i) of the lemma, policy π^* employs no unproductive worker and, therefore, π^* is feasible for Problem 1. For any initial state $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ and for any policy π feasible for Problem 1, we note that π is also feasible for Problem 2, and we let $C^0_{\pi,1}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ and $C^0_{\pi,2}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ respectively be the infinite-horizon total expected discounted cost of policy π in Problem 1 and 2 and

we observe that $C^0_{\pi,1}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) = C^0_{\pi,2}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$. Because π^* is optimal for Problem 2 and feasible for Problem 1, we obtain that $C^0_{\pi^*,1}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) = C^0_{\pi^*,2}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) \leq C^0_{\pi,2}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) = C^0_{\pi,1}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$ for all π feasible for Problem 1. Hence, π^* is also optimal for Problem 1.

Only if: Let π^* be optimal for Problem 1. Then, any policy, π, that is feasible for Problem 1 is feasible for Problem 2, and $C^0_{\pi,1}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) = C^0_{\pi,2}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$. By part (i) of the Lemma, we know that any policy π that is feasible for Problem 2 but not for Problem 1 cannot be optimal. Then, by optimality and feasibility we obtain that $C^0_{\pi^*,2}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) = C^0_{\pi^*,1}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) \leq C^0_{\pi,1}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) = C^0_{\pi,2}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$, and policy π^* is also optimal for Problem 2.

Proof of Lemma 5.2.

Suppose that $\pi \in \Pi$ is a policy for Problem 1 and that $\mathbb{E}[\Lambda_i] < \infty$ for all $i \in \mathcal{S}_0$. No policy for Problem 1 can use an employee after he has quit. Thus, the random variable $\Lambda_i(\pi)$ in (5.3) satisfies $0 \leq \Lambda_i(\pi) \leq \Lambda_i$ on every sample path, for all $i \in \mathcal{S}_0$. Suppose, by contradiction, that policy $\pi \in \Pi$ only employs $\kappa < \infty$ workers with some positive probability $\epsilon > 0$. Because $\pi \in \Pi$ we have

$$\mathbb{P}\left(\sum_{i=1}^{\kappa} \Lambda_i(\pi) \geq \zeta\right) \geq \epsilon \tag{5.22}$$

for all $\zeta \in \mathbb{R}$. Given that $0 \leq \Lambda_i(\pi) \leq \Lambda_i$ and that the Λ_i's are *iid*, Markov's inequality implies that $\mathbb{P}\left(\sum_{i=1}^{\kappa} \Lambda_i(\pi) \geq \zeta\right) \leq \mathbb{P}\left(\sum_{i=1}^{\kappa} \Lambda_i \geq \zeta\right) \leq \kappa\mathbb{E}[\Lambda_1]/\zeta$. Picking any $\zeta > \kappa\mathbb{E}[\Lambda_1]/\epsilon$ would contradict (5.22) from which we conclude that $\pi \notin \Pi$. Hence, each policy for Problem 1 employs an infinite number of workers with probability 1.

Properties of the Value Function and of the Gittins Index.

Lemma 5.13. *For each ν, n, and u, $V(\nu,n,u,m)$ is concave, non-decreasing and Lipschitz continuous in m, with Lipschitz constant equal to 1.*

Proof. We proceed by means of the Value Iteration Algorithm (see, e.g. Bertsekas and

Shreve, 1978, Section 9.5, Definition 9.10 and Proposition 9.14). Let $v^0(\nu, n, u, m) = 0$ for all $m \in \mathbb{R}$, and notice that v^0 is trivially nondecreasing, concave, and Lipschitz-1 continuous in m for each ν, n, and u. Assume that $v^{k-1}(\nu, n, u, m)$ is nondecreasing, concave, and Lipschitz-1 continuous in m for each ν, n, and u. Let

$$\begin{aligned} v^k(\nu, n, u, m) = \min \Big\{ m,\ & c_s u + c_h \mathbb{1}(n = 0) + \mathbb{E}[c(Z(\nu, n))] \\ & + \gamma(1 - q_n)\mathbb{E}[v^{k-1}(\beta(\nu, Z(\nu, n)), n + 1, 0, m)] \\ & + \gamma q_n \Big[c_q - c_s + v^{k-1}(\mathbb{1}_K, n + 1, 0, m) \Big] \Big\}, \end{aligned}$$

and notice that $c_s u + c_h \mathbb{1}(n = 0) + \mathbb{E}[c(Z(\nu, n))]$ is constant with respect to m, $\gamma(1 - q_n)\mathbb{E}[v^{k-1}(\beta(\nu, Z(\nu, n)), n + 1, 0, m)]$ is nondecreasing, concave, and Lipschitz-$\gamma(1 - q_n)$ continuous in m by the induction assumption and the fact that these properties are preserved when taking expectations.The induction assumption also yields that $\gamma q_n \big[c_q - c_s + v^{k-1}(\mathbb{1}_K, n+1, 0, m) \big]$ is nondecreasing, concave, and Lipschitz-γq_n continuous in m. Monotonicity and concavity are preserved under minimization, so we have that $v^k(\nu, n, u, m)$ is nondecreasing and concave in m.

To obtain that $v^k(\nu, n, u, m)$ is also Lipschitz-1 continuous in m the argument is similar, but a little more care is required. Given two Lipschitz functions h, h' with Lipschitz constants c_1, c_2 respectively, $\min\{h, h'\}$ is Lipschitz with constant $c_3 = \max\{c_1, c_2\}$. In our context, the left minimand is Lipschitz-1 continuous, and the right minimand is Lipschitz-γ continuous, with $\gamma < 1$, so that $v^k(\nu, n, u, m)$ is also Lipschitz-1 continuous in m. To conclude our argument, we let $k \to \infty$ so $v^k(\nu, n, u, m) \to V(\nu, n, u, m)$. □

Lemma 5.14. (i) *$HV(\nu, n, u, m) < m$ if and only if $M(\nu, n, u) < m$.* (ii) *$HV(\nu, n, u, m) > m$ if and only if $m < M(\nu, n, u)$.* (iii) *$HV(\nu, n, u, m) = m$ if and only if $m = M(\nu, n, u)$.*

Proof. We prove each of the three statements in turn. (i) If $M(\nu, n, u) < m$ then $V(\nu, n, u, m) < m$. In turn, $V(\nu, n, u, m) < m$ implies it is optimal not to retire so $HV(\nu, n, u, m) = V(\nu, n, u, m) < m$. If $HV(\nu, n, u, m) < m$, we have that $HV(\nu, n, u, m) =$

$V(\nu, n, u, m) < m$. Then the fact that $M(\nu, n, u) < m$ follows by the definition of the Gittins index (5.18), $M(\nu, n, u)$, and the fact that the Bellman equation (5.13), $V(\nu, n, u, m)$, is concave and non-decreasing in m with $V(\nu, n, u, m) \leq m$ for all m. (ii) It follows directly from the proof of (i) by reversing the inequalities. (iii) It follows combining claims (i) and (ii). □

Lemma 5.15. *For each ν, and n $M(\nu, n, 0) \leq M(\nu, n, 1)$.*

Proof. Because $c_s \geq 0$ it is immediate to see that $V(n, \nu, 0, m) \leq V(n, \nu, 1, m)$ for each m. Then, given the monotonicity property of the value function $V(\nu, n, u, m)$ in m for each given n, ν, u (Lemma 5.13), we have that $M(\nu, n, 0) = \sup\{m : \ V(n, \nu, 0, m) = m\} \leq \sup\{m : \ V(n, \nu, 1, m) = m\} = M(n, \nu, 1)$. □

Proof of Theorem 5.3

Given the initial state $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ such that $\nu_{i,0} \equiv \widehat{\nu}$, $n_{i,0} \equiv 0$, and $u_{i,0} \equiv 1$ for all $i \in \mathcal{S}_0$, we have that all workers have index $M(\widehat{\nu}, 0, 1) \equiv \widehat{m}$. Thus, at any time t there are at most t workers who have been employed, so that there are at most t indices with values different than $\widehat{m}$. Hence, for each $t = 0, 1, 2, \ldots$, the infimum in Theorem 5.3 is attained and the index policy described in Theorem 5.3 is well defined.

Recall that $c_s = 0$ by hypothesis. Then, the optimality equation (5.13) and the expected discounted cost of continuing (5.14) are constant with respect to u, i.e. $V(\nu, n, 0, m) = V(\nu, n, 1, m)$ and $HV(\nu, n, 0, m) = HV(\nu, n, 1, m)$ for all ν, n, m. We also have $M(\nu, n, 0) = M(\nu, n, 1)$, and the value of the Gittins index of a given worker is independent from that of other workers.

To prove Theorem 5.3, we now introduce some additional notation. We let $\pi(j)$ be the hiring and retention policy that begins by employing worker j and continues according to the index rule. We also let $\pi(i, j)$ be the policy that first employs worker i (with ability distribution $\nu_{i,0}$, experience $n_{i,0}$, and switching indicator $u_{i,0}$) as long as his Gittins index

does not exceed its original value, $M(\nu_{i,0}, n_{i,0}, u_{i,0})$. Policy $\pi(i,j)$ then employs worker j for at least one period, until j's index exceeds the original value of worker i's index, $M(\nu_{i,0}, n_{i,0}, u_{i,0})$. After employing worker i and j as described, policy $\pi(i,j)$ continues according to the index rule.

Lemmas 5.16-5.18 study the cost of the employment policies $\pi(i,j), \pi(j,i), \pi(i)$, and $\pi(j)$. The lemmas hold for any initial state, $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$, such that there are infinitely many workers, i, with $\nu_{i,0} = \widehat{\nu}$, $n_{i,0} = 0$, and $u_{i,0} = 1$.

Lemma 5.16. *If $M_i = M_j$ then $C^0_{\pi(i,j)}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) = C^0_{\pi(j,i)}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$.*

Proof. By construction, the infinite-horizon expected discounted cost of policy $\pi(i,j)$ is

$$\begin{aligned} C^0_{\pi(i,j)}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) &= \mathbb{E}\left[\sum_{t=0}^{\widetilde{\Lambda}_i(M_i)-1} \gamma^t C_{i,t} + \sum_{t=\widetilde{\Lambda}_i(M_i)}^{\widetilde{\Lambda}_i(M_i)+\widetilde{\Lambda}_j(M_i)-1} \gamma^t C_{j,t} + \sum_{t=\widetilde{\Lambda}_i(M_i)+\widetilde{\Lambda}_j(M_i)}^{\infty} \gamma^t C_{\pi(t),t}\right] \\ &= HV_i(M_i) + \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_i(M_i)}\right]\left\{-M_i + HV_j(M_i) - \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_j(M_i)}\right] M_i\right\} \\ &\quad + C^{\widetilde{\Lambda}_i(M_i)+\widetilde{\Lambda}_j(M_i)}_{\pi(i,j)}(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}') \end{aligned} \tag{5.23}$$

where $C^{\widetilde{\Lambda}_i(M_i)+\widetilde{\Lambda}_j(M_i)}_{\pi(i,j)}(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}')$ is the expected discounted (to $t = 0$) continuation cost of policy $\pi(i,j)$ after having employed worker i for $\widetilde{\Lambda}_i(M_i)$ periods, and worker j for $\widetilde{\Lambda}_j(M_i)$ periods. Because only workers i and j have been employed, the new state, $(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}')$, differs from $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ only in its ith and jth coordinates. Similarly,

$$\begin{aligned} C^0_{\pi(j,i)}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) &= HV_j(M_j) + \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_j(M_j)}\right]\left\{-M_j + HV_i(M_j) - \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_i(M_j)}\right] M_j\right\} \\ &\quad + C^{\widetilde{\Lambda}_j(M_j)+\widetilde{\Lambda}_i(M_j)}_{\pi(j,i)}(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}'). \end{aligned}$$

Because $M_i = M_j$ we have that, at time $\widetilde{\Lambda}_i(M_i) + \widetilde{\Lambda}_j(M_i)$, the continuation costs

$$C^{\widetilde{\Lambda}_i(M_i)+\widetilde{\Lambda}_j(M_i)}_{\pi(i,j)}(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}') \quad \text{and} \quad C^{\widetilde{\Lambda}_j(M_j)+\widetilde{\Lambda}_i(M_j)}_{\pi(j,i)}(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}')$$

are equal. Moreover, we can use Lemma 5.14 to obtain that $HV_j(M_i) = M_i = HV_i(M_i)$

and $HV_i(M_j) = M_j = HV_j(M_j)$ so that

$$C^0_{\pi(i,j)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) - C^0_{\pi(j,i)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) \\ = M_i - \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_i(M_i)}\right]\mathbb{E}\left[\gamma^{\widetilde{\Lambda}_j(M_i)}\right]M_i - M_i + \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_j(M_i)}\right]\mathbb{E}\left[\gamma^{\widetilde{\Lambda}_i(M_i)}\right]M_i = 0,$$

as desired. □

Lemma 5.17. *If $M_i = \inf_k M_k$ and $M_i < M_j$ then $C^0_{\pi(i,j)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) < C^0_{\pi(j)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$.*

Proof. Policy $\pi(j)$ employs worker j for the first period and then continues according to the index rule. After his first performance, worker j is retained as long as he is index minimal. When worker j is terminated, Lemma 5.16, tells us that we can choose policy $\pi(j)$ to employ worker i, and continuing with the index rule. Thus,

$$\begin{aligned} C^0_{\pi(j)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) &= HV_j(M_i) + \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_j(M_i)}\right]\left\{-M_i + HV_i(M_i) - \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_i(M_i)}\right]M_i\right\} \\ &\quad + C^{\widetilde{\Lambda}_j(M_i)+\widetilde{\Lambda}_i(M_i)}_{\pi(j)}(\boldsymbol{\nu}',\boldsymbol{n}',\boldsymbol{u}'), \end{aligned}$$

where $C^{\widetilde{\Lambda}_j(M_i)+\widetilde{\Lambda}_i(M_i)}_{\pi(j)}(\boldsymbol{\nu}',\boldsymbol{n}',\boldsymbol{u}')$ is the expected discounted continuation cost of policy $\pi(j)$ after having employed worker j for $\widetilde{\Lambda}_j(M_i)$ periods, and worker i for $\widetilde{\Lambda}_i(M_i)$ periods. The new state, $(\boldsymbol{\nu}',\boldsymbol{n}',\boldsymbol{u}')$, differs from $(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$ only in his jth and ith coordinates.

We now recall the representation (5.23) for the expected cost of policy $\pi(i,j)$, and we observe that the expected continuation costs $C^{\widetilde{\Lambda}_i(M_i)+\widetilde{\Lambda}_j(M_i)}_{\pi(i,j)}(\boldsymbol{\nu}',\boldsymbol{n}',\boldsymbol{u}')$ and $C^{\widetilde{\Lambda}_j(M_i)+\widetilde{\Lambda}_i(M_i)}_{\pi(j)}(\boldsymbol{\nu}',\boldsymbol{n}',\boldsymbol{u}')$ are equal. From Lemma 5.14 we know that $HV_i(M_i) = M_i$. Because $M_i < M_j$ we also have $M_i < HV_j(M_i)$. Then,

$$C^0_{\pi(i,j)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) - C^0_{\pi(j)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) = [M_i - HV_j(M_i)]\left(1 - \mathbb{E}\left[\gamma^{\widetilde{\Lambda}_i(M_i)}\right]\right) < 0,$$

as desired. □

Lemma 5.18. *If $M_i = \inf_k M_k$ and $M_i < M_j$ then $C^0_{\pi(i)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) < C^0_{\pi(j)}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$.*

Proof. Because $M_i = \inf_k M_k$ and $M_i < M_j$, Lemma 5.17 tells us that policy $\pi(i,j)$ strictly improves policy $\pi(j)$. We now argue that $\pi(i,j)$ can be improved by employing a Gittins index minimal worker at all times. The first worker that is employed by policy $\pi(i,j)$, i, is Gittins index minimal. At his termination, the state of the system changes from the initial $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ to $(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}')$, which differs only in the ith coordinate. After the employment of worker i, policy $\pi(i,j)$ prescribes the employment of worker j. Its continuation value then equals that of policy $\pi(j)$ when starting in state $(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}')$. Lemma 5.17 then tells us that if j is not Gittins index minimal at $(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}')$ then, it is strictly better to use the policy $\pi(\ell, j)$ where worker ℓ is such that $M_\ell = \inf_k M_k$, and $M_\ell < M_j$. Iterating on this reasoning we obtain that policy $\pi(i)$, the index policy, is strictly better than any index policy in that begins with a worker that is not index minimal. □

We are now ready to complete the proof of Theorem 5.3.

Proof of Theorem 5.3. "If:" Let π be any employment policy and consider the policy π^T such that $\pi^T(t) = \pi(t)$ for all $0 \le t < T$ and $\pi^T(t) = \pi^*(t)$ for $T \le t$, where π^* denotes the index rule. At any time T the system is in state $(\boldsymbol{\nu}', \boldsymbol{n}', \boldsymbol{u}')$ which is different from the initial $(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ in at most T coordinates. Thus, there are infinitely many workers whose state has never changed, and whose index equals $\widehat{m}$, so that policy π^T is well defined. Because the problem is discounted ($\gamma < 1$) and the one-period costs are uniformly bounded, we can pick any $\epsilon > 0$ and choose T so that $C^0_{\pi^T}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) - C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) < \epsilon$. Then, according to Lemma 5.18, we might improve policy π^T by employing a Gittins-index minimal worker at time $T-1$. Thus $C^0_{\pi^{T-1}}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) \le C^0_{\pi^T}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ and also $C^0_{\pi^{T-1}}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) - C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) < \epsilon$. Iterating back to $T = 1$ we have $C^0_{\pi^0}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) - C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) < \epsilon$, where π^0 is the index policy π^*. Because ϵ is arbitrary we then have $C^0_{\pi^0}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) \le C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$. Because the choice of policy π was also arbitrary, we can choose π to be any optimal policy so that $C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) \le C^0_{\pi^0}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}) \le C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$. Thus, the index policy π^0 is optimal too.

"Only if:" Let π be an optimal policy, and assume that π is *not* an index policy. Let T be

the first time at which π does not employ a Gittins-index minimal worker, and construct the policy $\widehat{\pi}$ such that $\widehat{\pi}(t) = \pi(t)$ for all $0 \leq t \leq T$ and $\widehat{\pi}(t) = \pi^*(t)$ for all $T < t$, where, as usual, π^* denotes the index policy. Because both π and π^* are optimal, policy $\widehat{\pi}$ is optimal too. However, by Lemma 5.18 we can strictly improve on policy $\widehat{\pi}$ by selecting an index minimal worker at time T, and by doing so we obtain that policies $\widehat{\pi}$ and π cannot be optimal, a contradiction. □

Proof of Corollary 5.5

To prove Corollary 5.5, let $C^0_\pi(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_h, c_s)$ be the cost function (5.9) that makes explicit the dependence on the training cost, c_h, and on the switching cost c_s. We know by Theorem 5.3 that $C^0_\pi(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_h, 0)$ is minimized if and only if π is an index policy. Similarly, the same happens for $C^0_\pi(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_s + c_h, 0)$ because we are just imposing a different training cost, $c_s + c_h$. For all policy $\pi \in \Pi$, we then have that

$$C^0_{\pi^*}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_s + c_h, 0) \leq C^0_\pi(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_s + c_h, 0) \leq c_s + C^0_\pi(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_h, c_s). \tag{5.24}$$

The first inequality holds by the optimality of policy π^*. The second inequality holds because the switching cost, c_s, is incurred every time the workers employed in two subsequent periods differ (not only at the first employment of a new worker). The second inequality is met with equality for all policies π that never recall previously employed workers.

"If:" We now show that if π is the index policy in Corollary 5.5, then $c_s + C^0_\pi(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$ achieves the lower bound (5.24). At time $t = 0$ all workers have the same index, $\widehat{m}$, and the employer chooses a worker, i, at random from the pool. Worker i is then employed for $\widetilde{\Lambda}_i(\widehat{m})$ periods, and his index $M_i(\nu_{i,\widetilde{\Lambda}_i(\widehat{m})}, n_{i,\widetilde{\Lambda}_i(\widehat{m})}, 0) > \widehat{m}$. Because worker i is not index minimal at time $\widetilde{\Lambda}_i(\widehat{m})$, another worker, j, is employed. This causes a transition of the state of worker i, from $(\nu_{i,\widetilde{\Lambda}_i(\widehat{m})}, n_{i,\widetilde{\Lambda}_i(\widehat{m})}, 0)$ to $(\nu_{i,\widetilde{\Lambda}_i(\widehat{m})+1}, n_{i,\widetilde{\Lambda}_i(\widehat{m})+1}, 1)$, with $\nu_{i,\widetilde{\Lambda}_i(\widehat{m})} = \nu_{i,\widetilde{\Lambda}_i(\widehat{m})+1}$, and $n_{i,\widetilde{\Lambda}_i(\widehat{m})} = n_{i,\widetilde{\Lambda}_i(\widehat{m})+1}$. By Lemma 5.15 we know that $M(n, \nu, 0) \leq M(n, \nu, 1)$ for each ν, n. Because worker i in state $(\nu_{i,\widetilde{\Lambda}_i(\widehat{m})}, n_{i,\widetilde{\Lambda}_i(\widehat{m})}, 0)$ has index exceeding $\widehat{m}$, the same happens

to worker i when in state $(\nu_{i,\widetilde{\Lambda}_i(\widehat{m})+1}, n_{i,\widetilde{\Lambda}_i(\widehat{m})+1}, 1)$.

Repeating this argument for all employed workers, we see that the transition of u from 0 to 1 only increases the indices of workers whose indices are greater than $\widehat{m}$ and, in turn does not change the dynamics of the index policy which then agrees with the index policy, π^*, used to achieve $C^0_{\pi^*}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_s + c_h, 0)$.

"Only if:" Assume that π is an optimal policy for $C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_h, c_s)$. From the "if" part of the proof, we know that an optimal π satisfies

$$C^0_{\pi^*}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_s + c_h, 0) = c_s + C^0_{\pi}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u}, c_h, c_s),$$

i.e. it achieves the lower bound (5.24). Then π is also an optimal policy for $C^0_{\pi^*}(\boldsymbol{\nu}, \boldsymbol{n}, \boldsymbol{u})$, and Theorem 5.3 tells us that π must be an index policy.

Proof of Corollary 5.7.

At $t = 0$, no worker has ever been employed and all the workers have Gittins index $\widehat{m}$. Then, the sampling process starts with a random selection of worker, i, from the stationary pool of candidates. Worker i is employed at all times, t, such that $M_i(\nu_{i,t}, n_{i,t}, u_{i,t}) = \inf_j\{M_j(\nu_{j,t}, n_{j,t}, u_{j,t})\} \leq \widehat{m}$. As soon as i is discarded, $M_i(\nu_{i,t}, n_{i,t}, u_{i,t}) > \widehat{m}$ and the sampling process starts again.

Proof of Corollary 5.8.

It follows immediately from Lemma 5.1 and Theorem 5.3.

Proof of Theorem 5.9.

Consider the retirement-option problem described in Section 5.3. By Lemma 5.14-(iii), we obtain $\widehat{m} = HV(\widehat{\nu}, 0, 1, \widehat{m})$, and we note that $HV(\widehat{\nu}, 0, 1, \widehat{m})$ is the total expected discounted cost of employing a productive worker, i, with ability distribution, $\nu_{i,0} = \widehat{\nu}$, experience

$n_{i,0} = 0$, and switching indicator $u_{i,0} = 1$ for at least one period followed by an optimal termination decision that depends on the retirement payment $\widehat{m}$. Recall now the definition of the optimal stopping time $\widetilde{\Lambda}(\nu, n, u, m)$ in (5.16) and the stopping-time representation for $HV(\nu, n, u, m)$ in (5.17). Thus

$$HV(\widehat{\nu}, 0, 1, \widehat{m}) = \mathbb{E}\left[c_s + c_h + \sum_{r=0}^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})-1} \gamma^r c(Z(\nu_r, r)) + \gamma^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})}\left[(c_q - c_s)\mathbb{1}_{Q_0} + \widehat{m}\right]\right].$$

Because $\widehat{m} = HV(\widehat{\nu}, 0, 1, \widehat{m})$, we obtain

$$\begin{aligned}\left(1 - \mathbb{E}\left[\gamma^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})}\right]\right)\widehat{m} \qquad & (5.25)\\ = \mathbb{E}\left[c_s + c_h + \sum_{r=0}^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})-1} \gamma^r c(Z(\nu_r, r)) + \gamma^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})}(c_q - c_s)\mathbb{1}_{Q_0}\right].\end{aligned}$$

At time $t = 0$ all workers $i \in \mathcal{S}_0$ have ability distribution, $\nu_{i,0} = \widehat{\nu}$, experience $n_{i,0} = 0$, and switching indicator $u_{i,0} = 1$. Theorem 5.3 tells us that worker i can be optimally retained at time t if and only if his Gittins-index is minimal, i.e. $M_i(\nu_{i,t}, n_{i,t}, u_{i,t}) \leq \widehat{m}$. Worker i stops being employed at time $\widetilde{\Lambda}_i(\widehat{\nu}, 0, 1, \widehat{m})$ either because he is terminated or he quits. Because all workers $i \in \mathcal{S}_0$ are identical, the sequence $\{\widetilde{\Lambda}_i \equiv \widetilde{\Lambda}_i(\widehat{\nu}, 0, 1, \widehat{m}),\ i = 1, 2, 3, \ldots\}$ is *iid*. Set $\widetilde{\Lambda}_0 \equiv 0$, recall that Λ_i is the time at which worker i becomes unproductive, and let $Q_{i,0} = \{\omega :\ \widetilde{\Lambda}_i(\widehat{\nu}, 0, 1, \widehat{m}) = \Lambda_i\}$ be the set of sample paths for which worker i quits before

he is terminated. Then,

$$\begin{aligned} c_s + \inf_{\pi\in\Pi} C_\pi^0(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) &= \mathbb{E}\Bigg[\sum_{k=1}^{\infty}\gamma^{\sum_{i=0}^{k-1}\widetilde{\Lambda}_i}\Bigg(\sum_{r=0}^{\widetilde{\Lambda}_k-1}\gamma^r\left\{(c_s+c_h)\mathbb{1}(r=0)+c\left(Z(\nu_r,r)\right)\right\} \\ &\qquad + \gamma^{\widetilde{\Lambda}_k}(c_q-c_s)\mathbb{1}_{Q_{k,0}}\Bigg)\Bigg] \\ &= \sum_{k=1}^{\infty}\mathbb{E}\left[\gamma^{\sum_{i=0}^{k-1}\widetilde{\Lambda}_i}\right]\mathbb{E}\Bigg[\sum_{r=0}^{\widetilde{\Lambda}_k-1}\gamma^r\left\{(c_s+c_h)\mathbb{1}(r=0)+c\left(Z(\nu_r,r)\right)\right\} \\ &\qquad + \gamma^{\widetilde{\Lambda}_k}(c_q-c_s)\mathbb{1}_{Q_{k,0}}\Bigg] \\ &= \widehat{m}\left(1-\mathbb{E}\left[\gamma^{\widetilde{\Lambda}_1}\right]\right)\sum_{k=1}^{\infty}\mathbb{E}\left[\gamma^{\sum_{j=0}^{k-1}\widetilde{\Lambda}_j}\right] \\ &= \widehat{m} \end{aligned} \tag{5.26}$$

where (5.26) follows from (5.25), and $\sum_{k=1}^{\infty}\mathbb{E}\left[\gamma^{\sum_{j=0}^{k-1}\widetilde{\Lambda}_j}\right] = 1+\mathbb{E}[\gamma^{\widetilde{\Lambda}_1}]+\mathbb{E}[\gamma^{\widetilde{\Lambda}_1}]^2+\ldots = \left(1-\mathbb{E}\left[\gamma^{\widetilde{\Lambda}_1}\right]\right)^{-1}$.

Proof of Proposition 5.10.

We first prove the following lemma that uses the notion of a likelihood ratio order (Shaked and Shanthikumar, 2007, Section 1.C). Suppose that X is a random variable with probability density function (pdf) f_X and that Y is a random variable with pdf f_Y. We write $X \leq_{\text{lr}} Y$ (X is stochastically smaller than Y in the likelihood ratio sense) if $f_Y(z)/f_X(z)$ increases in z over the union of the supports of X and Y.

Lemma 5.19. *Let $g:\mathbb{R}^3\to\mathbb{R}$ be such that for $A\sim\nu$, $\beta(\nu,z)([-\infty,a]) = \mathbb{P}\left(A\leq a|Z=z\right)$ is nondecreasing in z for any given ν. If, for any $a\leq a'$, $\xi_n(z|a')/\xi_n(z|a)$ is nondecreasing in z, then $V(\nu,n,u,m)\leq V(\nu',n,u,m)$, for any $\nu\leq_{\text{lr}}\nu'$, and for each given n,u,m.*

The monotonicity of the Bayes operator ensures that the Bayesian update implies that larger observations lead to stochastically larger posterior distributions in some sense. Notice also that, for several well-known families of distributions, the likelihood ratio comparison can be simply checked by comparing distribution parameters. Müller and Stoyan (2002, Table

1.1) proposes such comparison criteria for several continuous and discrete distributions.

Proof of Lemma 5.19. To show monotonicity of the value function (5.13) with respect to the likelihood ratio order, we proceed by means of the Value Iteration Algorithm (see, e.g. Bertsekas and Shreve, 1978, Section 9.5, Definition 9.10 and Proposition 9.14). We fix n, u, and m, and we let $v^0(\nu, n, u, m) = 0$ for all distributions ν. Trivially, we have that v^0 is lr-nondecreasing in ν. We then assume that $v^{k-1}(\nu, n, u, m) \leq v^{k-1}(\nu', n, u, m)$ for $\nu \leq_{\text{lr}} \nu'$, and we write

$$\begin{aligned} v^k(\nu, n, u, m) = \min\{m,\ & c_s u + c_h \mathbb{1}(n = 0) + \mathbb{E}[c(Z(\nu, n))] \\ & + \gamma(1 - q_n)\mathbb{E}[v^{k-1}(\beta(\nu, Z(\nu, n)), n + 1, 0, m)] \\ + \gamma q_n[c_q - c_s + v^{k-1}(\mathbb{1}_K, n + 1, 0, m)]\}. & \end{aligned} \tag{5.27}$$

In equation (5.27), we first notice that the quantity $c_s u + c_h \mathbb{1}(n = 0) + \gamma q_n[c_q - c_s + v^{k-1}(\mathbb{1}_K, n + 1, 0, m)]$ is independent, hence constant, with respect to ν and ν'.

Because $\xi_n(z|a')/\xi_n(z|a)$ is nondecreasing in z for any $a \leq a'$, the definition of the likelihood ratio order yields that $Z(a, n) \leq_{\text{lr}} Z(a', n)$. From Shaked and Shanthikumar (2007, Theorem 1.C.17) and the fact that $\nu \leq_{\text{lr}} \nu'$ we obtain that $Z(\nu, n) \leq_{\text{lr}} Z(\nu', n)$. We now also have that $\mathbb{E}[Z(\nu, n)] \leq \mathbb{E}[Z(\nu', n)]$ because the likelihood ratio order ($\leq_{\text{lr}}$) implies the usual stochastic order ($\leq_{\text{st}}$) (Shaked and Shanthikumar, 2007, Theorem 1.C.1).

Finally, by noting that $Z(\nu, n) \leq_{\text{lr}} Z(\nu', n)$ we obtain that

$$\beta(\nu, Z(\nu, n)) \leq_{\text{lr}} \beta(\nu, Z(\nu', n)) \leq_{\text{lr}} \beta(\nu', Z(\nu', n)).$$

The first ordering holds because $\beta(\nu, z)$ is nondecreasing in z (Shaked and Shanthikumar, 2007, Theorem 1.C.8). The second ordering holds because $\nu \leq_{\text{lr}} \nu'$ (Shaked and Shanthikumar, 2007, Example 1.C.58). Then, the induction assumption and the monotonicity property of the expected value yield that $\mathbb{E}[v^{k-1}(\beta(\nu, Z(\nu, n)), n + 1, 0, m)] \leq$

$\mathbb{E}[v^{k-1}(\beta(\nu', Z(\nu', n)), n+1, 0, m)]$. Thus, the right minimand in (5.27) is lr-nondecreasing in ν. The first minimand, m, is constant with respect to ν, so that $v^k(\nu, n, u, m) \leq v^k(\nu', n, u, m)$, provided that $\nu \leq_{\text{lr}} \nu'$. Repeated application of the Value Iteration Algorithm then yields $V(\nu, n, u, m) \leq V(\nu', n, u, m)$, for any $\nu \leq_{\text{lr}} \nu'$, as desired. □

Proof of Proposition 5.10. The posterior distribution of A has distribution $N(w_p, \sigma^2/p)$. The normal distribution has the monotone likelihood ratio property required by Lemma 5.19 (see, e.g. Müller and Stoyan, 2002, Table 1.1). An application of that lemma proves the desired monotonicity for V.

For the Gittins index we have the following. Given $\nu \sim N(w_p, \sigma^2/p)$ and $\nu' \sim N(w'_p, \sigma^2/p)$ with $w_p \leq w'_p$, we have that $\nu \leq_{\text{lr}} \nu'$, so $V(\nu, n, u, m) \leq V(\nu', n, u, m)$ for any n, u, and m. Then,

$$M(\nu, n, u) = \sup\{m : \ V(\nu, n, u, m) = m\} \leq \sup\{m : \ V(\nu', n, u, m) = m\} = M(\nu', n, u),$$

as desired. □

Proof of Theorem 5.12.

For each given ν, n, u and m, recall the definition of the stopping time $\widetilde{\Lambda}(\nu, n, u, m)$ in (5.16). Also, recall that Λ_i is the time at which i becomes unproductive, and let $Q_{i,0} = \{\omega : \ \widetilde{\Lambda}_i(\widehat{\nu}, 0, 1, \widehat{m}) = \Lambda_i\}$ be the set of sample paths for which worker i with state $(\widehat{\nu}, 0, 1)$ quits before he is terminated. At time $t = 0$ all workers $i \in \mathcal{S}_0$ have $\nu_{i,0} = \widehat{\nu}$, $n_{i,0} = 0$, and $u_{i,0} = 1$. From the proof of Theorem 5.9 we know that the sequence $\{\widetilde{\Lambda}_i(\widehat{\nu}, 0, 1, \widehat{m}), \ i = 1, 2, 3, \ldots\}$ is *iid*. The $\{\Lambda_i, \ i = 1, 2, 3, \ldots\}$ are also *iid*, so that the $Q_{i,0}$'s are *iid* too.

Because $q_{i,n} = q$ for all $i \in \mathcal{S}_0$ and all n, the proof of this result hinges on showing that

$$\mathbb{E}\left[\sum_{r=0}^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})-1} \gamma^{r+1} q\right] = \mathbb{E}\left[\gamma^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})} \mathbb{1}_{Q_0}\right], \tag{5.28}$$

which would imply that

$$\mathbb{E}\Bigg[\sum_{r=0}^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})-1}\gamma^r c(Z(\nu_r,r))+\gamma^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})}(c_q-c_s)\mathbb{1}_{Q_0}\Bigg] \tag{5.29}$$
$$=\mathbb{E}\left[\sum_{r=0}^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})-1}\gamma^r\left\{c(Z(\nu_r,r))+\gamma q(c_q-c_s)\right\}\right].$$

This would give an alternative representation for $C^0_{\pi^*}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$. Under the optimal employment policy

$$C^0_{\pi^*}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$$
$$=-c_s+\mathbb{E}\Bigg[\sum_{k=1}^{\infty}\gamma^{\sum_{i=0}^{k-1}\widetilde{\Lambda}_i}\Bigg(\sum_{r=0}^{\widetilde{\Lambda}_k-1}\gamma^r\left\{(c_s+c_h)\mathbb{1}(r=0)+c\left(Z(\nu_r,r)\right)\right\}+\gamma^{\widetilde{\Lambda}_k}(c_q-c_s)\mathbb{1}_{Q_{i,0}}\Bigg)\Bigg],$$

where the $\widetilde{\Lambda}_k\equiv\widetilde{\Lambda}_k(\widehat{\nu},0,1,\widehat{m})$ for all k, $\widetilde{\Lambda}_0\equiv 0$. Then, (5.29) allows us to write $C^0_{\pi^*}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$ as

$$C^0_{\pi^*}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u})$$
$$=-c_s+\mathbb{E}\left[\sum_{k=1}^{\infty}\gamma^{\sum_{i=0}^{k-1}\widetilde{\Lambda}_i}\sum_{r=0}^{\widetilde{\Lambda}_k-1}\gamma^r\left\{(c_s+c_h)\mathbb{1}(r=0)+c\left(Z(\nu_r,r)\right)+\gamma q(c_q-c_s)\right\}\right]$$
$$=-c_s+\inf_{\pi\in\Pi}\left\{\mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t\left\{(c_s+c_h)\mathbb{1}(n_{\pi(t),t}=0)+c\left(Z(\nu_{\pi(t),t},n_{\pi(t),t})\right)+\gamma q(c_q-c_s)\right\}\Bigg]\right\}.$$

The quantity $\gamma q(c_q-c_s)$ is a shifting constant that does not affect the minimization problem, so we have

$$C^0_{\pi^*}(\boldsymbol{\nu},\boldsymbol{n},\boldsymbol{u}) \tag{5.30}$$
$$=-c_s+\inf_{\pi\in\Pi}\left\{\mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t\left\{(c_s+c_h)\mathbb{1}(n_{\pi(t),t}=0)+c\left(Z(\nu_{\pi(t),t},n_{\pi(t),t})\right)\right\}\Bigg]\right\}+\frac{\gamma q(c_q-c_s)}{1-\gamma}.$$

The solution to the minimization problem on the right hand side is the same as the solution

to that minimization problem if the training cost is $c_s + c_h$ and the switching and quitting costs are set equal to 0. As a consequence $M_i(n_{i,t}, \nu_{i,t}, u_{i,t}, c_h, c_s, c_q) < \widehat{m}(c_h, c_s, c_q)$ if and only if $M_i(n_{i,t}, \nu_{i,t}, u_{i,t}, c_h + c_s, 0, 0) < \widehat{m}(c_h + c_s, 0, 0)$ for all $t \geq 0$. To complete our argument, we then need to prove (5.28). The left-hand side satisfies

$$\mathbb{E}\left[\sum_{r=0}^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})-1} \gamma^{r+1} q\right] = \sum_{r=1}^{\infty} \gamma^r q \mathbb{P}(\widetilde{\Lambda}(\widehat{\nu}, 0, 1, \widehat{m}) \geq r), \tag{5.31}$$

and the right-hand side satisfies

$$\begin{aligned}\mathbb{E}\left[\gamma^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})} \mathbb{1}_{Q_0}\right] &= \mathbb{E}\left[\gamma^{\widetilde{\Lambda}(\widehat{\nu},0,1,\widehat{m})} \mathbb{1}(\widetilde{\Lambda}(\widehat{\nu}, 0, 1, \widehat{m}) = \Lambda)\right] \\ &= \sum_{r=1}^{\infty} \gamma^r \mathbb{P}(\Lambda = r, \widetilde{\Lambda}(\widehat{\nu}, 0, 1, \widehat{m}) = r). \end{aligned} \tag{5.32}$$

By using the shorthand $\widetilde{\Lambda} \equiv \widetilde{\Lambda}(\widehat{\nu}, 0, 1, \widehat{m})$, recalling that $\widetilde{\Lambda} \stackrel{d}{=} 1 + \Lambda(\pi^*)$, and using the definition for the quitting probability, q, in (5.4) we have

$$\begin{aligned} q\mathbb{P}(\widetilde{\Lambda} \geq r) &= \mathbb{P}(L_{r-1} = 1 | \Lambda(\pi^*) \geq r - 1)\mathbb{P}(\widetilde{\Lambda} \geq r) \\ &= \mathbb{P}(L_{r-1} = 1 | \widetilde{\Lambda} \geq r)\mathbb{P}(\widetilde{\Lambda} \geq r) \\ &= \mathbb{P}(L_{r-1} = 1, \widetilde{\Lambda} \geq r), \end{aligned}$$

where the last equality follows from the definition of conditional probability. Recall from (5.11) that $\mathbb{P}(L_{r-1} = 1, \widetilde{\Lambda} \geq r) = \mathbb{P}(\Lambda = r, \widetilde{\Lambda} \geq r)$, and because $\Lambda = r$ implies $\widetilde{\Lambda} \leq r$ we also have $\mathbb{P}(\Lambda = r, \widetilde{\Lambda} \geq r) = \mathbb{P}(\Lambda = r, \widetilde{\Lambda} = r)$, which in turn implies $q\mathbb{P}(\widetilde{\Lambda} \geq r) = \mathbb{P}(\Lambda = r, \widetilde{\Lambda} = r)$, just as needed in (5.31) and (5.32) to complete the proof of (5.28).

BIBLIOGRAPHY

Aksin, Z., M. Armony, and V. Mehrotra. 2007. "The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research". *Production and Operations Management* 16 (6): 665–688.

Arlotto, A., R. W. Chen, L. A. Shepp, and J. M. Steele. 2011. "Online selection of alternating subsequences from a random sample". *J. Appl. Probab.* 48 (4): 1114–1132.

Arlotto, A., S. E. Chick, and N. Gans. 2012. "Optimal Hiring and Rentention Policies for Heterogeneous Workers who Learn". *Working paper – University of Pennsylvania*.

Arlotto, A., N. Gans, and S. E. Chick. 2010. "Optimal employee retention when inferring unknown learning curves". In *Proc. 2010 Winter Simulation Conference*, Edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1178–1188. IEEE, Inc.

Arlotto, A., N. Gans, and J. M. Steele. 2012. "Markov Decision Problems where Means Bound Variances". *Working paper – University of Pennsylvania*.

Arlotto, A., and J. M. Steele. 2011. "Optimal Sequential Selection of a Unimodal Subsequence of a Random Sequence". *Combinatorics, Probability and Computing* 20 (06): 799–814.

Arlotto, A., and J. M. Steele. 2012. "Optimal On-Line Selection of an Alternating Subsequence: A Central Limit Theorem". *Working paper – University of Pennsylvania*.

Asawa, M., and D. Teneketzis. 1996. "Multi-armed bandits with switching penalties". *IEEE Trans. Automat. Control* 41 (3): 328–348.

Baik, J., P. Deift, and K. Johansson. 1999. "On the distribution of the length of the longest increasing subsequence of random permutations". *J. Amer. Math. Soc.* 12 (4): 1119–1178.

Bailey, C. D. 1989. "Forgetting and the learning curve: a laboratory study". *Management Science* 35 (3): 340–352.

Banks, J. S., and R. K. Sundaram. 1992. "Denumerable-armed bandits". *Econometrica* 60 (5): 1071–1096.

Banks, J. S., and R. K. Sundaram. 1994. "Switching Costs and the Gittins Index". *Econometrica* 62 (3): pp. 687–694.

Baykal-Gürsoy, M., and K. W. Ross. 1992. "Variability sensitive Markov decision processes". *Math. Oper. Res.* 17 (3): 558–571.

Bergemann, D., and J. Välimäki. 2001. "Stationary multi-choice bandit problems". *Journal of Economic Dynamics and Control* 25 (10): 1585 – 1594.

Bertsekas, D. P., and S. E. Shreve. 1978. *Stochastic optimal control*, Volume 139. New York: Academic Press Inc.

Bollobás, B., and G. Brightwell. 1992. "The height of a random partial order: concentration of measure". *Ann. Appl. Probab.* 2 (4): 1009–1018.

Bollobás, B., and S. Janson. 1997. "On the length of the longest increasing subsequence in a random permutation". In *Combinatorics, geometry and probability (Cambridge, 1993)*, 121–128. Cambridge: Cambridge Univ. Press.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. "Statistical analysis of a telephone call center: a queueing-science perspective". *J. Amer. Statist. Assoc.* 100 (469): 36–50.

Bruss, F. T., and F. Delbaen. 2001. "Optimal rules for the sequential selection of monotone subsequences of maximum expected length". *Stochastic Process. Appl.* 96 (2): 313–342.

Bruss, F. T., and F. Delbaen. 2004. "A central limit theorem for the optimal selection process for monotone subsequences of maximum expected length". *Stochastic Process. Appl.* 114 (2): 287–311.

Bruss, F. T., and J. B. Robertson. 1991. ""Wald's lemma" for sums of order statistics of i.i.d. random variables". *Adv. in Appl. Probab.* 23 (3): 612–623.

Carillo, J. E., and C. Gaimon. 2000. "Improving Manufacturing Performance Through Process Change and Knowledge Creation". *Management Science* 46 (2): 265–288.

Chan, C. W., and V. F. Farias. 2009. "Stochastic depletion problems: effective myopic policies for a class of dynamic optimization problems". *Math. Oper. Res.* 34 (2): 333–350.

Chung, F. R. K. 1980. "On unimodal subsequences". *J. Combin. Theory Ser. A* 29 (3): 267–279.

Coffman, Jr., E. G., L. Flatto, and R. R. Weber. 1987. "Optimal selection of stochastic intervals under a sum constraint". *Adv. in Appl. Probab.* 19 (2): 454–473.

Dada, M., and K. N. Srikanth. 1990. "Monopolistic Pricing and the Learning Curve: An Algorithmic Approach". *Operations Research* 38 (4): 656–666.

Derman, C., G. J. Lieberman, and S. M. Ross. 1975. "A stochastic sequential allocation model". *Operations Res.* 23 (6): 1120–1130.

Efron, B., and C. Stein. 1981. "The jackknife estimate of variance". *Ann. Statist.* 9 (3): 586–596.

Erdős, P., and G. Szekeres. 1935. "A combinatorial problem in geometry". *Compositio Math.* 2:463–470.

Farias, V. F., and R. Madan. 2011. "The Irrevocable Multi-Armed Bandit Problem". *Operations Research* 59 (2): 383–399.

Feinberg, E. A., and J. Fei. 2009. "An inequality for variances of the discounted rewards". *J. Appl. Probab.* 46 (4): 1209–1212.

Filar, J. A., L. C. M. Kallenberg, and H.-M. Lee. 1989. "Variance-penalized Markov decision processes". *Math. Oper. Res.* 14 (1): 147–161.

Frazier, P., W. Powell, and S. Dayanik. 2009. "The knowledge-gradient policy for correlated normal beliefs". *INFORMS J. Comput.* 21 (4): 599–613.

Freeman, P. R. 1983. "The secretary problem and its extensions: a review". *Internat. Statist. Rev.* 51 (2): 189–206.

Gaimon, C. 1997. "Planning Information Technology-Knowledge Worker Systems". *Management Science* 43 (9): 1308–1328.

Gaimon, C., G. F. Ozkan, and K. Napoleon. 2011. "Dynamic Resource Capabilities: Managing Workforce Knowledge with a Technology Upgrade". *Organization Science* 22 (6): 1560–1578.

Gans, N., G. Koole, and A. Mandelbaum. 2003. "Telephone Call Centers: Tutorial, Review, and Research Prospects". *Manufacturing & Service Operations Management* 5 (2): 79–141.

Gans, N., N. Liu, A. Mandelbaum, H. Shen, and H. Ye. 2010. "Service times in call centers: agent heterogeneity and learning with some operational consequences". In *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown*, Volume 6, 99–123. Inst. Math. Statist.

Gans, N., and Y.-P. Zhou. 2002. "Managing Learning and Turnover in Employee Staffing". *Operations Research* 50 (6): 991–1006.

Gittins, J. C. 1989. *Multi-armed bandit allocation indices*. Chichester, England: John Wiley & Sons.

Gittins, J. C., and D. M. Jones. 1974. "A dynamic allocation index for the sequential design of experiments". In *Progress in statistics (European Meeting Statisticians, Budapest, 1972)*, 241–266. Colloq. Math. Soc. János Bolyai, Vol. 9. Amsterdam: North-Holland.

Gnedin, A. V. 1999. "Sequential selection of an increasing subsequence from a sample of random size". *J. Appl. Probab.* 36 (4): 1074–1085.

Gnedin, A. V. 2000. "Sequential selection of an increasing subsequence from a random sample with geometrically distributed sample-size". In *Game theory, optimal stopping, probability and statistics*, Volume 35 of *IMS Lecture Notes Monogr. Ser.*, 101–109. Beachwood, OH: Inst. Math. Statist.

Goldberg, M. S., and A. E. Touw. 2003. *Statistical methods for learning curves and cost analysis*. INFORMS. Topics in Operations Research Series.

Houdré, C., and R. Restrepo. 2010. "A Probabilistic Approach to the Asymptotics of the Length of the Longest Alternating Subsequence". *Electron. J. Combin.* 17 (1): Research Paper 168, 1–19.

Huselid, M. A. 1995. "The impact of human resource practices on turnover, productivity, and corporate financial performance". *Academy of Management Journal* 38 (3): 635 – 672.

Jones, G. L. 2004. "On the Markov chain central limit theorem". *Probab. Surv.* 1:299–320.

Jovanovic, B. 1979. "Job Matching and the Theory of Turnover". *Journal of Political Economy* 87 (5): 972.

Jun, T. 2004. "A survey on the bandit problem with switching costs". *De Economist* 152:513–541.

Kellerer, H., U. Pferschy, and D. Pisinger. 2004. *Knapsack problems*. Berlin: Springer-Verlag.

Lapré, M. A., A. S. Mukherjee, and L. N. Van Wassenhove. 2000. "Behind the Learning Curve: Linking Learning Activities to Waste Reduction". *Management Science* 46 (5): 597–611.

Lugosi, G. 2009. *Concentration-of-measure inequalities*. available on-line at `http:www.econ.upf.edu/~lugosi/anu.pdf`.

March, J. G. 1991. "Exploration and Exploitation in Organizational Learning". *Organization Science* 2 (1): pp. 71–87.

Martello, S., and P. Toth. 1990. *Knapsack problems*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Chichester: John Wiley & Sons Ltd. Algorithms and computer implementations.

Mazzola, J. B., and K. F. McCardle. 1996. "A Bayesian Approach to Managing Learning-Curve Uncertainty". *Management Science* 42 (5): 680–692.

Mazzola, J. B., and K. F. McCardle. 1997. "The Stochastic Learning Curve: Optimal Production in the Presence of Learning-Curve Uncertainty". *Operations Research* 45 (3): 440–450.

Müller, A., and D. Stoyan. 2002. *Comparison methods for stochastic models and risks*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd.

Nagypál, E. 2007. "Learning by Doing vs. Learning About Match Quality: Can We Tell Them Apart?". *Review of Economic Studies* 74:537–566.

Nembhard, D. A. 2001. "Heuristic approach for assigning workers to tasks based on individual learning rates". *International Journal of Production Research* 39:1955–1968(14).

Nembhard, D. A., and N. Osothsilp. 2002. "Task complexity effects on between-individual learning/forgetting variability". *International Journal of Industrial Ergonomics* 29:297–306(10).

Nembhard, D. A., and M. V. Uzumeri. 2000a. "Experiential learning and forgetting for manual and cognitive tasks". *International Journal of Industrial Ergonomics* 25:315–326(12).

Nembhard, D. A., and M. V. Uzumeri. 2000b, Aug. "An individual-based description of learning within an organization". *IEEE Transactions on Engineering Management* 47 (3): 370–378.

Niño-Mora, J. 2008. "A faster index algorithm and a computational study for bandits with switching costs". *INFORMS J. Comput.* 20 (2): 255–269.

Papastavrou, J. D., S. Rajagopalan, and A. J. Kleywegt. 1996. "The Dynamic and Stochastic Knapsack Problem with Deadlines". *Management Science* 42 (12): 1706–1718.

Pinker, E. J., and R. A. Shumsky. 2000. "The Efficiency-Quality Trade-Off of Cross-Trained Workers". *Manufacturing & Service Operations Management* 2 (1): 32–48.

Pisano, G. P., R. M. Bohmer, and A. C. Edmondson. 2001. "Organizational Differences in Rates of Learning: Evidence from the Adoption of Minimally Invasive Cardiac Surgery". *Management Science* 47 (6): 752–768.

Prastacos, G. P. 1983. "Optimal Sequential Investment Decisions under Conditions of Uncertainty". *Management Science* 29 (1): 118–134.

Puterman, M. L. 1994. *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc. A Wiley-Interscience Publication.

Rhee, W., and M. Talagrand. 1991. "A note on the selection of random variables under a sum constraint". *J. Appl. Probab.* 28 (4): 919–923.

Samuels, S. M., and J. M. Steele. 1981. "Optimal sequential selection of a monotone sequence from a random sample". *Ann. Probab.* 9 (6): 937–947.

Shafer, S. M., D. A. Nembhard, and M. V. Uzumeri. 2001. "The Effects of Worker Learning, Forgetting, and Heterogeneity on Assembly Line Productivity". *Management Science* 47 (12): 1639–1653.

Shaked, M., and J. G. Shanthikumar. 2007. *Stochastic orders*. Springer Series in Statistics. New York: Springer.

Shen, H. 2003. *Estimation, confidence intervals and nonparametric regression for problems involving lognormal distributions.* Ph.D. Thesis, University of Pennsylvania.

Shen, H., and L. D. Brown. 2006. "Non-parametric modelling for time-varying customer service time at a bank call centre". *Appl. Stoch. Models Bus. Ind.* 22 (3): 297–311.

Sobel, M. J. 1982. "The variance of discounted Markov decision processes". *J. Appl. Probab.* 19 (4): 794–802.

Sobel, M. J. 1994. "Mean-Variance Tradeoffs in an Undiscounted MDP". *Operations Research* 42 (1): 175–183.

Stanke, M. 2004. "Sequential selection of random vectors under a sum constraint". *J. Appl. Probab.* 41 (1): 131–146.

Stanley, R. P. 2007. "Increasing and decreasing subsequences and their variants". In *International Congress of Mathematicians. Vol. I*, 545–579. Eur. Math. Soc., Zürich.

Stanley, R. P. 2008. "Longest alternating subsequences of permutations". *Michigan Math. J.* 57:675–687. Special volume in honor of Melvin Hochster.

Stanley, R. P. 2010. "A survey of alternating permutations". *Contemp. Math.* 531:165–196.

Steele, J. M. 1981. "Long unimodal subsequences: a problem of F. R. K. Chung". *Discrete Math.* 33 (2): 223–225.

Steele, J. M. 1986. "An Efron-Stein inequality for nonsymmetric statistics". *Ann. Statist.* 14 (2): 753–758.

Sundaram, R. K. 2005. "Generalized bandit problems". In *Social choice and strategic decisions*, Studies in Choice and Welfare, 131–162. Berlin: Springer.

Talluri, K. T., and G. J. van Ryzin. 2004. *The theory and practice of revenue management.* International Series in Operations Research & Management Science, 68. Boston, MA: Kluwer Academic Publishers.

Uzumeri, M., and D. Nembhard. 1998. "A population of learners: A new way to measure organizational learning". *Journal of Operations Management* 16 (5): 515 – 528.

White, D. J. 1988. "Mean, variance, and probabilistic criteria in finite Markov decision processes: a review". *J. Optim. Theory Appl.* 56 (1): 1–29.

Whitt, W. 2006. "The Impact of Increased Employee Retention on Performance in a Customer Contact Center". *Manufacturing Service Operations Management* 8 (3): 235–252.

Whittle, P. 1980. "Multi-armed bandits and the Gittins index". *J. Roy. Statist. Soc. Ser. B* 42 (2): 143–149.

Widom, H. 2006. "On the limiting distribution for the length of the longest alternating sequence in a random permutation". *Electron. J. Combin.* 13 (1): Research Paper 25, 1–7.

Wiersma, E. 2007. "Conditions That Shape the Learning Curve: Factors That Increase the Ability and Opportunity to Learn". *Management Science* 53 (12): 1903–1915.

Yelle, L. E. 1979. "The learning curve: historical review and comprehnsive survey". *Decision Sciences* 10 (2): 302–328.