Markov Decision Problems where Means bound Variances

Alessandro Arlotto
The Fuqua School of Business; Duke University; Durham, NC, 27708, U.S.A.;
aa249@duke.edu

Noah Gans
OPIM Department; The Wharton School; University of Pennsylvania; Philadelphia, PA, 19104, U.S.A.;
gans@wharton.upenn.edu

J. Michael Steele
Statistics Department; The Wharton School; University of Pennsylvania; Philadelphia, PA, 19104, U.S.A.;
steele@wharton.upenn.edu

We identify a rich class of finite-horizon Markov decision problems (MDPs) for which the variance of the optimal total reward can be bounded by a simple linear function of its expected value. The class is characterized by three natural properties: *reward non-negativity and boundedness*, *existence of a do-nothing action*, and *optimal action monotonicity*. These properties are commonly present and typically easy to check. Implications of the class properties and of the variance bound are illustrated by examples of MDPs from operations research, operations management, financial engineering, and combinatorial optimization.

Key words: Markov decision problems, variance bounds, optimal total reward
MSC2000 subject classification: Primary: 90C40, 90C39; Secondary: 60G42
OR/MS subject classification: dynamic programming; probability
History: first version: February 12, 2013; this version: March 11, 2014.

1. Looking at More than Means

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, for many MDPs, the expected value of $R_n(\pi_n^*)$ is well understood. Still, just knowing the mean of $R_n(\pi_n^*)$ leaves much that is unknown, and, given the extensive literature on MDPs, it is striking that one seldom has a substantial understanding of the distribution of $R_n(\pi_n^*)$. Even the variance of $R_n(\pi_n^*)$ often goes unstudied.

This situation deserves to be addressed since in many MDPs the reward $R_n(\pi_n^*)$ has a direct economic interpretation, and any well-founded judgment about a policy needs to take into account the riskiness (or uncertainty) of the reward. Our main goal here is to identify a substantial class of MDPs for which one has general, explicit bounds on the variance of $R_n(\pi_n^*)$. Specifically, we characterize 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 — or, in some instances, a simple affine function of the expectation. Useful consequences of this bound include practical constraints on the riskiness of the realized reward and a straightforward weak law of large numbers for $R_n(\pi_n^*)$.

Our main result offers positive encouragement for MDP modelers whose objective is to maximize the expected total reward over a finite time horizon with n periods. If the MDP satisfies three natural properties — reward non-negativity and boundedness, existence of a do-nothing action, and optimal action monotonicity — then the total reward that one obtains is (probably) close to what one expects, provided that n is sufficiently large. Thus, one has an *ex-ante* justification for viewing the expected total reward as a reliable objective function.

1.1. An Informative Example

To fix ideas and to build intuition, we first consider a sequential knapsack problem. We view the knapsack capacity $C \in (0,\infty)$ as given, and we sequentially consider n items with sizes $Y_1, Y_2, \ldots, Y_n$. Moreover, we assume that the item sizes are independent non-negative random variables with a common distribution F, and, for specificity, we assume that F is regular at 0 in the sense that there are constants $A > 0$ and $\alpha > 0$ such that $F(x) \sim Ax^\alpha$ as $x \to 0$. (Here, given two real-valued functions f and g, we write $f(x) \sim g(x)$ as $x \to x_0$, and we say that f and g are *asymptotically equivalent* as $x \to x_0$ if $\lim_{x\to x_0} f(x)/g(x) = 1$.) In the simplest case, when the item sizes are uniformly distributed on $[0,1]$, we have $A = 1$, $\alpha = 1$, and $F(x) = x$ for $x \in [0,1]$.

At time $t \in \{1,2,\ldots,n\}$, when a newly presented item of size Y_t is first seen, the decision maker must decide to include or exclude the item from the knapsack. In the version of the problem that we consider here, the decision maker's goal is to maximize the expected number of items that can be included without the sum of the sizes of the accepted items exceeding a capacity constraint.

We let $\Pi(n)$ denote the set of all non-anticipating Markov deterministic knapsack policies, and for any policy $\pi \in \Pi(n)$ we let $\tau_i \in \{1,\ldots,n\}$ denote the index of the ith item that is chosen for inclusion in the knapsack. Here by *deterministic* we just mean that $\Pi(n)$ does not include any randomized decision rules, and by *non-anticipating* we mean that decisions to include or exclude newly presented items can use only the information (or history) of the selection process up to the decision time. More formally, to say that π is non-anticipating is the same as saying that each τ_i is a stopping time with respect to the increasing sequence of σ-fields $\mathcal{F}_t = \sigma\{Y_1, Y_2, \ldots, Y_t\}$, $1 \le t \le n$.

The reward attained by the policy π is the number of inclusions, so, in terms of the stopping times, we have

$$R_n(\pi) = \max\left\{k : 1 \le \tau_1 < \tau_2 < \cdots < \tau_k \le n \text{ and } \sum_{i=1}^{k} Y_{\tau_i} \le C\right\}.$$

Classical results from dynamic programming (Bertsekas and Shreve, 1978, Corollary 8.5.1) assure us that for each n there is a Markov deterministic policy that is optimal within the set of all non-anticipating policies, i.e. there is a $\pi_n^* \in \Pi(n)$ such that

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

Coffman, Flatto and Weber (1987) proved that for this sequential knapsack problem the optimal Markov deterministic policy is also unique, and they showed that

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

This asymptotic relation was subsequently refined by an explicit upper bound in Bruss and Robertson (1991) and by an explicit lower bound in Rhee and Talagrand (1991). The sequential knapsack problem is a leading example of an MDP for which there is an almost complete understanding of the expected value of the reward provided by the optimal policy π_n^*.

1.2. First Example of the Variance Bound

As a consequence of the general variance bound that is given below in Theorem 1, one also has a variance bound for the sequential knapsack problem:

$$\mathrm{Var}\left[R_n(\pi_n^*)\right] \le \mathbb{E}\left[R_n(\pi_n^*)\right] \quad \text{for all } A > 0,\, \alpha > 0,\, 0 < C < \infty, \text{ and } 1 \le n < \infty, \tag{2}$$

and from this bound one quickly obtains a weak law of large numbers for $R_n(\pi_n^*)$. Specifically, from the asymptotic result for the mean (1), the variance bound (2), and Chebyshev's inequality one finds that

$$n^{-1/(1+\alpha)} R_n(\pi_n^*) \quad \text{converges in probability to } \left[AC\alpha^{-\alpha}(\alpha+1)^{\alpha}\right]^{1/(1+\alpha)} \qquad \text{as } n \to \infty.$$

Here, we should note that there is a strategic nuance to the bound (2). The policy π_n^* is determined by optimizing the expected reward functional $\pi \mapsto \mathbb{E}[R_n(\pi)]$ over all $\pi \in \Pi(n)$, and, since the optimality criterion focuses unilaterally on the *expected reward*, there would seem to be no *a priori* connection between $\mathbb{E}\left[R_n(\pi_n^*)\right]$ and the variance $\mathrm{Var}[R_n(\pi_n^*)]$. What prevents the mean-focused optimal policy π_n^* from greatly inflating the variance $\mathrm{Var}[R_n(\pi_n^*)]$ just to eke out a modest increment to the mean $\mathbb{E}[R_n(\pi_n^*)]$? This is a possibility that seems perfectly feasible. Nevertheless, there is a substantial class of natural problems for which the mean-focused optimal policies are never so foul.

1.3. Tools, Proofs, and Further Examples

The class of MDPs that are of concern here share three characteristics. One is the *existence of a do-nothing action*, which is the general analog of not accepting an item in the sequential knapsack problem. The other two are *reward non-negativity and boundedness* and *optimal action monotonicity*. These order properties are easily checked in concrete problems, but the general definitions require some careful notation that we develop in Sections 3 and 4. In Section 4, we also state our main result, which we prove in Section 5.

Sections 6 and 7 discuss examples and counterexamples. In Section 6, we give an example of an MDP that does not have optimal action monotonicity and for which the variance bound fails, while in Section 7, we present examples of MDPs that satisfy our three natural properties. Finally, in Section 8 we underscore some open problems.

2. A Brief Review of MDPs that Attend to Moments

The vast majority of work on Markov decision problems takes a *risk-neutral* perspective where one seeks to optimize an expected total, possibly discounted, reward, or the long-run average expected reward per time period. Nevertheless, there are numerous investigations that take a *risk-aware* perspective where the optimization criterion incorporates some measure of uncertainty such as the variance of the reward. White (1988) provides a useful review of earlier work that takes such a point of view.

There are also several investigations that consider the possibility of mean-variance tradeoffs in average reward models. Specifically, Sobel (1994) considers *stationary* (time-homogeneous) policies that are Pareto optimal with respect to the steady-state mean and variance they generate, and, in the same settings, Chung (1994) develops an algorithm for identifying the Pareto optima in the unichain model. In closely related work, Sobel (1985) investigates stationary policies that maximize the steady-state ratio of the mean to the standard deviation.

There are some natural alternatives to the Pareto optimality framework. For example, Kawai (1987) considers variance minimization subject to a constraint on the mean, and numerous investigations have considered a variance penalty in the objective function, e.g. Filar, Kallenberg and Lee (1989), Baykal-Gürsoy and Ross (1992), and Huang and Kallenberg (1994). More recently, Haskell and Jain (2012) studied Markov decision problems subject to stochastic domination constraints.

There are some common elements to these investigations. First, essentially all consider average reward models for which a stationary policy is optimal. Moreover, most of these investigations focus on the variance of the long-run reward, rather than the limiting variance of the total reward (see Sobel, 1994, p. 178). From a probabilistic standpoint, the latter is usually more appropriate,

while the former is almost always more tractable. As Huang and Kallenberg (1994) observe (p. 434, note 2), these two measures of uncertainty are not easily related.

Risk-aware optimization criteria have also been considered when the objective is to maximize the expected total reward over a finite time horizon, or the expected total discounted reward over an infinite time horizon. For instance, Jaquette (1972; 1973) proposes a method for identifying policies with minimum variance among the set of policies that maximize expected rewards. Ruszczyński (2010) studies Markov risk measures to formulate risk-adverse Markov decision problems; Mannor and Tsitsiklis (2013) study the computational complexity of finite-horizon Markov decision problems where the performance measure includes both the mean and the variance of the cumulative reward.

Sobel (1982) considers the total discounted reward in an infinite-horizon model where the decision maker maximizes total expected discounted rewards, and, with motivations parallel to our own, he proves a noteworthy closed-form formula for the variance of the optimal total discounted reward. In this setting a stationary policy is optimal, and Chung and Sobel (1987) further characterize the distribution function of the discounted reward.

Recent work of Feinberg and Fei (2009) can also be reinterpreted in discrete-time to give a useful relation between the variance of the optimal total reward in an infinite-horizon discounted problem (with discount factor β) and the total reward in the analogous problem with an independent random horizon with the geometric distribution (with parameter $1-\beta$). The expected total rewards are equal in each problem formulation, but the variance of the former is smaller.

Finally, there are numerous instances in the theory of MDPs where one uses bounds on the total expected reward to quantify the effectiveness of suboptimal policies, such as approximate linear programming, martingale duality, and information relaxation (Brown, Smith and Sun, 2010; Desai, Farias and Moallemi, 2012). Here one might also reasonably include the extensive theory of prophet inequalities (Hill and Kertz, 1992) even though most of this work is not framed in the language of MDPs.

3. A General MDP Framework

We now consider a general discrete-time Markov decision problem with $n < \infty$ decision times (or periods) indexed by $t = 1, 2, \ldots, n$. By $\mathcal{X}$ we denote a set that we call the *state space*, and, at each decision time t, the decision maker is assumed to know the current state $x \in \mathcal{X}$ of the system. Also, at time t, the decision maker is assumed to know the current value y of an exogenous sequence $\{Y_t : 1 \le t \le n\}$ of independent $\mathcal{Y}$-valued random variables (or vectors) with a known sequence of distributions $\mathcal{D}_Y \equiv \{F_t : 1 \le t \le n\}$ that do not depend on the state.

Given a pair $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and a time $1 \le t \le n$, we let $\mathcal{A}(t,x,y)$ denote the set of actions available to the decision maker. For each $1 \le t \le n$, we consider the set of *admissible* state-action pairs

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

and we define the *action space* by setting $\mathcal{A} = \bigcup_{t,x,y} \mathcal{A}(t,x,y)$. The set of available actions $\mathcal{A}(t,x,y)$ is specified by taking into account the history of the system up to time t and the state pair (x,y). Thus, any available action $a \in \mathcal{A}(t,x,y)$ is non-anticipating, i.e. it is determined completely by what is known to the decision maker at time t.

After action $a \in \mathcal{A}(t,x,y)$ is chosen at time t, the decision maker also sees the realization w of another random variable (or vector) $W_t^{x,a}$ with known distribution $G_t^{x,a}$, that may depend on the state value x, the action chosen a, and the decision time t, but that is independent of the past observations, $W_1, \ldots, W_{t-1}$. We assume that the support of $W_t^{x,a}$ is contained in a set $\mathcal{W}$ for all $1 \le t \le n$, $x \in \mathcal{X}$, and $a \in \mathcal{A}$. We also assume that the random variables Y_t and $W_t^{x,a}$ are independent for each $1 \le t \le n$, $x \in \mathcal{X}$ and $a \in \mathcal{A}$, and we let $\mathcal{D}_W$ denote the family of distributions $\{G_t^{x,a} : 1 \le t \le n,\, x \in \mathcal{X},\, \text{and } a \in \mathcal{A}\}$. In what follows, we drop the superscript x from $W_t^{x,a}$ for economy, and the dependence on the state will be implicit.

If action $a \in \mathcal{A}(t,x,y)$ is chosen at time t and one has the realization $W_t^a = w$, then the decision maker receives the real-valued reward $r(t,x,y,a,w)$, and the state of the system moves from x to $f(t,x,y,a,w) \in \mathcal{X}$. The *reward function* $r : \{1,\ldots,n\} \times \Gamma_t \times \mathcal{W} \to \mathbb{R}$ and the *state-transition function* $f : \{1,\ldots,n\} \times \Gamma_t \times \mathcal{W} \to \mathcal{X}$ are assumed to be deterministic functions that are known to the decision maker. We also allow for discounting of the one-period rewards accrued over time, and we let $0 < \beta \le 1$ be the discount factor. As usual, we assume that the sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{W}$ are Polish spaces and the functions r and f are Borel measurable (see, e.g. Bertsekas and Shreve, 1978, Section 8.1).

A sequence $(A_1, A_2, \ldots, A_n)$ of non-anticipating actions such that $A_t \in \mathcal{A}(t,\cdot,\cdot)$ is called a *policy* π of length n. Given the state $X_1 = \bar{x}$ at time $t = 1$ and a policy $\pi = (A_1, A_2, \ldots, A_n)$, then the state values $\{X_t : 1 < t \le n\}$ are determined by the recursion

$$X_{t+1} = f(t, X_t, Y_t, A_t, W_t^A), \quad 1 < t \le n, \tag{3}$$

where Y_t is the t'th random variable of the process $(Y_1, \ldots, Y_n)$, $A_t \in \mathcal{A}(t, X_t, Y_t)$ is the action taken at time t when the state pair is (X_t, Y_t), and W_t^A is the t'th element of the sequence $(W_1^A, \ldots, W_n^A)$, where, as noted earlier, the distribution of each W_t^A might depend on the state, X_t, and on the action chosen, A_t.

If we now let $\Pi(n)$ be the set of Markov deterministic policies for our general MDP, then the cumulative (discounted) reward gained by the policy $\pi \in \Pi(n)$ up to and including time k is given by the random sum

$$R_k(\pi) = \sum_{t=1}^{k} \beta^{t-1} r(t, X_t, Y_t, A_t, W_t^A), \qquad 1 \le k \le n,$$

and our main goal is to understand the variance of $R_n(\pi_n^*)$ when $\pi_n^* = (A_1^*, A_2^*, \ldots, A_n^*)$ is any Markov deterministic policy that maximizes the total expected reward; i.e.

$$\mathbb{E}[R_n(\pi_n^*)] = \sup_{\pi \in \Pi(n)} \mathbb{E}[R_n(\pi)]. \tag{4}$$

Here, we limit our analysis to Markov deterministic policies, but such policies are often optimal within the larger class of non-anticipating policies. (see, e.g. Bertsekas and Shreve, 1978, Proposition 8.5.)

In addition to the defining relation (4), there are several representations for the optimal expected reward $\mathbb{E}[R_n(\pi_n^*)]$, and it is particularly useful to make explicit the dependence of $\mathbb{E}[R_n(\pi_n^*)]$ on the initial state $\bar{x}$. If $\pi_n^* = (A_1^*, A_2^*, \ldots, A_n^*)$ is an optimal policy, then for $1 \le t \le n$, we define the sequence of *value functions* $v_t : \mathcal{X} \to \mathbb{R}$ by setting

$$v_t(x) = \mathbb{E}\left[\sum_{s=t}^{n} \beta^{s-t} r(s, X_s, Y_s, A_s^*, W_s^*) \,\middle|\, X_t = x \right], \tag{5}$$

so $v_t(x)$ represents the *expected reward to-go* that the decision maker collects from periods t through n under the optimal policy when the state at time t is x. In this notation we have

$$\mathbb{E}[R_n(\pi_n^*)] = v_1(\bar{x}),$$

and the optimality principle of dynamic programming gives us a natural way to compute $\mathbb{E}[R_n(\pi_n^*)]$. Specifically, for $1 \le t \le n$, we have the Bellman equation

$$v_t(x) = \int_{\mathcal{Y}} \Big\{ \sup_{a \in \mathcal{A}(t,x,y)} \mathbb{E}\left[r(t, x, y, a, W_t^a) + \beta v_{t+1}\left(f(t, x, y, a, W_t^a) \right) \right] \Big\} \, dF_t(y), \tag{6}$$

so backward recursion determines the value $v_1(\bar{x}) = \mathbb{E}[R_n(\pi_n^*)]$ if one starts by setting $v_{n+1}(x) = 0$ for all $x \in \mathcal{X}$.

The Bellman equation (6) is determined by two nested integrations, the inner one with respect to the distribution of W_t^*, and the outer one with respect to the distribution of Y_t. This form reflects the general structure of the MDPs we study here, in which the realization $Y_t = y$ becomes available before the decision maker chooses an optimal action $a^* \in \mathcal{A}(t, x, y)$, while the realization $W_t^* = w$ becomes available after action a^* is chosen. In general, when the decision maker chooses an optimal

action $a^* \in \mathcal{A}(t,x,y)$, he does not know the exact reward he earns, and he does not know the exact system state after the action is chosen. Rather, he knows the distribution of the one-period reward, $r(t,x,y,a^*,W_t^*)$, as well as the distribution of the optimal successor state, $X_{t+1} = f(t,x,y,a^*,W_t^*)$.

Our framework also allows for studying MDPs in which the only uncertainty that is realized at time t happens before the action is chosen (i.e., no W_t^a is relevant to the model), or MDPs in which the only uncertainty that matters is realized after the action is chosen (i.e., no Y_t is relevant). The knapsack problem of Section 1.1 is an example of the former. Several other examples are discussed in Section 7.

4. Bounding the Variance by the Mean: Sufficient Conditions

We can now isolate a class of Markov decision problems for which one can bound the variance of $R_n(\pi_n^*)$ by an explicit linear function of its mean. The class is determined by three natural properties that are common in MDPs.

Property 1 (Non-negative and Bounded Rewards). *There is a constant $K < \infty$ such that*

$$0 \le r(t,x,y,a,w) \le K \quad \text{for all } (x,y,a,w) \in \Gamma_t \times \mathcal{W} \text{ and } 1 \le t \le n.$$

Property 2 (Existence of a Do-nothing Action). *For each time $1 \le t \le n$ and pair (x,y), the set of actions $\mathcal{A}(t,x,y)$ includes an action a^0 such that*

$$f(t,x,y,a^0,w) = x, \quad \text{for all } w \in \mathcal{W}. \tag{7}$$

Action a^0 is called a do-nothing *action. Moreover, the expected reward to-go, $v_{t+1}(x)$, that one obtains by selecting the do-nothing action satisfies the inequality*

$$-r(t,x,y,a^*,w) \le r(t,x,y,a^*,w) + \beta v_{t+1}\left(f(t,x,y,a^*,w)\right) - \beta v_{t+1}(x), \quad \text{for all } w \in \mathcal{W}, \tag{8}$$

where $a^* \in \mathcal{A}(t,x,y)$ is an optimal action that achieves the supremum in (6).

The existence of a do-nothing action requires two simultaneous conditions. The condition given by (7) allows for the state of the system at time $t+1$ to be equal to the state of the system at time t. The condition given by (8) provides us with an inequality that integrates into what one would obtain in expectation from the optimality of action $a^* \in \mathcal{A}(t,x,y)$. In fact, by optimality we know that the sum of the one-period reward and the expected reward to-go of the do-nothing action satisfies the inequality

$$\mathbb{E}\left[r(t,x,y,a^0,W_t^0) + \beta v_{t+1}(x)\right] \le \mathbb{E}\left[r(t,x,y,a^*,W_t^*) + \beta v_{t+1}(f(t,x,y,a^*,W_t^*))\right],$$

which, together with the non-negativity of the reward function (Property 1), implies

$$0 \le \mathbb{E}\left[r(t,x,y,a^*,W_t^*) + \beta v_{t+1}(f(t,x,y,a^*,W_t^*)) - \beta v_{t+1}(x)\right].$$

Thus, equation (8) is a sample-path constraint that limits the extent for which the integrand on the right-hand side can be negative. At this point, it is also useful to note that the optimality of action $a^* \in \mathcal{A}(t,x,y)$ immediately implies (8) in MDPs in which the one-period rewards and the state-transition functions are not affected by the random sequence $W_1^*, \ldots, W_n^*$. The sequential knapsack problem discussed in Section 1.1 is an important example with such property. Several others are discussed in Section 7.

Our third criterion complements the pointwise bound (8) in a natural way.

Property 3 (Optimal Action Monotonicity). *For each time $1 \le t \le n$ and state $x \in \mathcal{X}$ one has the inequality*

$$v_{t+1}(f(t,x,y,a^*,w)) \le v_{t+1}(x) \tag{9}$$

for all $y \in \mathcal{Y}$, $w \in \mathcal{W}$, and any optimal action $a^* \in \mathcal{A}(t,x,y)$.

The existence of a do-nothing action tells us that it is always possible for the state of the system at time $t+1$ to be the same as it was at time t, so the right side of the inequality (9) is always meaningful. Obviously, if the do-nothing action is optimal, then one has $f(t,x,y,a^*,w) = x$ and (9) becomes an equality.

Inequality (9) has an intuitive interpretation in the common case of an MDP without ties, i.e. an MDP where there is a unique optimal action at each instance. In such a problem, if one has optimal action monotonicity, then the decision maker will never choose an action that changes the state of the system *unless* the decision maker gains a positive immediate reward for the action he chooses.

In the common situation in which one has reward non-negativity and boundedness (Property 1), the existence of a do-nothing action (Property 2), and optimal action monotonicity (Property 3), then one can prove an easy and effective variance bound.

Theorem 1 (Variance Bound). *Suppose that a Markov decision problem satisfies reward non-negativity and boundedness, the existence of a do-nothing action, and optimal action monotonicity. If $\pi_n^* \in \Pi(n)$ is any Markov deterministic policy such that*

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

then

$$\mathrm{Var}[R_n(\pi_n^*)] \le K\, \mathbb{E}[R_n(\pi_n^*)], \tag{10}$$

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

To gain intuition on the meaning of the variance bound in Theorem 1, we note that if the optimal one-period rewards, $\{r(t, X_t, Y_t, A_t^*, W_t^*) : 1 \le t \le n\}$, were independent, then one would immediately have the variance inequality

$$\mathrm{Var}[R_n(\pi_n^*)] = \sum_{t=1}^{n} \mathrm{Var}[\beta^{t-1} r(t, X_t, Y_t, A_t^*, W_t^*)] \le \sum_{t=1}^{n} \mathbb{E}[\{\beta^{t-1} r(t, X_t, Y_t, A_t^*, W_t^*)\}^2] \le K\, \mathbb{E}[R_n(\pi_n^*)].$$

The quantity on the right-hand side above is obtained assuming independence of the one-period rewards, and it equals our variance bound in Theorem 1. With Theorem 1, we obtain the same bound, but without assuming a condition of independence, a property that is not commonly present in the reward process of an MDP.

Theorem 1 also yields an immediate measure of the dispersion of the optimal total reward $R_n(\pi_n^*)$. Specifically, it gives us a bound on the coefficient of variation:

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

Here K bounds the *one-period reward* and $\mathbb{E}[R_n(\pi_n^*)]$ is the *multi-period optimal expected reward* which typically approaches infinity as $n \to \infty$. Hence, for the typical MDP that satisfies our three structural properties, the coefficient of variation converges to zero as $n \to \infty$.

The variance bound (10) and Chebyshev's inequality also provide easy estimates of concentration for the distribution of the optimal total reward. Specifically, for any $\epsilon > 0$, Chebyshev's inequality and the variance bound (10) tell us that

$$\mathbb{P}\left(|R_n(\pi_n^*) - \mathbb{E}[R_n(\pi_n^*)]| > \epsilon\right) \le \epsilon^{-2}\, K\, \mathbb{E}[R_n(\pi_n^*)]. \tag{11}$$

If we take $\lambda > 1$ and set $\epsilon = \lambda \{K\, \mathbb{E}[R_n(\pi_n^*)]\}^{1/2}$, then we have

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

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 worth detailing as a corollary.

Corollary 1 (Weak Law for Optimal Total Rewards with Large Horizon).
Suppose that a Markov decision problem satisfies reward non-negativity and boundedness, the existence of a do-nothing action, and optimal action monotonicity. If $\pi_n^* \in \Pi(n)$ is any optimal Markov deterministic policy and if $\mathbb{E}[R_n(\pi_n^*)] \to \infty$ as $n \to \infty$, then

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

This corollary is good news for variability-adverse decision makers. In the common case where $\mathbb{E}[R_n(\pi_n^*)] \to \infty$ as $n \to \infty$, it says that the reward realized by the optimal strategy will (with increasingly high probability) behave like its mean.

5. Variance Bounds and the Proof of Theorem 1

The proof of Theorem 1 begins by noting that the Bellman equation (6) leads one to a useful martingale that captures all of the information we need to bound the variance of the optimal total reward. The main task is to check that one can bound the size of martingale differences with help from our key properties: (1) reward non-negativity and boundedness, (2) existence of a do-nothing action, and (3) optimal action monotonicity.

LEMMA 1 (BELLMAN MARTINGALE). *For $0 \le t \le n$, the process defined by*

$$M_t = R_t(\pi_n^*) + \beta^t v_{t+1}(X_{t+1})$$

is a martingale with respect to the natural filtration $\mathcal{F}_t = \sigma\{Y_1, W_1^*, Y_2, W_2^*, \ldots, Y_t, W_t^*\}$ and the trivial σ-field $\mathcal{F}_0$.

Proof of Lemma 1. First, by the definition of $R_t(\pi_n^*)$ and the state-transition recursion (3), one sees that M_t is $\mathcal{F}_t$-measurable, i.e. it is determined by $(Y_1, W_1^*, Y_2, W_2^*, \ldots, Y_t, W_t^*)$. By reward boundedness (Property 1) and the finiteness of the horizon n, the process $\{M_t : 0 \le t \le n\}$ is also bounded.

We now need to verify the martingale property $\mathbb{E}[M_{t+1}|\mathcal{F}_t] = M_t$. Since we have $M_{t+1} = R_{t+1}(\pi_n^*) + \beta^{t+1} v_{t+2}(X_{t+2})$, we immediately obtain that

$$\mathbb{E}[M_{t+1}|\mathcal{F}_t] = \mathbb{E}[R_{t+1}(\pi_n^*) + \beta^{t+1} v_{t+2}(X_{t+2})|\mathcal{F}_t].$$

We now recall that $R_{t+1}(\pi_n^*) = R_t(\pi_n^*) + \beta^t r(t+1, X_{t+1}, Y_{t+1}, A_{t+1}^*, W_{t+1}^*)$ and note that $R_t(\pi_n^*)$ is $\mathcal{F}_t$-measurable, so we obtain

$$\mathbb{E}[M_{t+1}|\mathcal{F}_t] = R_t(\pi_n^*) + \beta^t \mathbb{E}[r(t+1, X_{t+1}, Y_{t+1}, A_{t+1}^*, W_{t+1}^*) + \beta v_{t+2}(X_{t+2})|\mathcal{F}_t]. \tag{12}$$

Here, we also have that the state X_{t+1} is $\mathcal{F}_t$-measurable since $X_{t+1} = f(t, X_t, Y_t, A_t^*, W_t^*)$, so, together with the principle of optimality of dynamic programming, we obtain that

$$\mathbb{E}[r(t+1, X_{t+1}, Y_{t+1}, A_{t+1}^*, W_{t+1}^*) + \beta v_{t+2}(X_{t+2})|\mathcal{F}_t] = \mathbb{E}[v_{t+1}(X_{t+1})|\mathcal{F}_t] = v_{t+1}(X_{t+1}).$$

Hence, it follows from (12) that

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

confirming that the process $\{M_t : 0 \le t \le n\}$ is a martingale. $\square$

Using the notation of Section 3, we have the initial state $X_1 = \bar{x}$, so, for the initial and terminal values of M_t, we have

$$M_0 = v_1(\bar{x}) = \mathbb{E}[R_n(\pi_n^*)] \qquad \text{and} \qquad M_n = R_n(\pi_n^*).$$

Recalling the definition of the state-transition function f, we obtain the optimal successor state $X_{t+1} = f(t, X_t, Y_t, A_t^*, W_t^*)$, and, for each $1 \le t \le n$, we have the martingale difference sequence

$$d_t = M_t - M_{t-1} = \beta^{t-1} r(t, X_t, Y_t, A_t^*, W_t^*) + \beta^t v_{t+1}(X_{t+1}) - \beta^{t-1} v_t(X_t). \tag{13}$$

By telescoping the sum and using the orthogonality of the martingale differences (see, e.g. Williams, 1991, Section 12.1) we find

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

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

At each time $1 \le t \le n$, the decision maker can always exercise the do-nothing action (Property 2) to obtain a one-period reward equal to $r(t, X_t, Y_t, A_t^0, W_t^0)$ and an expected reward to-go equal to $\beta v_{t+1}(X_t)$. If we now add and subtract $\beta^t v_{t+1}(X_t)$ in (13), then we can write the martingale difference d_t as

$$d_t = \beta^{t-1}\{B_t + C_t\},$$

where we define B_t and C_t by setting

$$B_t = \beta v_{t+1}(X_t) - v_t(X_t) \qquad \text{and} \qquad C_t = r(t, X_t, Y_t, A_t^*, W_t^*) + \beta v_{t+1}(X_{t+1}) - \beta v_{t+1}(X_t). \tag{14}$$

Since $X_t = f(t, X_{t-1}, Y_{t-1}, A_{t-1}^*, W_{t-1}^*)$, we see that X_t and B_t are $\mathcal{F}_{t-1}$-measurable, and our representation for d_t gives us

$$\mathbb{E}[d_t^2 \,|\, \mathcal{F}_{t-1}] = \beta^{2(t-1)}\{B_t^2 + 2B_t\mathbb{E}[C_t \,|\, \mathcal{F}_{t-1}] + \mathbb{E}[C_t^2 \,|\, \mathcal{F}_{t-1}]\}. \tag{15}$$

Since $0 = \mathbb{E}[d_t \,|\, \mathcal{F}_{t-1}] = \beta^{t-1}\{B_t + \mathbb{E}[C_t \,|\, \mathcal{F}_{t-1}]\}$ we have $\mathbb{E}[C_t \,|\, \mathcal{F}_{t-1}] = -B_t$, and we obtain from (15) that

$$\mathbb{E}[d_t^2 \,|\, \mathcal{F}_{t-1}] = \beta^{2(t-1)}\{\mathbb{E}[C_t^2 \,|\, \mathcal{F}_{t-1}] - B_t^2\} \le \beta^{2(t-1)}\mathbb{E}[C_t^2 \,|\, \mathcal{F}_{t-1}]. \tag{16}$$

We next check that the following bounds hold:

$$-r(t, X_t, Y_t, A_t^*, W_t^*) \le C_t \le r(t, X_t, Y_t, A_t^*, W_t^*). \tag{17}$$

To prove the first inequality of (17), we recall that condition (8) of Property 2 gives us the bound

$$-r(t, x, y, a^*, w) \le r(t, x, y, a^*, w) + \beta v_{t+1}(f(t, x, y, a^*, w)) - \beta v_{t+1}(x), \quad \text{for all } y \in \mathcal{Y} \text{ and } w \in \mathcal{W}.$$

The substitution $(x, y, a^*, w) \hookleftarrow (X_t, Y_t, A_t^*, W_t^*)$ and the definition (14) of C_t then give us $-r(t, X_t, Y_t, A_t^*, W_t^*) \le C_t$.

To prove the second inequality of (17), we note that optimal action monotonicity implies $\beta v_{t+1}(X_{t+1}) - \beta v_{t+1}(X_t) \le 0$, so we also obtain the desired upper bound directly from the definition of C_t.

Reward boundedness (Property 1) and (17) now give us

$$C_t^2 \le r(t, X_t, Y_t, A_t^*, W_t^*)^2 \le K\, r(t, X_t, Y_t, A_t^*, W_t^*), \tag{18}$$

so, when we take conditional expectations, we obtain

$$\mathbb{E}[C_t^2 \,|\, \mathcal{F}_{t-1}] \le K\, \mathbb{E}[r(t, X_t, Y_t, A_t^*, W_t^*) \,|\, \mathcal{F}_{t-1}].$$

Finally, we recall the bound (16), take total expectations, and sum to conclude that

$$\mathrm{Var}[R_n(\pi_n^*)] \le K\, \mathbb{E}\left[\sum_{t=1}^{n} \beta^{t-1} r(t, X_t, Y_t, A_t^*, W_t^*)\right] = K\, \mathbb{E}[R_n(\pi_n^*)], \tag{19}$$

as needed. □

Remark 1. Our three crucial properties and the decomposition $d_t = \beta^{t-1}\{B_t + C_t\}$, $1 \le t \le n$, also combine nicely to imply that the martingale $\{M_t : 0 \le t \le n\}$ has bounded differences, and $|d_t| \le 2K$. To see this, first note from (17) that $-K \le -r(t, X_t, Y_t, A_t^*, W_t^*) \le C_t \le r(t, X_t, Y_t, A_t^*, W_t^*) \le K$. Next, recall that we have the representation $\mathbb{E}[C_t \,|\, \mathcal{F}_{t-1}] = -B_t$, so the bounds on C_t also give us $-K \le B_t \le K$. Taken together, these inequalities give us the uniform bound $|d_t| = \beta^{t-1}|B_t + C_t| \le 2K$.

The bounded difference property allows us to obtain concentration bounds for the optimal total reward, $R_n(\pi_n^*)$. Specifically, the Azuma–Hoeffding inequality (see, e.g. Boucheron, Lugosi and Massart, 2013, Section 1.1) tells us that

$$\mathbb{P}\left(|R_n(\pi_n^*) - \mathbb{E}[R_n(\pi_n^*)]| > \epsilon\right) \le 2 \exp\left\{-\frac{\epsilon^2}{8K^2 n}\right\}, \tag{20}$$

which we can compare with the concentration bound (11) obtained by Chebyshev's inequality. If the expected optimal total reward satisfies the growth condition

$$n^{-1/2}\, \mathbb{E}[R_n(\pi_n^*)] \to \infty \qquad \text{as} \qquad n \to \infty, \tag{21}$$

then an application of (20) also implies our Corollary 1, but one should note that (21) is quite restrictive. For example, it does not hold for the simple knapsack problem of Section 1.1 when the item sizes have the uniform distribution on $[0, 1]$.

REMARK 2. We can also relax the assumption that there is a uniform bound on the reward function $r(t,\cdot)$, $1 \le t \le n$. The argument given here can be repeated almost word for word provided that one assumes an L^2-L^1 bound of the form

$$\mathbb{E}[r^2(t, X_t, Y_t, A_t^*, W_t^*)] \le K\,\mathbb{E}[r(t, X_t, Y_t, A_t^*, W_t^*)] \quad \text{for all } 1 \le t \le n. \tag{22}$$

The only change in the proof of Theorem 1 is that one uses (22) when one takes expectations in (18). The bound (22) holds rather widely; in particular, it holds when the rewards have exponentially bounded tail probabilities: $P(r(t, X_t, Y_t, A_t^*, W_t^*) \ge \lambda) \le A\exp(-B\lambda)$ for some $A > 0$ and $B > 0$ and all $\lambda \ge 0$.

6. When Optimal Action Monotonicity Fails: A Counterexample

It is natural to anticipate that if any one of the three conditions of Theorem 1 were to fail, then the variance bound (10) may also fail. In particular, one can use a modification of the sequential knapsack problem of Section 1.1 to construct a novel MDP that illustrates the joint failure of optimal action monotonicity and the variance bound (2).

We take a horizon $n = 2$, a capacity $C \in (0, 1/3)$, and independent random variables Y_1 and Y_2 that are uniformly distributed on $[0,1]$. We write the action space as $\mathcal{A} = \{0,1\}$, where $a = 0$ means we do not include the item in the knapsack and $a = 1$ means that we include it. The state x is the amount of available capacity, so for any pair (x,y) and $t \in \{1,2\}$, the set of actions available is $\mathcal{A}(t,x,y) = \{0, \mathbb{1}(y \le x)\}$; in particular, if $x < y$ then $\mathcal{A}(t,x,y)$ is the singleton $\{0\}$. For the reward function we make the obvious choice $r(t,x,y,a) = a$, but in our state-transition function we introduce a novel twist and take

$$f(t,x,y,a) = \begin{cases} x & \text{if } a = 0 \\ 3x & \text{if } a = 1. \end{cases}$$

Thus, we have modified the knapsack problem so that the "remaining capacity" is tripled when we accept an item. Since we have the short horizon $n = 2$, this capacity boost materializes in a useful way only if we accept the first item. For this MDP, reward non-negativity and boundedness are trivial, and $a = 0$ is our do-nothing action. It remains to check that we have failure of optimal action monotonicity and failure of the variance bound (2).

We first check the failure of the variance bound (2). Here, it is easy to verify that the unique optimal Markov deterministic policy $\pi_2^* = (A_1^*, A_2^*)$ is just the greedy policy: accept any item that is feasible. In our notation we then have $(A_1^*, A_2^*) = (\mathbb{1}(Y_1 \le X_1), \mathbb{1}(Y_2 \le X_2))$, where $X_1 = C < 1/3$, so for the optimal policy π_2^*, we have a simple representation of the total reward

$$R_2(\pi_2^*) = \mathbb{1}(Y_1 \le C) + \mathbb{1}(Y_2 \le X_2).$$

Explicit computations now give

$$\mathbb{E}\,[R_2(\pi_2^*)] = 2(C + C^2) \qquad \text{and} \qquad \text{Var}\,(R_2(\pi_2^*)) = 2(C + C^2) + 2C^2(1 - 4C - 2C^2).$$

From the solution of the quadratic equation we find

$$\text{Var}\,(R_2(\pi_2^*)) > \mathbb{E}\,[R_2(\pi_2^*)] \qquad \text{for all } 0 < C < (-2 + \sqrt{6})/2 \approx 0.225,$$

and this confirms the violation of the variance bound (2).

It remains to check that our MDP also violates optimal action monotonicity (Property 3). Here we first note that if $t = 2$ and $X_2 = x$, then

$$v_2(x) = \mathbb{E}\,[\mathbb{1}(Y_2 \le x)] = x \quad \text{for any } x \in [0, 3C]. \tag{23}$$

Now, if at time $t = 1$ one has the state x and an item size $y \le x$, then the optimal action is $a^* = 1$ and our transition function gives us $x^* = f(1, x, y, a^*) = 3x$, so using (23) for x^* gives us

$$v_2(x^*) = 3x > v_2(x) = x,$$

which confirms the violation of optimal action monotonicity.

The counterexample of this section provides an MDP in which the reward process under the optimal policy exhibits substantial positive correlation. The same principle can be used to construct counterexamples with any horizon length that violate the variance bound (10) at different rates. For instance, by independent regeneration of the counterexample discussed here, one obtains an MDP with variance within a constant factor from the upper bound given by Theorem 1. It is also possible to encounter MDPs in which the variance of the optimal total reward grows to infinity at a rate that is faster than the rate at which the optimal mean grows to infinity. The simplest such example is a trivial MDP with only one action available at each decision time and with a reward process that is given by a dependent Bernoulli process. The reader is referred to James, James and Qi (2008, p. 2341) for an explicit example of such a process.

7. Positive Examples: Four that Illustrate Many

It is remarkably easy to find MDPs with reward non-negativity and boundedness, a do-nothing action, and optimal action monotonicity. Examples occur in operations research, operations management, financial engineering, and combinatorial optimization. Here, we focus on four examples that are illustrative of many others.

The examples in Sections 7.2, 7.3, and 7.5 consider MDPs in which all of the within-period uncertainty is realized before the decision maker chooses an optimal action (i.e., with no W_t^a that

matters). The discussion of stochastic depletion problems in Section 7.4 considers MDPs in which the within-period uncertainty realizes after the decision maker chooses an optimal action (i.e., with no Y_t that matters), allowing for problems with stochastic state-transitions.

To facilitate our discussion, we first remark that optimal action monotonicity is immediately satisfied if easy monotonicity conditions on the value function and on the evolution of the state space hold.

7.1. Optimal Action Monotonicity: Sufficient Conditions

Optimal action monotonicity (Property 3) requires that, for each $1 \le t \le n$ and each state $x \in \mathcal{X}$, the expected reward to-go that the decision maker earns after choosing an optimal action is smaller than the expected reward to-go obtained if he chooses the do-nothing action. In formulae, if $x^* \equiv f(t, x, y, a^*, w)$ is the optimal successor state for $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, then optimal action monotonicity requires that

$$v_{t+1}(x^*) \le v_{t+1}(x).$$

This property is easily verified provided that one of the two following sets of sufficient conditions holds.

Sufficient Conditions. *Suppose that the state space $\mathcal{X}$ is a subset of a finite-dimensional Euclidean space equipped with a partial order $\preceq$. Then, optimal action monotonicity (Property 3) is implied by* either one *of the following sets of conditions.*

1. *For each $1 \le t \le n$,* (i) *the map $x \mapsto v_t(x)$ is* non-decreasing *(so $x \preceq x'$ implies $v_t(x) \le v_t(x')$); and* (ii) *for $y \in \mathcal{Y}$, $w \in \mathcal{W}$, and each optimal action $a^* \in \mathcal{A}(t, x, y)$ one has $f(t, x, y, a^*, w) \equiv x^* \preceq x$.*
2. *For each $1 \le t \le n$,* (i) *the map $x \mapsto v_t(x)$ is* non-increasing *(so $x \preceq x'$ implies $v_t(x') \le v_t(x)$); and* (ii) *for $y \in \mathcal{Y}$, $w \in \mathcal{W}$, and each optimal action $a^* \in \mathcal{A}(t, x, y)$ one has $x \preceq x^* \equiv f(t, x, y, a^*, w)$.*

Both sets of sufficient conditions include monotonicity properties of the value functions and the evolution of the system state over time. Such properties are common and often easy-to-prove, as we discuss next.

7.2. Dynamic and Stochastic Knapsack Problems

First it is useful to see how the dynamic knapsack problem of Section 1.1 has been generalized. Specifically, Papastavrou, Rajagopalan and Kleywegt (1996) consider a knapsack with capacity $0 < C < \infty$, and a finite horizon n. For each time period, $1 \le t \le n$, they assume that a new item is offered in period t with probability $p > 0$. Moreover, to each arriving item there is an associated

pair (U, Z) of random variables, where U is the size of the arriving item "to be packed", and Z is the reward that one earns if the item is included in the knapsack.

Here, one assumes that the joint size-reward pairs (U_t, Z_t), $1 \le t \le n$, are independent with common distribution $F(u, z) = \mathbb{P}(U \le u, Z \le z)$ whose support is $\mathbb{R}_+ \times [0, K]$ where $K < \infty$. As usual, an arriving item can be accepted only if its size is not larger than the remaining capacity of the knapsack, and the objective is to maximize 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 $1-p$ we fail to have a new arrival, 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 (u, z) becomes known to the decision maker. With probability $1 - F(x, K)$ the freshly observed size u exceeds the remaining capacity; the arriving item cannot be accepted, and one is again left with the expected reward to-go, $v_{t+1}(x)$. Finally, if $u \le x$, then it is feasible to accept the arriving item, and one chooses the action that yields the largest sum of the one-period reward and the 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 $z + v_{t+1}(x-u)$.

Assembling these observations, we see that for each time $1 \le t \le n$ and for each level of remaining capacity $x \in [0, C]$, the Bellman equation is given by

$$v_t(x) = \{1 - p\, F(x, K)\}\, v_{t+1}(x) + p \int_{[0,x]\times[0,K]} \max\{v_{t+1}(x), z + v_{t+1}(x-u)\}\, dF(u, z) \qquad (24)$$

with the boundary conditions

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

We also note that the Bellman equation (24) is a special case of (6) since there is no random quantity that affects the state-transition function of any given action.

Here, reward non-negativity and boundedness and existence of do-nothing action are immediate. To check optimal action monotonicity, note that the remaining capacity X_t is a non-increasing function of t under any feasible policy and that the value function $v_t(x)$ is non-decreasing in x (cf. Papastavrou, Rajagopalan and Kleywegt, 1996, Lemma 1). Taken together these two properties immediately imply optimal action monotonicity (see Sufficient Conditions 1), so we have the variance bound (10) for the knapsack problem of Papastavrou, Rajagopalan and Kleywegt (1996).

Remark 3 (Single-Resource and Network Capacity Control). A multidimensional generalization of the knapsack problem discussed here is given by optimal network capacity

control. In the basic version of this problem (cf. Talluri and van Ryzin, 2004, Section 3.2), a network has m resources, and a firm sells ℓ products.

Each product is a "bundle" of a subset of the m resources. At each time $1 \le t \le n$, a decision maker is presented with an arriving customer who offers prices $\boldsymbol{Y}_t = (y_{t,1}, y_{t,2}, \ldots, y_{t,\ell})^\top$ for each of the ℓ products. The decision maker needs to decide which products to sell to maximize total expected revenues over the whole selling horizon. The sale of product bundle j at price y_j implies consumption of one unit of each of the resources that constitute product j. Formally, this can be expressed with a $m \times \ell$ incidence matrix $\boldsymbol{B} = [b_{ij}]_{\substack{i=1,\ldots,m \\ j=1,\ldots,\ell}}$ where the ij-entry $b_{ij} = 1$ if product bundle j includes resource i, and $b_{ij} = 0$ otherwise.

The state of the system is given by a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^\top$ of resource capacities, which decreases over time as selling decisions are made. We also use the m-dimensional zero vector, $\mathbf{0}$, to denote the system with no remaining capacity. Given this problem description, one can construct the Bellman equation for this problem as a multidimensional analog to (24). Specifically, if one lets $\mathcal{A}(\boldsymbol{x}) = \{\boldsymbol{a} \in \{0,1\}^\ell : \boldsymbol{B}\boldsymbol{a} \le \boldsymbol{x}\}$ be the set of available selling decisions when the resource capacities are given by $\boldsymbol{x}$, then one obtains a sequence of value functions $\{v_t(\cdot) : 1 \le t \le n\}$ that satisfies the recursive equation

$$v_t(\boldsymbol{x}) = \mathbb{E}\left[\max_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{x})} \{\boldsymbol{Y}_t^\top \boldsymbol{a} + v_{t+1}(\boldsymbol{x} - \boldsymbol{B}\boldsymbol{a})\}\right], \tag{25}$$

together with boundary conditions

$$v_t(\mathbf{0}) = 0, \quad \text{for } 1 \le t \le n, \qquad \text{and} \qquad v_{n+1}(\boldsymbol{x}) = 0, \quad \text{for all } \boldsymbol{x}.$$

It is easy to prove that the value functions given by (25) are non-decreasing in the vector of resource capacities, $\boldsymbol{x}$. Optimal action monotonicity (Property 3) is then immediately verified by appealing to Sufficient Conditions 1, while Properties 1 and 2 follow trivially form the problem definition. Hence, the variance bound (10) also holds for the network capacity control problem.

The special case with just a single resource (i.e. with $m = 1$ and $\ell = 1$ in the set-up described above) has been extensively studied in the literature (cf. Talluri and van Ryzin, 2004, Section 2.5.1). The variance bound (10) holds also for this special case, and we find that optimal total revenues have relatively small variability, a fact that usefully complements earlier work on risk-sensitive capacity-control for revenue management (Barz and Waldmann, 2007).

An important implication of the variance bound (10) for the revenue management problems discussed here is that it also implies a weak law for the optimal total revenues scaled by the total expected revenues that one obtains by implementing a fixed-bid-price heuristic, rather than the optimal policy (see Talluri and Ryzin, 1998). More generally, such a weak law also holds when the scaling is given by any (asymptotically optimal) mean-field approximation to the dynamic revenue management problem.

7.3. Investment Problems with Stochastic Opportunities

Derman, Lieberman and Ross (1975) and Prastacos (1983) study a sequential investment problem with initial capital C. At each time $1 \le t \le n$, an investment opportunity arises independently with probability $0 < p \le 1$, and, at the time of arrival, the investor gets to see the "quality," $Y_t = y$, of the investment. The investor then determines an amount, a, to invest in the opportunity, and the investment generates a return, $r(y,a)$, that is modeled as a deterministic, non-negative, non-decreasing, bounded function of the pair (y,a), continuous and differentiable at every point. One also takes $r(y,0) = 0$ for all y, so nothing ventured, nothing gained.

To derive the Bellman equation of this problem suppose that at time t the investor has an amount of capital x. With probability $1-p$ no investment opportunity arrives, no capital is invested, and the investor is left with just the opportunity to collect the expected return over periods $t+1$ through n, $v_{t+1}(x)$. Alternatively, with probability p, an investment opportunity presents itself, the investor sees its quality $Y_t = y$ and chooses an investment amount $a \le x$ to maximize the return function $a \mapsto r(y,a) + v_{t+1}(x-a)$. For $1 \le t \le n$, the investor's Bellman equation is then given by

$$v_t(x) = (1-p)v_{t+1}(x) + p \int \max_{0 \le a \le x} \{r(y,a) + v_{t+1}(x-a)\}\, dF(y), \tag{26}$$

with attending boundary conditions

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

Here, the Bellman equation (26) is also a special case of (6) as no uncertainty is realized after the decision maker chooses an optimal action.

Reward non-negativity and boundedness follow from our assumptions on $r(\cdot,\cdot)$, and the option of taking $a = 0$ gives us an appropriate do-nothing action. The map $x \mapsto v_t(x)$ is non-decreasing in x for all $1 \le t \le n$ (cf. Prastacos, 1983, Theorem 2.1), and the remaining capital X_t is non-increasing under any feasible policy. These observations confirm optimal action monotonicity (see Sufficient Conditions 1), so we have the variance bound (10).

Several natural extensions of this problem still have our three required properties. In particular, one can accommodate time-dependent opportunity probabilities $\{p_t : 1 \le t \le n\}$ or time-dependent investment quality distributions $\{F_t : 1 \le t \le n\}$.

7.4. Stochastic Depletion Problems

In a stochastic depletion problem (cf. Chan and Farias, 2009) a decision maker obtains a reward for depleting a given collection of items over a finite time horizon of n periods. The items available belong to one of M different types, indexed by m, and there are at most $\bar{x}_m$ items of type m,

$1 \le m \le M$. The state of the system at time $1 \le t \le n$ is given by a vector $\boldsymbol{x} = (x_{t,1}, \ldots, x_{t,m})$ of remaining capacities for each item type. The decision maker must choose an action a from a set of feasible actions $\mathcal{A}(t, \boldsymbol{x})$, and the action causes the depletion of $\boldsymbol{W}_t^a = (W_{t,1}^a, \ldots, W_{t,m}^a)$ items, where $\boldsymbol{W}_t^a$ is a $\{\times_{m=1}^M [0, x_{t,m}]\}$-valued random vector with known probability distribution $G_t^{\boldsymbol{x},a}$.

The choice of action $a \in \mathcal{A}(t, \boldsymbol{x})$ at time t when the state of the system is given by $\boldsymbol{x}$ generates a non-negative one-period reward $r(t, \boldsymbol{x}, a, \boldsymbol{W}_t^a)$, and the state at time $t+1$ is given by $\boldsymbol{X}_{t+1} = \boldsymbol{x} - \boldsymbol{W}_t^a$. Here, the realization of the depleting vector $\boldsymbol{W}_t^a$ becomes available only after action a is chosen. The decision maker seeks to maximize total expected rewards over the whole time horizon, and we have the Bellman equation

$$v_t(\boldsymbol{x}) = \sup_{a \in \mathcal{A}(t,\boldsymbol{x})} \mathbb{E}\left[r(t, \boldsymbol{x}, a, \boldsymbol{W}_t^a) + v_{t+1}(\boldsymbol{x} - \boldsymbol{W}_t^a)\right],$$

where, as usual, the backwards induction begins by setting $v_{n+1}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$. Here, one should note that the Bellman equation above is a special case of (6); specifically, in a stochastic depletion problem there is no exogenous Y_t that is realized before the decision maker chooses at time t an optimal action.

Chan and Farias (2009) identify two properties of many stochastic depletion problems, *value function monotonicity* and *immediate rewards*, that ensure that myopic policies are so-called 2-approximations; that is, their expected rewards are within a factor of two of optimality. The first property is equivalent to our set of Sufficient Conditions 1 for optimal action monotonicity, while the second is a stricter version of our condition (8) that is required by the existence of a do-nothing action. Thus, if the reward function is also uniformly bounded and there exists an action $a^0 \in \mathcal{A}(t, \boldsymbol{x})$ such that $\boldsymbol{W}_t^0$ equals the zero vector for all $1 \le t \le n$ and all $\boldsymbol{x}$, then the variance bound (10) holds for this class of stochastic depletion problems. As noted by Chan and Farias (2009), several challenging dynamic optimization problems of practical interest are stochastic depletion problems, and our analysis also applies to many of these.

7.5. Monotone and Unimodal Subsequences

A combinatorial optimization problem that satisfies our three properties is the sequential selection of a monotone subsequence first studied by Samuels and Steele (1981). In this problem, a decision maker sequentially views n independent and identically distributed random variables $Y_1, Y_2, \ldots, Y_n$ with uniform distribution on $[0, 1]$. The goal of the decision maker is to select a subsequence $\{Y_{\tau_1} < Y_{\tau_2} < \cdots < Y_{\tau_k} : 1 \le \tau_1 < \tau_2 < \cdots < \tau_k \le n\}$ of maximal expected length.

For each decision time $1 \le t \le n$, the state of the system can be represented with the value, X_t, of the last observation selected prior to time t, and, by convention, we set $X_1 \equiv 0$. For $X_t = x$, the Bellman equation is then given by

$$v_t(x) = x v_{t+1}(x) + \int_x^1 \max\{v_{t+1}(x), 1 + v_{t+1}(y)\}\, dy, \tag{27}$$

with $v_{n+1}(x) = 0$ for all $x \in [0,1]$.

To derive the Bellman equation (27) we first suppose that at time t the value of the last observation selected is equal to x. With probability x the arriving observation, Y_t, is smaller than x and cannot be selected, and we are left with the expected reward to-go, $v_{t+1}(x)$. With probability $1-x$, the arriving observation $Y_t = y$ exceeds the value of the last observation selected, and we can choose the action that yields the largest sum of the one-period reward and the expected reward to-go. Not selecting yields $v_{t+1}(x)$ while selecting yields $1 + v_{t+1}(y)$.

In this MDP, we have rewards that are non-negative and uniformly bounded by one. Any time we see an observation we can reject it, leaving the value of the last observation selected unchanged and receiving zero reward. So, we know that reward non-negativity and boundedness, as well as the existence of a do-nothing action are satisfied. The definition of the monotone subsequence problem guarantees that the selected values $\{X_t : 1 \le t \le n\}$ are a non-decreasing sequence, and, by induction on the Bellman equation, one can check that the value function $v_t(\cdot)$ is strictly decreasing on $[0,1]$. Thus, optimal action monotonicity is also verified (see Sufficient Conditions 2), and we have the variance bound (10).

For the monotone subsequence problem, the upper bound on the variance was observed by Arlotto and Steele (2011) where special features of the selection problem were used to prove a complementary lower bound of the same order. Specifically, in this context, one has

$$(1/3)\mathbb{E}[R_n(\pi_n^*)] - 2 \le \mathrm{Var}[R_n(\pi_n^*)] \le \mathbb{E}[R_n(\pi_n^*)] \qquad \text{for all } n \ge 1.$$

The expected value is known to satisfy $\mathbb{E}[R_n(\pi_n^*)] \sim \sqrt{2n}$ as $n \to \infty$, so the two inequalities above imply that the variance $\mathrm{Var}[R_n(\pi_n^*)]$ has the same order as the mean $\mathbb{E}[R_n(\pi_n^*)]$.

There are several related combinatorial problems for which one has the variance bound (10). For example, it holds for the multidimensional monotone subsequence problem studied by Baryshnikov and Gnedin (2000) and the unimodal subsequence problem studied by Arlotto and Steele (2011).

8. Conclusions

In the literature on Markov decision problems, it is uncommon to consider the *distributional properties* of the optimal total reward. Nevertheless, in many situations the economic value of an optimal solution cannot be judged without some understanding of more than just its expected value. At a minimum, one needs some understanding of the variance of the solution. Here, we have isolated a class of MDPs for which we prove that the variance of the optimal total reward is relatively small. Moreover, the class is characterized by three natural properties that are often easily verified.

Clearly, one could ask for more. In particular, it would be useful to know when one can provide a lower bound on the variance to complement the upper bound given by Theorem 1. In some special

cases — such as the one mentioned in Section 7.5 — there is a complementary lower bound that is of the same order as the upper bound, but, so far, it has not been possible to give general criteria for this useful situation.

Ideally, one could also ask for limit theorems for distribution of the optimal total reward, but this problem is usually intractable because of the strong time dependence of the optimal policy. Even for the classic knapsack problem of Section 1.1 many basic questions remain open; for example, the asymptotic behavior of the variance is unknown.

Nevertheless, when the time dependence of the optimal policy is not overly strong, it is sometimes possible to characterize the asymptotic behavior of the variance — or even to obtain a limit theorem for the distribution of the total reward. One idea is to try to approximate the optimal policy with a stationary policy that is easier to analyze. For example, this approach was used in Arlotto, Chen, Shepp and Steele (2011) and Arlotto and Steele (2014) to study the optimal sequential selection of an alternating subsequence from a sequence of n independent and identically distributed observations. One can also develop the distributional limit theory of MDPs by focusing on those problems with a more intrinsic stationary formulation, such as MDPs with Poisson arrivals, or a geometric number of arrivals, or with infinite-horizon discounting. For example, Bruss and Delbaen (2001; 2004) study a Poissonized version of the monotone subsequence problem of Section 7.5, and they obtain both precise asymptotics for the variance and a central limit theorem.

Passage to a more stationary formulation almost always brightens the prospects for precise asymptotic analysis, but, if one is really interested in the finite horizon problem, there is still work to be done. Somehow the findings in the stationary formulation need to be translated into results for the finite horizon problem. This task is analogous to classical Tauberian theory (see, e.g. Korevaar, 2004), but so far there are no general results — or even precise conjectures.

Acknowledgments

The authors are grateful to an Associate Editor and two anonymous referees whose comments helped us to significantly improve the paper.

This material is based upon work supported by the National Science Foundation under Grant No. CMMI-0800645. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

References

Arlotto, A., Chen, R. W., Shepp, L. A. and Steele, J. M. (2011), 'Online selection of alternating subsequences from a random sample', *J. Appl. Probab.* **48**(4), 1114–1132.

Arlotto, A. and Steele, J. M. (2011), 'Optimal sequential selection of a unimodal subsequence of a random sequence', *Combinatorics, Probability and Computing* **20**(06), 799–814.

Arlotto, A. and Steele, J. M. (2014), 'Optimal online selection of an alternating subsequence: a central limit theorem', *Adv. in Appl. Probab.* **46**(2).

Baryshnikov, Y. M. and Gnedin, A. V. (2000), 'Sequential selection of an increasing sequence from a multidimensional random sample', *Ann. Appl. Probab.* **10**(1), 258–267.

Barz, C. and Waldmann, K.-H. (2007), 'Risk-sensitive capacity control in revenue management', *Math. Methods Oper. Res.* **65**(3), 565–579.

Baykal-Gürsoy, M. and Ross, K. W. (1992), 'Variability sensitive Markov decision processes', *Math. Oper. Res.* **17**(3), 558–571.

Bertsekas, D. P. and Shreve, S. E. (1978), *Stochastic optimal control: the discrete time case*, Vol. 139 of *Mathematics in Science and Engineering*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, NY.

Boucheron, S., Lugosi, G. and Massart, P. (2013), *Concentration Inequalities*, Oxford University Press, Oxford. A Nonasymptotic Theory of Independence.

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

Bruss, F. T. and Delbaen, F. (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 Delbaen, F. (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 Robertson, J. B. (1991), '"Wald's lemma" for sums of order statistics of i.i.d. random variables', *Adv. in Appl. Probab.* **23**(3), 612–623.

Chan, C. W. and Farias, V. F. (2009), 'Stochastic depletion problems: effective myopic policies for a class of dynamic optimization problems', *Math. Oper. Res.* **34**(2), 333–350.

Chung, K.-J. (1994), 'Mean-variance tradeoffs in an undiscounted mdp: The unichain case', *Operations Research* **42**(1), pp. 184–188.

Chung, K. J. and Sobel, M. J. (1987), 'Discounted MDPs: distribution functions and exponential utility maximization', *SIAM J. Control Optim.* **25**(1), 49–62.

Coffman, Jr., E. G., Flatto, L. and Weber, R. R. (1987), 'Optimal selection of stochastic intervals under a sum constraint', *Adv. in Appl. Probab.* **19**(2), 454–473.

Derman, C., Lieberman, G. J. and Ross, S. M. (1975), 'A stochastic sequential allocation model', *Operations Res.* **23**(6), 1120–1130.

Desai, V. V., Farias, V. F. and Moallemi, C. C. (2012), Bounds for markov decision processes, *in* 'Reinforcement Learning and Approximate Dynamic Programming for Feedback Control', F.L. Lewis and D. Liu eds., Wiley-IEEE Press.

Feinberg, E. A. and Fei, J. (2009), 'An inequality for variances of the discounted rewards', *J. Appl. Probab.* **46**(4), 1209–1212.

Filar, J. A., Kallenberg, L. C. M. and Lee, H.-M. (1989), 'Variance-penalized Markov decision processes', *Math. Oper. Res.* **14**(1), 147–161.

Haskell, W. B. and Jain, R. (2012), 'Stochastic dominance-constrained Markov decision processes', *arXiv e-Print 1206.4568* .

Hill, T. P. and Kertz, R. P. (1992), A survey of prophet inequalities in optimal stopping theory, *in* 'Strategies for sequential search and selection in real time (Amherst, MA, 1990)', Vol. 125 of *Contemp. Math.*, Amer. Math. Soc., Providence, RI, pp. 191–207.

Huang, Y. and Kallenberg, L. C. M. (1994), 'On finding optimal policies for Markov decision chains: a unifying framework for mean-variance-tradeoffs', *Math. Oper. Res.* **19**(2), 434–448.

James, B., James, K. and Qi, Y. (2008), 'Limit theorems for correlated Bernoulli random variables', *Statist. Probab. Lett.* **78**(15), 2339–2345.

Jaquette, S. C. (1972), 'Markov decision processes with a new optimality criterion: small interest rates', *Ann. Math. Statist.* **43**, 1894–1901.

Jaquette, S. C. (1973), 'Markov decision processes with a new optimality criterion: discrete time', *Ann. Statist.* **1**, 496–505.

Kawai, H. (1987), 'A variance minimization problem for a Markov decision process', *European J. Oper. Res.* **31**(1), 140–145.

Korevaar, J. (2004), *Tauberian theory: a century of developments*, Vol. 329 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, Springer-Verlag, Berlin.

Mannor, S. and Tsitsiklis, J. N. (2013), 'Algorithmic aspects of mean-variance optimization in Markov decision processes', *European J. Oper. Res.* **231**(3), 645–653. Shorter version in *Proceedings of the 28th International Conference on Machine Learning*, 177–184, 2011.

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

Prastacos, G. P. (1983), 'Optimal sequential investment decisions under conditions of uncertainty', *Management Science* **29**(1), 118–134.

Rhee, W. and Talagrand, M. (1991), 'A note on the selection of random variables under a sum constraint', *J. Appl. Probab.* **28**(4), 919–923.

Ruszczyński, A. (2010), 'Risk-averse dynamic programming for Markov decision processes', *Math. Program.* **125**(2, Ser. B), 235–261.

Samuels, S. M. and Steele, J. M. (1981), 'Optimal sequential selection of a monotone sequence from a random sample', *Ann. Probab.* **9**(6), 937–947.

Sobel, M. J. (1982), 'The variance of discounted Markov decision processes', *J. Appl. Probab.* **19**(4), 794–802.

Sobel, M. J. (1985), 'Maximal mean/standard deviation ratio in an undiscounted MDP', *Oper. Res. Lett.* **4**(4), 157–159.

Sobel, M. J. (1994), 'Mean-variance tradeoffs in an undiscounted mdp', *Operations Research* **42**(1), 175–183.

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

Talluri, K. T. and van Ryzin, G. J. (2004), *The theory and practice of revenue management*, International Series in Operations Research & Management Science, 68, Kluwer Academic Publishers, Boston, MA.

White, D. J. (1988), 'Mean, variance, and probabilistic criteria in finite Markov decision processes: a review', *J. Optim. Theory Appl.* **56**(1), 1–29.

Williams, D. (1991), *Probability with martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge.