J. Appl. Prob. **48,** 954–967 (2011)
Printed in England

SENSITIVITY ANALYSIS IN MARKOV DECISION PROCESSES WITH UNCERTAIN REWARD PARAMETERS

CHIN HON TAN [*] [**] AND
JOSEPH C. HARTMAN,[*] *University of Florida*

Abstract

Sequential decision problems can often be modeled as Markov decision processes. Classical solution approaches assume that the parameters of the model are known. However, model parameters are usually estimated and uncertain in practice. As a result, managers are often interested in how estimation errors affect the optimal solution. In this paper we illustrate how sensitivity analysis can be performed directly for a Markov decision process with uncertain reward parameters using the Bellman equations. In particular, we consider problems involving (i) a single stationary parameter, (ii) multiple stationary parameters, and (iii) multiple nonstationary parameters. We illustrate the applicability of this work through a capacitated stochastic lot-sizing problem.

Keywords: Markov decision process; dynamic programming; sensitivity analysis

2010 Mathematics Subject Classification: Primary 90C40
Secondary 90C39; 90C31

1. Introduction

The Markov decision process (MDP) framework has been used by researchers to model a variety of sequential decision problems because it can account for the dynamics of a complex system and handle a wide range of reward functions. An MDP is defined by a set of states, with a set of potential actions associated with each state. Classical solution approaches assume that the parameters of the model, including rewards, transition probabilities, and the discount factor, are known (see [15], [16], or [26]). However, these are often estimated and uncertain in practice. For example, it is difficult to quantify the cost of not having an item in the store upon the arrival of a customer (stock-out cost).

White and El-Deib [25] identified optimal policies for some realization of the imprecise parameters, termed nondominated policies, for an MDP with imprecise rewards. Harmanec [7] studied a similar problem where the imprecision was defined in the transition probabilities, rather than the rewards. However, a decision maker can only implement a single policy in practice. One approach is to assume that the imprecision is resolved in the most pessimistic scenario (see [9] and [14]). This is often referred to as the max–min policy. However, it has been highlighted that max–min policies can be overly conservative and may not be practical in reality [22].

Managers are often interested in how an optimal solution changes with deviations in the model parameters. The typical approach to answering this question is to solve the problem

Received 6 May 2011; revision received 20 July 2011.
[*] Postal address: Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall, PO Box 116595, Gainesville, FL 32611-6595, USA.
[**] Email address: chinhon@ufl.edu

for different values of the uncertain parameter, but this can be very time consuming when the problem is large. For example, Topaloglu and Powell [20] were interested in the benefits of adding an extra vehicle or load in a dynamic fleet management model. Sandvik and Thorlund-Petersen [17] were interested in the conditions where there is at most one critical risk tolerance value, such that the knowledge of whether the individual's risk tolerance is above or below that value is sufficient for identifying the preferred decision.

In this paper we consider an MDP where rewards are expressed as affine functions of uncertain parameters. Problems of this form abound in the MDP literature, including the lot-sizing problem [13, pp. 151–159], the equipment replacement problem [18], the sequential search problem [10], and various resource allocation problems (see [3], [4], and [6]). Bounds on the perturbations in the state values for a given policy are computed in [12]. We are interested in the maximum range parameters are allowed to vary such that a policy remains optimal. Hopp [8] derived bounds on the minimum perturbations in the future state values required to change the current optimal decision (i.e. at time 0) and extended the results to perturbations in the rewards at each state. Our model allows for dependencies between the uncertainties in the rewards associated with different actions and states. Another important difference is that we are not deriving bounds, but computing the actual range of values our parameters are allowed to vary.

A single-parameter analysis provides insight on the stability of the solution with respect to a particular parameter. However, estimation errors can exist for multiple parameters. Wendell [24] proposed finding a tolerance level which indicates the maximum percentage parameters are allowed to vary from their base value such that the optimal basis of a linear program remains optimal. In this paper we illustrate how the maximum tolerance can be obtained for our MDP when multiple uncertain parameters are allowed to vary simultaneously. In addition, we allow these parameters to be nonstationary.

The primary contributions of this paper to the literature on MDPs and sensitivity analysis are as follows.

1. We obtain the range in which a single parameter and multiple parameters are allowed to vary while maintaining the optimality of the current solution (Propositions 1 and 2).

2. We illustrate how the maximum allowable tolerance can be computed when uncertain parameters are nonstationary (Proposition 3) and show that it cannot be greater than the allowable tolerance of the stationary problem (Theorem 3).

3. We derive the conditions where the tolerances of the stationary and nonstationary rewards problems are the same (Corollary 2) and the conditions where they differ (Theorem 4). In particular, we show that, under mild assumptions, the tolerances of lot-sizing problems with uncertain ordering costs and backlog penalties differ when the maximum allowable tolerance is associated with an action that changes the reorder point (Theorem 5).

This paper is organized as follows. In the next section we describe our stationary rewards model and illustrate how single-parameter sensitivity analysis can be performed for this problem. In addition, we demonstrate how the maximum allowable tolerance can be computed when the uncertain parameters are allowed to vary simultaneously. Next, we illustrate how the maximum allowable tolerance can be computed for the nonstationary rewards problem. In Section 4 we study and discuss the difference in the maximum allowable tolerance of the two problems. We conclude with a short summary and future research directions.

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

2. Stationary rewards

Consider a finite state, finite action, infinite horizon MDP. Let S and $A(s)$ denote the set of states and the set of actions available with $s \in S$, respectively. Each $a_s \in A(s)$ transitions the system from state s to state s' with probability $P^{a_s}(s')$. Let $\tilde{r}^{a_s}$ denote the reward associated with a_s, expressed as an affine function of N uncertain parameters:

$$\tilde{r}^{a_s} = \lambda_0^{a_s} + \boldsymbol{\lambda}^{a_s}\tilde{\boldsymbol{x}}.$$

Here $\lambda_0^{a_s}$ is some known constant, $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N)^\top$ represents the uncertain parameters, and $\boldsymbol{\lambda}^{a_s} = (\lambda_1^{a_s}, \lambda_2^{a_s}, \ldots, \lambda_N^{a_s})$ represents the respective known coefficients. We assume that an estimation of $\tilde{x}_i$ is available for $i = 1, 2, \ldots, N$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)^\top$ denote the vector of estimated parameter values. In addition, let r^{a_s} denote the estimated reward associated with a_s:

$$r^{a_s} = \lambda_0^{a_s} + \boldsymbol{\lambda}^{a_s}\boldsymbol{x}.$$

Let $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_N)^\top$ denote the corresponding estimation error, where

$$\rho_i = \frac{\tilde{x}_i - x_i}{x_i},$$

and let $\boldsymbol{\Delta}^{a_s} = (\Delta_1^{a_s}, \Delta_2^{a_s}, \ldots, \Delta_N^{a_s})$ denote the corresponding coefficient, where

$$\Delta_i^{a_s} = \lambda_i^{a_s} x_i,$$

such that $\tilde{r}^{a_s}$ can be re-expressed as

$$\tilde{r}^{a_s} = r^{a_s} + \boldsymbol{\Delta}^{a_s}\boldsymbol{\rho}.$$

As in sensitivity analysis, we are interested in the stability of the solution obtained using the estimated parameters $\boldsymbol{x}$. In this section we obtain the relationship between the estimation errors and the total reward received. In addition, we compute the range of error values where the current solution remains optimal. Here, we consider a problem where the value of $\boldsymbol{\rho}$ is uncertain but stationary. Since the rewards are stationary (i.e. do not vary with time), there must exist a stationary optimal policy where the action is determined solely by the state of the process (see [16, pp. 146–158]). Let γ and Π denote the periodic discount factor and the set of all possible stationary policies, respectively. We assume that r^{a_s} is bounded and $\gamma < 1$ to ensure that the value function is finite. Let $\pi(s) \in A(s)$ denote the action that is taken at state s under $\pi \in \Pi$, and let $V_s^\pi(\boldsymbol{\rho})$ denote the value function of state s under policy π for a given $\boldsymbol{\rho}$. The value function of a state can be expressed by the following recursive equation:

$$V_s^\pi(\boldsymbol{\rho}) = r^{\pi(s)} + \boldsymbol{\Delta}^{\pi(s)}\boldsymbol{\rho} + \gamma \sum_{s' \in S} P^{\pi(s)}(s') V_{s'}^\pi(\boldsymbol{\rho}) \quad \text{for all } s, \pi. \tag{1}$$

Note that $V_s^\pi(\boldsymbol{\rho})$ depends on the value functions of the other states. Hence, it is convenient to express (1) in matrix form. Let $\boldsymbol{V}^\pi(\boldsymbol{\rho}) = (V_1^\pi(\boldsymbol{\rho}), V_2^\pi(\boldsymbol{\rho}), \ldots, V_{|S|}^\pi(\boldsymbol{\rho}))^\top$, $\boldsymbol{r}^\pi = (r^{\pi(1)}, r^{\pi(2)}, \ldots, r^{\pi(|S|)})^\top$,

$$\boldsymbol{\Delta}^\pi = \begin{pmatrix} \Delta_1^{\pi(1)} & \Delta_2^{\pi(1)} & \cdots & \Delta_N^{\pi(1)} \\ \Delta_1^{\pi(2)} & \Delta_2^{\pi(2)} & \cdots & \Delta_N^{\pi(2)} \\ \vdots & \ddots & \ddots & \vdots \\ \Delta_1^{\pi(|S|)} & \Delta_2^{\pi(|S|)} & \cdots & \Delta_N^{\pi(|S|)} \end{pmatrix},$$

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

and

$$\boldsymbol{P}^{\pi} = \begin{pmatrix} P^{\pi(1)}(1) & P^{\pi(1)}(2) & \cdots & P^{\pi(1)}(|S|) \\ P^{\pi(2)}(1) & P^{\pi(2)}(2) & \cdots & P^{\pi(2)}(|S|) \\ \vdots & \ddots & \ddots & \vdots \\ P^{\pi(|S|)}(1) & P^{\pi(|S|)}(2) & \cdots & P^{\pi(|S|)}(|S|) \end{pmatrix}.$$

We can express $\boldsymbol{V}^{\pi}(\boldsymbol{\rho})$ as

$$\begin{aligned} \boldsymbol{V}^{\pi}(\boldsymbol{\rho}) &= \boldsymbol{r}^{\pi} + \boldsymbol{\Delta}^{\pi}\boldsymbol{\rho} + \gamma \boldsymbol{P}^{\pi}\boldsymbol{V}^{\pi}(\boldsymbol{\rho}), \\ (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})\boldsymbol{V}^{\pi}(\boldsymbol{\rho}) &= \boldsymbol{r}^{\pi} + \boldsymbol{\Delta}^{\pi}\boldsymbol{\rho}, \\ \boldsymbol{V}^{\pi}(\boldsymbol{\rho}) &= (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})^{-1}(\boldsymbol{r}^{\pi} + \boldsymbol{\Delta}^{\pi}\boldsymbol{\rho}), \\ \boldsymbol{V}^{\pi}(\boldsymbol{\rho}) &= (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})^{-1}\boldsymbol{r}^{\pi} + (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})^{-1}\boldsymbol{\Delta}^{\pi}\boldsymbol{\rho}, \\ \boldsymbol{V}^{\pi}(\boldsymbol{\rho}) &= \boldsymbol{V}^{\pi}(\boldsymbol{0}) + (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})^{-1}\boldsymbol{\Delta}^{\pi}\boldsymbol{\rho}. \end{aligned} \tag{2}$$

Let $\boldsymbol{V}(\boldsymbol{\rho}) = \max_{\pi} \boldsymbol{V}^{\pi}(\boldsymbol{\rho})$. For a given $\boldsymbol{\rho}$ (including $\boldsymbol{\rho} = \boldsymbol{0}$), the policy that maximizes $\boldsymbol{V}^{\pi}(\boldsymbol{\rho})$ can be obtained through value and/or policy iteration approaches (see [16, pp. 158–195]). Let $\widetilde{\pi}$ denote a policy that maximizes $\boldsymbol{V}^{\pi}(\boldsymbol{0})$. It follows from (2) that, within the region where $\widetilde{\pi}$ is the optimal policy, the marginal change in $\boldsymbol{V}(\boldsymbol{\rho})$ is $(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widetilde{\pi}})^{-1}\boldsymbol{\Delta}^{\widetilde{\pi}}$.

In linear programs, sensitivity analysis is performed by deriving a set of necessary and sufficient conditions for optimality based on the reduced cost of each variable and finding the range of values for which these conditions hold [1, pp. 307–314]. In theory, MDPs can be formulated as linear programs [11] and the allowable $\boldsymbol{\rho}$ values can be obtained by applying results from parametric linear programming (see [5] and [23]) on the dual of the associated linear program (see [19]). However, the set of necessary and sufficient conditions (i.e. Bellman equations) is readily available for MDPs [2]. We state the Bellman equations for this problem, re-express them in a compact form, and use this form to obtain the maximum allowable error for the single-parameter and multiple-parameter problems in Sections 2.1 and 2.2, respectively.

Let $\boldsymbol{P}^{a_s} = (P^{a_s}(1), P^{a_s}(2), \ldots, P^{a_s}(|S|))$. Note that $\boldsymbol{P}^{\pi(s)}$ is the sth row of $\boldsymbol{P}^{\pi}$. The Bellman equations for the stationary rewards problem are

$$V_s(\boldsymbol{\rho}) = \max_{a_s \in A(s)} \{r^{a_s} + \boldsymbol{\Delta}^{a_s}\boldsymbol{\rho} + \gamma \boldsymbol{P}^{a_s}\boldsymbol{V}(\boldsymbol{\rho})\} \quad \text{for all } s \in S,$$

and $\widetilde{\pi}$ is optimal if and only if

$$r^{\widetilde{\pi}(s)} + \boldsymbol{\Delta}^{\widetilde{\pi}(s)}\boldsymbol{\rho} + \gamma \boldsymbol{P}^{\widetilde{\pi}(s)}\boldsymbol{V}^{\widetilde{\pi}}(\boldsymbol{\rho}) \geq r^{a_s} + \boldsymbol{\Delta}^{a_s}\boldsymbol{\rho} + \gamma \boldsymbol{P}^{a_s}\boldsymbol{V}^{\widetilde{\pi}}(\boldsymbol{\rho}) \quad \text{for all } s \in S,\ a_s \in A(s). \tag{3}$$

Define

$$c^{\widetilde{\pi},a_s} = r^{\widetilde{\pi}(s)} - r^{a_s} + \gamma(\boldsymbol{P}^{\widetilde{\pi}(s)} - \boldsymbol{P}^{a_s})(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widetilde{\pi}})^{-1}\boldsymbol{r}^{\widetilde{\pi}}$$

and

$$\boldsymbol{b}^{\widetilde{\pi},a_s} = \boldsymbol{\Delta}^{\widetilde{\pi}(s)} - \boldsymbol{\Delta}^{a_s} + \gamma(\boldsymbol{P}^{\widetilde{\pi}(s)} - \boldsymbol{P}^{a_s})(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widetilde{\pi}})^{-1}\boldsymbol{\Delta}^{\widetilde{\pi}}.$$

Note that $c^{\widetilde{\pi},a_s}$ is the marginal decrease in the estimated reward that results from a single perturbation of the action at s, while $\boldsymbol{b}^{\widetilde{\pi},a_s}$ is the marginal change in the estimation error that results from that action perturbation. Using our definitions of $c^{\widetilde{\pi},a_s}$ and $\boldsymbol{b}^{\widetilde{\pi},a_s}$, (3) can be rewritten as

$$c^{\widetilde{\pi},a_s} + \boldsymbol{b}^{\widetilde{\pi},a_s}\boldsymbol{\rho} \geq 0 \quad \text{for all } s \in S,\ a_s \in A(s). \tag{4}$$

Let H denote the region where $\widetilde{\pi}$ is optimal:

$$H = \{\boldsymbol{\rho} \colon V^{\widetilde{\pi}}(\boldsymbol{\rho}) \geq V^{\pi}(\boldsymbol{\rho}) \text{ for all } \pi \in \Pi\}.$$

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

Theorem 1. *Given (2), H is closed and convex.*

Proof. It follows from (4) that H is the intersection of closed half-spaces. Hence, H is closed and convex.

2.1. Single-parameter sensitivity analysis

In single-parameter sensitivity analysis we are interested in the set of ρ_i values where $\widetilde{\pi}$ remains optimal when $\rho_{j\neq i} = 0$. It follows from Theorem 1 that there exist constants $\rho_i^{(l)}, \rho_i^{(u)} \in \mathbb{R}$ such that $\widetilde{\pi}$ remains optimal when $\rho_{j\neq i} = 0$ and $\rho_i \in [\rho_i^{(l)}, \rho_i^{(u)}]$.

Proposition 1. *Given (2),*

$$\rho_i^{(l)} = \begin{cases} \infty, & b_i^{\widetilde{\pi},a_s} \leq 0 \text{ for all } s \in S,\ a_s \in A(s), \\ \max\limits_{b_i^{\widetilde{\pi},a_s} > 0 \text{ for all } s,a_s} \left\{ -\dfrac{c^{\widetilde{\pi},a_s}}{b_i^{\widetilde{\pi},a_s}} \right\}, & \text{otherwise} \end{cases}$$

and

$$\rho_i^{(u)} = \begin{cases} -\infty, & b_i^{\widetilde{\pi},a_s} \geq 0 \text{ for all } s \in S,\ a_s \in A(s), \\ \min\limits_{b_i^{\widetilde{\pi},a_s} < 0 \text{ for all } s,a_s} \left\{ -\dfrac{c^{\widetilde{\pi},a_s}}{b_i^{\widetilde{\pi},a_s}} \right\}, & \text{otherwise.} \end{cases}$$

Proof. Let $b_i^{\widetilde{\pi},a_s}$ denote the ith entry of $\boldsymbol{b}^{\widetilde{\pi},a_s}$. Setting $\rho_{j\neq i} = 0$ and using our definitions of $c^{\widetilde{\pi},a_s}$ and $b_i^{\widetilde{\pi},a_s}$, we obtain the following necessary and sufficient optimality conditions from (4):

$$\rho_i \geq -\frac{c^{\widetilde{\pi},a_s}}{b_i^{\widetilde{\pi},a_s}} \quad \text{when} \quad b_i^{\widetilde{\pi},a_s} > 0 \tag{5}$$

and

$$\rho_i \leq -\frac{c^{\widetilde{\pi},a_s}}{b_i^{\widetilde{\pi},a_s}} \quad \text{when} \quad b_i^{\widetilde{\pi},a_s} < 0. \tag{6}$$

Given that these hold for all values of ρ_i at optimality, the extreme values are of interest and the proposition follows from (5) and (6).

Next, we illustrate how single-parameter sensitivity analysis can be conducted for a capacitated stochastic lot-sizing problem with uncertain ordering cost and backlog penalty. The interested reader is referred to [13] for an introduction on lot-sizing problems.

Example 1. (*Capacitated stochastic lot-sizing problem.*) Consider a lot-sizing problem where the probability distribution of demand in each period is stationary and given by $\mathrm{P}(D = 0) = \mathrm{P}(D = 1) = \mathrm{P}(D = 2) = \frac{1}{3}$. Each item sells for \$150. The inventory capacity of the system is three and backlogging is allowed. There are a total of five states, $S = \{-1, 0, 1, 2, 3\}$. The index of each state represents the amount of inventory that is available at the beginning of the period. Orders are placed at the start of the period and an order must be placed if there is no inventory. Hence, the set of feasible actions (i.e. order quantity) for each state is $A(-1) = \{2, 3, 4\}$, $A(0) = \{1, 2, 3\}$, $A(1) = \{0, 1, 2\}$, $A(2) = \{0, 1\}$, and $A(3) = \{0\}$. We assume that the stock arrives at the end of the period in which it is ordered and the demand is also realized at the end of the period. The production cost of each item is \$20 and the holding cost of each item is \$5 per period. The value of the ordering cost and backlog penalty are

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

unclear, but believed to be \$40 and \$100, respectively. We analyze each uncertain parameter independently such that $N = 1$ for each case. The problem extends across an infinite horizon with a per-period discount factor of 0.9.

Let ρ_1 and ρ_2 denote the estimation error in the order cost and backlog penalty, respectively. It follows that the expected reward associated with $\pi(s)$ is

$$r^{\pi(s)} = \begin{cases} (0.9)(150)\left(\frac{1}{3}+\frac{2}{3}\right) - 20\pi(s) - 40 - 100 & \text{if } s = -1, \\ (0.9)(150)\left(\frac{1}{3}+\frac{2}{3}\right) - 5s - 20\pi(s) - 40I_{\pi(s)} & \text{otherwise,} \end{cases}$$

and

$$\boldsymbol{\Delta}^{\pi} = \begin{pmatrix} -I_{\pi(-1)} & -100 \\ -I_{\pi(0)} & 0 \\ -I_{\pi(1)} & 0 \\ -I_{\pi(2)} & 0 \\ -I_{\pi(3)} & 0 \end{pmatrix},$$

where

$$I_{\pi(s)} = \begin{cases} 40 & \text{if } \pi(s) \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The transition probabilities are defined by

$$P^{a_s}(s') = \begin{cases} \frac{1}{3} & \text{if } -2 \leq s' - (s + a_s) \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Solving the MDP described above with the policy iteration approach, we obtain $\widetilde{\pi} = (4, 3, 2, 0, 0)$ and $\boldsymbol{V}^{\widetilde{\pi}}(\mathbf{0}) = (\$723.5, \$843.5, \$858.5, \$908, \$928.5)^{\top}$. It follows from (2) that the marginal change in $\boldsymbol{V}^{\widetilde{\pi}}(\boldsymbol{\rho})$ is $(-196, -196, -196, -168, -156)^{\top}\rho_1 + (-100, 0, 0, 0, 0)^{\top}\rho_2$.

There are two uncertain parameters in this problem. We analyze the optimal region of $\widetilde{\pi}$ with respect to ρ_1 by setting $\rho_2 = 0$. The corresponding $c^{\widetilde{\pi},a_s}$ and $b_i^{\widetilde{\pi},a_s}$ values of each action are listed in Table 1. At state -1, the optimal decision is to order four units when $\boldsymbol{\rho} = \mathbf{0}$.

TABLE 1: Single-parameter sensitivity analysis and τ.

s	a_s	$c^{\widetilde{\pi},a_s}$	$b_1^{\widetilde{\pi},a_s}$	$b_2^{\widetilde{\pi},a_s}$	$-c^{\widetilde{\pi},a_s}/b_1^{\widetilde{\pi},a_s}$	$-c^{\widetilde{\pi},a_s}/b_2^{\widetilde{\pi},a_s}$	τ_{s,a_s}
−1	2	40.85	20.4	30	−2.00	−1.36	0.81
−1	3	5.5	12	0	−0.46	—	0.46
−1	4	0	0	0	—	—	—
0	1	40.85	20.4	30	−2.00	−1.36	0.81
0	2	5.5	12	0	−0.46	—	0.46
0	3	0	0	0	—	—	—
1	0	0.85	−19.6	30	0.04	−0.03	0.02
1	1	5.5	12	0	−0.46	—	0.46
1	2	0	0	0	—	—	—
2	0	0	0	0	—	—	—
2	1	34.5	28	0	−1.23	—	1.23
3	0	0	0	0	—	—	—

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

However, if the ordering cost decreases by more than 46% (with ρ_2 remaining 0), ordering three units will result in higher expected profits than ordering four units. In addition, the decision to order four units at state -1 is better than that of ordering two units so long as the ordering cost does not decline by 200% of its estimated value (i.e. $\$40(1 - 2.00) = -\$40$). Assuming that the ordering cost must be nonnegative, it follows that $a_{-1} = 2$ is suboptimal when $\rho_2 = 0$. It follows from Table 1 that $\rho^{(l)} = \max\{-2.00, -0.46, -1.23, -\infty\} = -0.46$ and $\rho^{(u)} = \min\{0.04, \infty\} = 0.04$. Hence, $\widetilde{\pi}$ remains optimal for all $\rho_1 \in [-0.46, 0.04]$ given that $\rho_2 = 0$.

In a similar fashion, we find that $\widetilde{\pi}$ remains optimal for all $\rho_2 \in [-0.03, \infty]$, given that $\rho_1 = 0$, from the results in Table 1.

2.2. Tolerance approach

When there are multiple uncertain parameters, Wendell [24] proposed finding a tolerance level τ, where τ is the maximum ratio uncertain parameters are allowed to vary from their base value such that the optimal basis of a linear program remains optimal. Note that τ is, by definition, nonnegative. Geometrically speaking, this entails finding the largest hypercube that is contained in the critical region (i.e. H). Wendell showed that the maximum allowable tolerance, which we denote by τ^*, can be obtained by finding the maximum tolerance with respect to each constraint independently. Following a similar approach, we illustrate how the tolerance level can be computed for an MDP with uncertain rewards. Let τ_{s,a_s} denote the maximum tolerance allowable by (4) for state s and action a_s:

$$\tau_{s,a_s} = \max\{y\colon c^{\widetilde{\pi},a_s} + \boldsymbol{b}^{\widetilde{\pi},a_s}\boldsymbol{\rho} \geq 0 \text{ and } |\rho_i| \leq y \text{ for } i = 1, 2, \ldots, N\}.$$

Proposition 2. *Given (2),*

$$\tau^* = \min_{s,a_s} \frac{c^{\widetilde{\pi},a_s}}{\sum_{i=1}^{N} |b_i^{\widetilde{\pi},a_s}|}.$$

Proof. To find the maximum allowable tolerance for each constraint expressed by (4), we consider the worst case scenario where

$$\rho_i = \begin{cases} \tau_{s,a_s} & \text{if } b_i^{\widetilde{\pi},a_s} \leq 0, \\ -\tau_{s,a_s} & \text{otherwise.} \end{cases}$$

Since $c^{\widetilde{\pi},a_s}$ is, by definition, nonnegative, we obtain from (4) the following expression for τ_{s,a_s}:

$$\tau_{s,a_s} = \frac{c^{\widetilde{\pi},a_s}}{\sum_{i=1}^{N} |b_i^{\widetilde{\pi},a_s}|}.$$

Since τ^* cannot be larger than any of the individual tolerances τ_{s,a_s},

$$\tau^* = \min_{s,a_s} \tau_{s,a_s} = \min_{s,a_s} \frac{c^{\widetilde{\pi},a_s}}{\sum_{i=1}^{N} |b_i^{\widetilde{\pi},a_s}|}.$$

We reconsider Example 1, allowing for simultaneous perturbations in the ordering cost and backlog penalty, as follows.

Example 2. (*Tolerance approach.*) Consider Example 1 again. When the ordering cost and backlog penalty are allowed to perturb simultaneously, it follows from Table 1 and Proposition 2

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

that $\tau^* = 0.02$ and is associated with the action $a_1 = 0$. This implies that $\widetilde{\pi}$ will remain optimal so long as the ordering cost and backlog penalty do not deviate from their current estimates by more than 2%. In particular, it is suboptimal to order when $s = 1$ if we underestimate the ordering cost and overestimate the penalty cost by more than 2% each.

3. Nonstationary rewards

In this section we consider the nonstationary rewards problem where uncertain parameters are allowed to vary at each period. Let υ denote the tolerance for the nonstationary rewards problem. In addition, let $\boldsymbol{\omega}$ represent the estimation error in the nonstationary reward problem, where $\omega_{s,i,t}$ denotes the value of ρ_i at state s at period t. We say that $\boldsymbol{\omega}$ is stationary if $\omega_{s,i,t_1} = \omega_{s,i,t_2}$ for all t_1, t_2, s, and i. If $\boldsymbol{\omega}$ is stationary, ρ_i depends only on the state that the process is in and we denote the value of ρ_i at state s by $\omega_{s,i}$. Let $\Omega_\upsilon^{\mathrm{NS}}$ and $\Omega_\upsilon^{\mathrm{ST}}$ denote the sets containing all nonstationary $\boldsymbol{\omega}$ and stationary $\boldsymbol{\omega}$ for a given υ, respectively. Let $P_{s,t}^{\pi}(i)$ denote the probability of being in state i at time t under policy π given initial state s. The value function of state s given π and $\boldsymbol{\omega}$ is

$$V_s^{\pi}(\boldsymbol{\omega}) = \sum_{t=0}^{\infty} \sum_{i \in S} \gamma^t P_{s,t}^{\pi}(i) \left[r^{\pi(i)} + \sum_{j=1}^{N} \Delta_j^{\pi(i)} \omega_{i,j,t} \right]. \tag{7}$$

We say that $\widetilde{\pi}$ is υ-optimal if

$$V_s^{\widetilde{\pi}}(\boldsymbol{\omega}) \geq V_s^{\pi}(\boldsymbol{\omega}) \quad \text{for all } s \in S,\ \pi \in \Pi,\ \boldsymbol{\omega} \in \Omega_\upsilon^{\mathrm{NS}}. \tag{8}$$

Since there are infinitely many elements in $\Omega_\upsilon^{\mathrm{NS}}$, it is impossible to evaluate the υ-optimality of a policy with condition (8). Theorem 2 below highlights that we can limit our analysis to stationary $\boldsymbol{\omega}$.

Theorem 2. *Given (7), $\widetilde{\pi}$ is υ-optimal if and only if*

$$V_s^{\widetilde{\pi}}(\boldsymbol{\omega}) \geq V_s^{\pi}(\boldsymbol{\omega}) \quad \textit{for all } s \in S,\ \pi \in \Pi,\ \boldsymbol{\omega} \in \Omega_\upsilon^{\mathrm{ST}}. \tag{9}$$

Proof. First, we prove that $\widetilde{\pi}$ is υ-optimal if condition (9) holds. We prove this by contradiction. Assume that condition (9) holds and that $\widetilde{\pi}$ is not υ-optimal. This implies that there must exist some $\boldsymbol{\omega}' \in \Omega_\upsilon^{\mathrm{NS}} \setminus \Omega_\upsilon^{\mathrm{ST}}$, $s \in S$, and $\pi' \in \Pi$ such that $V_s^{\widetilde{\pi}}(\boldsymbol{\omega}') - V_s^{\pi'}(\boldsymbol{\omega}') < 0$. It follows from (7) that

$$V_s^{\widetilde{\pi}}(\boldsymbol{\omega}') - V_s^{\pi'}(\boldsymbol{\omega}') = \sum_{t=0}^{\infty} \sum_{i \in S} \gamma^t \Bigg[P_{s,t}^{\widetilde{\pi}}(i) r^{\widetilde{\pi}(i)} - P_{s,t}^{\pi'}(i) r^{\pi'(i)} + \sum_{j=1}^{N} (P_{s,t}^{\widetilde{\pi}}(i) \Delta_j^{\widetilde{\pi}(i)} - P_{s,t}^{\pi'}(i) \Delta_j^{\pi'(i)}) \omega_{i,j,t} \Bigg].$$

We construct a stationary $\boldsymbol{\omega}''$ by setting

$$\omega''_{i,j} = \begin{cases} \upsilon & \text{if } P_{s,t}^{\widetilde{\pi}}(i) \Delta_j^{\widetilde{\pi}(i)} - P_{s,t}^{\pi'}(i) \Delta_j^{\pi'(i)} < 0, \\ -\upsilon & \text{otherwise.} \end{cases}$$

Note that $\boldsymbol{\omega}'' \in \Omega_\upsilon^{\mathrm{ST}}$. In addition, $V_s^{\widetilde{\pi}}(\boldsymbol{\omega}'') - V_s^{\pi'}(\boldsymbol{\omega}'') \leq V_s^{\widetilde{\pi}}(\boldsymbol{\omega}') - V_s^{\pi'}(\boldsymbol{\omega}') < 0$, contradicting condition (9). Therefore, condition (9) implies that $\widetilde{\pi}$ is υ-optimal. The proof for the reverse direction is straightforward and follows from the observation that $\Omega_\upsilon^{\mathrm{ST}} \subseteq \Omega_\upsilon^{\mathrm{NS}}$.

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

Theorem 2 provides a set of conditions that can be used to evaluate the υ-optimality of a policy. However, the number of policies in Π can grow rapidly with the size of the problem. Corollary 1 below provides a more compact set of conditions. Let $\boldsymbol{V}^{\pi}(\boldsymbol{\omega}) = (V_1^{\pi}(\boldsymbol{\omega}), V_2^{\pi}(\boldsymbol{\omega}), \ldots, V_{|S|}^{\pi}(\boldsymbol{\omega}))^{\top}$ and $\boldsymbol{\omega}_s = (\omega_{s,1}, \omega_{s,2}, \ldots, \omega_{s,N})^{\top}$.

Corollary 1. *Given (7), $\widetilde{\pi}$ is υ-optimal if and only if the following Bellman equations are satisfied:*

$$r^{\widetilde{\pi}(s)} - r^{a_s} + (\boldsymbol{\Delta}^{\widetilde{\pi}(s)} - \boldsymbol{\Delta}^{a_s})\boldsymbol{\omega}_s + \gamma(\boldsymbol{P}^{\widetilde{\pi}(s)} - \boldsymbol{P}^{a_s})\boldsymbol{V}^{\widetilde{\pi}}(\boldsymbol{\omega}) \geq 0$$

for all $s \in S$, $a_s \in A(s)$, and $\boldsymbol{\omega} \in \Omega_{\upsilon}^{\mathrm{ST}}$.

Proof. For a given $\boldsymbol{\omega}$, the Bellman equations are necessary and sufficient:

$$V_s^{\widetilde{\pi}}(\boldsymbol{\omega}) \geq V_s^{\pi}(\boldsymbol{\omega})$$

for all $s \in S$ and $\pi \in \Pi$ if and only if

$$r^{\widetilde{\pi}(s)} - r^{a_s} + (\boldsymbol{\Delta}^{\widetilde{\pi}(s)} - \boldsymbol{\Delta}^{a_s})\boldsymbol{\omega}_s + \gamma(\boldsymbol{P}^{\widetilde{\pi}(s)} - \boldsymbol{P}^{a_s})\boldsymbol{V}^{\widetilde{\pi}}(\boldsymbol{\omega}) \geq 0$$

for all $s \in S$ and $a_s \in A(s)$. This implies that

$$V_s^{\widetilde{\pi}}(\boldsymbol{\omega}) \geq V_s^{\pi}(\boldsymbol{\omega})$$

for all $s \in S$, $\pi \in \Pi$, and $\boldsymbol{\omega} \in \Omega_{\upsilon}^{\mathrm{ST}}$ if and only if

$$r^{\widetilde{\pi}(s)} - r^{a_s} + (\boldsymbol{\Delta}^{\widetilde{\pi}(s)} - \boldsymbol{\Delta}^{a_s})\boldsymbol{\omega}_s + \gamma(\boldsymbol{P}^{\widetilde{\pi}(s)} - \boldsymbol{P}^{a_s})\boldsymbol{V}^{\widetilde{\pi}}(\boldsymbol{\omega}) \geq 0$$

for all $s \in S$, $a_s \in A(s)$, and $\boldsymbol{\omega} \in \Omega_{\upsilon}^{\mathrm{ST}}$, and the corollary follows from Theorem 2.

Next, we illustrate how υ^*, the maximum allowable tolerance for the nonstationary problem, can be obtained from Corollary 1. Let $(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})_{s,i}^{-1}$ denote the entry in the sth row and ith column of the matrix $(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})^{-1}$. For stationary $\boldsymbol{\omega}$, the value function of a state can be expressed as

$$V_s^{\pi}(\boldsymbol{\omega}) = \sum_{i \in S} (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi})_{s,i}^{-1} \left(r^{\pi(i)} + \sum_{j=1}^{N} \Delta_j^{\pi(i)} \omega_{i,j} \right). \tag{10}$$

Define

$$f_{i,j}^{\widetilde{\pi},a_s} = \begin{cases} (1 + G_s^{\widetilde{\pi},a_s})\Delta_j^{\widetilde{\pi}(s)} - \Delta_j^{a_s} & \text{if } i = s, \\ G_i^{\widetilde{\pi},a_s}\Delta_j^{\widetilde{\pi}(i)} & \text{otherwise,} \end{cases}$$

where $G_i^{\widetilde{\pi},a_s}$ denotes the ith entry of $\gamma(\boldsymbol{P}^{\widetilde{\pi}(s)} - \boldsymbol{P}^{a_s})(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widetilde{\pi}})^{-1}$. Note that $f_{i,j}^{\widetilde{\pi},a_s}$ is the marginal change in the estimation error associated with state i and parameter j that result from the action perturbation a_s. Substituting (10) into Corollary 1 and using our definitions of $c^{\widetilde{\pi},a_s}$ and $f_{i,j}^{\widetilde{\pi},a_s}$ defines the following optimality conditions:

$$c^{\widetilde{\pi},a_s} + \sum_{i \in S} \sum_{j=1}^{N} f_{i,j}^{\widetilde{\pi},a_s} \omega_{i,j} \geq 0 \quad \text{for all } s \in S,\ a_s \in A(s),\ \boldsymbol{\omega} \in \Omega_{\upsilon}^{\mathrm{NS}}. \tag{11}$$

Proposition 3. *Given (7),*

$$\upsilon^* = \min_{s,a_s} \frac{c^{\widetilde{\pi},a_s}}{\sum_{j=1}^{N} d_j^{\widetilde{\pi},a_s}}.$$

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

TABLE 2: Computing υ.

s	a_s	$c^{\widetilde{\pi},a_s}$	$d_1^{\widetilde{\pi},a_s}$	$d_2^{\widetilde{\pi},a_s}$	υ_{s,a_s}
−1	2	40.85	20.4	30	0.81
−1	3	5.5	12	0	0.46
−1	4	0	0	0	—
0	1	40.85	20.4	30	0.81
0	2	5.5	12	0	0.46
0	3	0	0	0	—
1	0	0.85	60.4	30	0.01
1	1	5.5	12	0	0.46
1	2	0	0	0	—
2	0	0	0	0	—
2	1	34.5	52	0	0.66
3	0	0	0	0	—

Proof. Let υ_{s,a_s} denote the maximum tolerance allowable by (11) for state s and action a_s:

$$\upsilon_{s,a_s} = \max\left\{ y : c^{\widetilde{\pi},a_s} + \sum_{i\in S}\sum_{j=1}^{N} f_{i,j}^{\widetilde{\pi},a_s}\omega_{i,j} \geq 0 \text{ and } |\omega_{i,j}| \leq y \text{ for } i \in S,\ j = 1, 2, \ldots, N \right\}.$$

Similar to the stationary reward case, we consider the worst case scenario and obtain the following expression for υ_{s,a_s} from (11):

$$\upsilon_{s,a_s} = \frac{c^{\widetilde{\pi},a_s}}{\sum_{j=1}^{N} d_j^{\widetilde{\pi},a_s}},$$

where $d_j^{\widetilde{\pi},a_s} = \sum_{i\in S} |f_{i,j}^{\widetilde{\pi},a_s}|$. Hence, the maximum allowable tolerance for the nonstationary rewards problem is

$$\upsilon^* = \min_{s,a_s} \upsilon_{s,a_s} = \min_{s,a_s} \frac{c^{\widetilde{\pi},a_s}}{\sum_{j=1}^{N} d_j^{\widetilde{\pi},a_s}}.$$

We re-examine our capacitated stochastic lot-sizing problem in Example 3 below by allowing the ordering cost and backlog penalty to vary over time.

Example 3. (*Nonstationary rewards.*) Consider Example 1 again. When the ordering cost and backlog penalty are allowed to perturb simultaneously at each period, it follows from Table 2 and Proposition 3 that $\upsilon^* = 0.01$ and is associated with the action $a_1 = 0$. This implies that $\widetilde{\pi}$ remains optimal so long as the ordering cost and backlog penalty do not deviate from their current estimates by more than 1% across all periods.

4. Tolerance gap

In the introduction we claimed that the maximum allowable tolerance obtained under the assumption that the parameters are stationary may be overly optimistic if the values of the uncertain parameters vary across the horizon. In this section we provide a formal proof for this statement and highlight the conditions where the tolerances of the stationary and nonstationary rewards problems are the same and the conditions where they differ.

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

If the current decision $\widetilde{\pi}(s)$ is replaced by a_s, $|b_j^{\widetilde{\pi},a_s}|$ and $d_j^{\widetilde{\pi},a_s}$ represent the marginal changes in the estimation error of the stationary and nonstationary rewards problems, respectively. Recall that τ_{s,a_s} and υ_{s,a_s} denote the maximum allowable tolerance for state s and action a_s in the stationary and nonstationary problems, respectively. We denote by τ^* and υ^* the maximum allowable tolerance for the stationary and nonstationary problems, respectively. It follows from Propositions 2 and 3 that $\tau^* = \upsilon^*$ if $|b_j^{\widetilde{\pi},a_s}| = d_j^{\widetilde{\pi},a_s}$ for all a_s and j. If that is not true, the allowable tolerances of the stationary and nonstationary rewards problems may differ. In particular, $\tau^* > \upsilon^*$ if $|b_j^{\widetilde{\pi},a_s}| < d_j^{\widetilde{\pi},a_s}$ for some s and a_s, where $\tau_{s,a_s} = \tau^*$.

Lemma 1. *It holds that* $|b_j^{\widetilde{\pi},a_s}| \leq d_j^{\widetilde{\pi},a_s}$.

Proof. Note that $b_j^{\widetilde{\pi},a_s}$ and $d_j^{\widetilde{\pi},a_s}$ are the sums of the values and absolute values of $f_{i,j}^{\widetilde{\pi},a_s}$ across all states, respectively. Hence, it follows that

$$|b_j^{\widetilde{\pi},a_s}| = \left| \sum_{i \in S} f_{i,j}^{\widetilde{\pi},a_s} \right| \leq \sum_{i \in S} |f_{i,j}^{\widetilde{\pi},a_s}| = d_j^{\widetilde{\pi},a_s}.$$

Theorem 3. *It holds that* $\tau^* \geq \upsilon^*$.

Proof. The proof follows directly from Lemma 1 and Propositions 2 and 3.

Lemma 1 highlights that $|b_j^{\widetilde{\pi},a_s}|$ cannot be greater than $d_j^{\widetilde{\pi},a_s}$. Theorem 3 states that the maximum allowable tolerance obtained for a stationary rewards problem is at least as great as (i.e. more optimistic) the maximum allowable tolerance obtained when the stationary reward parameters assumption is relaxed.

Lemma 2. *It holds that* $|b_j^{\widetilde{\pi},a_s}| < d_j^{\widetilde{\pi},a_s}$ *if and only if there exist* $s_1, s_2 \in S$ *where* $f_{s_1,j}^{\widetilde{\pi},a_s}$ *is positive and* $f_{s_2,j}^{\widetilde{\pi},a_s}$ *is negative.*

Proof. If there exist $s_1, s_2 \in S$ where $f_{s_1,j}^{\widetilde{\pi},a_s}$ is positive and $f_{s_2,j}^{\widetilde{\pi},a_s}$ is negative,

$$|b_j^{\widetilde{\pi},a_s}| = |\sum_{i \in S} f_{i,j}^{\widetilde{\pi},a_s}| < \sum_{i \in S} |f_{i,j}^{\widetilde{\pi},a_s}| = d_j^{\widetilde{\pi},a_s}.$$

If $|b_j^{\widetilde{\pi},a_s}| < d_j^{\widetilde{\pi},a_s}$, $|\sum_{i \in S} f_{i,j}^{\widetilde{\pi},a_s}| < \sum_{i \in S} |f_{i,j}^{\widetilde{\pi},a_s}|$ and there must exist $s_1, s_2 \in S$ where $f_{s_1,j}^{\widetilde{\pi},a_s}$ is positive and $f_{s_2,j}^{\widetilde{\pi},a_s}$ is negative.

Theorem 4. *It holds that* $\tau_{s,a_s} > \upsilon_{s,a_s}$ *if and only if there exist some parameter* j *and states* $s_1, s_2 \in S$ *where* $f_{s_1,j}^{\widetilde{\pi},a_s}$ *is positive and* $f_{s_2,j}^{\widetilde{\pi},a_s}$ *is negative.*

Proof. If there exist some parameter j and states $s_1, s_2 \in S$ where $f_{s_1,j}^{\widetilde{\pi},a_s}$ is positive and $f_{s_2,j}^{\widetilde{\pi},a_s}$ is negative, it follows from Lemma 2 that $|b_j^{\widetilde{\pi},a_s}| < d_j^{\widetilde{\pi},a_s}$. Since $|b_j^{\widetilde{\pi},a_s}| \leq d_j^{\widetilde{\pi},a_s}$ for all j (Lemma 1), $\tau_{s,a_s} > \upsilon_{s,a_s}$.

If $\tau_{s,a_s} > \upsilon_{s,a_s}$, there exists some parameter j where $|b_j^{\widetilde{\pi},a_s}| < d_j^{\widetilde{\pi},a_s}$ and it follows from Lemma 2 that there exist states $s_1, s_2 \in S$ where $f_{s_1,j}^{\widetilde{\pi},a_s}$ is positive and $f_{s_2,j}^{\widetilde{\pi},a_s}$ is negative.

Theorem 4 provides a set of necessary and sufficient conditions for there to be a difference in the tolerances between the stationary and nonstationary rewards problems. In particular, the allowable tolerance associated with action a_s differs if a_s increases and decreases the effect of parameter j on the rewards associated with two different states. In Example 4 below we

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

illustrate how this condition can be checked without performing the actual computations. Next, we present a set of conditions where $|b_j^{\widetilde{\pi},a_s}| = d_j^{\widetilde{\pi},a_s}$.

Corollary 2. *It holds that* $|b_j^{\widetilde{\pi},a_s}| = d_j^{\widetilde{\pi},a_s}$ *if there exists* $k \in S$ *where* $\Delta_j^{\pi(i)} = 0$ *for all* $i \neq k$ *and* $\pi \in \Pi$.

Proof. If $\Delta_j^{\pi(i)} = 0$ for all $i \neq k$ and $\pi \in \Pi$, there can be at most one nonzero $f_{i,j}^{\widetilde{\pi},a_s}$ term and it follows from Lemma 1 and Lemma 2 that $|b_j^{\widetilde{\pi},a_s}| = d_j^{\widetilde{\pi},a_s}$.

Corollary 2 provides a sufficient condition for $|b_j^{\widetilde{\pi},a_s}| = d_j^{\widetilde{\pi},a_s}$. In particular, when the contribution of ρ_i to the value function is restricted to just a single state, the impact of ρ_i on the allowable tolerance is the same, regardless of whether ρ_i is stationary or not. The validity of Theorems 3 and 4 and Corollary 2 are illustrated in the following example.

Example 4. (*Tolerance gap.*) In Examples 2 and 3, we obtained $\tau^* = 0.02$ and $\upsilon^* = 0.01$, respectively. This result is consistent with Theorem 3, which states that $\tau^* \geq \upsilon^*$.

Example 2 highlights that $\tau_{1,0} = \tau^*$; τ^* will be strictly greater than υ^* if there is a difference in the allowable tolerance associated with not ordering when the inventory is one. Under $\widetilde{\pi}$, the optimal action is to order two units. If no order is placed, an ordering cost is avoided and the probability of entering state 1 remains the same. Hence, $f_{1,1}^{\widetilde{\pi},0}$ is negative. However, the probability of entering state -1 is increased (i.e. $G_{-1}^{\widetilde{\pi},0}$ is negative). Since $\Delta_1^{\widetilde{\pi}(-1)}$ is also negative, $f_{-1,1}^{\widetilde{\pi},0}$ is positive. Hence, it follows from Theorem 4 that $\tau^* > \upsilon^*$.

Since $\Delta_2^{\pi(s)} = 0$ for all $s \geq 0$, it follows from Corollary 2 that $|b_2^{\widetilde{\pi},a_s}| = d_2^{\widetilde{\pi},a_s}$ for all a_s. This is validated by comparing the values of $|b_2^{\widetilde{\pi},a_s}|$ and $d_2^{\widetilde{\pi},a_s}$ in Tables 1 and 2, respectively. This implies that the impact of ρ_2 on the allowable tolerances for the stationary and nonstationary rewards problems are the same and that any reduction in the allowable tolerance of the nonstationary problem is due to the relaxation of the stationary assumption on ρ_1.

In Example 1 we found that the optimal policy is to bring the inventory level up to three whenever the inventory drops below two. We refer to this as an order-up-to policy. Next, we will show that $\tau^* > \upsilon^*$ if the action associated with τ^* changes the reorder point for general lot-sizing problems under mild assumptions.

Consider a lot-sizing problem where p_i denotes the probability that demand is i. We assume that p_i is stationary (i.e. remains the same across the horizon). There is a constant lead time, a discount factor $0 < \gamma < 1$, linear production cost, and a convex holding cost. In addition, each order incurs an uncertain ordering cost and each backlog item incurs an uncertain penalty cost. As in Example 1, we model the uncertainties in the ordering cost and backlog penalty by ρ_1 and ρ_2, respectively. The objective is to find the policy that minimizes the long-run expected costs.

This problem can be formulated as an MDP where the states represent the amount of inventory available and the actions represent the amount of inventory to order. It is well known that an optimal order-up-to policy exists for this problem for a given ρ_1 and ρ_2 [21].

Theorem 5. *For the lot-sizing problem described above,* $\tau^* > \upsilon^*$ *if there exists* $\tau_{s,a_s} = \tau^*$, *where* a_s *changes the reorder point and* $p_i < 1/2\gamma$ *for all* i.

Proof. First, we consider the case where $\widetilde{\pi}(s) > 0$ and $a_s = 0$. If no order is placed at s, there must exist some state $s' < s$ where $G_{s'}^{\widetilde{\pi},a_s}$ is negative (i.e. the probability of entering s' is increased as a result of a_s). Since $\widetilde{\pi}$ is an order-up-to policy, $\Delta_1^{a_{s'}}$ is negative and $f_{s',1}^{\widetilde{\pi},0}$ is positive. Recall that $G_s^{\widetilde{\pi},a_s}$ is the discounted expected number of subsequent visits to state s

https://doi.org/10.1239/jap/1324046012 Published online by Cambridge University Press

under a_s and $\widetilde{\pi}$. Since $p_i < 1/2\gamma$, $G_s^{\widetilde{\pi},a_s} < 1$. Since $G_s^{\widetilde{\pi},a_s} < 1$, $\Delta_j^{\widetilde{\pi}(s)} < 0$, and $\Delta_j^{a_s} = 0$, $f_{s,1}^{\widetilde{\pi},0}$ is negative. Hence, it follows from Theorem 4 that $\tau^* > \upsilon^*$.

When $\widetilde{\pi}(s) = 0$ and $a_s > 0$, there must exist some state $s' < s$ where $G_{s'}^{\widetilde{\pi},a_s}$ is positive and $f_{s',1}^{\widetilde{\pi},0}$ is negative. Since $\Delta_j^{\widetilde{\pi}(s)} = 0$ and $\Delta_j^{a_s} < 0$, $f_{s,1}^{\widetilde{\pi},0}$ is positive. Hence, it follows from Theorem 4 that $\tau^* > \upsilon^*$.

Theorem 5 highlights that a tolerance gap exists for lot-sizing problems where the action associated with τ^* changes the reorder point when the p_i are bounded from above by $1/2\gamma$. Since $\gamma < 1$, the upper bound is greater than 0.5 for all problems and is, in our opinion, a reasonable assumption for most practical problems.

5. Summary

In this paper we examined how sensitivity analysis can be performed directly for an MDP with uncertain rewards. For the single-parameter problem, we illustrated how the optimal region of a policy can be obtained by considering the region in which the current policy is optimal with respect to each possible action (Proposition 1). When the uncertain parameters are allowed to vary simultaneously, we computed the maximum allowable error in the estimated values such that the current solution remains optimal (Proposition 2), and illustrated how the maximum allowable tolerance can be computed when the uncertain parameters are nonstationary (Proposition 3) by showing that it is sufficient to consider a subset of possible estimation errors (Theorem 2). In addition, we highlighted that the maximum allowable tolerance of the stationary problem is at least as great as that of the nonstationary problem (Theorem 3), and derived the conditions where the tolerances of the stationary and nonstationary problems are the same (Corollary 2) and the conditions where they differ (Theorem 4).

This work was motivated by the fact that rewards are often estimated and uncertain in practice. In this paper we considered a capacitated stochastic lot-sizing problem where the ordering costs and backlog penalties are uncertain. Other sequential problems that involve uncertain rewards include equipment replacement (e.g. uncertain salvage value), medical decision making (e.g. value of a human life), and dynamic assignment (e.g. value of a task).

In this paper we considered MDPs where rewards are expressed as affine functions of uncertain parameters. One extension is to consider rewards involving more general functions. Another extension is to consider changes in other model parameters, including transition probabilities, the discount factor, and horizon. In this paper we highlighted the conditions where stationary uncertain parameter assumptions lead to overly optimistic tolerance levels for a general lot-sizing problem under mild assumptions (Theorem 5). Another area of further research is to identify conditions where this is true for other sequential decision problems.

Acknowledgements

The authors are grateful to an anonymous referee whose remarks greatly improved the presentation of this work. Support from the National Science Foundation, under grant number CMMI-0813671, is gratefully acknowledged.

References

[1] Bazaraa, M. S., Jarvis, J. J. and Sherali, H. D. (2005). *Linear Programming and Network Flows*, 3rd edn. John Wiley, Hoboken, NJ.

[2] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

[3] Charalambous, C. and Gittins, J. C. (2008). Optimal selection policies for a sequence of candidate drugs. *Adv. Appl. Prob.* **40,** 359–376.
[4] Erkin, Z. *et al.* (2010). Eliciting patients' revealed preferences: an inverse Markov decision process approach. *Decision Anal.* **7,** 358–365.
[5] Gal, T. and Greenberg, H. J. (1997). *Advances in Sensitivity Analysis and Parametric Programming*. Kluwer, Dordrecht.
[6] Glazebrook, K. D., Ansell, P. S., Dunn, R. T. and Lumley R. R. (2004). On the optimal allocation of service to impatient tasks. *J. Appl. Prob.* **41,** 51–72.
[7] Harmanec, D. (2002). Generalizing Markov decision processes to imprecise probabilities. *J. Statist. Planning Infer.* **105,** 199–213.
[8] Hopp, W. J. (1988). Sensitivity analysis in discrete dynamic programming. *J. Optimization Theory Appl.* **56,** 257–269.
[9] Iyengar, G. N. (2005). Robust dynamic programming. *Math. Operat. Res.* **30,** 257–280.
[10] Lim, C., Bearden, J. N. and Smith, J. C. (2006). Sequential search with multiattribute options. *Decision Anal.* **3,** 3–15.
[11] Manne, A. S. (1960). Linear programming and sequential decisions. *Manag. Sci.* **6,** 259–267.
[12] Mitrophanov A. Y., Lomsadze A. and Borodovsky M. (2005). Sensitivity of hidden Markov models. *J. Appl. Prob.* **42,** 632–642.
[13] Muckstadt, J. A. and Sapra A. (2010). *Principles of Inventory Management*. Springer, New York.
[14] Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operat. Res.* **53,** 780–798.
[15] Powell, W. B. (2007). *Approximate Dynamic Programming*. John Wiley, Hoboken, NJ.
[16] Puterman, M. L. (1994). *Markov Decision Processes. Discrete Stochastic Dynamic Programming*. John Wiley, New York.
[17] Sandvik, B. and Thorlund-Petersen, L. (2010). Sensitivity analysis of risk tolerance. *Decision Anal.* **7,** 313–321.
[18] Tan, C. H. and Hartman, J. C. (2010). Equipment replacement analyis with an uncertain finite horizon. *IIE Trans.* **42,** 342–353.
[19] Tan, C. H. and Hartman, J. C. (2011). Sensitivity analysis and dynamic programming. In *Wiley Encyclopedia of Operations Research and Management Science*, ed. J. J. Cochran, John Wiley, New York.
[20] Topaloglu, H. and Powell, W. B. (2007). Sensitivity analysis of a dynamic fleet management model using approximate dynamic programming. *Operat. Res.* **55,** 319–331.
[21] Veinott, A. F., Jr. and Wagner, H. M. (1965). Computing optimal (s, S) inventory policies. *Manag. Sci.* **11,** 525–552.
[22] Wallace, S. W. (2000). Decision making under uncertainty: is sensitivity analysis of any use? *Operat. Res.* **48,** 20–25.
[23] Ward, J. E. and Wendell, R. E. (1990). Approaches to sensitivity analysis in linear programming. *Ann. Operat. Res.* **27,** 3–38.
[24] Wendell, R. E. (1985). The tolerance approach to sensitivity analysis in linear programming. *Manag. Sci.* **31,** 564–578.
[25] White, C. C. and El-Deib, H. K. (1986). Parameter imprecision in finite state, finite action dynamic programs. *Operat. Res.* **34,** 120–129.
[26] White, D. J. (1993). *Markov Decision Processes*. John Wiley, Chichester.