Markov Decision Processes with Uncertain Transition Rates: Sensitivity and Max–min Control *

Suresh Kalyanasundaram[†] Edwin K. P. Chong[‡] Ness B. Shroff [§]

Abstract

Solution techniques for Markov decision problems rely on exact knowledge of the transition rates, which may be difficult or impossible to obtain. In this paper, we consider Markov decision problems with uncertain transition rates represented as compact sets. We first consider the problem of sensitivity analysis where the aim is to quantify the range of uncertainty of the average per-unit-time reward given the range of uncertainty of the transition rates. We then develop solution techniques for the problem of obtaining the max-min optimal policy, which maximizes the worst-case average per-unit-time reward. In each of these problems, we distinguish between systems that can have their transition rates chosen independently and those where the transition rates depend on each other. Our solution techniques are applicable to Markov decision processes with fixed but unknown transition rates and to those with time-varying transition rates.

Keywords: Markov decision processes; sensitivity; max-min control; call admission control; policy iteration.

1 Introduction

Markov decision process (MDP) methods are used in dynamic probabilistic systems to make sequential decisions that optimize an appropriate objective [1, 2, 3]. In MDPs, the system

*This research was supported in part by the National Science Foundation through ANI-9805441, 0099137-ANI, 0098089-ECS, ANI-0207892, and ANI-0207728.

[†]Motorola India Electronics Limited, No. 66/1, Plot 5, Bagmane Techpark, C. V. Raman Nagar Post, Bangalore 560 093, India. E-mail: Suresh.Kalyanasundaram@motorola.com. Phone: 91-80-26012115. Fax. 91-80-25343100.

[‡]Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523-1373, USA. E-mail: echong@engr.colostate.edu. Phone: 1-970-491-7858. Fax: 1-970-491-2249.

[§]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285, USA. E-mail: shroff@ecn.purdue.edu. Phone: 1-765-494-3471 Fax: 1-765-494-3358.

can be in a finite number of states and the decision-maker has a choice of several actions in each of those states. Once an action has been chosen in each state, the system evolves as a continuous-time Markov chain (CTMC) and accumulates reward with time. (While the discussion in this paper is written for continuous-time MDPs, the case of discrete-time MDPs can also be handled similarly.) The rate of reward accrual depends on the state of the system and the choice of control action in that state. In other words, associated with each state (and possibly depending on the control action chosen in that state), there is a reward accrual rate. The reward accumulates additively with time. In Markov decision problems, we wish to determine a policy (or equivalently, an action for each state) that optimizes the chosen objective. The chosen objective could either be discounted total reward or average per-unit-time reward. In the discounted total reward criterion, future rewards are discounted by a certain discount factor, which captures the notion of future rewards being less important than those accrued at the present time. In the average reward criterion problems, the objective is to determine a policy that maximizes the expected per-unit-time reward. (In our case, we wish to *maximize* the chosen objective because we associate "rewards" with each state. Equivalently, some authors associate "costs" with each state and *minimize* the chosen objective [4].) The choice of which objective to use depends on the specific problem being solved. For the admission control example discussed in Section 3, we find that the average per-unit-time reward is a more natural objective to use. In many problems, we find that the choice of discount factor has to be determined rather arbitrarily, and because the optimal policy might depend on the discount factor, the use of the discounted total reward criterion may not be appropriate.

Efficient methods exist to solve MDP problems, including value iteration, policy iteration, and linear programming [1]. The existence of such efficient methods have made MDPs practically attractive and has resulted in their use in several areas, including telecommunications [5, 6, 7]. However, a drawback of these solution techniques is that they require exact knowledge of the transition rates of the underlying Markov chain for every policy choice. The exact value of these parameters can be very difficult or impossible to obtain. We describe such a scenario in Section 3, where we discuss the admission control problem in communication networks. In this example, the transition rates are the call arrival rates and the call holding durations of user classes — it would be impossible for the network to know these parameters beforehand. In this work, we develop techniques to address this problem of lack of knowledge of the exact value of the transition rates. We first consider the problem of sensitivity analysis, where we analyze the sensitivity of a decision that is based on a particular choice for the values of the transition rates. We then develop robust decision schemes using which we can design policies for the worst-case and best-case scenarios.

The problem of lack of knowledge of the transition rates has been recognized by other

authors [8, 9, 10]. All these authors consider the case of total discounted reward criterion. In [10], a value iteration technique has been developed for analyzing the sensitivity of decisions and for designing robust decision schemes. The authors of [8] develop a policy iteration technique to obtain robust decisions. In [9], the authors consider the criterion of discounted total reward with the uncertainty in the transition rates given as a finite number of linear inequalities. In this work, we consider continuous-time MDPs with average reward optimization criterion. We introduce the notion of dependence among transition rates, and distinguish between systems with independent transition rates and those with dependent transition rates. For both these types of systems, we develop optimization and policy iteration techniques to perform sensitivity analysis and to design robust decisions for the case of average reward criterion in continuous-time MDPs. Based on our new classification, all previous work cited above is for the special case of systems with independent transition rates.

Our contributions in this work are as follows. We introduce the notion of dependence among transition rates, and consider the case of systems with independent transition rates and those with dependent transition rates. While other authors have considered the case of maximization of discounted total reward, we consider maximization of average per-unit-time reward as the optimization criterion. We develop policy iteration and optimization approaches for solving the sensitivity analysis and max-min problems. Using our optimization approach to the sensitivity analysis problem, we show that we can obtain bounds on the performance of systems with dependent transition rates by analyzing the system with independent transition rates that results by eliminating the dependence among transition rates in the original problem.

This paper is structured as follows. In Section 2, we introduce the notation used in this paper and formulate the problems to be addressed in this work. In Section 3, we present an example from the call admission problem in communication networks, which illustrates the problems described in Section 2. Sections 4 and 5 consider the sensitivity analysis problems for systems with independent and dependent transition rates, respectively. In Sections 6 and 7 we develop techniques for obtaining the max-min optimal policy for systems with independent and dependent transition rates, respectively. In Section 8, we illustrate the methods developed in this paper by presenting numerical results for the call admission control problem presented in Section 3.

2 Overview and Problem Formulation

In this section, we introduce the notation used in this paper and formulate the problems to be addressed. Throughout this paper, our description is for infinite-horizon continuous-time MDPs with average reward per unit time as the optimization criterion. Extensions of

this work to discrete-time MDPs and for the discounted total reward criterion can be made in a straightforward manner. We only consider finite-state and finite-action spaces in our work. We first give a brief description of continuous-time MDPs. A continuous-time MDP is characterized as follows:

- At any time t, the system can be in any one of the states in some finite state space S.
- In each state $i \in S$, the decision-maker can choose an action from the finite action space K_i.
- When the system is in state $i \in S$ and when action $u \in K_i$ is chosen, the system accrues reward at the rate of r_i^u per unit time.
- When the system is in state $i \in S$ and when action $u \in K_i$ is chosen, the transition rate to state $j \in S$, $j \neq i$ is fixed at α_{ij}^u. Note that this implies that the system will stay in state i for an exponentially distributed duration of time with mean $1/\sum_{j \in S, j \neq i} \alpha_{ij}^u$ and state $j \in S$, $j \neq i$ is the next state visited by the system with probability $\alpha_{ij}^u / \sum_{k \in S, k \neq i} \alpha_{ik}^u$. Henceforth, to avoid clutter, we omit the condition $j \neq i$ when we write α_{ij}, but with the understanding that the condition $j \neq i$ is indeed required. Also all summations of the form $\sum_{j \in S} \alpha_{ij}$ are also carried out only over all states $\{j \in S, j \neq i\}$.
- The optimization criterion is the maximization of the average reward per unit time. The objective is to determine a sequence of actions as the system evolves that maximizes the appropriate criterion.

A *control policy* is defined as a sequence of state-to-action maps specifying, at each transition time and given the entire history of state visits and actions chosen, which action is chosen for the state visited. Let $\mathcal{F}$ denote the set of all control policies. This definition of control policies is very general and includes those that depend on the time at which the control action is chosen as well as those that depend on the entire history of states visited and actions chosen. For example, it includes the control policy that specifies that during the first visit to state i a specific control action has to be chosen, but starting from the second visit an alternate control action has to be chosen. Control policies that are independent of the history of states and actions given the current state are known as *Markov control policies*. Markov control policies that are also independent of the time of decision-making (and depend only on the current state of the system) are called *stationary policies*. To characterize the set of all stationary policies, we define d_i^u as the probability that we choose control action $u \in K_i$ when the system is in state $i \in S$. The d_i^u (for $u \in K_i$ and $i \in S$) are variables whose

values the decision-maker can set. It is clear that

$$\sum_{u \in K_i} d_i^u = 1 \text{ for } i \in S,$$

and that

$$d_i^u \geq 0 \text{ for } u \in K_i \text{ and } i \in S.$$

A stationary policy is the following set of probability distributions, one for each feasible state:

$$\{d_i^u : u \in K_i \text{ and } i \in S\}.$$

Implicit in the above notation is that the choice of control action does not depend on any time index. In other words, successive visits to a state will employ the same probability distribution in choosing the control action. Let $\mathcal{G}$ denote the set of all stationary policies.

An important subset of stationary policies are *pure policies*. A control policy is called a pure policy if it is stationary and for each $i \in S$ there is an action $u \in K_i$ such that $d_i^u = 1$. Let $\mathcal{H}$ denote the set of all pure policies. Clearly, $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$. Also, it should be clear from the above description that a pure policy can be represented as a mapping from the state space to the action space, i.e., for any pure policy, there is a map h such that $h(i) \in K_i$ is unique and represents the action chosen by the pure policy in state i.

Given that a pure policy $h \in \mathcal{H}$ is employed, the sequence of states visited by the system forms a CTMC with transition rates $\{\alpha_{ij}(h)\}$ such that

$$\alpha_{ij}(h) = \alpha_{ij}^{h(i)}. \tag{1}$$

Under pure policy h, reward is accumulated at the rate of $r_i^{h(i)}$ per unit time when the system is in state i. Similarly, given that a stationary policy $g \in \mathcal{G}$ is employed, the sequence of states visited by the system forms a CTMC with transition rates $\{\alpha_{ij}(g)\}$ given by

$$\alpha_{ij}(g) = \sum_{u \in K_i} \alpha_{ij}^u d_i^u(g), \tag{2}$$

where we have used $d_i^u(g)$ to denote the probabilities d_i^u associated with stationary policy g.

Let $Z_f(t)$ for $t \geq 0$ be the reward accumulated by the system until time t when policy $f \in \mathcal{F}$ is employed. The expected reward per unit time when the initial state of the system is i is given by

$$R_i(f) = \lim_{t \to \infty} E_f \left[\frac{Z_f(t)}{t} \middle| X(0) = i \right],$$

where $E_f(\cdot)$ denotes the expectation operator given that policy $f \in \mathcal{F}$ is employed, and $\{X(t); t \geq 0\}$ represents the state of the system at time t. A policy f is optimal if it achieves the largest value of $R(f)$. Let us fix a stationary policy $g \in \mathcal{G}$, and let $\pi_j(g)$ be the limiting

(or the steady-state) probability that the CTMC is in state j. It is known that if g gives rise to a single recurrent chain, then the term $R_i(g)$ reduces to the following equation and is independent of the initial state i [3].

$$R_i(g) = R(g) = \sum_{j \in S} \pi_j(g) r_j(g) \text{ for all } i, \tag{3}$$

where $r_j(g)$ is the mean reward accrual rate in state j. The interpretation of the above equation is that, since $\pi_j(g)$ represents the fraction of time that the system spends in state j, the average return per unit time is given by the weighted average of the expected reward rates in each state with the weights being the fraction of time spent in that state. It is also known that if every pure policy $h \in \mathcal{H}$ results in a CTMC with a single recurrent class plus a set (possibly empty) of transient states, then there exists an optimal pure policy (see [11, 12]).

We now formulate the problems that we will be addressing in this work. As mentioned earlier, the decision-maker may not have exact knowledge of the transition rates α_{ij}^u of the MDP, either because they are fixed but unknown or because they vary with time. The discussion that follows assumes fixed but unknown transition rates, but we will show later that our techniques can be applied even if the transition rates are time-varying. If the transition rates are fixed but unknown, estimation of these quantities is only a partial solution. Indeed, estimation with a finite number of samples always leads to estimation errors. Moreover, there exists a conflict between estimation and choosing the optimal decision — if we aim to estimate the transition rates accurately we may no longer be choosing the optimal decision [13, 14]. Note that, in our formulation's most general form, the estimation of a transition rate α_{ij}^u for a certain action $u \in K_i$ yields no information about the transition rate $\alpha_{ij}^{u'}$ for some other action $u' (\neq u) \in K_i$.

In sensitivity analysis problems, we assume that the decision-maker has irrevocably chosen a stationary policy $g \in \mathcal{G}$. (It is known that there exists an optimal stationary policy even for MDPs with constraints [15]. Therefore, our approach is quite general and can be used for sensitivity analysis of MDPs with constraints also.) In MDPs, this fixes the transition rates at $\alpha_{ij}(g)$ as given in Eq. (2). But, in our case, there is uncertainty in the original transition rates α_{ij}^u (and hence in the transition rates $\alpha_{ij}(g)$). However, the decision-maker is certain of the set over which the actual $\alpha_{ij}(g)$ ranges. The decision-maker now wishes to know the range of uncertainty of the expected reward per unit time given the range of uncertainty of the $\alpha_{ij}(g)$.

We distinguish between two cases. In the first case, the transition rates $\alpha_{ij}(g)$ may be chosen independently of each other, and in the second, they cannot be chosen independently of each other. We refer to systems that have the former property as systems with *independent* transition rates, and the latter as those with *dependent* transition rates. Unlike the case of systems with dependent transition rates, in the former case the range of uncertainty of each

of the transition rates can be specified independently of those of any other. This distinction is motivated by the observation that the real-world phenomena with which the transition rates are associated may not all be different. For example, in the admission control problem described in Section 3, several of the transition rates $\alpha_{ij}(g)$ are the call arrival rates of a particular class of calls. Generally, this quantity cannot have a different value in each state, but should all have the same value. In systems with dependent transition rates, when a certain value is chosen for a particular transition rate, the range of values that other transition rates can take are affected by this choice. In contrast, in systems with independent transition rates, the choice of a certain value for a particular transition rate does not affect the range of values that other transitions can take. More formally, in the case of independent transition rates, the decision-maker can separate the set over which the $\alpha_{ij}(g)$ range as follows:

$$\alpha_{ij}(g) \in \Omega_{ij}(g) \text{ for } i, j \in S.$$

But for the case of systems with dependent transition rates, the decision-maker merely knows that

$$\{\alpha_{ij}(g)\}_{i,j} \in \Omega(g),$$

where $\Omega(g)$ is a set in $|S|(|S|-1)$ dimensional space. Recently, in [16], the authors have considered placing constraints similar to the ones in the case of dependent transition rates on the action choices in different states. The authors call these correlated actions. The discussion in [16] also states that standard policy iteration cannot be applied to the case of correlated actions. The sensitivity analysis problem with dependent transition rates can be viewed as an MDP with correlated actions where the action choice does not affect the reward rate. In this work, we enhance the standard policy iteration algorithm so that it can be applied to the case of dependent transition rates. A suitably modified version of our algorithm for dependent transition rates can be applied to the case of correlated actions described in [16] as well.

We can state the sensitivity analysis problem as follows. (In sensitivity analysis problems, the policy $g \in \mathcal{G}$ is fixed. Therefore, to reduce notational complexity, we do not explicitly denote the dependence of the transition rates and rewards on the choice of the policy g.)

Sensitivity analysis problem with independent transition rates: Given a CTMC $\{X(t); t \geq 0\}$ with finite state space S, reward accrual rates r_i for $i \in S$, and fixed but unknown transition rates α_{ij} such that $\alpha_{ij} \in \Omega_{ij}$ for $i, j \in S$, determine the maximum and minimum expected reward per unit time. We only develop solution methods for the following minimization problem (the technique to maximize average per-unit-time reward is a straightforward modification):

$$\underset{\alpha_{ij} \in \Omega_{ij}}{\text{minimize}} \quad \text{(Average per-unit-time reward)}. \tag{4}$$

Sensitivity analysis problem with dependent transition rates: Given a CTMC $\{X(t); t \geq 0\}$ with finite state space S, reward accrual rates r_i for $i \in S$, and fixed but unknown transition rates α_{ij} such that $\{\alpha_{ij}\}_{i,j} \in \Omega$, determine the maximum and minimum expected reward per unit time. As in the independent case, we only develop techniques for the following problem:

$$\underset{\{\alpha_{ij}\}_{i,j} \in \Omega}{\text{minimize}} \quad \text{(Average per-unit-time reward)}. \tag{5}$$

We next describe the notions of max-min policies and max-max policies. Such policies have been described in [8, 10] as well (for the case of independent transition rates). The idea of max-min policies is that the decision-maker is uncertain about the transition rates and wishes to make a decision taking the pessimistic view that the transition rates are chosen to minimize the average per-unit-time reward. Therefore, the decision-maker will choose a policy that maximizes this minimum average per-unit-time reward. The maximization is done over all policies and the minimization is done over the transition rates over which the decision-maker has no control. Similarly, with a max-max policy the decision-maker makes a decision assuming that the transition rates are chosen to maximize the average per-unit-time reward. In this paper, we only describe algorithms for the max-min case. The max-max case can be handled similarly.

As before, we distinguish between two cases, depending on whether or not the transition rates can be chosen independently. The task is to determine a policy $f \in \mathcal{F}$ that is optimal in the max-min sense. Formally, we formulate the max-min problem as follows.

Max-min optimal policy with independent transition rates: Let a system have finite state space S and a choice of actions from the finite action space K_i in each state $i \in S$. Also, let the reward accrual rate be r_i^u for $i \in S$ and $u \in K_i$. The transition rates for the system α_{ij}^u are fixed but unknown such that $\alpha_{ij}^u \in \Omega_{ij}^u$ for $i, j \in S$ and $u \in K_i$. The problem is to determine a policy $f \in \mathcal{F}$ that maximizes the worst-case expected reward per unit time (determined by the choice of $\alpha_{ij}(f) \in \Omega_{ij}(f)$). In other words,

$$\max_{f \in \mathcal{F}} \min_{\alpha_{ij}(f) \in \Omega_{ij}(f)} \text{(Average per-unit-time reward)}.$$

Max-min optimal policy with dependent transition rates: In this case, the problem formulation is the same as that of systems with independent transition rates, except that the transition rates cannot be chosen independently. Therefore, we have the condition that $\{\alpha_{ij}(f)\}_{i,j} \in \Omega(f)$. The problem can be stated as follows:

$$\max_{f \in \mathcal{F}} \min_{\{\alpha_{ij}(f)\}_{i,j} \in \Omega(f)} \text{(Average per-unit-time reward)}.$$

3 Illustrative Example: Call Admission Control

In this section, we provide an example from the call admission problem in telecommunications, which will help illustrate the problems formulated in the last section. Variants of the same problem have been considered by other authors (e.g., [5, 6]). We consider the scenario where calls belonging to different classes are to be supported over a single link with finite capacity. Assume that there are a total of M classes and that a class-i user requires e_i bandwidth units (BWUs). Let the capacity of the link be C BWUs. We also assume that calls belonging to class i, $i = 1, 2, \ldots, M$, arrive according to an independent Poisson process with rate λ_i, and that the call holding duration of class-i calls are i.i.d. (independent and identically distributed) exponential with mean $1/\mu_i$. Given the above framework, the problem is to determine an admission control policy (decision on whether or not to admit a call) such that the long-run average throughput is maximized. Throughput is the time average of the total amount of bandwidth used by the users admitted to the system. Note that the average per-unit-time reward has an intuitively appealing interpretation for this problem in the sense that it can be seen as the average link utilization or throughput. The discounted total reward criterion does not have such an appealing interpretation. Moreover, the choice of the discount factor to use for the discounted total reward criterion would have to be made in an arbitrary fashion and the optimal policy corresponding to the discounted total reward criterion would depend on the value of the discount factor. Therefore, the discounted total reward is not a very meaningful objective to use for this problem.

The above problem can be formulated as an MDP. We do this by characterizing the state space, the action space, the reward rate, and the transition rates of the problem. We denote the state of the system by a vector of M components $(n_1, n_2, \ldots, n_M)$ such that n_i represents the number of active class-i users in the system. Then the set S of all feasible states is

$$S = \left\{ \boldsymbol{n} = (n_1, n_2, \ldots, n_M) : \sum_{i=1}^{M} n_i e_i \leq C, n_i \geq 0 \text{ and } n_i \text{ integer for } i = 1, 2, \ldots, M \right\}. \quad (6)$$

The set S includes all possible states such that the combined bandwidth requirement of the active users does not exceed the link bandwidth. An action is denoted by a vector $\boldsymbol{u}$ such that $\boldsymbol{u} = (u_1, u_2, \ldots, u_M)$ and $u_i = 0$ or 1 with $u_i = 0$ implying we block any call belonging to class i, and $u_i = 1$ implying we accept any call belonging to class i. We define the set $K_{\boldsymbol{n}}$ of all allowed control actions in state $\boldsymbol{n}$ as follows:

$$\begin{aligned} K_{\boldsymbol{n}} \;=\; & \{(u_1, u_2, \ldots, u_M) : u_i = 0 \text{ or } 1 \text{ for } i = 1, 2, \ldots, M \text{ and} \\ & \boldsymbol{n} + \boldsymbol{\delta}_i u_i \in S \text{ for } i = 1, 2, \ldots, M\}, \end{aligned}$$

where $\boldsymbol{\delta}_i$ is a vector of M components with zeros in all except the i-th position where it has a one. Note that an action $\boldsymbol{u}$ is allowed in state $\boldsymbol{n}$ only if there is enough idle capacity in

the system to admit a call belonging to any of the traffic classes that the action $\boldsymbol{u}$ decides to admit. The set $K_{\boldsymbol{n}}$ contains precisely those actions. For each state $\boldsymbol{n} \in S$, the set $K_{\boldsymbol{n}}$ has at most 2^M elements.

When the system stays in state $\boldsymbol{n} = (n_1, n_2, \ldots, n_M)$, and when action $\boldsymbol{u} \in K_{\boldsymbol{n}}$ is chosen, the following is the per-unit-time reward accrued:

$$r_{\boldsymbol{n}}^{\boldsymbol{u}} = \sum_{i=1}^{M} n_i e_i. \tag{7}$$

We note that the reward in state $\boldsymbol{n}$ is independent of the chosen action in that state. The reward $r_{\boldsymbol{n}}^{\boldsymbol{u}}$ is merely the amount of bandwidth in use in state $\boldsymbol{n}$. It follows that the long-run expected reward per unit time is the throughput of the system. If the system is in state $\boldsymbol{n}$ and action $\boldsymbol{u}$ is chosen, then the next state of the system is chosen according to the rates $\alpha_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}}$, obtained as follows:

$$\alpha_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}} = \begin{cases} \lambda_i & \text{if } \boldsymbol{p} = \boldsymbol{n} + \boldsymbol{\delta}_i \text{ and } \boldsymbol{u}_i = 1 \\ n_i \mu_i & \text{if } \boldsymbol{p} = \boldsymbol{n} - \boldsymbol{\delta}_i \text{ and } \boldsymbol{p} \in S \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Note that the control action $\boldsymbol{u}$ does not have any effect on call termination.

We now present numerical results for a simple case to illustrate how the optimal decision depends on the values of λ_i and μ_i. For the numerical results, we assume that there are two classes of calls and that the total capacity of the link is $C = 4$ BWUs. We also assume that class-1 calls require 1 BWU and class-2 calls require 2 BWUs, i.e., $e_1 = 1$ BWU and $e_2 = 2$ BWUs. The call holding durations for both classes of calls are assumed exponential with mean $1/\mu = 60$ seconds. We fix the arrival rate of class-2 calls and treat λ_1 as a parameter. In Table 1, we show the probabilities with which we should choose actions in a given state to maximize throughput. We put an 'x' for the probabilities of actions in "don't care" states. For example, for $\lambda_1 = .05$, we find that state (1,0) is a "don't care" state. This is because state (1,0) is a transient state, since the optimal policy is to accept only class-2 calls in all states. From the results in the table we see that the optimal decision depends on the λ_i and μ_i values. For example, for very low values of λ_1, the optimal policy is to accept only class-2 calls. This can be explained as follows. Since the arrival rate of class-2 calls is very high compared to that of class-1 calls, it pays to wait a little longer for a class-2 call and increase the throughput rather than accept a class-1 call. For $\lambda_1 = 1$ and $\lambda_2 = .5$, the optimal policy is to admit any call as long as there is idle capacity in the system while for $\lambda_1 = .44$ and $\lambda_2 = .5$, the optimal policy admits only class-2 calls in state (0,1) and admits all classes of calls in other states.

An important consequence of the dependence of the optimal decision on the parameters of the system is that the admission controller is expected to have exact knowledge of these

parameters. In practice, it may be extremely difficult or impossible for the admission controller to have exact knowledge of these quantities. The problems formulated in the last section will be useful in this context. The admission control problem is also an illustration of a problem that has dependent transition rates. Notice that due to the dependence of the transition rates on λ_i and μ_i as seen from Eq. (8), the decision-maker cannot express the range of uncertainty of the transition rates independently. Instead, the decision-maker can only decide on the range of uncertainty of the λ_i and the μ_i, and the range of uncertainty of the transition rates are automatically fixed. The motivation for considering systems with dependent transition rates should be clear from this example.

With the useful insights gained from the admission control example, we now return to the problems formulated in Section 2 and proceed to develop solution techniques for them.

4 Sensitivity Analysis with Independent Transition Rates

4.1 Optimization Approach

In sensitivity analysis problems, a stationary policy $g \in \mathcal{G}$ is fixed. Therefore, as mentioned earlier, we do not explicitly denote the dependence of the transition rates and reward rates on the policy choice g. Given this fixed stationary policy, the state of the system evolves as a CTMC.

In sensitivity analysis problems, we wish to determine the minimum average reward per unit time, given that $\alpha_{ij} \in \Omega_{ij}$, where Ω_{ij} is the range of uncertainty of the transition rate α_{ij}. We will require that the Ω_{ij} be compact.

From Eq. (3), it follows that our objective is to minimize $\sum_{i \in S} \pi_i r_i$ given that $\alpha_{ij} \in \Omega_{ij}$. We know that the steady state probabilities π_i have to satisfy the following conditions:

$$\pi_i \sum_{j \in S} \alpha_{ij} = \sum_{j \in S} \pi_j \alpha_{ji} \text{ for } i \in S$$

$$\sum_{i \in S} \pi_i = 1$$

$$\pi_i \geq 0 \text{ for } i \in S.$$

Therefore, the sensitivity analysis problem with independent transition rates can be formulated as follows:

$$\underset{\{\alpha_{ij}, \pi_i\}}{\text{minimize}} \sum_{i \in S} \pi_i r_i$$

		$\lambda_1 = .05, \lambda_2 = .5$	$\lambda_1 = 1, \lambda_2 = .5$	$\lambda_1 = .44, \lambda_2 = .5$
State (0,0)	Action (1,1)	0	1	1
	Action (1,0)	0	0	0
	Action (0,1)	1	0	0
	Action (0,0)	0	0	0
State (0,1)	Action (1,1)	0	1	0
	Action (1,0)	0	0	0
	Action (0,1)	1	0	1
	Action (0,0)	0	0	0
State (0,2)	Action (0,0)	1	1	1
State (1,0)	Action (1,1)	x	1	1
	Action (1,0)	x	0	0
	Action (0,1)	x	0	0
	Action (0,0)	x	0	0
State (1,1)	Action (1,0)	x	1	1
	Action (0,0)	x	0	0
State (2,0)	Action (1,1)	x	1	1
	Action (1,0)	x	0	0
	Action (0,1)	x	0	0
	Action (0,0)	x	0	0
State (2,1)	Action (0,0)	x	1	1
State (3,0)	Action (1,0)	x	1	1
	Action (0,0)	x	0	0
State (4,0)	Action (0,0)	x	1	1

Table 1: $d_{\boldsymbol{n}}^{\boldsymbol{u}}$ values for the scheme that maximizes throughput

$$\begin{aligned} \text{subject to} \quad & \pi_i \textstyle\sum_{j\in S} \alpha_{ij} = \sum_{j\in S} \pi_j \alpha_{ji} \text{ for } i \in S \\ & \textstyle\sum_{i\in S} \pi_i = 1 \\ & \pi_i \geq 0 \text{ for } i \in S \\ & \alpha_{ij} \in \Omega_{ij} \text{ for } i, j \in S. \end{aligned} \tag{9}$$

Because we assume the sets Ω_{ij} are compact, the minimum is guaranteed to exist by Weierstrass's theorem [17] because all the constraints of the above problem together form a compact constraint set.

We have thus formulated our sensitivity analysis problem as an optimization problem involving the variables α_{ij} and π_i. Note that each α_{ij} can take values in the set Ω_{ij} independently of any other α_{ij}. Note also that a particular choice of values for the α_{ij} uniquely determines π_i for each $i \in S$. Therefore, it is only the α_{ij} that are the actual decision variables in this problem, but we introduce the additional variables π_i for notational simplicity.

The problem in Eq. (9) can be interpreted as an MDP with a compact action space. But the choice of control action in this case does not affect the reward rate r_i. The choice of control action merely affects the transition rates α_{ij}.

A further simplification of the above problem is possible if we make the assumption that the sets Ω_{ij} are closed intervals. In this case, we can write $\Omega_{ij} = [\underline{\alpha}_{ij}, \overline{\alpha}_{ij}]$, where $\underline{\alpha}_{ij}$ and $\overline{\alpha}_{ij}$ are the minimum and maximum values that α_{ij} can take, respectively. With this assumption and by writing $x_{ij} = \pi_i \alpha_{ij}$ we can formulate the optimization problem in Eq. (9) as a linear program as follows:

$$\begin{aligned} & \underset{\{x_{ij}, \pi_i\}}{\text{minimize}} \ \sum_{i\in S} \pi_i r_i \\ \text{subject to} \quad & \textstyle\sum_{j\in S} x_{ij} = \sum_{j\in S} x_{ji} \text{ for } i \in S \\ & \textstyle\sum_{i\in S} \pi_i = 1 \\ & \pi_i \geq 0 \text{ for } i \in S \\ & \pi_i \underline{\alpha}_{ij} \leq x_{ij} \leq \pi_i \overline{\alpha}_{ij} \text{ for } i, j \in S. \end{aligned} \tag{10}$$

Note that to write the last constraint we might have carried out division by zero, if $\pi_i = 0$. But this does not pose a difficulty because $\pi_i = 0$ implies that state i is a transient state, and the value of α_{ij} for $j \in S$ are "don't care" values. (Because the state space S is finite, only positive recurrent and transient states exist. Therefore, when $\pi_i = 0$ it is guaranteed that state i is transient.) In other words, although the last constraint in the problem in Eq. (10) does not convey the meaning in the last constraint of Eq. (9) when $\pi_i = 0$, it does not pose a difficulty because state i is then a transient state and transition rates from transient states can arbitrarily be set to any value without affecting the solution. Also note that if all possible choices of the transition rates $\{\alpha_{ij}\}$ lead to the state space S forming an irreducible CTMC, then it is guaranteed that $\pi_i > 0$ [18].

Because the transition rates are fixed but unknown, we can associate a transition rate value with each state and search for such an optimal transition rate value. Our solution techniques (the optimization and policy iteration approaches) search for such an optimal transition rate value. If the transition rates are time-varying and if their values as a function of time is not known, then we need to determine the optimal transition rate value for each state and each visit to the state. (While the transition rates are time-varying, the sets Ω_{ij} are assumed to be time-invariant.) But if the sets Ω_{ij} are compact, it has been shown that there exists an optimal pure policy under fairly general conditions [19, 20]. Note that the sensitivity analysis problem with time-varying and unknown transition rates is itself an MDP with the action space in state i being $\prod_{j \in S} \Omega_{ij}$. Thus if the conditions under which an optimal pure policy exists are satisfied, then we can use the techniques developed in this paper (the optimization and policy iteration approaches) for solving the sensitivity analysis problem with time-varying transition rates.

4.2 Policy Iteration Approach

We now describe a policy-iteration equivalent to the above optimization solution approach. The policy iteration technique we describe in this section searches for the optimal solution only among pure policies that choose the same transition rate during each visit to a state. The policy iteration technique developed here is similar to the one in [8].

Algorithm: Policy iteration solution to the sensitivity analysis problem with independent transition rates.

1. Select any feasible set of transition rates $\{\alpha_{ij}\}$ (i.e., satisfying the condition $\alpha_{ij} \in \Omega_{ij}$ for all i and j) and some $k \in S$.

2. We introduce a set of auxiliary variables ν_i, $i \in S$, and use the set $\{\alpha_{ij}\}$ to solve for ν_i, $i \in S$, with $\nu_k = 0$ using the following system of equations:

$$R + \nu_i \sum_{j \in S} \alpha_{ij} = r_i + \sum_{j \in S} \alpha_{ij} \nu_j \text{ for } i \in S. \tag{11}$$

3. For each $i \in S$, find α_{ij} for $j \in S$ such that $\alpha_{ij} \in \Omega_{ij}$, and minimizes

$$r_i + \sum_{j \in S} \alpha_{ij} \nu_j - \sum_{j \in S} \alpha_{ij} \nu_i. \tag{12}$$

   Let $\{\alpha'_{ij}\}$ be the set of transition rates obtained that minimize Eq. (12) for each $i \in S$. If $\alpha'_{ij} = \alpha_{ij}$ for each i and j, then R is the minimum average reward per unit time, and $\{\alpha_{ij}\}$ is the set of transition rates that yields this minimum value. Otherwise, return to Step 2 with the new set of $\{\alpha'_{ij}\}$.

In the above algorithm, R represents the expected reward per unit time for that particular choice of $\{\alpha_{ij}\}$. To see this, multiply Eq. (11) by π_i and sum over all $i \in S$ to get

$$R\sum_{i\in S}\pi_i + \sum_{i\in S}\pi_i\nu_i\sum_{j\in S}\alpha_{ij} = \sum_{i\in S}\pi_i r_i + \sum_{i\in S}\sum_{j\in S}\pi_i\alpha_{ij}\nu_j.$$

Now using $\sum_{i\in S}\pi_i = 1$ and $\pi_i\sum_{j\in S}\alpha_{ij} = \sum_{j\in S}\pi_j\alpha_{ji}$ on the left hand side of the above equation we have

$$R + \sum_{i\in S}\sum_{j\in S}\nu_i\pi_j\alpha_{ji} = \sum_{i\in S}\pi_i r_i + \sum_{i\in S}\sum_{j\in S}\pi_i\alpha_{ij}\nu_j. \tag{13}$$

The second term on the right hand side can be written as $\sum_{i\in S}\sum_{j\in S}\nu_i\pi_j\alpha_{ji}$ by interchanging i and j. But this is the same as the second term on the left in Eq. (13). We thus have the following relation for R:

$$R = \sum_{i\in S}\pi_i r_i, \tag{14}$$

which is the same as the expression for the average reward per unit time R in Eq. (3).

We prove that the algorithm above converges in a finite number of steps. For this, we first show two propositions. We first prove that the algorithm has the descent property.

Proposition 1 *The policy iteration algorithm for sensitivity analysis has the descent property; i.e., if $R^{(k)}$ for $k = 1, 2, \ldots$ is the sequence of R values obtained from the above algorithm, then $R^{(k+1)} \leq R^{(k)}$ for all k.*

Proof: Let $\{\alpha_{ij}\}$ be the transition rates obtained from an iteration of the algorithm, and let $\{\nu_i\}$ and R be the values obtained from Eq. (11) for this choice of $\{\alpha_{ij}\}$. Let $\{\alpha'_{ij}\}$, $\{\nu'_i\}$, and R' be the corresponding values obtained from the next iteration. Assume that $\alpha_{ij} \neq \alpha'_{ij}$ for at least some i and j. Otherwise, the descent property trivially holds.

Using Eq. (11) we have, for each $i \in S$,

$$R - R' = \sum_{j\in S}\alpha_{ij}\nu_j - \nu_i\sum_{j\in S}\alpha_{ij} - \sum_{j\in s}\alpha'_{ij}\nu'_j + \nu'_i\sum_{j\in S}\alpha'_{ij}.$$

Adding and subtracting the term $\sum_{j\in S}\alpha'_{ij}\nu_j - \nu_i\sum_{j\in S}\alpha'_{ij}$ to the above equation and writing

$$\theta_i = \sum_{j\in S}\alpha_{ij}\nu_j - \nu_i\sum_{j\in S}\alpha_{ij} - \left[\sum_{j\in S}\alpha'_{ij}\nu_j - \nu_i\sum_{j\in S}\alpha'_{ij}\right], \tag{15}$$

we have

$$R - R' = \theta_i + \sum_{j\in S}\alpha'_{ij}(\nu_j - \nu'_j) - (\nu_i - \nu'_i)\sum_{j\in S}\alpha'_{ij}.$$

From the definition of θ_i it is clear that $\theta_i \geq 0$ for all $i \in S$ and that $\theta_i > 0$ for at least one $i \in S$. This is because $\{\alpha'_{ij}\}$ minimizes the expression in Eq. (12) and also because of

the assumption that $\{\alpha_{ij}\}$ and $\{\alpha'_{ij}\}$ are not identical. Now substituting $\Delta R = R - R'$ and $\Delta\nu_i = \nu_i - \nu'_i$ for each $i \in S$ we have

$$\Delta R + \Delta\nu_i \sum_{j\in S} \alpha'_{ij} = \theta_i + \sum_{j\in S} \alpha'_{ij}\Delta\nu_j.$$

The above equation is similar in form to Eq. (11), and following an argument similar to the derivation of Eq. (14) we have $\Delta R = \sum_{i\in S} \pi_i\theta_i$. But, as already seen, $\theta_i \geq 0$ for each $i \in S$ and $\theta_i > 0$ for at least one $i \in S$. Using the fact that $\pi_i \geq 0$ for all $i \in S$, it follows that $\Delta R \geq 0$ or $R \geq R'$. In particular, if θ_i is strictly positive for a recurrent state under the choice $\{\alpha'_{ij}\}$, then $R > R'$. Note that a recurrent state implies $\pi_i > 0$. Therefore, the policy iteration algorithm has the descent property. □

Next, we prove that if the algorithm terminates, then there exists no feasible $\{\alpha_{ij}\}$ that yields a smaller expected reward per unit time R.

Proposition 2 *If the algorithm terminates, then there cannot exist a feasible set of* $\{\alpha_{ij}\}$ *that yields a smaller expected reward per unit time* R.

Proof: The proof is by contradiction. Suppose the algorithm terminates with $\{\alpha_{ij}\}$ and there exists $\{\alpha'_{ij}\}$ different from $\{\alpha_{ij}\}$ that yields a smaller R value. In other words, if R and R' are the values obtained from $\{\alpha_{ij}\}$ and $\{\alpha'_{ij}\}$, respectively using Eq. (11), then $R' < R$. From Eq. (11) we can obtain the following expression for $R - R'$ using steps similar to the ones in the proof of Proposition 1:

$$R - R' = \theta_i + \sum_{j\in S} \alpha'_{ij}(\nu_j - \nu'_j) - (\nu_i - \nu'_i)\sum_{j\in S} \alpha'_{ij},$$

where θ_i is the same as that given in Eq. (15). But $\theta_i \leq 0$ for all $i \in S$ because $\{\alpha_{ij}\}$ yields the minimizer of the term $\sum_{j\in S} \alpha_{ij}\nu_j - \sum_{j\in S} \alpha_{ij}\nu_i$. Using the same reasoning as in the derivation of Eq. (14), we have

$$\Delta R = R - R' = \sum_{i\in S} \pi_i\theta_i.$$

But $\Delta R \leq 0$ because $\theta_i \leq 0$ for all $i \in S$. Hence, this is a contradiction because we assumed that $R' < R$. □

Using the above two propositions, we can show the convergence of our algorithm to the optimal solution in a finite number of steps, using the assumption that the sets Ω_{ij} are compact.

Theorem 1 *The policy iteration algorithm for sensitivity analysis converges to the optimal solution in a finite number of steps.*

Proof: Because the sets Ω_{ij} for each $i, j \in S$ are compact, there exist $\{\underline{\alpha}_{ij}\}$ and $\{\overline{\alpha}_{ij}\}$ such that

$$\underline{\alpha}_{ij} = \min\{x : x \in \Omega_{ij}\}$$

$$\overline{\alpha}_{ij} = \max\{x : x \in \Omega_{ij}\}$$

for each $i, j \in S$. Next note that each time we obtain the solution to the optimization problem in Eq. (12), each α_{ij} takes one of two values. The optimal solution would either be $\underline{\alpha}_{ij}$ or $\overline{\alpha}_{ij}$ depending on whether $\nu_j - \nu_i$ is positive or negative. By convention, let us choose $\underline{\alpha}_{ij}$ as the solution of Eq. (12) if $\nu_j - \nu_i = 0$. Thus the algorithm has only a finite number of points (at most $2^{|S|(|S|-1)}$) to visit.

Also, before termination, the algorithm cannot return to the same set of $\{\alpha_{ij}\}$. To show this, we need another property of the policy iteration algorithm that if there is no positive descent in the R value and if the algorithm has not terminated, then $\nu_i' < \nu_i$ for some transient state i under $\{\alpha_{ij}'\}$. This property can be proved using an argument similar to the proof in [12] (Vol. 2, page 216). Proposition 1 together with the above property guarantees that no set of R and ν_i values can be repeated. This implies that no set of $\{\alpha_{ij}\}$ can be repeated because a set of $\{\alpha_{ij}\}$ uniquely determines a set of R and ν_i values. Thus the algorithm has a finite number of points to visit, and during each iteration it visits a different point. Therefore, the algorithm terminates in a finite number of steps. By Proposition 2, the set of $\{\alpha_{ij}\}$ with which the algorithm terminates is the optimal set of transition rates. □

5 Sensitivity Analysis with Dependent Transition Rates

5.1 Optimization Approach

Following our optimization solution to the sensitivity analysis problem with independent transition rates, we can formulate the problem of finding the minimum expected reward per unit time when there is dependence among transition rates as follows:

$$\begin{aligned} \underset{\{\alpha_{ij}, \pi_i\}}{\text{minimize}} \quad & \sum_{i \in S} \pi_i r_i \\ \text{subject to} \quad & \pi_i \textstyle\sum_{j \in S} \alpha_{ij} = \sum_{j \in S} \pi_j \alpha_{ji} \text{ for } i \in S \\ & \textstyle\sum_{i \in S} \pi_i = 1 \\ & \pi_i \geq 0 \text{ for } i \in S \\ & \{\alpha_{ij}\} \in \Omega, \end{aligned} \tag{16}$$

where the set Ω is assumed compact as before to guarantee the existence of the minimizer. Note that, unlike in the case of independent transition rates, we cannot choose the values of α_{ij} independently. Thus the constraint that $\{\alpha_{ij}\} \in \Omega$ cannot be separated into individual constraints on the α_{ij}.

If the transition rates are fixed but unknown we can search for the optimal transition rate value associated with each state. If the transition rates are time-varying and if they are known to vary slow enough that the dependence between the transition rates is not lost, we can use the above technique (and the policy iteration technique in the following section) to determine the optimal transition rate value for each state. On the other hand, if the transition rates vary fast enough with time, then the dependence between the transition rates will be lost in the time duration needed for the system to move to a state with a dependent transition rate. Thus if the transition rates vary fast enough, then this problem can be treated as one with independent transition rates, and the techniques developed in Section 4 can be used.

We now prove a useful proposition with which we can approximate the performance of a system with dependent transition rates, by assuming independent transition rates. We first define the sets Ω'_{ij} such that for every point in Ω its (i,j)-th component α_{ij} belongs to Ω'_{ij} for each i and j. The Ω'_{ij} for each i and j can be obtained as follows. Let

$$\alpha_{ij}^{min} = \min\{\alpha_{ij} : \alpha_{ij} \in \Omega\}$$

and

$$\alpha_{ij}^{max} = \max\{\alpha_{ij} : \alpha_{ij} \in \Omega\}.$$

Then $\Omega'_{ij} = [\alpha_{ij}^{min}, \alpha_{ij}^{max}]$. In other words, the Ω'_{ij} are chosen such that the "box" defined by them completely contains Ω, i.e., $\Omega \subset \prod_{i,j} \Omega'_{ij}$, where $\prod_{i,j} \Omega'_{ij}$ is the cartesian product of the sets Ω'_{ij}. Because the set Ω is bounded, we can choose the sets Ω'_{ij} for all i and j such that they are compact.

Proposition 3 *Let $\underline{R}^{dep}$ and $\overline{R}^{dep}$ be the optimal average rewards per unit time of the minimization (as given in Eq. (5)) and maximization (with minimization in Eq. (5) replaced by maximization) sensitivity analysis problems for a system with dependent transition rates. Then the corresponding problem with independent transition rates obtained by replacing the constraint $\{\alpha_{ij}\} \in \Omega$ with the condition $\alpha_{ij} \in \Omega'_{ij}$ for $i, j \in S$ with $\Omega \subset \prod_{i,j} \Omega'_{ij}$ has minimum and maximum average rewards per unit time $\underline{R}^{ind}$ and $\overline{R}^{ind}$ such that $\underline{R}^{ind} \leq \underline{R}^{dep} \leq \overline{R}^{dep} \leq \overline{R}^{ind}$.*

Proof: Consider the following optimization problem:

$$\underset{\{\alpha_{ij}, \pi_i\}}{\text{minimize}} \sum_{i \in S} \pi_i r_i$$

$$\begin{aligned} \text{subject to} \quad & \pi_i \sum_{j \in S} \alpha_{ij} = \sum_{j \in S} \pi_j \alpha_{ji} \text{ for } i \in S \\ & \sum_{i \in S} \pi_i = 1 \\ & \pi_i \geq 0 \text{ for } i \in S \\ & \alpha_{ij} \in \Omega'_{ij} \text{ for } i, j \in S, \end{aligned} \tag{17}$$

which is obtained by replacing the constraint $\{\alpha_{ij}\} \in \Omega$ with the condition $\alpha_{ij} \in \Omega'_{ij}$ for $i, j \in S$. The above problem has independent transition rates, and let its optimal solution be $\underline{R}^{ind}$. Because $\Omega \subset \prod_{i,j} \Omega'_{ij}$, the constraint set in the optimization problem in Eq. (16) is a subset of that in the problem in Eq. (17). Therefore, $\underline{R}^{ind} \leq \underline{R}^{dep}$.

Similarly, it follows by changing the problem into one of maximization that $\overline{R}^{dep} \leq \overline{R}^{ind}$. And because $\underline{R}^{dep} \leq \overline{R}^{dep}$ the result follows. □

The above proposition gives us a means of getting bounds on the performance of the system with dependent transition rates. We merely replace the constraint set Ω by the constraint set $\prod_{i,j} \Omega'_{ij}$ and treat the problem as one with independent transition rates. This gives us bounds within which the performance of the system with dependent transition rates will lie. The above proposition is especially useful because efficient techniques like linear programming (see Eq. (10)) exist to solve the problem of sensitivity analysis with independent transition rates. Note that the problem of sensitivity analysis with dependent transition rates cannot be formulated as a linear program. Also, to get a good approximation of the performance of the sensitivity analysis problem with dependent transition rates, the sets Ω'_{ij} have to be chosen so that they tightly bound the original set Ω, i.e., the set $\prod_{i,j} \Omega'_{ij}$ should be the smallest "box" containing the set Ω.

5.2 Policy Iteration Approach

We now describe a policy iteration solution to the problem.

Algorithm: Policy iteration solution to the sensitivity analysis problem with dependent transition rates.

1. Select any feasible set of transition rates $\{\alpha_{ij}\}$ and some $k \in S$.

2. Use these α_{ij} to solve for ν_i, $i \in S$, with $\nu_k = 0$ using the following system of equations:

$$R + \nu_i \sum_{j \in S} \alpha_{ij} = r_i + \sum_{j \in S} \alpha_{ij} \nu_j \text{ for } i \in S. \tag{18}$$

3. Find $\{\alpha^*_{ij}\}$, optimizing the following problem:

$$\underset{\{\alpha'_{ij}, \pi'_i\}}{\text{maximize}} \sum_{i \in S} \pi'_i \left\{ \sum_{j \in S} \nu_j (\alpha_{ij} - \alpha'_{ij}) + \nu_i \sum_{j \in S} (\alpha'_{ij} - \alpha_{ij}) \right\}$$

$$\begin{aligned} \text{subject to} \quad & \pi_i' \textstyle\sum_{j \in S} \alpha_{ij}' = \sum_{j \in S} \pi_j' \alpha_{ji}' \text{ for } i \in S \\ & \textstyle\sum_{i \in S} \pi_i' = 1 \\ & \pi_i' \geq 0 \text{ for } i \in S \\ & \{\alpha_{ij}'\} \in \Omega. \end{aligned} \tag{19}$$

If $\alpha_{ij}^* = \alpha_{ij}$ for each i and j, then R is the minimum average reward per unit time, and $\{\alpha_{ij}\}$ is the set of transition rates that yields this minimum value. Otherwise return to Step 2 with $\alpha_{ij} = \alpha_{ij}^*$ for $i, j \in S$.

Note that, unlike the case where the transition rates are independent, we cannot decide the value of transition rates independently in each state. Thus Step 3 of the algorithm becomes more complicated because we have to decide on a set of $\{\alpha_{ij}\}$ that results in descent in the average per-unit-time reward. Step 3 of the algorithm for independent transition rates cannot be used because the choice of α_{ij} for $j \in S$ that results from this step for a particular i might not result in descent, because the transition rates from other states $k \neq i$ are also affected as a result of this choice. Notice that the term $\sum_{j \in S} \nu_j(\alpha_{ij} - \alpha_{ij}') + \nu_i \sum_{j \in S}(\alpha_{ij}' - \alpha_{ij})$ is the same as θ_i defined in Eq. (15). Therefore, the optimization problem in Step 3 of the above algorithm maximizes the magnitude of descent that was obtained in the proof of Proposition 1. Using techniques similar to those in the proof of Propositions 1 and 2 we can prove the following propositions.

Proposition 4 *The policy iteration algorithm for sensitivity analysis with dependent transition rates has the descent property, i.e., if $R^{(k)}$ for $k = 1, 2, \ldots$ is the sequence of R values obtained from the above algorithm, then $R^{(k+1)} \leq R^{(k)}$ for all k.*

Proposition 5 *If the policy iteration algorithm for systems with dependent transition rates terminates, then there cannot exist a feasible set of $\{\alpha_{ij}\}$ that yields a smaller expected reward per unit time R.*

6 Max-min Optimal Policy with Independent Transition Rates

In this section, we develop techniques to obtain the max-min optimal policy for a system with independent transition rates. We first make an assumption that guarantees the existence of an optimal policy among the set of pure policies. Note that if the action space in each state is finite, then the total number of distinct pure policies is given by $\Pi_{i \in S} |K_i|$, where we have used $|K_i|$ to denote the number of elements in the set K_i. For a given pure policy $h \in \mathcal{H}$,

let the solution to the sensitivity analysis problem be $\{\alpha_{ij}^*(h)\}$. It has been shown (see [11]) that if every pure policy h gives rise to a single recurrent class plus a set (possibly empty) of transient states, then there exists an optimal pure policy. We assume that every pure policy $h \in \mathcal{H}$, along with the pessimistic choice of transition rates $\{\alpha_{ij}^*(h)\}$, gives rise to a CTMC with a single recurrent class. With this assumption, we are justified in restricting our search for optimal policies among the set of all pure policies $\mathcal{H}$.

The problem of finding the max-min optimal policy with independent transition rates can be stated as follows:

$$\begin{aligned} &\max_{h\in\mathcal{H}} \min_{\{\alpha_{ij}(h),\pi_i(h)\}} \sum_{i\in S} \pi_i(h) r_i(h) \\ \text{subject to} \quad &\pi_i(h) \textstyle\sum_{j\in S} \alpha_{ij}(h) = \sum_{j\in S} \pi_j(h)\alpha_{ji}(h) \text{ for } i \in S \\ &\textstyle\sum_{i\in S} \pi_i(h) = 1 \\ &\pi_i(h) \geq 0 \text{ for } i \in S \\ &\alpha_{ij}(h) \in \Omega_{ij}(h) \text{ for } i,j \in S, \end{aligned} \tag{20}$$

where $\Omega_{ij}(h)$ for each i, j, and h are assumed compact to guarantee the existence of the optimizer. The maximization is done over all policies and the minimization is done over the range of uncertainty in the transition rates.

We now describe a policy iteration technique to obtain the optimal max-min policy in a system with independent transition rates. The policy iteration technique is a systematic search technique that searches for the optimal max-min policy among all pure policies.

Algorithm: Policy iteration solution to the max-min problem with independent transition rates.

1. Select a pure policy $h = \{h(i) : i \in S\}$ where $h(i) \in K_i$ and some $k \in S$. (Policy h is a mapping from the state space to the action space.)

2. Select a feasible set of transition rates $\alpha_{ij}(h)$. (Note that $\alpha_{ij}(h)$ can be written as $\alpha_{ij}^{h(i)}$.)

3. Use these $\alpha_{ij}(h)$ to solve for $\nu_i(h)$, $i \in S$, with $\nu_k(h) = 0$ using the following system of equations:

$$R(h) + \nu_i(h) \sum_{j\in S} \alpha_{ij}(h) = r_i(h) + \sum_{j\in S} \alpha_{ij}(h)\nu_j(h) \text{ for } i \in S, \tag{21}$$

   where $r_i(h) = r_i^{h(i)}$.

4. For each $i \in S$ find $\alpha_{ij}(h)$, $j \in S$, such that $\alpha_{ij}(h) \in \Omega_{ij}(h)$ and minimizing

$$r_i(h) + \sum_{j\in S} \alpha_{ij}(h)\nu_j(h) - \sum_{j\in S} \alpha_{ij}(h)\nu_i(h). \tag{22}$$

(Note that $\Omega_{ij}(h)$ can also be written as $\Omega_{ij}^{h(i)}$.) Let $\{\alpha_{ij}^{'}(h)\}$ be the set of transition rates obtained that minimizes Eq. (22) for each $i \in S$. If $\alpha_{ij}^{'}(h) = \alpha_{ij}(h)$ for each i and j, then go to Step 5 with $\alpha_{ij}^{*}(h) = \alpha_{ij}(h)$ as the pessimistic choice of transition rates for policy h. Otherwise return to Step 3 with $\alpha_{ij}(h) = \alpha_{ij}^{'}(h)$.

5. For each state $i \in S$, find alternatives $h^{'}(i)$ that maximize for $u \in K_i$

$$\min_{\alpha_{ij}^{u} \in \Omega_{ij}^{u}} r_i^u + \sum_{j \in S} \alpha_{ij}^{u} \nu_j^{*}(h) - \sum_{j \in S} \alpha_{ij}^{u} \nu_i^{*}(h),$$

where $\nu_i^{*}(h)$ are the values obtained from Step 3 for the pessimistic choice of transition rates $\{\alpha_{ij}^{*}(h)\}$. If the policy changes, return to Step 2 with $h(i) = h^{'}(i)$ for $i \in S$. Otherwise, terminate with h as the optimal max-min policy.

Note that Steps 3 and 4 are the same as those in the sensitivity analysis problem. With the assumption that the Ω_{ij}^{u} are compact for each i, j, and u, this part of the algorithm terminates in a finite number of steps as shown in Theorem 1. Next note that the number of pure policies is finite because of the finite action and state spaces. Therefore, the algorithm terminates in a finite number of steps. To prove that the above algorithm converges to the optimal policy we follow an approach similar to the one for the sensitivity analysis problem. We first prove that the algorithm results in a sequence of policies whose worst-case average per-unit-time reward increases. We then show that if the algorithm terminates, then there exists no better policy. Because the algorithm terminates in a finite number of steps, we would have proved the convergence of our algorithm to the optimal max-min policy. The techniques that we use in our proof are similar to those in [8].

Proposition 6 *The policy iteration algorithm for obtaining the optimal max-min policy for a system with independent transition rates has the descent property; i.e., if h and $h^{'}$ ($h^{'} \neq h$) are two successive policies obtained in that order from Step 5, and if $R^{*}(h)$ and $R^{*}(h^{'})$ are the corresponding worst-case values, then $R^{*}(h^{'}) \geq R^{*}(h)$.*

Proof: Let $\{\alpha_{ij}^{*}(h)\}$ and $\{\alpha_{ij}^{*}(h^{'})\}$ represent the optimal values obtained from Step 4 for policies h and $h^{'}$, respectively. In other words, these are the worst-case transition rates for policies h and $h^{'}$. Then $R^{*}(h)$, $\{\nu_i^{*}(h)\}$, $R^{*}(h^{'})$, and $\{\nu_i^{*}(h^{'})\}$ are obtained from the following equations:

$$R^{*}(h) + \nu_i^{*}(h) \sum_{j \in S} \alpha_{ij}^{*}(h) = r_i(h) + \sum_{j \in S} \alpha_{ij}^{*}(h) \nu_j^{*}(h) \text{ for } i \in S$$

$$R^{*}(h^{'}) + \nu_i^{*}(h^{'}) \sum_{j \in S} \alpha_{ij}^{*}(h^{'}) = r_i(h^{'}) + \sum_{j \in S} \alpha_{ij}^{*}(h^{'}) \nu_j^{*}(h^{'}) \text{ for } i \in S.$$

Therefore,

$$R^*(h^{'}) - R^*(h) + \nu_i^*(h^{'}) \sum_{j \in S} \alpha_{ij}^*(h^{'}) - \nu_i^*(h) \sum_{j \in S} \alpha_{ij}^*(h) =$$
$$r_i(h^{'}) - r_i(h) + \sum_{j \in S} \alpha_{ij}^*(h^{'}) \nu_j^*(h^{'}) - \sum_{j \in S} \alpha_{ij}^*(h) \nu_j^*(h). \tag{23}$$

Adding and subtracting the terms $\nu_i^*(h) \sum_{j \in S} \alpha_{ij}^*(h^{'}) + \sum_{j \in S} \alpha_{ij}^*(h^{'}) \nu_j^*(h)$ and by making the substitution

$$\theta_i = r_i(h^{'}) - r_i(h) + \sum_{j \in S} \alpha_{ij}^*(h^{'})(\nu_j^*(h) - \nu_i^*(h)) - \sum_{j \in S} \alpha_{ij}^*(h)(\nu_j^*(h) - \nu_i^*(h)),$$

we have the following expression:

$$R^*(h^{'}) - R^*(h) + (\nu_i^*(h^{'}) - \nu_i^*(h)) \sum_{j \in S} \alpha_{ij}^*(h^{'}) = \theta_i + \sum_{j \in S} \alpha_{ij}^*(h^{'})(\nu_j^*(h^{'}) - \nu_j^*(h)).$$

Making the substitution $\Delta R = R^*(h^{'}) - R^*(h)$ and $\Delta \nu_i = \nu_i^*(h^{'}) - \nu_i^*(h)$, we have

$$\Delta R + \Delta \nu_i \sum_{j \in S} \alpha_{ij}^*(h^{'}) = \theta_i + \sum_{j \in S} \alpha_{ij}^*(h^{'}) \Delta \nu_j.$$

Using a method similar to the derivation of Eq. (14), we get

$$\Delta R = \sum_{i \in S} \pi_i \theta_i.$$

It is clear that

$$r_i(h^{'}) + \sum_{j \in S} \alpha_{ij}^*(h^{'})(\nu_j^*(h) - \nu_i^*(h)) \geq \tag{24}$$
$$\min_{\alpha_{ij}(h^{'}) \in \Omega_{ij}(h^{'})} \left\{ r_i(h^{'}) + \sum_{j \in S} \alpha_{ij}(h^{'})(\nu_j^*(h) - \nu_i^*(h)) \right\} \geq r_i(h) + \sum_{j \in S} \alpha_{ij}^*(h)(\nu_j^*(h) - \nu_i^*(h)).$$

The first of the inequalities follows from the fact that $\{\alpha_{ij}^*(h^{'})\}$ is one specific element of the set over which the minimization is done in the term in the middle. The second of the inequalities follows because $h^{'}$ was obtained from Step 5 and satisfied this property. The difference between the first and the last terms in the above equation is θ_i, and thus we have $\theta_i \geq 0$ for all $i \in S$ and $\theta_i > 0$ for some $i \in S$. Because $\pi_i \geq 0$ for all $i \in S$, it follows that $\Delta R \geq 0$ or $R^*(h^{'}) \geq R^*(h)$. In particular, if $\theta_i > 0$ for a recurrent state for the choice of policy $h^{'}$ and transition rates $\{\alpha_{ij}^*(h^{'})\}$, then $R^*(h^{'}) > R^*(h)$. □

We now prove that if the max-min policy iteration algorithm terminates, then there exists no better policy.

Proposition 7 *If the policy iteration algorithm to obtain the optimal max-min policy terminates, then there cannot exist a policy that yields a larger worst-case expected reward per unit time.*

Proof: The proof is by contradiction. Assume that h is the policy with which the algorithm terminates, and let there exist a policy $h^{'}$ that yields a larger worst-case reward per unit time $R^*(h^{'})$. Then it follows that

$$R^*(h) + \nu_i^*(h) \sum_{j \in S} \alpha_{ij}^*(h) = r_i(h) + \sum_{j \in S} \alpha_{ij}^*(h) \nu_j^*(h) \text{ for } i \in S$$

$$R^*(h^{'}) + \nu_i^*(h^{'}) \sum_{j \in S} \alpha_{ij}^*(h^{'}) = r_i(h^{'}) + \sum_{j \in S} \alpha_{ij}^*(h^{'}) \nu_j^*(h^{'}) \text{ for } i \in S,$$

where $R^*(h)$ and $R^*(h^{'})$ are the worst-case per-unit-time rewards for policies h and $h^{'}$, respectively, and $\{\alpha_{ij}^*(h)\}$ and $\{\alpha_{ij}^*(h^{'})\}$ are the set of transition rates that yield these worst-case values. The assumption that $h^{'}$ yields a larger worst-case per-unit-time reward means that $R^*(h^{'}) > R^*(h)$.

Let $\{\alpha_{ij}(h^{'})\}$ be any feasible set of transition rates, and let $R(h^{'})$ and $\nu_i(h^{'})$ be the solution to the following system of equations:

$$R(h^{'}) + \nu_i(h^{'}) \sum_{j \in S} \alpha_{ij}(h^{'}) = r_i(h^{'}) + \sum_{j \in S} \alpha_{ij}(h^{'}) \nu_j(h^{'}) \text{ for } i \in S.$$

Since $\{\alpha_{ij}^*(h^{'})\}$ minimizes the per-unit-time reward under policy $h^{'}$, we have $R^*(h^{'}) \leq R(h^{'})$. Now if $h^{'}$ is a better policy it follows that $R^*(h) < R(h^{'})$ for any feasible $\{\alpha_{ij}(h^{'})\}$. But this is a contradiction.

To see this, suppose $\{\alpha_{ij}(h^{'})\}$ results from the minimization in Step 5 of the algorithm. But since the algorithm terminated with policy h, it follows from Step 5 that for all $i \in S$

$$r_i(h) + \sum_{j \in S} \alpha_{ij}^*(h) \nu_j^*(h) - \sum_{j \in S} \alpha_{ij}^*(h) \nu_i^*(h) \geq r_i(h^{'}) + \sum_{j \in S} \alpha_{ij}(h^{'}) \nu_j^*(h) - \sum_{j \in S} \alpha_{ij}(h^{'}) \nu_i^*(h). \quad (25)$$

Denote

$$\theta_i = r_i(h^{'}) - r_i(h) + \sum_{j \in S} (\alpha_{ij}(h^{'}) - \alpha_{ij}^*(h)) \nu_j^*(h) - \sum_{j \in S} (\alpha_{ij}(h^{'}) - \alpha_{ij}^*(h)) \nu_i^*(h).$$

And $\theta_i \leq 0$ for all $i \in S$ from Eq. (25). But

$$R(h^{'}) - R^*(h) + \nu_i(h^{'}) \sum_{j \in S} \alpha_{ij}(h^{'}) - \nu_i^*(h) \sum_{j \in S} \alpha_{ij}^*(h) =$$
$$r_i(h^{'}) - r_i(h) \sum_{j \in S} \alpha_{ij}(h^{'}) \nu_j(h^{'}) - \sum_{j \in S} \alpha_{ij}^*(h) \nu_j^*(h) \text{ for } i \in S.$$

Adding and subtracting $\sum_{j \in S} \alpha_{ij}(h')(\nu_j^*(h) - \nu_i^*(h))$, we have

$$R(h') - R^*(h) + (\nu_i(h') - \nu_i^*(h)) \sum_{j \in S} \alpha_{ij}(h') = \theta_i + \sum_{j \in S} \alpha_{ij}(h')(\nu_j(h') - \nu_j^*(h)).$$

Putting $\Delta R = R(h') - R^*(h)$ and $\Delta \nu_i = \nu_i(h') - \nu_i^*(h)$, and following an approach similar to the derivation of Eq. (14), we have

$$\Delta R = \sum_{i \in S} \pi_i(h') \theta_i.$$

Because $\theta_i \leq 0$ and $\pi_i(h') \geq 0$ for all $i \in S$, we have $\Delta R \leq 0$ or $R(h') \leq R^*(h)$ which gives us the contradiction. The proof is thus complete. □

Using the above two propositions and the assumption that $\Omega_{ij}(h)$ is compact for all $i, j \in S$ and all policies h, we can prove that the policy iteration algorithm for obtaining the max-min policy converges in a finite number of steps.

Theorem 2 *The policy iteration algorithm to obtain the max-min optimal policy converges to the optimal solution in a finite number of steps.*

Proof: As shown in the proof of Theorem 1, since the sets $\Omega_{ij}(h)$ are compact for every feasible policy h and $i, j \in S$, the inner iteration in Steps 3 and 4 converges in a finite number of steps to the optimal set of transition rates for each policy.

Because of the descent property of the algorithm, the same policy cannot be revisited in Step 5 unless the algorithm has terminated. The number of possible pure policies is also finite because of the finiteness of the state and action spaces. Therefore, the algorithm terminates in a finite number of steps. As a consequence of Proposition 7, the policy with which the algorithm terminates is the optimal max-min policy. □

7 Max-min Optimal Policy with Dependent Transition Rates

In this section, we consider the problem of determining a max-min optimal policy for the case where there is dependence among transition rates. As in the previous section, we restrict our search for optimal policies among all pure policies. For this purpose, we make the assumption that each pure policy $h \in \mathcal{H}$ along with the pessimistic choice of transition rates $\{\alpha_{ij}^*(h)\}$ gives rise to a CTMC with a single recurrent class along with a set (possibly empty) of transient states.

From the descriptions in Sections 5 and 6, it should be clear that we can formulate the problem of obtaining the max-min optimal policy with dependent transition rates as follows:

$$\begin{aligned} \max_{h\in\mathcal{H}} \min_{\{\alpha_{ij}(h),\pi_i(h)\}} & \sum_{i\in S} \pi_i(h) r_i(h) \\ \text{subject to} \quad \pi_i(h) \textstyle\sum_{j\in S} \alpha_{ij}(h) &= \textstyle\sum_{j\in S} \pi_j(h)\alpha_{ji}(h) \text{ for } i \in S \\ \textstyle\sum_{i\in S} \pi_i(h) &= 1 \\ \pi_i(h) &\geq 0 \text{ for } i \in S \\ \{\alpha_{ij}(h)\} &\in \Omega(h). \end{aligned} \tag{26}$$

Note that the last constraint has been modified so that we no longer have independent constraints on the individual transition rates. Instead, we have a constraint set $\Omega(h)$ within which the transition rates take values for a given policy $h \in \mathcal{H}$.

We now describe a policy iteration solution to the problem of finding an optimal max-min policy when the system has dependent transition rates.

Algorithm: Policy iteration solution to the max-min problem with dependent transition rates.

1. Select a pure policy $h = \{h(i) : i \in S\}$ where $h(i) \in K_i$ and some state $k \in S$.

2. Select a feasible set of transition rates $\alpha_{ij}(h)$.

3. Use these $\alpha_{ij}(h)$ to solve for $\nu_i(h)$, $i \in S$ with $\nu_k(h) = 0$ using the following system of equations:

$$R(h) + \nu_i(h) \sum_{j\in S} \alpha_{ij}(h) = r_i(h) + \sum_{j\in S} \alpha_{ij}(h)\nu_j(h) \text{ for } i \in S. \tag{27}$$

4. Find $\alpha^*_{ij}(h)$, $i, j \in S$ such that it is the optimal solution of the following problem:

$$\begin{aligned} \underset{\{\alpha'_{ij}(h),\pi'_i(h)\}}{\text{maximize}} & \sum_{i\in S} \pi'_i(h) \left\{ \sum_{j\in S} \nu_j(h)(\alpha_{ij}(h) - \alpha'_{ij}(h)) + \nu_i(h) \sum_{j\in S} (\alpha'_{ij}(h) - \alpha_{ij}(h)) \right\} \\ \text{subject to} \quad \pi'_i(h) \textstyle\sum_{j\in S} \alpha'_{ij}(h) &= \textstyle\sum_{j\in S} \pi'_j(h)\alpha'_{ji}(h) \text{ for } i \in S \\ \textstyle\sum_{i\in S} \pi'_i(h) &= 1 \\ \pi'_i(h) &\geq 0 \text{ for } i \in S \\ \{\alpha'_{ij}(h)\} &\in \Omega(h). \end{aligned} \tag{28}$$

   If $\alpha^*_{ij}(h) = \alpha_{ij}(h)$ for each i and j, then go to Step 5 with $\alpha^*_{ij}(h) = \alpha_{ij}(h)$ for each i and j as the pessimistic choice of transition rates for policy h. Otherwise return to Step 3 with $\alpha_{ij}(h) = \alpha^*_{ij}(h)$ for $i, j \in S$.

5. For state $i \in S$, find an alternative $h^{'}(i)$ (which together with $h(j)$ for $j \neq i$ defines a policy $h^{'}$) that maximizes the objective function value for the following optimization problem.

$$\begin{array}{ll} \underset{\{\alpha_{ij}(h^{'}),\pi_i(h^{'})\}}{\text{minimize}} & \sum_{i\in S} \pi_i(h^{'}) \Big\{ r_i(h^{'}) - r_i(h) + \sum_{j\in S} \nu_j^*(h)(\alpha_{ij}(h^{'}) - \alpha_{ij}^*(h)) + \\ & \qquad \nu_i^*(h) \sum_{j\in S} (\alpha_{ij}^*(h) - \alpha_{ij}(h^{'})) \Big\} \\ \text{subject to} & \pi_i(h^{'}) \sum_{j\in S} \alpha_{ij}(h^{'}) = \sum_{j\in S} \pi_j(h^{'}) \alpha_{ji}(h^{'}) \text{ for } i \in S \\ & \sum_{i\in S} \pi_i(h^{'}) = 1 \\ & \pi_i(h^{'}) \geq 0 \text{ for } i \in S \\ & \{\alpha_{ij}(h^{'})\} \in \Omega(h^{'}), \end{array} \tag{29}$$

where $\nu_i^*(h)$ are the values obtained from Step 3 for the pessimistic choice of transition rates $\{\alpha_{ij}^*(h)\}$. If no such action $h^{'}(i)$ can be found, go to state $i+1$ and repeat Step 5. If such an action $h^{'}(i)$ can be found return to Step 2 with $h(i) = h^{'}(i)$ and all other $h(j)$ unchanged. If the policy does not change for any state $i \in S$ terminate with $h = \{h(i) : i \in S\}$ as the optimal max-min policy.

Note that in the above algorithm, unlike the case of systems with independent transition rates, we cannot determine an alternate action for each state simultaneously. We find an alternate action in one state and return to Step 3 to obtain the optimal worst-case R and ν_i values. The algorithm terminates when there is no alternate action in any of the states that yields a positive objective function value for the optimization problem in Eq. (29).

Using techniques similar to the ones in the proofs of Proposition 6 and 7, we can prove the following results about the policy iteration algorithm.

Proposition 8 *The policy iteration algorithm for obtaining the optimal max-min policy for systems with dependent transition rates has the property that if h and $h^{'}$ ($h^{'} \neq h$) are two successive policies obtained in that order from Step 5, and if $R^*(h)$ and $R^*(h^{'})$ are the corresponding worst-case values, then $R^*(h^{'}) \leq R^*(h)$.*

Proposition 9 *If the policy iteration algorithm for obtaining the optimal max-min policy with dependent transition rates terminates, then there cannot exist a policy that yields a larger worst-case expected reward per unit time.*

8 Numerical Examples

In this section, we present numerical results for the call admission control problem described in Section 3 when there is uncertainty in the transition rates. Figure 1 shows the range of

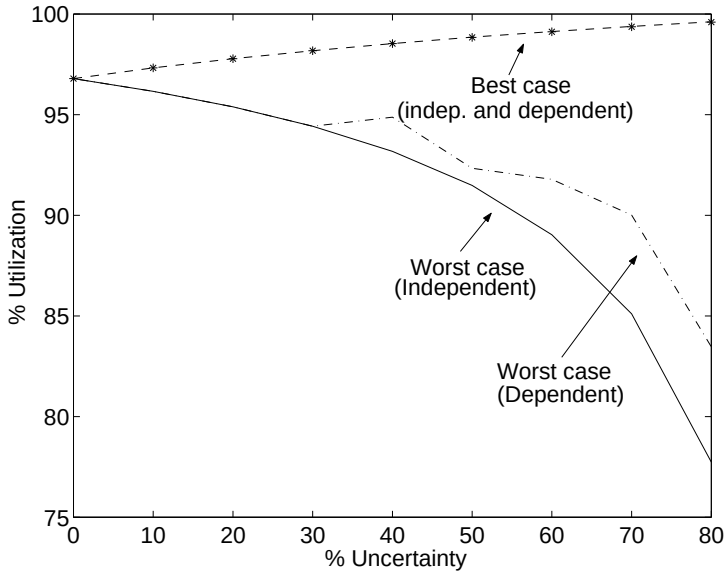

Figure 1: Sensitivity of the admission control decision to uncertainties in the parameters

the throughput values given the percentage uncertainty in the transition rates. (We assume that all the parameters λ_1, λ_2, and μ have the same percentage uncertainty.) Figure 1 is for the scenario where the channel capacity is 5 BWUs with two classes of calls. Class-1 calls require 1 BWU and class-2 calls require 2 BWUs. The nominal mean call holding durations of both classes of calls is fixed at 60 seconds. The nominal value of the arrival rate of class-1 calls is 0.45 calls/second and that of class-2 calls is 0.5 calls/second. A percentage uncertainty of $\delta\%$ in the class-1 call arrival rate implies that it can take any value between $0.45(1 - \delta/100)$ and $0.45(1 + \delta/100)$. As shown in Proposition 3, we find that the range of uncertainty of the throughput with dependent transition rates is in between that obtained with independent transition rates. Note that for the case of independent transition rates, the same quantity, say λ_1, can take different values in different states of the system. Thus the values obtained by assuming dependent transition rates model the call admission control problem exactly, if the system has fixed but unknown transition rates. If the parameters of the call admission control problem are time-varying but are slowly time-varying, i.e., at a much slower time-scale than that of state transitions, then the system with dependent transition rates will more accurately reflect the actual performance range. However, if there is fast variation in the parameters of the problem, assuming independent transition rates will be more accurate. As remarked earlier, assuming independent transition rates makes the task of obtaining the range of throughput values computationally easier.

In Figure 2, we plot the worst-case performances of the policy that assumes the nominal values to be the actual values, and the max-min optimal policy that computes the policy that maximizes the worst-case performance given the percentage uncertainty in the parameters.

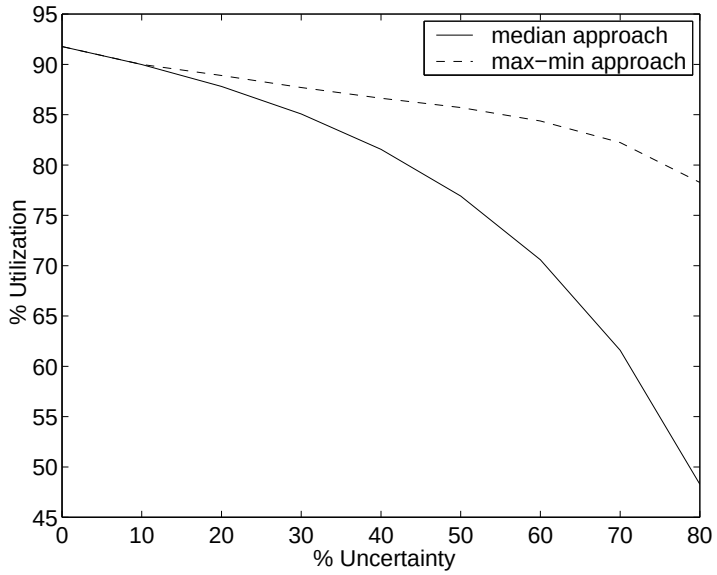

Figure 2: Performance of the max-min and median optimal policies as a function of percentage uncertainty. The values plotted here are their corresponding worst-case performances given the percentage uncertainty.

We label the scheme that assumes the nominal values as the "median policy" because it assumes the actual value of the transition rates to be the midpoint of the range of possible values. Figure 2 is for the scenario where there are two classes of calls, with e_1 and e_2 being 5 and 9 BWUs, respectively. The channel capacity C is assumed to be 10 BWUs. We assume the nominal value of the call arrival rate of class-1 calls to be 0.2 calls/second, and that of class-2 calls to be 1 call/second. The nominal mean call holding duration of both classes of calls is 60 seconds. As expected, the worst-case performance of the max-min optimal policy is better than that of the median policy, which assumes the nominal values to be the actual values. The performance shown here is for the case of independent transition rates. Figure 2 clearly illustrates that when there is uncertainty in the transition rates, designing a policy assuming the nominal values for these parameters can result in poor performance compared to what can be achieved by a robust scheme. Given the range of uncertainty of the transition rates, if the robust scheme is employed, the performance of the system can be no worse than that obtained in Figure 2.

In Figure 3, we plot the best-case performance of the max-max and the median policies. As expected, we find that the best-case performance of the max-max policy is better than that of the median policy, because the max-max policy is designed to maximize the best-case performance of the system. Figure 3 is for the same scenario as that of Figure 2 except that the nominal value of the arrival rate of class-1 calls is 0.1 calls/second. The values of the other parameters are the same as that in Figure 2. Figure 3 illustrates that the best-case

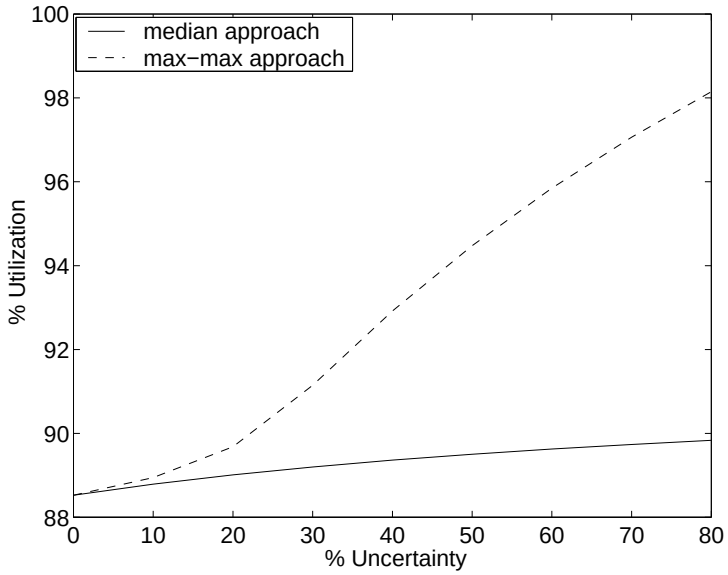

Figure 3: Performance of the max-max and median optimal policies as a function of percentage uncertainty. The values plotted here are their corresponding best-case performances given the percentage uncertainty.

performance of the median scheme can be poor compared to that obtained from the max-max policy.

9 Conclusions

In this paper, we considered infinite-horizon continuous-time MDPs with uncertain transition rates. We investigated the implications of lack of knowledge of the transition rates by considering two different problems. In the first problem, that of sensitivity analysis, the decision-maker wishes to know the range of uncertainty of the per-unit-time reward given the range of uncertainty of the transition rates. In the second problem, that of robust control, the decision-maker wants to make a decision that maximizes the worst-case per-unit-time reward (as determined by the choice of transition rates) given the range of uncertainty of the transition rates. In both these problems, we distinguished between systems with independent transition rates and those with dependent transition rates. In systems with independent transition rates, the transition rates in each state can be chosen independently of their values in other states. But this is not possible when there is dependence among the transition rates. We provide two solution techniques to each of these problems: an optimization-problem technique and a policy iteration algorithm. Our techniques can be applied to MDPs that have fixed but unknown transition rates, and to those with time-varying transition rates. We illustrated our algorithm with numerical examples from the call

admission control problem in telecommunication networks.

References

[1] H. Mine and S. Osaki, *Markovian Decision Processes.* American Elsevier Publishing Company Inc., 1970.

[2] S. M. Ross, *Applied Probability Models with Optimization Applications.* Holden-Day, 1970.

[3] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley and Sons, 1994.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic Programming.* Athena Scientific, 1996.

[5] K. W. Ross and D. H. K. Tsang, "Optimal circuit access policies in an ISDN environment: A Markov decision approach," *IEEE Transactions on Communications*, vol. 37, no. 9, pp. 934–939, September 1989.

[6] J. M. Hyman, A. A. Lazar, and G. Pacifici, "A separation principle between scheduling and admission control for broadband switching," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 4, pp. 605–616, May 1993.

[7] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," *Wireless Networks*, vol. 3, no. 1, pp. 29–41, March 1997.

[8] J. K. Satia and R. E. Lave, Jr., "Markovian decision processes with uncertain transition probabilities," *Operations Research*, vol. 21, pp. 728–740, 1973.

[9] C. C. White and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 43, pp. 739–749, 1994.

[10] R. Givan, S. Leach, and T. Dean, "Bounded parameter Markov decision processes," *Artificial Intelligence*, vol. 122, no. 1–2, pp. 71–109, 2000. Also available from http://dynamo.ecn.purdue.edu/∼givan.

[11] C. Derman, *Finite State Markovian Decision Processes.* Academic Press, 1970.

[12] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vols. I and II.* Athena Scientific, 1995.

[13] M. Sato, K. Abe, and H. Takeda, "An asymptotically optimal learning controller for finite Markov chains with unknown transition probabilities," *IEEE Transactions on Automatic Control*, vol. 30, no. 11, pp. 1147–1149, November 1985.

[14] M. Sato, K. Abe, and H. Takeda, "Learning control of finite Markov chains with unknown transition probabilities," *IEEE Transactions on Automatic Control*, vol. 27, pp. 502–505, April 1982.

[15] K. W. Ross, "Randomized and past-dependent policies for Markov decision processes with multiple constraints," *Operations Research*, vol. 37, no. 3, pp. 474–477, May-June 1989.

[16] X.-R. Cao and H.-T. Fang, "Gradient-based policy iteration: An example," in *Proceedings of the 41st International Conference on Decision and Control*, (Las Vegas, Nevada), pp. 3367–3371, December 2002.

[17] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*. John Wiley and Sons, 1996.

[18] E. Cinlar, *Introduction to Stochastic Processes*. Prentice Hall, 1974.

[19] M. Yasuda, "Semi-Markov decision processes with countable state space and compact action space," *Bulletin of Mathematical Statistics*, vol. 18, pp. 35–54, 1978.

[20] P. J. Schweitzer, "On undiscounted Markovian decision process with compact action spaces," *Operations Research (RAIRO)*, vol. 19, no. 1, pp. 71–86, February 1985.