with probability \( \epsilon > 0 \) to the goal, and costs \( \epsilon(D + 1) \). A simple calculation shows that \( J^*(s) = D + \epsilon > D \), even though reaching the goal from \( s \) is possible. Moreover, notice that this semantic paradox cannot be resolved just by increasing the penalty \( D \), because the cost of \( s \) will always exceed the dead-end penalty by \( \epsilon \).

Therefore, we change the semantics of the finite-penalty model as follows. Whenever the agent reaches any state with the expected cost of reaching the goal equaling \( D \) or greater, the agent simply pays the penalty \( D \) and "gives up", i.e., the process stops. Intuitively, this setting describes scenarios where the agent can put a price on how desirable reaching the goal is. For instance, in the example from the introduction involving a concert in another city, paying the penalty corresponds to deciding not to go to the concert, i.e., foregoing the pleasure the agent would have derived from attending the performance.

The benefit of putting a "cap" on any state’s cost as described above is that the cost of a state under any policy becomes finite, formally defined as

\[
J^f(s) = \min_{\pi} \left\{ D, \mathbb{E} \left[ \sum_{t=0}^{\infty} C(S_t^s, A_t^s) \right] \right\} \tag{5}
\]

It can be shown that for an improper SSP, there exists an optimal policy \( \pi^* \), one that satisfies

\[
\pi^*(s) = \arg \min_{\pi} J^f(s) \forall s \in \mathcal{S} \tag{6}
\]

As we show shortly, we can find such a policy using the expected-cost analysis similar to that for ordinary SSP MDPs. The intuitions just described motivate the fSSPUD model, defined at the end of this section.

The second way of dealing with dead ends we consider in this paper is to view them as truly irrecoverable situations and assign \( D = \infty \) for visiting them. As a motivation, recall the example of planning a climb to the top of Mount Everest. Since dead ends here cannot be avoided with certainty and the penalty of visiting them is \( \infty \), comparing policies based on the expected cost of reaching the goal breaks down — they all have an infinite expected cost. Instead, we would like to find a policy that maximizes the probability of reaching the goal and whose expected cost over the trajectories that reach the goal is the smallest.

To describe this policy evaluation criterion more precisely, let \( S_t^{s'} \) be a random variable denoting a distribution over states \( s' \) for which \( P^{s'}(s_t) > 0 \) and in which the MDP may end up if policy \( \pi \) is executed starting from state \( s \) for \( t \) steps. That is, \( S_t^{s'} \) differs from the variable \( S_t^s \) used previously by considering only states from which \( \pi \) can reach the goal. Using the \( S_t^{s'} \) variables, we can mathematically evaluate \( \pi \) with two ordered criteria by defining the cost of a state as an ordered pair

\[
J^\pi(s) = (P^\pi(s), [J^\pi|P^\pi](s)) \tag{7}
\]

where \([J^\pi|P^\pi](s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} C(S_t^{s'}, A_t^{s'}) \right] \tag{8}\)

Specifically, we write \( \pi(s) \prec \pi'(s) \), meaning \( \pi' \) is preferable to \( \pi \) at \( s \), whenever \( J^\pi(s) \prec J^{\pi'}(s) \), i.e., when either \( P^\pi(s) < P^{\pi'}(s) \), or \( P^\pi(s) = P^{\pi'}(s) \) and \([J^\pi|P^\pi](s) > [J^\pi'|P^{\pi'}](s) \). Notice that the second criterion is used conditionally, only if two policies are equal in terms of the probability of reaching the goal, since maximizing this probability is the foremost priority. Note also that if \( P^\pi(s) = P^{\pi'}(s) = 0 \), then both \([J^\pi|P^\pi](s) \) and \([J^\pi'|P^{\pi'}](s) \) are ill-defined. However, since that means that neither \( \pi \) nor \( \pi' \) can reach the goal from \( s \), we define \([J^\pi|P^\pi](s) = [J^\pi'|P^{\pi'}](s) = 0 \) for such cases, and hence \( J^\pi(s) = J^{\pi'}(s) \).

As in the finite-penalty case, we can demonstrate that there exists a policy \( \pi^* \) that is at least as large as all others at all states under the \( \prec \)-ordering above, and hence optimal, i.e.,

\[
\pi^*(s) = \arg \max_{\pi} J^\pi(s) \forall s \in \mathcal{S} \tag{9}
\]

We are now ready to capture the above intuitions in a definition of the SSPUD MDP class and its subclasses fSSPUD and iSSPUD:

**Definition** An SSP with Unavoidable Dead Ends (SSPUD) MDP is a tuple \( \langle S, A, T, C, G, D, s_0 \rangle \), where \( S, A, T, C, G, \) and \( s_0 \) are as in the SSP MDP definition, \( D \in \mathbb{R}^+ \cup \{\infty\} \) is a penalty incurred if a dead-end state is visited. In a SSPUD MDP, every improper policy must incur an infinite expected cost as defined by Eq. 1 at all states from which it can’t reach the goal with probability 1.

If \( D < \infty \), the MDP is called an fSSPUD MDP, and its optimal solution is a policy \( \pi^* \) satisfying \( \pi^*(s) = \arg \min_{\pi} J^\pi(s) \forall s \in \mathcal{S} \).

If \( D = \infty \), the MDP is called an iSSPUD MDP, and its optimal solution is a policy \( \pi^* \) satisfying \( \pi^*(s) = \arg \max_{\pi} J^\pi(s) \forall s \in \mathcal{S} \).

Our iSSPUDE class is related to multi-objective MDPs, which model problems with several competing objectives, e.g., total time, monetary cost, etc. [Chatterjee et al., 2006, Wukuta, 1995]. Their solutions are Pareto-sets of all non-dominated policies. Unfortunately, such solutions are impractical due to high computational requirements. Moreover, maximizing the probability of goal achievement converts the problem into a GSPPD and hence cannot be easily included in those models. A related criterion has also been studied in robotics [Koenig and Liu, 2002].

**5 The Case of a Finite Penalty**

Equation 5 tells us that for an fSSPUD instance, the cost of any policy at any state is finite. Intuitively, this implies