

# Existence and Finiteness Conditions for Risk-Sensitive Planning: First Results

**Yaxin Liu** ([yxliu@cc.gatech.edu](mailto:yxliu@cc.gatech.edu))  
College of Computing  
Georgia Institute of Technology

**Sven Koenig** ([skoenig@usc.edu](mailto:skoenig@usc.edu))  
Computer Science Department  
University of Southern California

## Abstract

Decision-theoretic planning with risk-sensitive planning objectives is important for building autonomous agents or decision-support agents for real-world applications. However, this line of research has been largely ignored in the artificial intelligence and operations research communities since planning with risk-sensitive planning objectives is much more complex than planning with risk-neutral planning objectives. To remedy this situation, we develop conditions that guarantee the existence and finiteness of the expected utilities of the total plan-execution reward for risk-sensitive planning with totally observable Markov decision process models. In case of Markov decision process models with both positive and negative rewards our results hold for stationary policies only, but we conjecture that they can be generalized to hold for all policies.

## Introduction

Decision-theoretic planning is important since real-world applications need to cope with uncertainty. Many decision-theoretic planners use totally observable Markov decision process (MDP) models from operations research (Puterman 1994) to represent planning problems under uncertainty. However, most of them minimize the expected total plan-execution cost or, synonymously, maximize the expected total reward (MER). This planning objective and similar simplistic planning objectives often do not take the preferences of human decision makers sufficiently into account, for example, their risk attitudes in planning domains with huge wins or losses of money, equipment or human life. This means that they are not well suited for real-world planning domains such as space applications (Zilberstein *et al.* 2002), environmental applications (Blythe 1997), and business applications (Goodwin, Akkiraju, & Wu 2002). In this paper, we provide a first step towards a comprehensive foundation of risk-sensitive planning. In particular, we develop sets of conditions that guarantee the existence and finiteness of the expected utilities when maximizing the expected utility (MEU) of the total reward for risk-sensitive planning with totally observable Markov decision process models and non-linear utility functions.

## Risk Attitudes and Utility Theory

Human decision makers are typically risk-sensitive and thus do not maximize the expected total reward in planning domains with huge wins or losses. Table 1 shows an example

Table 1: An Example of Risk Sensitivity

	Probability	Reward	Expected Reward	Utility	Expected Utility
Choice 1	50%	\$10,000,000	\$5,000,000	-0.050	-0.525
	50%	\$ 0		-1.000	
Choice 2	100%	\$ 4,500,000	\$4,500,000	-0.260	-0.260

for which many human decision makers prefer Choice 2 over Choice 1 even though its expected total reward is lower. They are risk-averse and thus accept a reduction in expected total reward for a reduction in variance. Utility theory (von Neumann & Morgenstern 1944) suggests that this behavior is rational because human decision makers maximize the expected utility of the total reward. Utility functions map total rewards to the corresponding finite utilities and capture the risk attitudes of human decision makers (Pratt 1964). They are strictly monotonically increasing in the total reward. Linear utility functions characterize risk-neutral human decision makers, while non-linear utility functions characterize risk-sensitive human decision makers. In particular, concave utility functions characterize risk-averse human decision makers (“insurance holders”), and convex utility functions characterize risk-seeking human decision makers (“lottery players”). For example, if a risk-averse human decision maker has the concave exponential utility function  $U(w) = -0.9999997^w$  and thus associates the utilities shown in Table 1 with the total rewards of the two choices, then Choice 2 maximizes their expected utility and should thus be chosen by them. On the other hand, MER planners choose Choice 1, and the human decision maker would thus be extremely unhappy with them with 50 percent probability.

## Markov Decision Process Models

We study decision-theoretic planners that use MDPs to represent probabilistic planning problems. Formally, an MDP is a 4-tuple  $(S, A, P, r)$  of a state space  $S$ , an action space  $A$ , a set of transition probabilities  $P$ , and a set of finite (immediate) rewards  $r$ . If an agent executes action  $a \in A$  in state  $s \in S$ , then it incurs reward  $r(s, a, s')$  and transitions to state  $s' \in S$  with probability  $P(s'|s, a)$ . An MDP is called finite if its state space and action space are both finite. We assume throughout this paper that the MDPs are finite since decision-theoretic planners typically use finite MDPs.

The number of time steps that a decision-theoretic planner

plans for is called its (planning) horizon. A history at time step  $t$  is the sequence  $h_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t)$  of states and actions from the initial state to the current state. The set of all histories at time step  $t$  is  $H_t = (S \times A)^t \times S$ . A trajectory is an element of  $H_\infty$  for infinite horizons and  $H_T$  for finite horizons, where  $T \geq 1$  denotes the last time step of the finite horizon.

Decision-theoretic planners determine a decision rule for every time step within the horizon. A decision rule determines which action the agent should execute in its current state. A deterministic history-dependent (HD) decision rule at time step  $t$  is a mapping  $d_t : H_t \rightarrow A$ . A randomized history-dependent (HR) decision rule at time step  $t$  is a mapping  $d_t : H_t \rightarrow \mathcal{P}(A)$ , where  $\mathcal{P}(A)$  is the set of probability distributions over  $A$ . Markovian decision rules are history-dependent decision rules whose actions depend only on the current state rather than the complete history at the current time step. A deterministic Markovian (MD) decision rule at time step  $t$  is a mapping  $d_t : S \rightarrow A$ . A randomized Markovian (MR) decision rule at time step  $t$  is a mapping  $d_t : S \rightarrow \mathcal{P}(A)$ . A policy  $\pi$  is a sequence of decision rules  $d_t$ , one for every time step  $t$  within the horizon. We use  $\Pi^K$  to denote the set of all policies whose decision rules all belong to the same class, where  $K \in \{\text{HR}, \text{HD}, \text{MR}, \text{MD}\}$ . The set of all possible policies  $\Pi$  is the same as  $\Pi^{\text{HR}}$ . Decision-theoretic planners typically determine stationary policies. A Markovian policy  $\pi \in \Pi$  is stationary if  $d_t = d$  for all time steps  $t$ , and we write  $\pi(s) = d(s)$ . We use  $\Pi^{\text{SD}}$  to denote the set of all deterministic stationary (SD) policies and  $\Pi^{\text{SR}}$  to denote the set of all randomized stationary (SR) policies. The state transitions resulting from stationary policies are determined by Markov chains. A state of a Markov chain and thus also a state of an MDP under a stationary policy is called recurrent iff the expected number of time steps between visiting the state is finite. A recurrent class is a maximal set of states that are recurrent and reachable from each other. These concepts play an important role in the proofs of our results.

## Planning Objectives

MEU planners determine policies that maximize the expected utility of the total reward for a given utility function  $U$ . One difference between the MER and MEU objective is that the MEU objective can result in planning problems that are not decomposable, which makes it impossible to use the divide-and-conquer principle (such as dynamic programming) to efficiently find policies that maximize the expected utility. Therefore, we need to re-examine the basic properties of decision-theoretic planning when switching from the MER to the MEU objective.

For planning problems with finite horizons  $T$ , the expected utility of the total reward starting from  $s \in S$  under  $\pi \in \Pi$  is

$$v_{U,T}^\pi(s) = E^{s,\pi} \left[ U \left( \sum_{t=0}^{T-1} r_t \right) \right],$$

where the expectation  $E^{s,\pi}$  is taken over all possible trajectories. The expected utilities exist (= are well-defined) under all policies and are bounded because the number of trajectories is finite for finite MDPs. MEU planners then need to determine the maximal expected utilities of the total reward  $v_{U,T}^*(s) = \sup_{\pi \in \Pi} v_{U,T}^\pi(s)$  and a policy that achieves them. The

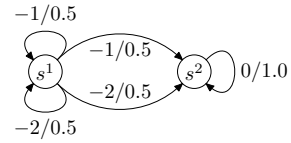


Figure 1: Example 1

maximal expected utilities exist and are finite because the expected utilities are bounded under all policies.

For planning problems with infinite horizons, the expected utility of the total reward starting from  $s \in S$  under policy  $\pi$  is

$$v_U^\pi(s) = \lim_{T \rightarrow \infty} v_{U,T}^\pi(s) = \lim_{T \rightarrow \infty} E^{s,\pi} \left[ U \left( \sum_{t=0}^{T-1} r_t \right) \right]. \quad (1)$$

The expected utilities exist iff the limit converges on the extended real line, that is, the limit results in a finite number, positive infinity or negative infinity. MEU planners then need to determine the maximal expected utilities of the total reward  $v_U^*(s) = \sup_{\pi \in \Pi} v_U^\pi(s)$  and a policy that achieves them. To simplify our terminology, we refer to the expected utilities  $v_U^\pi(s)$  for all  $s \in S$  as the values under policy  $\pi \in \Pi$  and to the maximal expected utilities  $v_U^*(s)$  for all  $s \in S$  as the optimal values. A policy is optimal if its values equal the optimal values for all states. A policy is  $K$ -optimal if it is optimal and it is in the class of policies  $\Pi^K$ , where  $K \in \{\text{HR}, \text{HD}, \text{MR}, \text{MD}, \text{SR}, \text{SD}\}$ .

## Discussion of Assumptions

So far, we have motivated why decision-theoretic planners should maximize the values for non-linear utility functions. The kinds of MDPs that decision-theoretic planners typically use tend to have goal states that need to get reached (Boutilier, Dean, & Hanks 1999). The MDP in Figure 1 gives an example. Its transitions are labeled with their rewards followed by their probabilities. The rewards of the two actions in state  $s^1$  are negative because they correspond to the costs of the actions. State  $s^2$  is the goal state, in which only one action can be executed, namely an action with zero reward whose execution leaves the state unchanged. The optimal value of a state then corresponds to the largest expected utility of the plan-execution cost for reaching the goal state from the given state. To achieve generality, however, we do not make any assumptions in this paper about the structure of the MDPs or their rewards. For example, we do not make any assumptions about how the structure of the MDPs and their rewards encode the goal states. Neither do we make any assumptions about whether all of the rewards are positive, negative or zero. We avoid such assumptions because MDPs can mix positive rewards (which model, for example, rewards for reaching a goal state) and negative rewards (which model, for example, action costs).

The results of this paper would be trivial if we used discounting, that is, assumed that a reward obtained at some time step is worth only a fraction of the same reward obtained one time step earlier. Discounting is a way of modeling interest on investments. In our case, there is typically no way to invest resources and thus no reason to use discounting. This is fortunate because discounting makes it difficult to find optimal policies for the MEU objective even

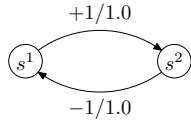


Figure 2: Example 2

though it guarantees that the expected utilities exist and are finite (White 1987). For example, all optimal policies can be non-stationary for the MEU objective if the utility function is exponential and discounting is used (Chung & Sobel 1987), which makes it very difficult to find an optimal policy. On the other hand, there always exists an SD-optimal policy for the MEU objective if the utility function is exponential and discounting is not used (Ávila-Godoy 1999; Cavazos-Cadena & Montes-de-Oca 2000). Thus, there is an advantage to not using discounting for the MEU objective. This advantage does not exist for the MER objective because there always exists an SD-optimal policy for the MER objective whether discounting is used or not (Puterman 1994).

### Existence and Finiteness Conditions

In this paper, we study conditions that guarantee that the values of all states exist and are finite for the MEU objective.

It is important that the optimal values exist since MEU planners determine a policy that achieves the optimal values. There are cases where the optimal values do not exist, as the MDP in Figure 2 illustrates. An agent that starts in state  $s^1$  receives the following sequence of rewards for its only policy:  $+1, -1, +1, -1, \dots$ , and consequently the following sequence of total rewards:  $+1, 0, +1, 0, \dots$ , which oscillates. Thus, the limit in Formula (1) does not exist for any utility function under this policy, and the optimal value of state  $s^1$  does not exist either. A similar argument holds for state  $s^2$  as well.

It is also important that the optimal values be finite. There are cases where the optimal values are not finite. The MDP in Figure 1 illustrates such a case and the problem that it poses for MEU planners. The MDP has two SD policies. Policy  $\pi_1$  assigns the top action to state  $s^1$ , and policy  $\pi_2$  assigns the bottom action to state  $s^1$ . The values of both states are the same under both policies and utility function  $U(w) = -0.5^w$ , namely

$$v_U^{\pi_1}(s^1) = \sum_{t=1}^{\infty} \left[ -0.5^{(-1)^t} \cdot 0.5^t \right] = - \sum_{t=1}^{\infty} 1 = -\infty,$$

$$v_U^{\pi_2}(s^1) = \sum_{t=1}^{\infty} \left[ -0.5^{(-2)^t} \cdot 0.5^t \right] = - \sum_{t=1}^{\infty} 2^t = -\infty,$$

and  $v_U^{\pi_1}(s^2) = v_U^{\pi_2}(s^2) = -1$ . Thus, the optimal value of state  $s^1$  exists but is negative infinity. All trajectories have identical probabilities under both policies, but the total reward and thus also the utility of each trajectory is larger under policy  $\pi_1$  than under policy  $\pi_2$ . Thus, policy  $\pi_1$  should be preferred over policy  $\pi_2$  for all utility functions. Policy  $\pi_2$  of this example shows that a policy that achieves the optimal values and thus is optimal according to our definition is not always the best one. The problem is that the values of the states under policy  $\pi_1$  are guaranteed to dominate the values of the states under policy

$\pi_2$ , but they are guaranteed to strictly dominate the values of the states under policy  $\pi_2$  only if the optimal values are finite. This example also shows that the optimal values are not guaranteed to be finite even if all policies reach a goal state with probability one. Furthermore, the optimal values of both states are, for example, finite for the utility function  $U(w) = w$  and thus any policy that achieves the optimal values is indeed the best one for this utility function, which shows that the problem can exist for some utility functions but not others, depending on whether the optimal values are finite for the given MDP.

## Existing Results

The easiest way to guarantee that the optimal values exist and are finite is to impose conditions that guarantee that the values of all states exist for all policies and are finite. We first review such conditions that have been obtained primarily for MDPs with linear and exponential utility functions. We then use these results to identify similar conditions for more general MDPs.

### Linear Utility Functions

Linear utility functions characterize risk-neutral human decision makers. In this case, the MEU objective is the same as the MER objective. We omit the subscript  $U$  for linear utility functions.

**Positive MDPs** MDPs for which Condition 1 holds are called positive (Puterman 1994). The values exist for positive MDPs under all policies since  $v_T^\pi(s)$  is monotonic in  $T$ . Thus, the optimal values exist as well. They are finite if Condition 2 holds (Puterman 1994). In fact, the optimal values are finite even if  $\Pi$  is replaced with  $\Pi^{\text{SD}}$  in Condition 2, since there exists an SD-optimal policy for the MER objective (Puterman 1994). Note that the values of all recurrent states under a policy are zero if Condition 2 holds.

**Condition 1:** For all  $s, s' \in S$  and all  $a \in A$ ,  $r(s, a, s') \geq 0$ .

**Condition 2:** For all  $\pi \in \Pi$  and all  $s \in S$ ,  $v^\pi(s)$  is finite.

**Negative MDPs** MDPs for which Condition 3 holds are called negative (Puterman 1994). Similar to positive MDPs, the values exist for negative MDPs under all policies and thus the optimal values exist as well. The optimal values are finite if Condition 4 holds (Puterman 1994). Again, the optimal values are finite even if  $\Pi$  is replaced with  $\Pi^{\text{SD}}$  in Condition 4 since there exists an SD-optimal policy for the MER objective (Puterman 1994).

**Condition 3:** For all  $s, s' \in S$  and all  $a \in A$ ,  $r(s, a, s') \leq 0$ .

**Condition 4:** There exists  $\pi \in \Pi$  such that for all  $s \in S$ ,  $v^\pi(s)$  are finite.

**General MDPs** In general, MDPs can have both positive and negative (as well as zero) rewards. We define the positive part of a real number  $r$  to be  $r^+ = \max(r, 0)$  and its negative part to be  $r^- = \min(r, 0)$ . We then obtain the positive part of an MDP by replacing every reward of the MDP with its positive part. We use  $v^{+\pi}(s)$  to denote the values of the positive part of an MDP under policy  $\pi \in \Pi$ . We define the negative part of an MDP and the values  $v^{-\pi}(s)$  in an analogous way. The values exist under all policies if Condition 5 holds (Feinberg 2002). Thus, the optimal values exist as well if Condition 5 holds. They are finite if Condition 5 and Con-

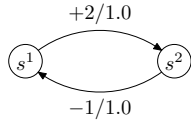


Figure 3: Example 3

dition 6 hold (Feinberg 2002). Again, the optimal values are finite even if  $\Pi$  is replaced with  $\Pi^{\text{SD}}$  in Condition 5 and Condition 6, since there exists an SD-optimal policy for the MER objective (Puterman 1994).

**Condition 5:** For all  $\pi \in \Pi$  and all  $s \in S$ ,  $v^{+\pi}(s)$  is finite.

**Condition 6:** There exists  $\pi \in \Pi$  such that for all  $s \in S$ ,  $v^{-\pi}(s)$  is finite.

Condition 7 is the weakest condition that we use in this paper. It is more general than Condition 1, Condition 3, and Condition 5, since, for example, Condition 1 implies that for all  $\pi \in \Pi$  and all  $s \in S$ ,  $v^{-\pi}(s) = 0$ . The values exist under all policies and  $v^{\pi}(s) = v^{+\pi}(s) + v^{-\pi}(s)$  for all  $s \in S$  and all  $\pi \in \Pi$  if Condition 7 holds (Puterman 1994). Thus, the optimal values exist as well if Condition 7 holds, but they are not guaranteed to be finite (Puterman 1994).

**Condition 7:** For all  $\pi \in \Pi$  and all  $s \in S$ , at least one of  $v^{+\pi}(s)$  and  $v^{-\pi}(s)$  is finite.

The MDPs in Figures 2 and 3 illustrate Condition 7. The MDP in Figure 2 does not satisfy Condition 7. The values of its states do not exist under its only policy  $\pi$ , as we have argued earlier. It is easy to see that  $v^{+\pi}(s^1) = +\infty$  and  $v^{-\pi}(s^1) = -\infty$ , which violates Condition 7 and illustrates that Condition 7 indeed rules out MDPs whose values do not exist under all policies. The MDP in Figure 3 is another MDP that does not satisfy Condition 7. The values of its states, however, exist under its only policy  $\pi'$ . For example, an agent that starts in state  $s^1$  receives the following sequence of rewards:  $+2, -1, +2, -1, \dots$ , and consequently the following sequence of total rewards:  $+2, +1, +3, +2, +4, +3, \dots$ , which converges toward positive infinity. Thus, the limit in Formula (1) exists under  $\pi'$ , and the value of state  $s^1$  thus exists as well under  $\pi'$ . However, it is easy to see that  $v^{+\pi'}(s^1) = +\infty$  and  $v^{-\pi'}(s^1) = -\infty$ , which violates Condition 7 and demonstrates that Condition 7 is not a necessary condition for the values to exist under all policies.

## Exponential Utility Functions

Exponential utility functions are the most widely used non-linear utility functions (Corner & Corner 1995). They are of the form  $U_{\text{exp}}(w) = \iota \gamma^w$ , where  $\iota = \text{sign} \ln \gamma$ . If  $\gamma > 1$ , then the exponential utility function is convex and characterizes risk-seeking human decision makers. If  $0 < \gamma < 1$ , then the utility function is concave and characterizes risk-averse human decision makers. We use the subscript  $\text{exp}$  instead of  $U$  for exponential utility functions and use  $\text{MEU}_{\text{exp}}$  instead of  $\text{MEU}$  to refer to the planning objective.

**Positive MDPs** The values exist for positive MDPs under all policies since  $v_{\text{exp},T}^{\pi}(s)$  is monotonic in  $T$ . Thus, the optimal values exist as well. They are finite if  $0 < \gamma < 1$  or if  $\gamma > 1$  and Condition 8 holds (Cavazos-Cadena & Montes-

de-Oca 2000). Again, the optimal values are finite even if  $\Pi$  is replaced with  $\Pi^{\text{SD}}$  in Condition 8 since there exists an SD-optimal policy for the  $\text{MEU}_{\text{exp}}$  objective (Cavazos-Cadena & Montes-de-Oca 2000).

**Condition 8:** For all  $\pi \in \Pi$  and all  $s \in S$ ,  $v_{\text{exp}}^{\pi}(s)$  is finite.

**Negative MDPs** The values exist for negative MDPs under all policies since  $v_{\text{exp},T}^{\pi}(s)$  is monotonic in  $T$ . Thus, the optimal values exist as well. They are finite if  $\gamma > 1$  or if  $0 < \gamma < 1$  and Condition 9 holds (Ávila-Godoy 1999). Again, the optimal values are finite even if  $\Pi$  is replaced with  $\Pi^{\text{SD}}$  in Condition 9 since there exists an SD-optimal policy for the  $\text{MEU}_{\text{exp}}$  objective (Ávila-Godoy 1999).

**Condition 9:** There exists  $\pi \in \Pi$  such that for all  $s \in S$ ,  $v_{\text{exp}}^{\pi}(s)$  is finite.

## Some Useful Lemmata

The following lemmata are key to proving Theorem 4, Theorem 6, Theorem 9, and Theorem 10. Because of the space limit, we state these lemmata and the following theorems without proof. Lemma 1 describes the behavior of the agent after entering a recurrent class, and Lemma 2 describes its behavior before entering a recurrent class. The idea of splitting trajectories according to whether the agent is in a recurrent class is key to the proofs of the results in following sections.

Lemma 1 states that one can only receive all non-negative rewards, all non-positive rewards or all zero rewards if one enters a recurrent class and Condition 7 holds.

**Lemma 1.** Assume that Condition 7 holds. Consider an arbitrary  $\pi \in \Pi^{\text{SR}}$  and an arbitrary  $s \in S$  that is recurrent under  $\pi$ . Let  $A_{\pi}(s) \subseteq A$  denote the set of actions whose probability is positive under the probability distribution  $\pi(s)$ .

- If  $v^{\pi}(s) = +\infty$ , then for all  $a \in A_{\pi}(s)$  and all  $s' \in S$  with  $P(s'|s, a) > 0$ ,  $r(s, a, s') \geq 0$ .
- If  $v^{\pi}(s) = -\infty$ , then for all  $a \in A_{\pi}(s)$  and all  $s' \in S$  with  $P(s'|s, a) > 0$ ,  $r(s, a, s') \leq 0$ .
- If  $v^{\pi}(s) = 0$ , then for all  $a \in A_{\pi}(s)$  and all  $s' \in S$  with  $P(s'|s, a) > 0$ ,  $r(s, a, s') = 0$ .

Lemma 2 concerns the well-known geometric rate of state evolution (Kemeny & Snell 1960) and its corollaries. We use this lemma, together with the fact that the rewards accumulate at a linear rate, to show that the limit in (1) converges on the extended real line under various conditions.

**Lemma 2.** For all  $\pi \in \Pi^{\text{SR}}$ , let  $R^{\pi}$  denote the set of recurrent states under  $\pi$ . Then for all  $s \in S$ , there exists  $0 < \rho < 1$  such that for all  $t \geq 0$ ,

- there exists  $a > 0$  such that  $P^{s,\pi}(s_t \notin R^{\pi}) \leq a\rho^t$ ,
- there exists  $b > 0$  such that  $P^{s,\pi}(s_t \notin R^{\pi}, s_{t+1} \in R^{\pi}) \leq b\rho^t$ , and
- for any recurrent class  $R_i^{\pi}$ , there exists  $c > 0$  such that  $P^{s,\pi}(s_t \notin R_i^{\pi}, s_{t+1} \in R_i^{\pi}) \leq c\rho^t$ ,

where  $P^{s,\pi}$  is a shorthand for the probability under  $\pi$  conditional on  $s_0 = s$ .

The MDP in Figure 4 illustrates Lemma 2. State  $s^2$  is the only recurrent state under its only policy if  $p > 0$ . For this policy  $\pi$  and all  $t \geq 0$ ,  $P^{s^1,\pi}(s_t \neq s^2) = (1-p)^t$  (which illustrates Lemma 2a) and  $P^{s^1,\pi}(s_t \neq s^2, s_{t+1} = s^2) = p(1-p)^t$  (which illustrates Lemma 2b and Lemma 2c).

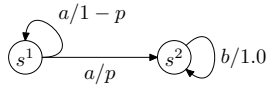


Figure 4: Example 4

## Exponential Utility Functions and General MDPs

We discuss convex and concave exponential utility functions separately for MDPs with both positive and negative rewards. We use  $v_{\text{exp}}^{+\pi}(s)$  to denote the values of the positive part of an MDP with exponential utility functions under policy  $\pi \in \Pi$  and  $v_{\text{exp}}^{+\ast}(s)$  to denote its optimal values. We define the values  $v_{\text{exp}}^{-\pi}(s)$  and  $v_{\text{exp}}^{-\ast}(s)$  in an analogous way.

### Convex Exponential Utility Functions

Theorem 4 shows that the values exist for convex exponential utility functions under all SR policies if Condition 10 holds. Condition 10 is analogous to Condition 7, and Lemma 3 relates them.

**Condition 10:** For all  $\pi \in \Pi$  and all  $s \in S$ , at least one of  $v_{\text{exp}}^{+\pi}(s)$  and  $v^{-\pi}(s)$  is finite.

**Lemma 3.** If  $\gamma > 1$ , Condition 10 implies Condition 7.

**Theorem 4.** Assume that Condition 10 holds and  $\gamma > 1$ . For all  $\pi \in \Pi^{\text{SR}}$  and all  $s \in S$ ,  $v_{\text{exp}}^{\pi}(s)$  exists.

However, it is still an open problem whether there exists an SD-optimal policy for MDPs with both positive and negative rewards for the  $\text{MEU}_{\text{exp}}$  objective. Therefore, it is currently unknown whether the optimal values exist.

Assume that Condition 10 holds and the optimal values exist. The optimal values are finite if for all  $\pi \in \Pi$  and all  $s \in S$ ,  $v_{\text{exp}}^{+\pi}(s)$  is finite. This is so because Condition 8 implies that for all  $s \in S$ ,  $v_{\text{exp}}^{+\ast}(s)$  is finite. Furthermore, for all  $\pi \in \Pi$  and all  $s \in S$ ,

$$v_{\text{exp},T}^{\pi}(s) = E^{s,\pi} \left[ U_{\text{exp}} \left( \sum_{t=0}^{T-1} r_t \right) \right] \leq E^{s,\pi} \left[ U_{\text{exp}} \left( \sum_{t=1}^{T-1} r_t^+ \right) \right] = v_{\text{exp},T}^{+\pi}(s).$$

Taking the limit as  $T$  approaches infinity shows that  $v_{\text{exp}}^{\pi}(s) \leq v_{\text{exp}}^{+\pi}(s) < +\infty$ . Therefore,  $v_{\text{exp}}^{\ast}(s) \leq v_{\text{exp}}^{+\ast}(s) < +\infty$ .

The MDP in Figure 4 illustrates Theorem 4. Its probabilities and rewards are parameterized, where  $p > 0$ . We will show that the values of both states under its only policy  $\pi$  exist for all parameter values if Condition 10 holds and  $\gamma > 1$ . Assume that the premise is true. We distinguish two cases: either  $v^{-\pi}(s^1)$  is finite or  $v^{-\pi}(s^1)$  is negative infinity.

If  $v^{-\pi}(s^1)$  is finite, then  $v^{-\pi}(s^2)$  is finite as well. If  $v^{-\pi}(s^2)$  is finite, then it is zero since state  $s^2$  is recurrent. If  $v^{-\pi}(s^2)$  is zero, then  $b \geq 0$ , that is, either  $b > 0$  or  $b = 0$ . If  $b > 0$ , then  $v_{\text{exp}}^{\pi}(s^1) = v_{\text{exp}}^{\pi}(s^2) = +\infty$ , that is, the values of both states exist. If  $b = 0$  (Case X), then  $v_{\text{exp}}^{\pi}(s^2) = 1$  and

$$v_{\text{exp}}^{\pi}(s^1) = \sum_{t=0}^{\infty} \gamma^{(t+1)a} p(1-p)^t.$$

If  $\gamma^a(1-p) < 1$ , then the above sum can be simplified to

$$v_{\text{exp}}^{\pi}(s^1) = \frac{p\gamma^a}{1 - \gamma^a(1-p)},$$

otherwise it is positive infinity. Thus, in either case, the values of both states exist.

If  $v^{-\pi}(s^1)$  is negative infinity, then  $v^{-\pi}(s^2)$  is negative infinity as well. If  $v^{-\pi}(s^2)$  is negative infinity, then  $b < 0$  and  $v_{\text{exp}}^{\pi}(s^2) = 0$ . We distinguish two cases: either  $a \leq 0$  or  $a > 0$ . If  $a \leq 0$ , then the total rewards of all trajectories are negative infinity and thus  $v_{\text{exp}}^{\pi}(s^1) = 0$ , that is, the values of both states exist. If  $a > 0$ , then  $\gamma^a(1-p) < 1$  because Condition 10 requires that  $v_{\text{exp}}^{+\pi}(s^1)$  be finite if  $v^{-\pi}(s^1)$  is negative infinity and the positive part of the MDP satisfies the conditions of Case X above. If the agent enters state  $s^2$  at time step  $t+1$ , then it receives reward  $b$  from then on. Thus,

$$\begin{aligned} v_{\text{exp},T}^{\pi}(s^1) &= \sum_{t=0}^{T-1} \gamma^{(t+1)a+(T-t-1)b} p(1-p)^t + \gamma^{Ta} (1-p)^T \\ &= p\gamma^{a-b} \frac{\gamma^{bT} - [\gamma^a(1-p)]^T}{1 - \gamma^{a-b}(1-p)} + [\gamma^a(1-p)]^T. \end{aligned}$$

Since  $\gamma > 1$  and  $\gamma^a(1-p) < 1$ , it holds that

$$v_{\text{exp}}^{\pi}(s^1) = \lim_{T \rightarrow \infty} v_{\text{exp},T}^{\pi}(s^1) = 0.$$

Thus, in all cases, Condition 10 indeed guarantees that the values of both states exist for the MDP in Figure 4.

### Concave Exponential Utility Functions

The results and proofs for concave exponential utility functions are analogous to the ones for convex exponential utility functions.

**Condition 11:** For all  $\pi \in \Pi$  and all  $s \in S$ , at least one of  $v^{+\pi}(s)$  and  $v_{\text{exp}}^{-\pi}(s)$  is finite.

**Lemma 5.** If  $0 < \gamma < 1$ , Condition 11 implies Condition 7.

**Theorem 6.** Assume that Condition 11 holds and  $0 < \gamma < 1$ . For all  $\pi \in \Pi^{\text{SR}}$  and all  $s \in S$ ,  $v_{\text{exp}}^{\pi}(s)$  exists.

Assume that Condition 11 holds and the optimal values exist. The optimal values are finite if there exists  $\pi \in \Pi$  such that for all  $s \in S$ ,  $v_{\text{exp}}^{-\pi}(s)$  is finite. This is so because for this  $\pi$  and all  $s \in S$ ,

$$v_{\text{exp},T}^{\pi}(s) = E^{s,\pi} \left[ U_{\text{exp}} \left( \sum_{t=0}^{T-1} r_t \right) \right] \geq E^{s,\pi} \left[ U_{\text{exp}} \left( \sum_{t=0}^{T-1} r_t^- \right) \right] = v_{\text{exp},T}^{-\pi}(s).$$

Taking the limit as  $T$  approaches infinity shows that  $v_{\text{exp}}^{\ast}(s) \geq v_{\text{exp}}^{\pi}(s) \geq v_{\text{exp}}^{-\pi}(s) > -\infty$ .

## General Utility Functions

We now consider non-linear utility functions that are more general than exponential utility functions. Such utility functions are necessary to model risk attitudes that change with the total reward.

### Positive and Negative MDPs

The values exist for positive and negative MDPs under all policies since  $v_{U,T}^{\pi}(s)$  is monotonic in  $T$ . Thus, the optimal values exist as well. Theorem 7 gives a condition under which the optimal values are finite for positive MDPs, and Theorem 8 gives a condition under which they are finite for negative MDPs.

**Theorem 7.** Assume that Condition 1 and Condition 8 hold for some  $\gamma > 1$ . If the utility function  $U$  satisfies  $U(w) = O(\gamma^w)$  as  $w \rightarrow +\infty$ , then for all  $\pi \in \Pi$  and all  $s \in S$ ,  $v_U^{\pi}(s)$  and  $v_U^{\ast}(s)$  are finite.

**Theorem 8.** Assume that Condition 3 and Condition 9 hold for some  $\pi \in \Pi$  and some  $\gamma$  with  $0 < \gamma < 1$ . If the utility function  $U$  satisfies  $U(w) = O(\gamma^w)$  as  $w \rightarrow -\infty$ , then for this  $\pi$  and all  $s \in S$ ,  $v_U^\pi(s)$  and  $v_U^*(s)$  are finite.

## General MDPs

The following sections suggest conditions that constrain MDPs with both positive and negative rewards and the utility functions to ensure that the values exist. The first part gives conditions that constrain the MDPs but not the utility functions. The second part gives conditions that constrain the utility functions but require the MDPs to satisfy only Condition 7. The last part, finally, gives conditions that mediate between the two extremes.

**Bounded Total Rewards** We first consider the case where the total reward is bounded. We use  $H^{s,\pi}$  to denote the set of trajectories starting from  $s \in S$  under policy  $\pi \in \Pi$ . We define  $w(h) = \sum_{t=0}^{\infty} r_t(h)$  for  $h \in H^{s,\pi}$ ,  $v_{\max}^\pi(s) = \max_{h \in H^{s,\pi}} w(h)$  and  $v_{\min}^\pi(s) = \min_{h \in H^{s,\pi}} w(h)$ . The total reward  $w(h)$  exists if Condition 7 holds. The optimal values exist and are finite if Condition 7 holds and for all  $s \in S$ ,

$$\sup_{\pi \in \Pi} v_{\max}^\pi(s) < +\infty \quad \text{and} \quad \inf_{\pi \in \Pi} v_{\min}^\pi(s) > -\infty.$$

These conditions are, for example, satisfied for acyclic MDPs if plan execution ends in absorbing states but are satisfied for some cyclic MDPs as well. Unfortunately, it can be difficult to check the conditions directly. However, the optimal values also exist and are finite if for all  $s \in S$ ,

$$\sup_{\pi \in \Pi} v_{\max}^{+\pi}(s) < +\infty \quad \text{and} \quad \inf_{\pi \in \Pi} v_{\min}^{-\pi}(s) > -\infty.$$

In fact, the optimal values also exist even if  $\Pi$  is replaced with  $\Pi^{\text{SD}}$  in this condition, which allows one to check the condition with a dynamic programming procedure.

**Bounded Utility Functions** We now consider the case where the utility functions are bounded. In this case,  $v_{U,T}^\pi(s)$  is bounded as  $T$  approaches infinity but the limit might not exist since the values can oscillate. Theorem 9 provides a condition that guarantees that the values exist under stationary policies.

**Theorem 9.** Assume that Condition 7 holds. If  $\lim_{w \rightarrow -\infty} U(w) = U^-$  and  $\lim_{w \rightarrow +\infty} U(w) = U^+$  with  $U^+ \neq U^-$  being finite, then for all  $\pi \in \Pi^{\text{SR}}$  and all  $s \in S$ ,  $v_U^\pi(s)$  exists and is finite.

**Linearly Bounded Utility Functions** Finally, we consider the case where the utility functions are bounded by linear functions. Theorem 10 shows that the values exist under conditions that are, in part, similar to those for the MER objective.

**Theorem 10.** Assume Condition 7 holds. If  $U(w) = O(w)$  as  $w \rightarrow \pm\infty$ , then for all  $\pi \in \Pi^{\text{SR}}$  and all  $s \in S$ ,  $v_U^\pi(s)$  exists.

## Conclusions and Future Work

We have discussed conditions that guarantee the existence and finiteness of the expected utilities of the total plan-execution reward for risk-sensitive planning with totally observable Markov decision process models. Our results are

only a first step towards a comprehensive foundation of risk-sensitive planning. In future work, we will study the existence of optimal and  $\epsilon$ -optimal policies, the structure of such policies, and basic computational procedures for obtaining them.

## Acknowledgments

This research was partly supported by NSF awards to Sven Koenig under contracts IIS-9984827 and IIS-0098807 and an IBM fellowship to Yaxin Liu. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsoring organizations, agencies, companies or the U.S. government.

## References

- Ávila-Godoy, M. G. 1999. *Controlled Markov Chains with Exponential Risk-Sensitive Criteria: Modularity, Structured Policies and Applications*. Ph.D. Dissertation, Department of Mathematics, University of Arizona.
- Blythe, J. 1997. *Planning under Uncertainty in Dynamic Domains*. Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University.
- Boutilier, C.; Dean, T.; and Hanks, S. 1999. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research* 11:1–94.
- Cavazos-Cadena, R., and Montes-de-Oca, R. 2000. Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces. *Mathematics Methods of Operations Research* 52:133–167.
- Chung, K.-J., and Sobel, M. J. 1987. Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM Journal of Control and Optimization* 35(1):49–62.
- Corner, J. L., and Corner, P. D. 1995. Characteristics of decisions in decision analysis practice. *The Journal of Operational Research Society* 46:304–314.
- Feinberg, E. A. 2002. Total reward criteria. In Feinberg, E. A., and Shwartz, A., eds., *Handbook of Markov Decision Processes: Methods and Applications*. Kluwer. chapter 6, 173–208.
- Goodwin, R. T.; Akkiraju, R.; and Wu, F. 2002. A decision-support system for quote-generation. In *Proceedings of the Fourteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-02)*, 830–837.
- Kemeny, J. G., and Snell, J. L. 1960. *Finite Markov Chains*. D. Van Nostrand Company.
- Pratt, J. W. 1964. Risk aversion in the small and in the large. *Econometrica* 32(1-2):122–136.
- Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.
- von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.
- White, D. J. 1987. Utility, probabilistic constraints, mean and variance of discounted rewards in Markov decision processes. *OR Spektrum* 9:13–22.
- Zilberstein, S.; Washington, R.; Bernstein, D. S.; and Mouaddib, A.-I. 2002. Decision-theoretic control of planetary rovers. In Beetz, M.; Hertzberg, J.; Ghallab, M.; and Pollack, M. E., eds., *Advances in Plan-Based Control of Robotic Agents*, volume 2466 of *Lecture Notes in Computer Science*. Springer. 270–289.