GUBS: a Utility-Based Semantic for Goal-Directed Markov Decision Processes

Valdinei Freire
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
Sao Paulo, Brazil
valdinei.freire@usp.br

Karina Valdivia Delgado*
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
Sao Paulo, Brazil
kvd@usp.br

ABSTRACT

A key question in stochastic planning is how to evaluate policies when a goal cannot be reached with probability one. Usually, in the Goal-Directed Markov Decision Process (GD-MDP) formalization, the outcome of a policy is summarized on two criteria: probability of reaching a goal state and expected cost to reach a goal state. The dual criterion solution considers a lexicography preference, by prioritizing probability of reaching the goal state and then minimizing the expected cost to goal. Some other solutions, consider only cost by using some math trick to guaranteed that every policy has a finite expected cost. In this paper we show that the lexicography solution does not allow a smooth trade-off between goal and cost, while the expected cost solution does not define a goal-semantic. We propose GUBS (Goals with Utility-Based Semantic), a new model to evaluate policies based on the expected utility theory; this model defines a trade-off between cost and goal by proposing an axiomatization for goal semantics in GD-MDPs. We show that our model can be solved by any continuous state MDP solver and propose an algorithm to solve a special class of GD-MDPs.

Keywords

Agent theories and models; Single agent planning; Markov Decision Process; Preferential Semantics; Utility Theory.

1. INTRODUCTION

Some research on probabilistic planning has focused on finding policies that maximize the probability of reaching a goal [10, 14, 4] or minimize the average accumulated costs if there exists a proper policy [1, 3, 2]. Other approaches optimize both criteria, maximizing the probability of reaching a goal and minimizing the average accumulated costs at the same time [13, 11]; these works consider a dual optimization criterion, which finds the cheapest policy among the policies that maximize goal probability. However, these dual criterion does not really establish a compromise between them; actually, the dual criterion prioritizes the probability of success and the performance is treated as secondary.

A key question is: should we prefer policies that maximize the probability to get the goal even at the expense of an increased expected cost, or those that minimize the expected cost even at the expense of a small increase in risk of failure? The river problem illustrate such a question.

Illustrative Example: The river problem. In the river problem, a person is on one side of a river and want to get to the other side. The person has two options. The first one is to walk y ($0 \leq y \leq 9,999$) meters in direction to the north and then swim across. And the second one is to walk 10,000 meters in direction to the north, where there is a bridge and then traverse it with probability 1 to reach the other side of the river. The action walk y is deterministic and the river is modeled as a grid $n \times m$, where n is the number of rows and m, the number of columns. In column 1 there is a dangerous waterfall, where the person has 100% chance of drowning. In other positions of the river $\langle x, y\rangle$, if the person decides to swim, there is 80% chance of success to get the position $\langle x+1, y\rangle$ and 20% chance of the person ends up in the position $\langle x, y-1\rangle$ because the flow of the river. The cost of the action walk is 1 unit. Note that, in this problem the more a person walks, the risk of falling into a waterfall is lesser.

If we consider the dual criterion, the optimal solution is to walk 10000 meters with a cost of 10000 units and then traverse the bridge because this policy guarantees the person to reach the other side of the river with a 100% of probability of success. This solution does not consider the cost at all since it prefers policies that maximize the probability to get the goal even at the expense of an increased expected cost.

This example motivates the need for a new model. In this paper, we argue that small increases in risk of failure should be accepted if a large decrease in expected cost can be obtained. Thus, to answer our previous question we need to define a trade-off between the expected cost and the probability to reach the goal. To model this compromise, this paper provides a preferential semantics for cost and goal states based on the expected utility theory. Additionally, we show how to solve GD-MDPs under these semantics.

2. FORMALIZATION

Consider a Goal-Directed MDP [1, 6] (GD-MDP) described by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, c, \mathcal{G}\rangle$ where:

- $\mathcal{S}$ is a set of states;

*The authors would like to thank the Sao Paulo Research Foundation (FAPESP) for the financial support (grant #2015/01587-0).

Appears in: *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

- $\mathcal{A}$ is a set of actions that can be performed at each period of decision $t \in \{0, 1, 2, \ldots\}$;
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a transition function that represents the probability of the system transits to a state $s' \in \mathcal{S}$ after the agent executes an action $a \in \mathcal{A}$ in a state $s \in \mathcal{S}$, i.e., $P(s_{t+1} = s' | s_t = s, a_t = a) = T(s, a, s')$;
- $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ is a cost function that represents the cost of taking an action $a \in \mathcal{A}$ when the process is in a state $s \in \mathcal{S}$, i.e., $c_t = c(s_t, a_t)$; and
- $\mathcal{G}$ is a set of absorbing (states with no outgoing transitions) called goal states.

The GD-MDP problem defines a discrete dynamic process. At any time t, the agent observes a state s_t, executes an action a_t, transits to a state s_{t+1} following T and pays a cost c_t. The process ends after reaching any goal state in $\mathcal{G}$.

Every finite history $h = s_0, a_0, c_0, s_1, a_1, c_1, \ldots, s_T$ can be resumed to a vector $(C_T, \beta_T) \in \mathbb{R} \times \{g, \neg g\}$, where g means a goal state was reached and $\neg g$ means no goal state was reached (dead-ends[1] included). If $s_T \in \mathcal{G}$ then $C_T = \sum_{t=0}^{T-1} c(s_t, a_t)$ and $\beta_T = g$, otherwise $C_T = \sum_{t=0}^{T-1} c(s_t, a_t)$ and $\beta_T = \neg g$

A decision-maker that prioritizes goals can be easily formalized for histories as follows.

DEFINITION 1. *A decision-maker prioritizes goals if the decision-maker presents the following preference:*

$$(C_T, g) \succ (C'_T, \neg g), \forall C_T, C'_T \in \mathbb{R}^+ \cup \infty, \quad (1)$$

$$(C_T, g) \succ (C'_T, g), \text{ if } C_T < C'_T \ \forall C_T, C'_T \in \mathbb{R}^+, \quad (2)$$

where $A \succ B$ means that the decision-maker prefers A to B, i.e., the decision-maker prefers a history that reaches the goal than other states, independently of the total cost; and if both histories reaches the goal, the decision-maker prefers the history with lower cost.

The Definition 1 defines preference among histories, but says nothing about how to define preference among policies. A non-stationary policy π maps partial histories h into an action a, i.e., $\pi : \mathcal{H} \to \mathcal{A}$, where $\mathcal{H}$ is all finite sequences of pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ and a current state s_t. Any policy π induces a probability distribution of complete histories; in the jargon of expected utility theory, a policy is a lottery. In this paper, we discuss how to setup preference regarding lotteries, i.e., distribution over histories.

3. GOAL-SEMANTIC IN THE LITERATURE

Before describing the main paper's contributions in more detail, we first review the goal semantic used in prior work.

In the GD-MDP setup, one way to evaluate the performance is computing the probability of reaching a goal [10, 14, 4]. GD-MDPs whose objective is to find policies that maximize the probability of reaching a goal are called MAXPROB [10].

Another way to evaluate the performance of GD-MDPs is the expected accumulated cost to reach the goal. However, if a proper policy (a policy that reaches the goal from any state with probability 1) does not exist, the expected accumulated cost to reach the goal is not well-defined [11].

Since the existence of a proper policy disallows the existence of dead ends, [11] introduces three new classes of GD-MDPs with different assumptions about the existence of these type of states: SSPADE, fSSPUDE and iSSPUDE. SSPADE has dead ends but they are avoidable if the agent acts optimally from the initial state. fSSPUPE has unavoidable dead ends but assumes that the agent can put a finite price (*penalty*) on running into a dead end. A *dual optimization criterion* is used to solve iSSPUDE, problems where dead ends are unavoidable and the cost of entering one is infinite. The dual optimization criterion [13, 11] finds the cheapest policy among the policies that maximize goal probability.

Finally, a common method to conciliate indefinite horizon and infinite horizon is the use of a discount factor. [15] proposes the use of *discounted cost criterion* to deal with problems with dead ends. However, [13] shows that this approach may not be appropriate in problems with complex cost structures.

Next we analyze the goal semantic used by [10, 11, 15]. To improve clarity, the proofs of theorems of this section are in the appendix.

3.1 MAXPROB Criterion

Let P_g^π be the goal probability function of a policy π, i.e. the probability of reaching the goal under policy π. The MAXPROB criterion evaluation considers that a policy π is better than a policy π' if and only if the probability of reaching a goal state is greater in the first one, i.e.,

$$P_g^\pi > P_g^{\pi'} \iff \pi \succ \pi'.$$

3.2 Dual Optimization Criterion

Let $\overline{C}_g^\pi$ be the expected cost to goal, i.e., the expected cost of histories that reach the goal state when following policy π. The dual criterion evaluation [13, 11] considers the following lexicography measure regarding lotteries:

(i) if $P_g^\pi > P_g^{\pi'}$, then $\pi \succ \pi'$; or

(ii) if $P_g^\pi = P_g^{\pi'}$ and $\overline{C}_g^\pi < \overline{C}_g^{\pi'}$, then $\pi \succ \pi'$.

Although this lexicography measure makes a GD-MDP a well-defined problem and guarantees condition in Definition 1, it may not represent properly the preference of decision-makers. For instance, the dual criterion measure may allow a huge increase in expected cost even if there is a small increase in the probability to reach the goal. This is formalized in the following theorem.

THEOREM 1. *There exists a GD-MDP $\mathcal{M}$ with policies π and π' and L arbitrarily large and $\delta > 0$ arbitrarily small such that:*

1. $P_g^\pi > P_g^{\pi'}$ *and* $P_g^\pi - P_g^{\pi'} < \delta$; *and*
2. $\overline{C}_g^\pi - \overline{C}_g^{\pi'} > L$.

3.3 Penalty to quit

Consider a GD-MDP augmented with an action q such that:

[1]A dead-end is a state from which reaching the goal is impossible.

1. $T(s,q,s') = \begin{cases} 1 & , \text{ if } s' = s_g \\ 0 & , \text{ if } s' \neq s_g \end{cases}$ where $s_g \in \mathcal{G}$; and

2. $c(s,q) = D$ for all $s \in \mathcal{S}$.

Action q can be seeing as a quit action, i.e., the decision-maker is willing to pay penalty D of giving up the chance to get to the goal. Then, an optimal policy is the one that reaches goal with probability 1, even if the policy makes use of the quit action q, and minimizes the expected cost $\overline{C}_D^{\pi} = \mathrm{E}\left[\sum_{t=0}^{\infty} c(s_t, a_t)\middle| \pi, s_0\right]$. Thus, the action q can be used to avoid dead ends by paying a finite cost.

This parameterized measure was proposed by [11]. A positive property of the penalty D is that there exists a large enough D such that the MAXPROB measure can be obtained. Additionally, if there exists proper policies, then the dual criterion measure can be obtained.

THEOREM 2. *Consider a GD-MDP with policies π and π' such that: (*i*) $P_g^{\pi} > P_g^{\pi'}$, or (*ii*) $P_g^{\pi} = P_g^{\pi'} = 1$ and $\overline{C}_g^{\pi} < \overline{C}_g^{\pi'}$; then, there exists a penalty D_0 such that $\overline{C}_D^{\pi} < \overline{C}_D^{\pi'}$ for any $D > D_0$.*

A negative property of the penalty is that given an arbitrary D, there exist and GD-MDP such that the MAXPROB measure can not be obtained.

THEOREM 3. *Consider an arbitrary penalty D, then there exists a GD-MDP with policies π and π' such that:*

1. $P_g^{\pi} > P_g^{\pi'}$; *and*

2. $\overline{C}_D^{\pi} > \overline{C}_D^{\pi'}$.

3.4 Discounted Cost Criterion

Given a discount factor $\gamma \in (0,1)$, a policy π can be simply evaluated by the expected accumulated discounted cost sum $\overline{C}_{\gamma}^{\pi} = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)|\pi, s_0\right]$. The discount factor γ guarantees that even if a history is infinite, the discounted cost sum is finite. As $\gamma \to 1$ the MAXPROB criterion can be obtained, additionally, if there exists proper policies, then the dual criterion measure can be obtained. Just like the penalty formulation, this criterion has similar positive and negative properties. This is formalized in the following theorems.

THEOREM 4. *Consider a GD-MDP with policies π and π' such that: (*i*) $P_g^{\pi} > P_g^{\pi'}$, or (*ii*) $P_g^{\pi} = P_g^{\pi'} = 1$ and $\overline{C}_g^{\pi} < \overline{C}_g^{\pi'}$; then, if, when reaching a goal, a reward R_g is considered, there exists a discount factor γ_0 such that $\overline{C}_{\gamma}^{\pi} < \overline{C}_{\gamma}^{\pi'}$ for any $\gamma > \gamma_0$.*

THEOREM 5. *Consider an arbitrary discount factor γ, then there exists a GD-MDP with policies π and π' such that:*

1. $P_g^{\pi} > P_g^{\pi'}$; *and*

2. $\overline{C}_{\gamma}^{\pi} > \overline{C}_{\gamma}^{\pi'}$.

4. GOAL-SEMANTIC BASED ON EXPECTED UTILITY THEORY

Theorem 1 shows that the dual optimization criterion, despite making a GD-MDP a well-defined problem, represents preference that may not represent a real decision-maker. Penalty and Discount measures model preference that make a trade-off between expected cost to goal and probability to goal and in the limit ($D \to \infty$ or $\gamma \to 1$) the MAXPROB measure can be obtained. However, D and γ cannot be set generically neither to obtain MAXPROB (Theorems 3 and 5), nor to prioritize goals (Definition 1). Next we define a new measure to evaluate policies based on Definition 1.

DEFINITION 2. *Consider the following utility function:*

$$U(C_T, \beta_T) = u(C_T) + K_g \mathbb{1}_{\beta_T = g}, \tag{3}$$

where $u(\cdot)$ is a utility function over cost, K_g is a constant utility for reaching the goal, and $\mathbb{1}_{cond}$ is the function that returns 1 when condition cond is true, and 0 otherwise. We call such a model GUBS*: Goal with Utility-Based Semantic, and a decision-maker follows* GUBS *model if a policy π is evaluated by:*

$$V^{\pi} = E[U(C_T, \beta_T)|\pi, s_0].$$

In the next sections we show theoretical results and how to solve GUBS model.

4.1 Theoretical Results

If function $u(C_T)$ and constant K_g are chosen appropriately, the conditions in Definition 1 can be guaranteed.

THEOREM 6. *The model* GUBS *guarantees the decision-maker prioritizes goals (Definition 1) if the following conditions are observed:*

1. $u : \mathbb{R} \to [U_{min}, U_{max}]$;

2. $u(C)$ *is strictly decreasing in C; and*

3. $K_g > U_{max} - U_{min}$.

PROOF. The condition $(C_T, g) \succ (C'_T, \neg g)$ is the same as $U(C_T, g) > U(C'_T, \neg g)$. Then, under the conditions in the theorem, we have:

$$\begin{aligned} U(C_T, g) &= u(C_T) + K_g \\ &> U_{min} + U_{max} - U_{min} \\ &= U_{max} \\ &\geq U(C'_T, \neg g). \end{aligned}$$

The condition $(C_T, g) \succ (C'_T, g)$ when $C_T < C'_T$ is the same as $U(C_T, g) > U(C'_T, g)$. Then, under the conditions in the theorem, we have $u(C_T) > u(C'_T)$, then:

$$U(C_T, g) = u(C_T) + K_g > u(C'_T) + K_g = U(C'_T, g).$$

□

Note that Theorem 6 gives a sufficient condition to guarantee the preference in Definition 1. Differently from penalty model and discount model, where the choice of discount factor γ or penalty D depends on the GD-MDP; K_g, U_{max} and U_{min} do not depend on the GD-MDP problem. Besides that, the following theorem also presents a result regarding policies.

THEOREM 7. *Consider a GD-MDP and two policies π and π' such that $P_g^{\pi} > P_g^{\pi'}$, then, under the* GUBS *model, $\pi' \succ \pi$ only if:*

$$\frac{P_g^{\pi'}}{P_g^{\pi}} > \frac{K_g}{U_{max} - U_{min} + K_g}.$$

PROOF. We have:

$$\begin{aligned} V^{\pi} &= \mathrm{E}[u(C_T) + K_g \mathbb{1}_{\beta_T = g} | \pi] \\ &= \mathrm{E}[u(C_T)|\pi] + K_g P_g^{\pi} \\ &\geq U_{min} + K_g P_g^{\pi}, \end{aligned}$$

and

$$\begin{aligned} V^{\pi'} &= \mathrm{E}[u(C_T) + K_g \mathbb{1}_{\beta_T = g} | \pi'] \\ &= \mathrm{E}[u(C_T)|\pi'] + K_g P_g^{\pi'} \\ &\leq P_g^{\pi'} U_{max} + (1 - P_g^{\pi'}) U_{min} + K_g P_g^{\pi'}. \end{aligned}$$

where the last equation was obtained by considering the best scenario constrained by P_g: when goal is reached the best utility ($K_g + U_{max}$) is obtained and if goal is not reached, an infinite cost must be payed, which gives the worst utility (U_{min}).

Since the condition $\pi' \succ \pi$ is the same as $V^{\pi'} > V^{\pi}$, we have:

$$\begin{aligned} P_g^{\pi'} U_{max} + (1 - P_g^{\pi'}) U_{min} + K_g P_g^{\pi'} &> U_{min} + K_g P_g^{\pi} \\ P_g^{\pi'} (U_{max} - U_{min} + K_g) &> K_g P_g^{\pi} \\ \frac{P_g^{\pi'}}{P_g^{\pi}} &> \frac{K_g}{U_{max} - U_{min} + K_g}. \end{aligned}$$

□

Theorem 7 guarantees that MAXPROB condition can be obtained by choosing K_g appropriately, but no K_g can obtain MAXPROP for any GD-MDP. However, differently from discount and penalty model, in the GUBS model it is possible: (*i*) to guarantee goal-prioritized preference (Definition 1), and (*ii*) to guarantee an arbitrary approximation to MAXPROB.

For example, consider an GD-MDP with two policies π and π' where $P_g^{\pi} = 0.9$ and $P_g^{\pi'} = 0.89$; $\overline{C}_g^{\pi} = 1000$ and $\overline{C}_g^{\pi'} = 1$. Then, in the dual criterion $\pi \succ \pi'$. In the GUBS model it is possible to set K_g that guarantees $\pi \succ \pi'$ independently of the GD-MDP structure, e.g., if $U_{max} - U_{min} = 1$ and $K_g > 89$. Note that K_g does not depend on the $\overline{C}_g^{\pi}, \overline{C}_g^{\pi'}$ values.

COROLLARY 1. *Consider a GD-MDP and two policies π and π' such that $P_g^{\pi} = 1$, then, under the model* GUBS *and $K_g = U_{max} - U_{min}$, $\pi' \succ \pi$ only if $P_g^{\pi'} \geq \frac{1}{2}$*

PROOF. This result follows from the direct substitution in the equation of Theorem 7. □

Corollary 1 suggests that if $P^{\pi} = 1$, then, to π' be better than π, the necessary condition is that $P^{\pi'} \geq \frac{1}{2}$, but the cost in π' must also be lesser than in π to compensate the difference between P^{π} and $P^{\pi'}$.

4.2 Algorithms

Although the model GUBS presents better semantic for goals, the optimal policy is not anymore stationary. To propose new algorithms to solve the model GUBS, we present two reformulations of the original problem: discrete-based cost and continuous-based cost.

DEFINITION 3 (DISCRETE COST). *Given a GD-MDP described by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, c, \mathcal{G} \rangle$ where $c : \mathcal{S} \times \mathcal{A} \to \mathbb{N}$, we restate the original GD-MDP problem into a new discrete cost MDP $\langle \mathcal{X}, \mathcal{A}', T', c', R'_g \rangle$, where:*

1. *$\mathcal{X} = \mathcal{S} \times \mathbb{N}$;*
2. *$\mathcal{A}' = \mathcal{A}$;*
3. *$T'(x, a, x') = T(s, a, s')$; where $x = (s, C)$ and $x' = (s', C + c(s, a))$;*
4. *$c'(x, a) = u(C + c(s, a)) - u(C)$, where $x = (s, C)$; and*
5. *$R'_g = K_g$ is the terminal reward for goal states.*

In this definition C is the accumulated cost and R'_g is the reward. $\mathcal{X}$ is the set of augmented states, that are the states with the accumulated cost C thus far. The immediate cost c' is defined considering the accumulated cost C and K_g is associated with the terminal reward for the goal states.

Under the Discrete Cost reformulation considering the cost c' in Definition 3, a history $h = s_0, a_0, c_0, s_1, a_1, c_1, \ldots, s_T$ where $s_T \in \mathcal{G}$ is evaluated by:

$$\begin{aligned} u(h) &= c'_0 + c'_1 + \ldots + c'_{T-1} + R'_g \\ &= [u(c_0) - u(0)] + [u(c_0 + c_1) - u(c_0)] + \\ &\quad + [u(c_0 + c_1 + c_2) - u(c_0 + c_1)] + \\ &\quad \ldots + [u(c_0 + \ldots + c_{T-1}) - u(c_0 + \ldots + c_{T-2})] \\ &\quad + K_g \\ &= u(c_0 + \ldots + c_{T-1}) - u(0) + K_g \\ &= u(C_T) - u(0) + K_g, \end{aligned}$$

since $u(0)$ is a constant, we have that preferences under the Discrete Cost reformulation is equivalent to the GUBS model.

The value iteration algorithm (Algorithm 1), solves the discrete-based cost problem. This algorithm has as inputs: a GD-MDP $\mathcal{M}$, the maximum cost C_{max}, the cost-utility function $u(\cdot)$ and the goal terminal reward K_g. Algorithm 1 computes for each possible accumulated cost the Q value for each pair (augmented-state, action) considering the cost c' in Definition 3 (Lines 4–7) and then computes the V value (Lines 8–10).

DEFINITION 4 (CONTINUOUS COST). *Given a GD-MDP described by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, c, \mathcal{G} \rangle$, we restate the original GD-MDP problem into a new continuous cost MDP $\langle \mathcal{X}, \mathcal{A}', T', c', R'_g \rangle$, where:*

1. *$\mathcal{X} = \mathcal{S} \times [U_{min}, U_{max}]$;*
2. *$\mathcal{A}' = \mathcal{A}$;*
3. *$T'(x, a, x') = T(s, a, s')$; where $x = (s, B)$ and $x' = (s', u(u^{-1}(B) + c(s, a)))$;*
4. *$c'(x, a, x') = B' - B$, where $x = (s, B)$ and $x' = (s', B')$; and*

Algorithm 1 Value Iteration for discrete-based cost GUBS model.

Input: GD-MDP $\mathcal{M}$, maximum cost C_{max}, utility function $u(\cdot)$, goal terminal reward K_g
Initialize: $V(s,C) = u(\infty) - u(C)\ \forall\ s \in \mathcal{S}, C \in \{C_{max}, \ldots, C_{max} + \max_{s,a} c(s,a)\}$; $V(s,C) = K_g\ \forall\ s \in \mathcal{G}, C \in \{0, 1, \ldots, C_{max} + \max_{s,a} c(s,a)\}$
for $C = C_{max}$ downto 0 **do**
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $c'(s,a) = u(C + c(s,a)) - u(C)$
    $Q(s,C,a) = c'(s,a) + \sum_{s' \in \mathcal{S} \cup \mathcal{G}} T(s,a,s') V(s', C + c(s,a))$
  end for
  for $s \in \mathcal{S}$ **do**
    $V(s,C) = \max_{a \in \mathcal{A}} Q(s,C,a)$
  end for
end for
return $V(\cdot)$

5. $R'_g = K_g$ is the terminal reward for goal states.

Continuous Cost reformulation is similar to Discrete Cost reformulation; whereas C represents the accumulated cost, B represents the utility of the accumulated cost, i.e $B = u(C)$ or $C = u^{-1}(B)$. We do not present an algorithm for Continuous Cost reformulation, but any hybrid discrete-continuous algorithm can be used to solve it.

5. EXPERIMENTS

We run experiments in the river problem introduced in Section 1 to show how parameters influence the models: *GUBS*, *Penalty to quit* and *Discount cost*.

5.1 Environment Setup

The river problem considers a grid world $N_x \times N_y$, where extremes in x coordinate ($x = 1$ and $x = N_x$) represents the river bank. The agent must cross the river, that can be made by: (i) swimming from any point of the river bank, or (ii) going along the river bank until a bridge at $y = N_y$. However, the river flows to a waterfall (in $y = 1$), where the agent can get trapped or death.

The initial state is in one side of the river and far from the bridge, $x_0 = 1$ and $y_0 = 2$, and the goal is in the other side of the river bank far from the bridge, $x_g = N_x$ and $y_g = 1$. Actions can be taken in any of the cardinal directions: N, S, E and W. If actions are taken on the river bank or in the bridge then transitions are deterministic to the cardinal directions; if actions are taken in the river then transitions are probabilistic and follows the chosen cardinal directions with probability $1 - P$ or follows down the river with probability P. The waterfall is modeled as dead-end states.

We set $N_x = 5$ and $N_y = 100$, whereas varying $P \in \{0.4, 0.6, 0.8\}$, and constant immediate cost 1. Clearly, the *Dual criterion* or MAXPROB criterion would always choose to cross the river by the bridge, which guarantees getting to the goal with probability 1 and cost to goal 201, independently of the probability P.

In the following experiments, we compare how each compared model trades-off *Cost to goal* and *Probability to goal*.

5.2 Models Setup

For the GUBS model we need to use a utility function $u : \mathbb{R} \to [U_{min}, U_{max}]$. In order to guarantee this, we choose the exponential utility function used by Risk Sensitive Markov Decision [9, 12], i.e., $u(C_T) = e^{-\lambda C_T}$ with the risk factor $\lambda = 0.1$. We choose $C_{max} = 1,000$ whereas varying 83 values for K_g between 0 and 30.

The *Penalty to quit* model has only one parameter: penalty D; we vary 72 values for D between 1 and 10,000. The *Discount cost* model has two parameters: discount γ and goal reward R_g; we vary 16 values for γ between 0.75 and 0.999999999; and set $R_g = 0$. Note that values D and γ are difficult to set, since D must be high and γ must be close to 1 to reach a high probability to goal. Because all dead-ends have the same immediate cost, as $\gamma \to 1$, the *Discount* criterion gets closer to MAXPROB independently of R_g.

Penalty to quit and *Discount cost* models are solved approximately by the value iteration algorithm; the policy for them is stationary; and *Probability to goal* and *Cost to goal* are evaluated exactly. Since the optimal policy in *GUBS* model is non-stationary; *Probability to goal* and *Cost to goal* are evaluated in C_{max}, being a lower bound in both cases.

5.3 Results

For each model, we compare the *Probability to goal* for different P values, *Cost to goal* for different P values, and *Probability to goal* versus *Cost to goal*. Regarding the parameters used by each model, we show the log of them in order to produce a better visualization of variation in *Probability to goal* and *Cost to goal*.

5.3.1 Probability to goal

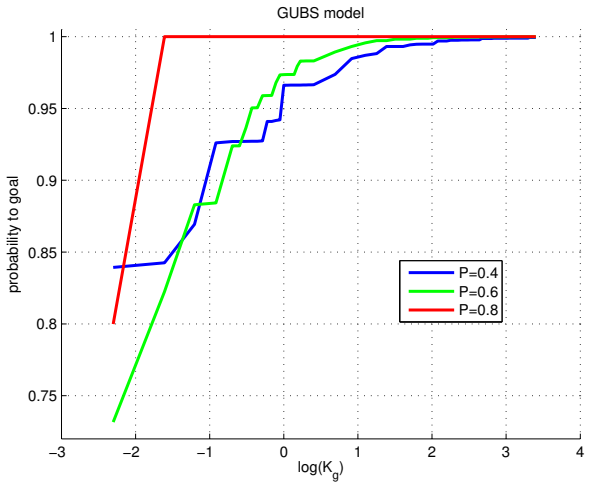
Figure 1 shows the *Probability to goal* under *GUBS*, *Discount cost* and *Penalty to quit* models. In the *GUBS* model, since $U_{max} = 1$ and $U_{min} = 0$, to attend condition in Equation 1, it is enough to set $K_g = 1$ ($\log(K_g) = 0$). Because we know there exists a proper policy rooted at the initial state in the river problem, Corollary 1 guarantees that if $K_g = U_{max} - U_{min}$ and π^* is optimal, then $P_g^{\pi^*} \geq 0.5$; in fact, Figure 1 shows that $P_g^{\pi^*} > 0.96$ for all of the considered scenarios when $K_g = 1$. Note that the proper policy rooted at the initial state (crossing river by the bridge) is optimal only when $P = 0.8$, but in the other scenarios any of the models obtain a probability to goal very near to 1.

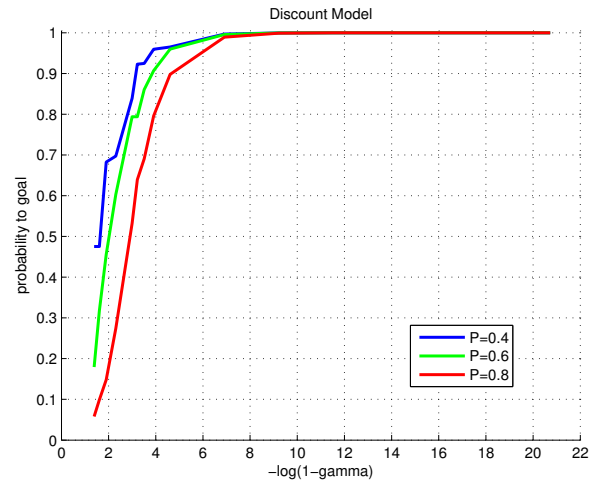
5.3.2 Cost to goal

Figure 2 shows the *Cost to goal* under *GUBS*, *Discount cost* and *Penalty to quit* models. In the *GUBS* and *Discount cost* models, when $P = 0.8$, it is possible to see the cost of the proper policy rooted at the initial state ($C_g^{\pi} = 201$). In all of the models, it is possible to see that *Cost to goal* increases indeterminately, eventually, obtaining MAXPROB solution, i.e. the proper policy rooted at the initial state.

5.3.3 Probability to goal vs. Cost to goal

Figure 3 (top) crosses the data in Figures 1 and 2 and shows the trade-off of each model regarding *Probability* and *Cost to goal*. As expected, the curves of environment with low valued of P presents higher *Probability to goal* for a fixed *Cost to goal*. With $P = 0.4$ and $P = 0.6$, all of the three models present a similar trade-off of *Probability* and *Cost to goal*, whereas with $P = 0.8$ the GUBS model presents an





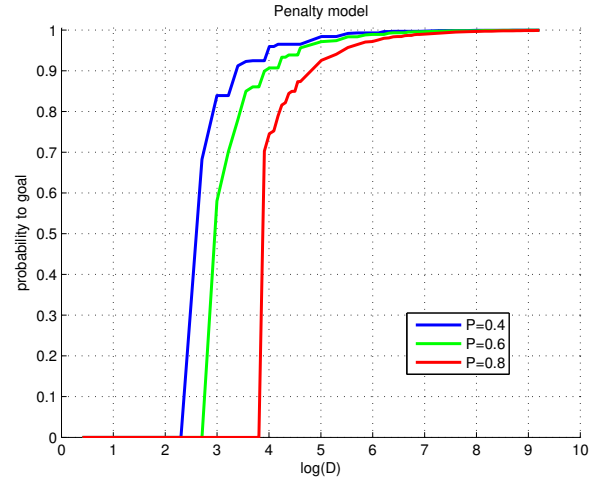
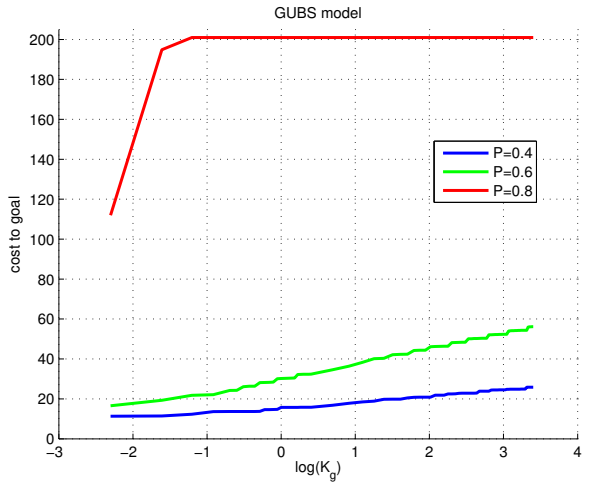
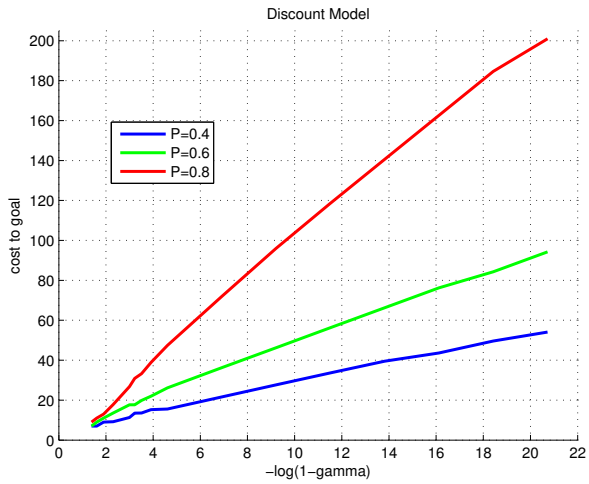


Figure 1: Probability to Goal considering *GUBS*, *Discount cost* and *Penalty to quit* models for different P values.





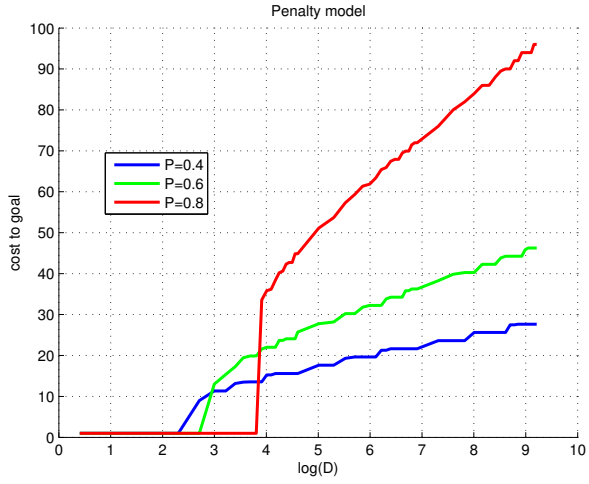


Figure 2: Cost to Goal considering *GUBS*, *Discount cost* and *Penalty to quit* models for different P values.

inferior trade-off, meaning that for a fixed *Cost to goal*, an inferior *Probability to goal* is obtained.

Because utility function $u(\cdot)$ presents risk-averse attitude, in GUBS model, expected cost may be higher. Such a difference appears when $P = 0.8$, the environment with the largest risk. In Figure 3 (bottom), we also run experiments with lesser risk attitude for GUBS by setting $\lambda \in 0.01, 0.07, 0.08, 0.09$. It can be seen that the smaller the risk (small λ), better the trade-off between *Probability* and *Cost to goal*.

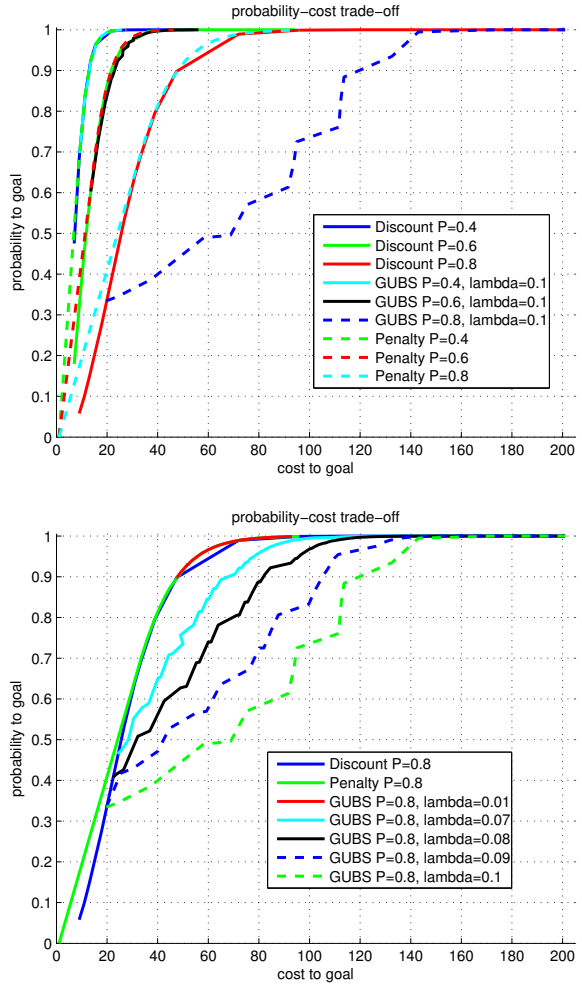


Figure 3: Probability and Cost to Goal for different P values (top) and different λ with $P = 0.8$ (bottom).

6. RELATED WORK SECTION

The finite penalty used in fSSPUDE is similar with the limits proposed in this paper. However the finite penalty proposed in [11] depends on the problem that must be solved and the limits proposed in this paper are independent of the problem. The semantics is well defined for the deterministic case, however, for the probabilistic case, the weights must be defined considering the preference of the decision maker. Different from [11], the motivation of this work is to deal with the trade-off between the expected cost and the probability to reach the goal.

Another class of MDPs that are related are Multi-objective MDPs. In these MDPs there are several competing objectives such as total time, monetary cost, latency, power [5].

A preferential semantics for goals is introduced in [16]. Our proposal is similar because our work is also based on the utility theory. However, [16] considers $(C_T, g) \succ (C_T, \neg g)$ instead of $(C_T, g) \succ (C'_T, \neg g)$.

In [7] the risk is defined as the probability of entering in undesirable or dangerous states. The objective is to find policies whose risk is smaller than some user-specified threshold (the risk we are willing to accept). The problem is formalized as a constrained MDP with two criteria that maximizes the value while the risk is kept below the threshold. They also propose a reinforcement learning algorithm based on weighted the original value and the risk. Different from our work, [7] is not based on utility theory.

GD-MDPs where the objective is to find a policy that maximizes the probability that the cumulative cost is less or equal than some user-defined cost threshold is studied by [8]. An optimal policy for these problem depends on the accumulated cost thus far. So, the agent can take riskier actions when the accumulated cost is small, but should avoid them when the accumulated cost is close to the threshold. [8] is similar to our because they also make a trade off between cost and risk, however this is also not based on utility theory.

7. CONCLUSION

In this paper, we propose a new model to GD-MDP: the GUBS model. First, it defines a semantic for goals based on utility functions; second, we show that GUBS model can show similar trade-off between probability to goal and coast to goal when compared to other models in the literature (*Discount* and *Penalty* models). Besides being based on a normative decision theory, parameters can be rationally set under GUBS model.

We analyze three models on the literature to show that they do not present the same properties that GUBS model and our experiments contribute to the understanding of our theoretical results. We also show how to solve GUBS model in the class of discrete cost GD-MDP, by presenting a new algorithm; and in the class of continuous cost GD-MDP, by presenting a reformulation of the GD-MDP that allows to any hybrid discrete-continuous state MDP algorithm to be used to solve GUBS models.

While *Discount*, *Penalty* and *Dual Criterion* models have stationary policies as optimal policies, the GUBS model has non-stationary ones. Although generic algorithms can be used to solve GUBS model, its structure of transitions between states may be used to formulate more efficient algorithms. The structure of the augmented MDP (original state and accumulated cost) presents two properties: 1) transition among original states does not depend on accumulated cost; and 2) increment in accumulated cost depends only on the original states.

APPENDIX

THEOREM 1. *Consider a GD-MDP and the dual criterion measure, then there exists a GD-MDP $\mathcal{M}$ with policies π and π' and L arbitrarily large and $\delta > 0$ arbitrarily small such that:*

1. $P_g^{\pi} > P_g^{\pi'}$ *and* $P_g^{\pi} - P_g^{\pi'} < \delta$; *and*

2. $\overline{C}_g^{\pi} - \overline{C}_g^{\pi'} > L$.

PROOF. Consider the MDP in Figure 7, where $s_g \in \mathcal{G}$; and policies π, where action a is chosen, and π', where action b is chosen. Then:

1. $P_g^{\pi} = 1$ and $P_g^{\pi'} = P$, and
2. $\overline{C}_g^{\pi} = c(s_0, a)$ and $\overline{C}_g^{\pi'} = c(s_0, b)$.

Choose $P > 1 - \delta$, $c(s_0, a) > L + 1$ and $c(s_0, b) = 1$, then the condition in the theorem is observed. □

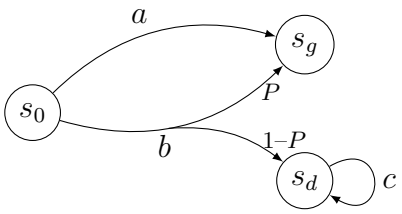


Figure 4: A simple example

THEOREM 2. *Consider a GD-MDP with policies π and π' such that: (*i*) $P_g^{\pi} > P_g^{\pi'}$, or (*ii*) $P_g^{\pi} = P_g^{\pi'} = 1$ and $\overline{C}_g^{\pi} < \overline{C}_g^{\pi'}$; then, there exists a penalty D_0 such that $\overline{C}_D^{\pi} < \overline{C}_D^{\pi'}$ for any $D > D_0$.*

PROOF. First, consider the case $P_g^{\pi} > P_g^{\pi'}$, then we:

$$\begin{aligned}\overline{C}_{D_0}^{\pi} &\leq \overline{C}_g^{\pi} P_g^{\pi} + (D_0 + \overline{C}_g^{\pi})(1 - P_g^{\pi}) \\ &= \overline{C}_g^{\pi} + D_0 - D_0 P_g^{\pi}\end{aligned}$$

If we choose

$$D_0 > \frac{\overline{C}_g^{\pi}}{P_g^{\pi} - P_g^{\pi'}}$$

we have:

$$\begin{aligned}\overline{C}_{D_0}^{\pi} &< \overline{C}_g^{\pi} + D_0 - D_0 P_g^{\pi'} - \overline{C}_g^{\pi} \\ &= D_0(1 - P_g^{\pi'}) \\ &\leq \overline{C}_{D_0}^{\pi'}.\end{aligned}$$

Now, consider the case $P_g^{\pi} = P_g^{\pi'} = 1$ and $\overline{C}_g^{\pi} < \overline{C}_g^{\pi'}$, then we have that:

1. $\overline{C}_D^{\pi} \leq \overline{C}_g^{\pi}$;
2. $\overline{C}_D^{\pi} \leq D$; and
3. $\overline{C}_D^{\pi}$ is continuous in D.

Then, since $\overline{C}_g^{\pi} < \overline{C}_g^{\pi'}$ there exists D_0 such that $\overline{C}_D^{\pi'} > \overline{C}_g^{\pi}$ for any $D > D_0$ and we have $\overline{C}_D^{\pi} \leq \overline{C}_g^{\pi} \leq \overline{C}_D^{\pi'}$. □

THEOREM 3. *Consider an arbitrary penalty D, then there exists a GD-MDP with policies π and π' such that:*

1. $P_g^{\pi} > P_g^{\pi'}$; *and*
2. $\overline{C}_D^{\pi} > \overline{C}_D^{\pi'}$.

PROOF. Consider again the MDP in figure 7, then:

1. $P_g^{\pi} = 1$ and $P_g^{\pi'} = P$;
2. $\overline{C}_D^{\pi} = \min\{c(s_0, a), D\}$; and
3. $\overline{C}_D^{\pi'} = \min\{c(s_0, b) + (1 - P)D, D)\}$.

Choose $c(s_0, a) > D$ and $c(s_0, b) < PD$, then we have:

$$\begin{aligned}\overline{C}_D^{\pi} &= \min\{c(s_0, a), D\} = D \\ &> (1 - P)D + c(s_0, b) \\ &\geq \min\{c(s_0, b) + (1 - P)D, D\} = \overline{C}_D^{\pi'}.\end{aligned}$$

□

THEOREM 4. *Consider a GD-MDP with policies π and π' such that: (*i*) $P_g^{\pi} > P_g^{\pi'}$, or (*ii*) $P_g^{\pi} = P_g^{\pi'} = 1$ and $\overline{C}_g^{\pi} < \overline{C}_g^{\pi'}$; then, if, when reaching a goal, a reward R_g is considered, there exists a discount factor γ_0 such that $\overline{C}_{\gamma}^{\pi} < \overline{C}_{\gamma}^{\pi'}$ for any $\gamma > \gamma_0$.*

PROOF. Consider the first scenario where $P_g^{\pi} > P_g^{\pi'}$ and define the probability $P_{g,n}^{\pi}$ of reaching the goal state before n time-steps by following policy π. Choose the following parameters:

1. n_0 so that $P_{g,n_0}^{\pi} > P_g^{\pi'}$, such n_0 there exists because $P_g^{\pi} > P_g^{\pi'}$ and the MDP is finite;
2. $\gamma > \left(\frac{P_g^{\pi'}}{P_{g,n_0}^{\pi}}\right)^{\frac{1}{n_0}}$ to obtain $\gamma^{n_0} P_{g,n_0}^{\pi} - P_g^{\pi'} > 0$; and
3. $R_g > \dfrac{\overline{C}_g^{\pi} P_g^{\pi} + \max\limits_{s \in \mathcal{S}, a \in \mathcal{A}}(c(s,a)) \dfrac{1}{1-\gamma}(1 - P_g^{\pi})}{\gamma^{n_0} P_{g,n_0}^{\pi} - P_g^{\pi'}}$ to obtain:

$$\begin{aligned}&- R_g \gamma^{n_0} P_{g,n_0}^{\pi} + \overline{C}_g^{\pi} P_g^{\pi} + \\ &\quad + \max_{s \in \mathcal{S}, a \in \mathcal{A}}(c(s,a)) \frac{1}{1-\gamma}(1 - P_g^{\pi}) < -R_g P_g^{\pi'}.\end{aligned}$$

Then we have:

$$\begin{aligned}\overline{C}_{\gamma}^{\pi} &\leq -R_g \gamma^{n_0} P_{g,n_0}^{\pi} + \overline{C}_g^{\pi} P_g^{\pi} \\ &\quad + \max_{s \in \mathcal{S}, a \in \mathcal{A}}(c(s,a)) \frac{1}{1-\gamma}(1 - P_g^{\pi}) \\ &< -R_g P_g^{\pi'} \leq \overline{C}_{\gamma}^{\pi'}.\end{aligned}$$

Now, consider the second scenario where $P_g^{\pi} = P_g^{\pi'} = 1$ and $\overline{C}_g^{\pi} < \overline{C}_g^{\pi'}$. Since $\lim_{\gamma \to 1} \overline{C}_{\gamma}^{\pi} = \overline{C}_g^{\pi}$ and $\lim_{\gamma \to 1} \overline{C}_{\gamma}^{\pi'} = \overline{C}_g^{\pi'}$, there exists γ_0 such that $\overline{C}_{\gamma}^{\pi} < \overline{C}_{\gamma}^{\pi'}$ for any $\gamma > \gamma_0$. □

THEOREM 5. *Consider an arbitrary discount factor γ, then there exists a GD-MDP with policies π and π' such that:*

1. $P_g^{\pi} > P_g^{\pi'}$; *and*
2. $\overline{C}_{\gamma}^{\pi} > \overline{C}_{\gamma}^{\pi'}$.

PROOF. Consider again the MDP in figure 7, then:

1. $P_g^{\pi} = 1$ and $P_g^{\pi'} = P$;
2. $\overline{C}_{\gamma}^{\pi} = c(s_0, a)$; and
3. $\overline{C}_{\gamma}^{\pi'} = c(s_0, b) + (1 - P)c(s_0, b)\frac{\gamma}{1-\gamma}$.

It is easy to choose $c(s_0, a)$, $c(s_0, b)$ and D such that $\overline{C}_{\gamma}^{\pi} > \overline{C}_{\gamma}^{\pi'}$. □

REFERENCES

[1] D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, Aug. 1991.

[2] B. Bonet and H. Geffner. Labeled RTDP: Improving the convergence of real-time dynamic programming. *Proceedings of International Conference on Automated Planning and Scheduling*, pages 12–21, 2003.

[3] B. Bonet and H. Geffner. mGPT: A probabilistic planner based on heuristic search. In *Journal of Artificial Intelligence Research*, volume 24, pages 933–944, 2005.

[4] A. Camacho, C. Muise, and S. A. McIlraith. From FOND to robust probabilistic planning: Computing compact policies that bypass avoidable deadends. In *The 26th International Conference on Automated Planning and Scheduling*, pages 65–69, 2016.

[5] K. Chatterjee, R. Majumdar, and T. A. Henzinger. *Markov Decision Processes with Multiple Objectives*, pages 325–336. Springer Berlin Heidelberg, 2006.

[6] H. Geffner and B. Bonet. A concise introduction to models and methods for automated planning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(1):1–141, 2013.

[7] P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *J. Artif. Intell. Res. (JAIR)*, 24:81–108, 2005.

[8] P. Hou, W. Yeoh, and P. Varakantham. Revisiting risk-sensitive MDPs: New algorithms and results. In *Proceedings of the Twenty-Fourth International Conference on Automated Planning and Scheduling (ICAPS)*, 2014.

[9] R. A. Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

[10] A. Kolobov, Mausam, D. Weld, and H. Geffner. *Heuristic search for generalized stochastic shortest path MDPs*, pages 130–137. 2011.

[11] A. Kolobov, Mausam, and D. S. Weld. A theory of goal-oriented MDPs with dead ends. In N. de Freitas and K. P. Murphy, editors, *UAI*, pages 438–447. AUAI Press, 2012.

[12] S. D. Patek. On terminating Markov decision processes with a risk averse objective function. *Automatica*, 37:1379–1386, 2001.

[13] F. Teichteil-Königsbuch. Stochastic safest and shortest path problems. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 1825–1831. AAAI Press, 2012.

[14] F. Teichteil-Königsbuch, U. Kuter, and G. Infantes. Incremental plan aggregation for generating policies in MDPs. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, AAMAS '10, pages 1231–1238, 2010.

[15] F. Teichteil-Königsbuch, V. Vidal, and G. Infantes. Extending classical planning heuristics to probabilistic planning with dead-ends. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, 2011.

[16] M. P. Wellman and J. Doyle. Preferential semantics for goals. In *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, AAAI'91, pages 698–703. AAAI Press, 1991.