# Probabilistic Planning with Non-Linear Utility Functions and Worst-Case Guarantees *

Stefano Ermon, Carla Gomes, Bart Selman
Department of Computer Science
Cornell University
{ermonste,gomes,selman}@cs.cornell.edu

Alexander Vladimirsky
Department of Mathematics
Cornell University
vlad@math.cornell.edu

## ABSTRACT

Markov Decision Processes are one of the most widely used frameworks to formulate probabilistic planning problems. Since planners are often risk-sensitive in high-stake situations, non-linear utility functions are often introduced to describe their preferences among all possible outcomes. Alternatively, risk-sensitive decision makers often require their plans to satisfy certain worst-case guarantees.

We show how to combine these two approaches by considering problems where we maximize the expected utility of the total reward subject to worst-case constraints. We generalize several existing results on the structure of optimal policies to the constrained case, both for finite and infinite horizon problems. We provide a Dynamic Programming algorithm to compute the optimal policy, and we introduce an admissible heuristic to effectively prune the search space. Finally, we use a stochastic shortest path problem on large real-world road networks to demonstrate the practical applicability of our method.

## Categories and Subject Descriptors

I.2 [**ARTIFICIAL INTELLIGENCE**]: Problem Solving, Control Methods, and Search

## General Terms

Algorithms, Theory

## Keywords

Planning, Utility Functions, Constraints

## 1. INTRODUCTION

Markov Decision Processes (MDPs) are one of the most widely used frameworks to formulate probabilistic planning problems. In these problems, the notion of risk is related to the fact that, given the stochastic nature of the problem, each policy can generally produce several possible outcomes, and some of them might reflect unsatisfactory performance. In many applications, such as space planning and natural resource management, it is critical to use performance metrics that allow the ability to manage the risk, i.e. a certain level of control over unfavorable outcomes [4, 18].

The problem of managing the risk has been studied extensively in artificial intelligence, operations research, and control theory. Many formulations have been proposed (see Section 2 for more details), among which decision theoretic planning and worst-case approaches are the two most widely used. The former is based on decision theory, more specifically, on the fact that decision makers accepting a small number of axioms always choose the course of actions that maximizes the expected utility of the total reward [16], where the specific form of the utility function describes the risk attitude of the planners. The latter is focused on providing deterministic guarantees for the plans by looking at worst-case realizations of the random processes involved.

In this paper, we show how to combine these two approaches by considering problems where the objective is to maximize the expected utility of the total reward subject to worst-case, linear constraints. For example, in the case of a linear utility function, we can maximize the expected total reward only among those policies whose reward is larger than a given threshold, even in the worst-case scenario. With a (non-linear) "step" utility function, we can maximize the probability of reaching a target reward level, while enforcing the worst-case constraint at the same time.

Our theoretical results extend previous work on MDPs with non-linear utility functions and show that the optimal policy for the constrained optimization problem is highly structured: it is deterministic, and even though generally not Markovian, it depends on the history only through the total accumulated reward. Therefore, an optimal policy can be represented (and approximated) much more effectively than general history-dependent policies. Furthermore, we show how to exploit the presence of worst-case constraints to define an admissible heuristic, which we use in a Dynamic Programming algorithm to speed up the policy search.

To demonstrate the practical applicability of our method, we consider stochastic shortest path problems as a special case of MDPs. We show that our algorithm scales to large real-world road networks, and it leads to plans that are significantly different from the ones obtained with traditional optimization criteria. We think this type of formulation can be particularly useful for time-dependent problems, such as the ones faced by the Green Driver App [2], where traffic light information are explicitly modeled. In fact, in this situation it is necessary to consider policies that are non-Markovian, even when given linear utility functions.

*Supported by NSF Grants 0832782 and DMS-1016150.

**Appears in:** *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

## 2. RELATED WORK

Decision theoretic planning has been studied extensively, mainly in artificial intelligence [18, 14], control theory, and operations research [10, 15, 1, 7]. Typically, monotonically non-decreasing utility functions, mapping total rewards to utility values, are used to to describe the preferences of the planner. In particular, exponential utility functions [10] are commonly used because they satisfy a separability property that allows an efficient Dynamic Programming solution. Recently, researchers have also considered planning problems with more general non-linear utility functions [12]. The results in this paper are related to that line of work (and we use a similar notation whenever possible), but with the novel introduction of worst-case constraints.

The most conservative approach to account for risk is worst case planning, where only the worst possible outcome is optimized. A generalization known as $\alpha$-value criterion is introduced in [9], where outcomes that happen with a probability smaller than $\alpha$ are not considered (when $\alpha = 0$, it is equivalent to the worst case). In this work, instead of optimizing the worst-case scenario, we introduce constraints that need to be satisfied by the plan, under all possible realizations of the randomness. The relationship of our approach with worst-case planning is discussed in detail below in Section 4.1.

There are several existing frameworks for constrained probabilistic planning problems in the literature. Many of them [1, 7] involve the maximization of an expected (discounted) reward subject to upper bounds on the total expected (discounted) costs. The main limitation of this approach is that upper bounding an expected value might provide a guarantee in terms of risk that is too weak, because it constitutes only a mild restriction on the possible outcomes (constraints are satisfied only on average, while our constraints are met by all possible realizations). The same holds for mean-variance analysis [17], where the problem is analyzed in terms of the tradeoff between expected value and variance of the total reward (either by imposing constraints on the variance, or associating a cost with it). The constrained formulation that is closest to our work is the sample-path constraint introduced in [15]. In [15], they consider time-average MDPs, with a reward and cost associated with each decision. The optimization problem is to maximize the expected average reward over all policies that meet the sample-path constraint, where a policy is said to meet the sample-path constraint if the time-average cost is below a specified threshold with probability one. Notice that also in this case the guarantee can be quite weak, because the constraint is imposed only on an averaged quantity. Finally, in [8] they derive and solve Dynamic Programming equations for two special types of worst-case constrained Stochastic Shortest path problems (in our formalism these correspond to $U_L$ and $U_{K,L}$ utility functions defined in section 6). We emphasize that the new framework in this paper is significantly more general and can be applied to general MDPs to provide worst-case performance guarantees with general non-linear utility functions.

## 3. PROBLEM DEFINITION

We consider probabilistic planning problems represented as Markov Decision Processes. Formally, an MDP is a tuple $(S, A, P, r)$ where $S$ is a set of states, $A$ is a set of actions, $P$ is a set of transition probabilities and $r : S \times A \times S \mapsto \mathbb{R}$ is an (immediate) reward function. If an agent executes an action $a \in A$ while in a state $s \in S$, then it receives an immediate reward $r(s, a, s')$ and it transitions to a new state $s' \in S$ with probability $P(s'|s, a)$. We denote by $A_s \subseteq A$ the set of actions available while the agent is in state $s$.

In this paper we consider *finite* MDP where both the state space $S$ and action space $A$ are finite sets.

**Policies.** Let the planning horizon $T$ be the (possibly infinite) number of time steps that the agent plans for. A *history* at time step $t$ is a sequence $h_t = (s_0, a_0, \cdots, s_{t-1}, a_{t-1}, s_t)$ of states and actions that leads from the initial state $s_0$ to state $s_t$ at time step $t$. The set of all histories at time step $t$ is denoted $H_t = (S \times A)^t \times S$.

In a probabilistic setting, a plan is represented by a *policy*, where a policy is a sequence of decision rules, one for each time step in the planning horizon. The most general decision rules are *randomized history-dependent* (HR), which are mappings $d_t : H_t \to P(A)$, where $P(A)$ is the set of probability distributions over the set of actions $A$. A history-dependent decision rule is called *Markovian* if it depends only on the current state $s_t$, while it is called *deterministic* if it deterministically choses the action to be taken. A policy is called *stationary* if $d_t = d$ for all time steps $t$ within the planning horizon and $d$ is a Markovian decision rule. We denote the class of *deterministic stationary* (SD) policies by $\Pi^{SD}$ and the class of *randomized stationary* policies by $\Pi^{SR}$.

**Utility Functions.** Let $w_T$ be the total reward received by the agent, that is the sum of all the immediate rewards accrued within the planning horizon

$$w_T = \sum_{t=0}^{T-1} r_t(s_t, a_t, s_{t+1}) \tag{1}$$

A standard approach to model the preferences of the planner among the possible realizations of $w_T$ is to use a monotonically non-decreasing utility function $U : \mathbb{R} \mapsto \mathbb{R}$, which maps total rewards to utility values. Decision theory suggests that decision makers accepting a small number of axioms always choose the course of actions that maximizes the expected utility of the total reward [16].

## 4. FINITE HORIZON PROBLEMS

First we consider planning problems where the planning horizon $T$ is finite, and we will later extend the results to the infinite horizon case. We define the *value* of a policy $\pi \in \Pi^{HR}$ from an initial state $s \in S$ as

$$v^{\pi}_{U,T}(s) = \mathbb{E}^{s,\pi}\left[U\left(\sum_{t=0}^{T} r_t\right)\right] = \mathbb{E}^{s,\pi}\left[U\left(w_T\right)\right] \tag{2}$$

which is the expected utility of the total reward $w_T$. For standard utility functions, the expected utilities exist and are finite because there is a finite number of possible finite trajectories in finite MDPs [14]. The optimal values

$$v^{*}_{U,T}(s) = \sup_{\pi \in \Pi} v^{\pi}_{U,T}(s) \tag{3}$$

exist since the *values* exist for every policy $\pi \in \Pi$. Properties of the optimal policy for this case have been studied in [12]. In particular, the optimal policy is deterministic and even though it is generally not Markovian, it depends on the history $h_t$ only through the accumulated reward $w_t$.

## 4.1 Worst-Case Constraints

Risk-sensitive planners are often interested in worst-case scenarios. With this perspective, a common approach is to look for a policy that maximizes the worst case performance (*game against nature*). Formally, in the max-min version of the problem, we seek to optimize

$$d_T^*(s) = \sup_{\pi \in \Pi} d_T^\pi(s) \tag{4}$$

where

$$d_T^\pi(s) = \min \{k | \mathbb{P}[w_T = k] > 0\} \tag{5}$$

is the worst-case realization of the total reward $w_T$ (the definition is well posed since $w_T$ is a discrete random variable with a finite sample space for a finite MDP).

In this paper, we consider situations where the planner wants to enforce linear worst case constraints on the total reward $w_T$ of the form

$$w_T > L \tag{6}$$

for all possible realizations of $w_T$ (equivalently, (6) has to hold almost surely, because $w_T$ has a finite sample space). Notice that the *game against nature* approach is equivalent to finding the largest value of $L$ such that condition (6) can be met.

In this paper we combine the problem of finding a policy that maximizes the expected utility, with the presence of linear worst-case constraints on the total reward. Formally, we wish to find

$$v_{U,T,L}^*(s) = \sup_{\pi \in \Pi(L)} v_{U,T}^\pi(s) \tag{7}$$

where the optimization is restricted to the set of policies $\Pi(L)$ whose corresponding total reward $w_T$ satisfies condition (6), for all possible realizations of $w_T$. The problem is well defined when the set $\Pi(L)$ is not empty, that is if and only if $L < d_T^*(s)$.

As an example, in the simplest case of a linear utility function $U(x) = x$, the objective is to maximize the total expected reward but only among those policies with a guaranteed lower bound $L$ on the total reward.

## 4.2 Extended-Value Utility Functions

We show that the constrained problem defined by Equation (7) can be solved by considering the original optimization problem defined by Equation (3) with a more general utility function. We introduce the concept of an *extended-value utility function*, which can be used to model linear worst case constraints on the total reward $w_T$. Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$ be the affinely extended real number system, which turns into a totally ordered set by defining $-\infty \leq a \leq +\infty$ for all $a \in \mathbb{R}$. We define an *extended-value utility function* a *monotonically nondecreasing* function $U : \mathbb{R} \to \overline{\mathbb{R}}$ that maps wealth levels to the corresponding utility values.

Let us consider the problem previously defined by Equations (2) and (3) in the more general case where $U$ is an *extended-value utility function*. Recall that $\mathbb{E}[X]$ exists and $\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]$ when $\mathbb{E}[X_+] < \infty$ or $\mathbb{E}[X_-] < \infty$, where $X_+$ and $X_-$ denote the positive and negative part of $X$, respectively. In the more general case of an *extended-value utility function*, the expected utilities defined by (2) exist (but are not always finite), because $\mathbb{E}[U(w_T)_+] < \infty$ since the number of trajectories is finite for finite MDPs.

Notice that agents acting as though they were maximizing expected *extended utility functions* satisfy the *Completeness* and *Transitivity* axioms of the Von Neumann-Morgenstern Utility Theorem [16], because the extended real number system is totally ordered and the order relation is transitive. Moreover, they satisfy the *Indipendence* axiom by the linearity of expectation. Notice however that they violate the *Continuity* axiom. In fact, given three *lotteries* such that $A \succeq B \succeq C$ ($A$ is preferred over $B$, and $B$ over $C$) and where the expected utilities of $A$ and $B$ are finite but the expected utility of $C$ is $-\infty$, there is no combination of $A$ and $C$ that gives an expected utility that is equal to the one of $B$.

The optimal values defined by (3) need not to be finite, but they are bounded from above since the total rewards are bounded and thus the expected utilities of the total reward for all policies are bounded from above as well. The following result holds:

LEMMA 1. *For any extended value utility function $U$, let $L = \sup\{w | U(w) = -\infty\}$. Then for any policy $\pi \in \Pi^{HR}$, $v_{U,T}^\pi(s) > -\infty$ if and only if $w_T > L$ almost surely.*

PROOF. If $P[w_T \leq L] > 0$, it follows $v_{U,T}^\pi(s) = -\infty$. If $w_T > L$ almost surely, then it follows by monotonicity that $v_{U,T}^\pi(s) > U(L) \geq -\infty$. □

As a corollary, if the optimal value is finite, then the worst case constraint $w_T > L$ is satisfied by the optimal plan.

Using Lemma 1, we show that we can solve the constrained problem defined by Equation (7) for a standard utility function $U$ by solving the unconstrained problem (3) with an *extended-value* utility function $U_e$ defined as follows

$$U_e(x) = \begin{cases} -\infty & x \leq L \\ U(x) & x > L \end{cases}$$

LEMMA 2. *For any utility function $U$ and lower bound $L$ such that $\Pi(L)$ is non empty, $v_{U,T,L}^*(s) = v_{U_e,T}^*(s)$.*

PROOF. Let $\pi'$ be the optimal policy for the constrained problem. Since $\pi' \in \Pi(L)$, we have $w_T > L$ almost surely and therefore

$$-\infty < v_{U,T,L}^*(s) = v_{U,T,L}^{\pi'}(s) = v_{U_e,T}^{\pi'}(s) \leq v_{U_e,T}^*(s)$$

Since $v_{U_e,T}^*(s) = v_{U_e,T}^{\pi^*}(s) > \infty$, by Lemma 1, $\pi^*$ satisfies the constraint (6) almost surely, so $\pi^* \in \Pi(L)$ and

$$v_{U,T,L}^*(s) \geq v_{U_e,T}^{\pi^*}(s) = v_{U_e,T}^*(s)$$

□

We focus now on characterizing the optimal policy for the unconstrained problem with an extended value utility function $U_e$ (we drop the subscript for compactness). In the rest of the paper, we will use subscripts to indicate the length of the planning horizon $T$ and the utility function $U$ used. We will use superscripts to indicate the policy $\pi$ used and, when relevant, the decision epoch $t$ the value refers to.

## 4.3 Optimality Conditions

We provide a characterization of the optimal policies for maximum expected utility planning problems by generalizing some results obtained in [12] to the more general case of extended value utility functions.

THEOREM 1. *Let $\pi = (d_0, \ldots, d_{T-1}) \in \Pi^{HR}$ be a policy. The values $v_{U,T}^{\pi,t}(h_t) = \mathbb{E}^{\pi}[U(w_T)|h_t]$ of a policy $\pi$ at time step $t$ given an history $h_t \in H_t$ satisfy*

$$v_{U,T}^{\pi,T}(h_T) = U(w_T), h_T \in H_T$$

$$v_{U,T}^{\pi,t}(h_t) = \sum_{a \in A_{s_t}} d_t(h_t, a) \sum_{s' \in S} P(s'|s_t, a) v_{U,T}^{\pi,t+1}(h_t \circ (a, s'))$$

*where $h_t \in H_t, 0 \leq t < T$, $\circ$ is the composition operator, and the last component of $h_t$ is $s_t$.*

PROOF. Similar to Theorem 4.1 in [12]. □

If we define the optimal values for a history $h_t$ as $v_{U,T}^{*,t}(h_t) = \sup_{\pi \in \Pi} v_{U,T}^{\pi,t}(h_t)$, then we have $v_{U,T}^{*,0}(s) = v_{U,T}^{*}(s)$ for any initial state $s$. Furthermore,

THEOREM 2. *The values $v_{U,T}^{*,t}(h_t)$ are the unique solutions to the optimality equations*

$$v_{U,T}^{*,T}(h_T) = U(w_T), h_T \in H_T$$

$$v_{U,T}^{*,t}(h_t) = \max_{a \in A_{s_t}} \sum_{s' \in S} P(s'|s_t, a) v_{U,T}^{*,t+1}(h_t \circ (a, s')) \quad (8)$$

*for $h_t \in H_t, 0 \leq t < T$ and where the last component of $h_t$ is $s_t$.*

PROOF. Similar to Theorem 4.2 in [12]. □

The above results also show that there exists a *deterministic* history-dependent optimal policy, that for an history $h_t$ chooses a maximizer in Eq. (8) as action. However, the policy might depend on the entire previous history $h_t$. In the following sections we use the state-augmentation approach to show that the policy has more structure, i.e. it depends on the history only through the total reward accumulated so far $w_t = \sum_{k=0}^{t-1} r_k(s_k, a_k, s_{k+1})$.

## 4.4 State Space Augmentation

A deeper characterization of the structure of the optimal policy can be obtained by considering a new *augmented* MDP where the state space is augmented with wealth levels (corresponding to the sum of accumulated rewards). As in [12], let

$$R = \{0\} \cup \{r(s, a, s') | P(s'|s, a) > 0, s, s' \in S, a \in A_s\}$$

be the set of possible rewards. Then the set $W^t$ of all possible wealth levels at time step $t$ is inductively defined as follows

$$W^0 = \{0\}, W^{t+1} = \{r + w | r \in R, w \in W_t\}$$

We consider an extended MDP where the augmented states space is $\langle S \rangle = (S \times W^0) \cup (S \times W^1) \cup \cdots \cup (S \times W^T)$. The actions available in an augmented state $\langle s \rangle = (s, w)$ are $A_{\langle s \rangle} = A_{(s,w)} = A_s$ for all wealth levels $w$. The transition probability from a state $\langle s \rangle = (s, w)$ to $\langle s \rangle' = (s', w')$ is

$$P(\langle s \rangle' | \langle s \rangle, a) = \begin{cases} P(s'|s, a) & \text{if } w' = w + r(s, a, s') \\ 0 & \text{otherwise} \end{cases}$$

All augmented rewards $r(\langle s \rangle, a, \langle s \rangle')$ are zero, and there is a terminal augmented reward $J(\langle s \rangle) = U(w)$ applicable at time $T$ for an augmented state $\langle s \rangle = (s, w)$. There is no utility function for the augmented model, so the value $\langle z \rangle_T^{\langle \pi \rangle}(s, w)$ of an augmented policy $\langle \pi \rangle$ from initial augmented state $(s, w)$ is given by the expected total augmented reward (equivalently, by the expected augmented terminal reward, since all other augmented rewards are zero).

Notice that the construction previously used in [12] cannot be used in our generalized case because it defines rewards in the augmented model as the difference of two utilities, which might not be well defined in for an extended value utility function. Notice also that the augmented MDP is still finite for a finite planning horizon $T$.

The original MDP and the augmented one are closely related, and intuitively the two underlying stochastic processes are equivalent. Formally, it can be shown as done in [12] that there is a 1-1 mapping between a history of the original model and a class of equivalent histories of the augmented model.

LEMMA 3. *For any wealth level $w$, and for any history of the original model $h_t = (s_0, a_0, s_1, \cdots, s_t) \in H_t$, the sequence $\phi_w(h_t) = \langle h \rangle_t = (\langle s \rangle_0, a_0, \langle s \rangle_1, \cdots, \langle s \rangle_t)$ is a history of the augmented model, where*

$$\langle s \rangle_k = (s_k, \tilde{w}_k) = (s_k, w + w_k), 0 \leq k \leq t$$

*Furthermore, for any history of the augmented model $\langle h \rangle_t = (\langle s \rangle_0, a_0, \langle s \rangle_1, \cdots, \langle s \rangle_t) \in \langle H \rangle_t$ where $\langle s \rangle_k = (s_k, \tilde{w}_k)$ for all $0 \leq k \leq t$, there exists a wealth level $w$ such that $\tilde{w}_k = w + w_k$ and the sequence $\psi(\langle h \rangle_t) = (s_0, a_0, s_1, \cdots, s_t)$ is a history of the original model.*

PROOF. Similar to Lemmas 4.3 and 4.4 in [12]. □

Similarly, using Lemma 3, for any policy in the original model $\pi = (d_0, d_1, \cdots, d_{T-1}) \in \Pi^{HR}$, we define a policy of the augmented model, $\Psi(\pi) = (\langle d \rangle_0, \langle d \rangle_1, \cdots, \langle d \rangle_{T-1})$, such that for all augmented histories $\langle h \rangle$, $\langle d \rangle_t (\langle h \rangle, a) = d_t(\psi(\langle h \rangle), a)$.

For any augmented policy $(\langle d \rangle_0, \langle d \rangle_1, \cdots, \langle d \rangle_{T-1})$, we define a policy in the original model, $\Phi_w = (d_0, d_1, \cdots, d_{T-1})$, such that for all histories $h \in H_t$,

$$d_t(h, a) = \langle d \rangle_t (\phi_w(h), a)$$

Furthermore, the values of the policies in the original and augmented model are closely related:

THEOREM 3. *For each policy $\pi \in \Pi^{HR}$ in the original MDP and for all states $s \in S$,*

$$\langle z \rangle_T^{\Psi(\pi)}(s, w) = \mathbb{E}^{s,\pi}\left[U(w + \sum_{t=0}^{T-1} r_t)\right]$$

*For each policy $\langle \pi \rangle$ in the augmented MDP, for each wealth level $w \in W$,*

$$\langle z \rangle_T^{\langle \pi \rangle}(s, w) = \mathbb{E}^{s, \Phi_w(\langle \pi \rangle)}\left[U(w + \sum_{t=0}^{T-1} r_t)\right]$$

PROOF. First, we need to prove the probabilistic equivalence of the stochastic processes induced by policies that correspond through the mappings $\Psi$ and $\Phi_w$. The proof is similar to the one of Theorem 4.5 in [12] and is omitted. Furthermore,

$$\langle z \rangle_T^{\Psi(\pi)}(s, w) = \mathbb{E}^{(s,w),\Psi(\pi)}\left[\sum_{t=0}^{T-1} \langle r \rangle_t + J(\widetilde{s}_T, \widetilde{w}_T)\right] =$$

$$\mathbb{E}^{(s,w),\Psi(\pi)}[U(\widetilde{w}_T)] = \mathbb{E}^{s,\pi}\left[U(w + \sum_{t=0}^{T-1} r_t)\right]$$

where $(\widetilde{s}_T, \widetilde{w}_T)$ is the terminal augmented state, because both policies produce equivalent random processes. For the second part,

$$\langle z \rangle_T^{\langle \pi \rangle}(s,w) = \mathbb{E}^{(s,w),\langle \pi \rangle}\left[\sum_{t=0}^{T-1} \langle r \rangle_t + J(\widetilde{s}_T, \widetilde{w}_T)\right] =$$

$$\mathbb{E}^{(s,w),\langle \pi \rangle}\left[J(\widetilde{s}_T, \widetilde{w}_T)\right] = \mathbb{E}^{s,\Phi_w(\langle \pi \rangle)}\left[U(w + \sum_{t=0}^{T-1} r_t)\right]$$

because of the probabilistic equivalence. □

Since the augmented MDP is a standard MDP, it is well known [3] that the optimal values $\langle z \rangle_T^*(s,w)$ exist (but are not necessarily finite) for all augmented states $(s,w)$. Moreover, there exists a Markovian, deterministic policy $\langle \pi \rangle_T^*$ that is optimal for the augmented model. We show that $\Phi_0(\langle \pi \rangle_T^*)$ is an optimal policy for the original MDP (optimality for the original MDP refers to the maximum expected utility criterion):

$$\begin{aligned} v_{U,T}^*(s) &= v_{U,T}^{\pi_T^*}(s) = \langle z \rangle_T^{\Psi(\pi_T^*)}(s,0) \\ &\leq \langle z \rangle_T^*(s,0) = \langle z \rangle_T^{\langle \pi \rangle_T^*}(s,0) = v_{U,T}^{\Phi_0(\langle \pi \rangle_T^*)}(s) \end{aligned}$$

Notice that the optimal policy $\Phi_0(\langle \pi \rangle_T^*)$ for the original model is not Markovian anymore. However, the dependency on the history $h_t$ is limited, since the decision rules $d_t$ only depend on the accumulated reward $w_t$.

It is also very important for the optimal values to be finite. Policies that do not meet the worst-case requirements (i.e., with infinite values for the expected utility) all have the same value, even though they might not perform equally badly.

## 4.5 Policy Computation and Pruning

The augmented problem previously described is a standard Markov Decision Process, where the objective is to maximize the total expected reward. Therefore, the optimal policy $\langle \pi \rangle_T^*$ can be computed using Dynamic Programming equations, as shown in Algorithm 1.

**Algorithm 1** Dynamic Programming equations for the augmented problem

$t \leftarrow T$
**for all** $s \in S$ **do**
  Initialize $\langle z \rangle_T^{*,T}(s,w) = U(w)$
**for** $t = T-1 \rightarrow 0$ **do**
  **for all** $s \in S$ **do**
    **for all** $w \in W^t$ **do**

$$\langle z \rangle_T^{*,t-1}(s,w) = \max_{a \in A_s} \sum_{s' \in S} P(s'|s,a)\left[\langle z \rangle_T^{*,t}(s', w + r(s,a,s'))\right]$$

Notice also that $\langle z \rangle_T^{*,t}(s,w) = -\infty$ whenever

$$w + d_{T-t}^*(s) \leq L \tag{9}$$

where $d_k^*(s)$ are the optimal max-min values defined by Equation (4). Intuitively, it means that $\langle z \rangle_T^{*,t}(s,w) = -\infty$ whenever we cannot meet the worst-case requirement, not even when optimizing the worst-case performance. Using condition (9), we can introduce additional pruning in a Forward Dynamic Programming algorithm, where the optimal values $\langle z \rangle_T^{*,t}(s,w)$ are recursively computed according to

$$\langle z \rangle_T^{*,t}(s,w) = \max_{a \in A_s} \sum_{s' \in S} P(s'|s,a)\left[\langle z \rangle_T^{*,t+1}(s', w + r(s,a,s'))\right]$$

with the two base cases:

$$\langle z \rangle_T^{*,t}(s,w) = \begin{cases} U(w) & \text{if } t = T \\ -\infty & \text{if } w + d_{T-t}^*(s) \leq L \end{cases}$$

once we precomputed the optimal max-min values $d_k^*(s)$, for all $s \in S$ and $0 \leq k \leq T$.

# 5. INFINITE HORIZON

For an infinite horizon planning problem, the *value* of a policy $\pi \in \Pi^{HR}$ is defined as

$$v_U^\pi(s) = \lim_{T \to \infty} v_{U,T}^\pi(s) = \lim_{T \to \infty} \mathbb{E}^{s,\pi}\left[U\left(\sum_{t=0}^{T} r_t\right)\right] \tag{10}$$

In general, the limit is neither guaranteed to exists nor to be finite, even in the standard case of a real-valued utility function [13].

However, in the special case of *Negative MDPs* (where all rewards are non-positive, i.e. $r(s,a,s') \leq 0$), we can prove the existence of the limit in Equation (10) in the general case of an extended value utility function. In fact, we already proved that the expectation exists for each $T$, and the existence of the limit derives from the monotonicity of the utility function $U$ and $w_T$.

As in the finite horizon case, it is crucial that the optimal values are finite, because otherwise plans cannot be compared in a meaningful way based on their expected utility. We therefore provide sufficient conditions that guarantee the finiteness of the optimal values.

We consider a special class of infinite horizon goal directed MDPs where there is a finite set of *goal states* $G \subseteq S$, where the agent stops to execute actions, and no longer receives rewards. Further, we restrict ourselves to the case of *negative MDPs*, where $r(s,a,s') \leq 0$.

## 5.1 Finiteness

Let's consider the infinite horizon version of the max-min problem previously described. Let

$$d^\pi(s) = \lim_{T \to \infty} d_T^\pi(s)$$

where $d_T^\pi(s)$ is defined according to Equation (5). Again, we can prove that the limit exists by monotonicity. Let $d^*(s) = \sup_\pi d^\pi(s)$ be the optimal worst-case value. By definition, under the optimal worst-case policy $\pi^{WC}$ we have $d^*(s) \leq w_T$ for all $T \geq 0$ and initial states $s$, so for each $T \geq 0$

$$U(d^*(s)) \leq v_{U,T}^{\pi^{WC}}(s) \leq v_{U,T}^*(s) \leq U(0) \tag{11}$$

It follows that if $U(d^*(s))$ is finite, then from (11) we have that $v_U^*(s)$ exists and is also finite. Intuitively, this condition means that the worst-case constraint encoded with the extended value utility function cannot be too restrictive, that is it cannot be more restrictive than what is possible to achieve using the optimal worst-case policy.

## 5.2 Properties of the Optimal Policy

We consider a *negative goal-directed MDP* that satisfies condition (11), with the additional condition $r(s,a,s') < 0$ for all $s \in S \setminus G$ (*strictly negative rewards*). Let

$$L = \sup\{w | U(w) = -\infty\} > -\infty$$

We show there exists an optimal policy for the Maximum Expected Utility objective that is *stationary*, deterministic and that depends on the history $h_t$ only through the accumulated reward $w_t$. Let

$$\overline{r} = \max_{s \in S \setminus G} \max_{a \in A_s} \max_{s' \in S} r(s,a,s') < 0,\ \overline{T} = \lceil L/|\overline{r}| \rceil + 1$$

The following result holds:

LEMMA 4. *For any policy $\pi \in \Pi^{HR}$ for the infinite horizon problem ($T = \infty$) with strictly negative rewards, for any state $s \in S$, and for all $T' \geq \overline{T}$,*

$$v^{\pi}_{U,T'}(s) = v^{\pi}_{U,\overline{T}}(s)$$

PROOF. Let $w_T$ be defined as in (1). By monotonicity, $v^{\pi}_{U,T'}(s) \leq v^{\pi}_{U,\overline{T}}(s)$. If $v^{\pi}_{U,\overline{T}}(s) = -\infty$, we are done. Otherwise, it must be $w_{\overline{T}} \geq L$ almost surely. By the definition of $\overline{T}$ and $L$, it must be the case that $s_{\overline{T}} \in G$ almost surely. This concludes the proof because $r_k(s_k, a_k, s_{k+1}) = 0$ almost surely for any $k \geq \overline{T}$, because we must have reached a goal state. □

Using Lemma 4 and taking limits, we have that for any policy $\pi \in \Pi^{HR}$ and for all initial states, $\lim_{T' \to \infty} v^{\pi}_{U,T'}(s) = v^{\pi}_{U,\overline{T}}(s)$. Therefore,

$$\sup_{\pi \in \Pi} \lim_{T' \to \infty} v^{\pi}_{U,T'}(s) = \sup_{\pi \in \Pi} v^{\pi}_{U,\overline{T}}(s) = v^{*}_{U,\overline{T}}(s)$$

which means that we can solve the infinite horizon problem by planning for a finite horizon of length $\overline{T}$, for instance using Algorithm 1. Using the results in Section 4 and Lemma 4, the optimal policy for the infinite horizon problem is deterministic and history-dependent (but the dependency on the history is only through the accumulated reward). Furthermore, we show it is stationary.

LEMMA 5. *For any state $s \in S$, and wealth level $w$ in the augmented MDP problem we have*

$$\langle z \rangle^{*,t_1}_{\overline{T}}(s,w) = \langle z \rangle^{*,t_2}_{\overline{T}}(s,w)$$

PROOF. Let $\langle \pi \rangle^{*}_{\overline{T}} = (\langle d_0 \rangle, \cdots, \langle d_{\overline{T}-1} \rangle)$ be the optimal policy for the augmented problem. By the optimality principle, $\langle z \rangle^{*,t}_{\overline{T}}(s',w') = \langle z \rangle^{*}_{\overline{T}-t}(s',w')$ . Without loss of generality, let $t_1 < t_2 \leq \overline{T}$. By previous theorems, monotonicity of $U$ and negative rewards assumption

$$\langle z \rangle^{*}_{\overline{T}-t_1}(s,w) = \mathbb{E}^{\langle s \rangle, \langle \pi \rangle^{*}_{\overline{T}-t_1}}[U(w + \sum_{k=0}^{T-1-t_1} r_k)] \leq$$

$$\mathbb{E}^{\langle s \rangle, \langle \pi \rangle^{*}_{\overline{T}-t_1}}[U(w + \sum_{k=0}^{T-1-t_2} r_k)] \leq \langle z \rangle^{*}_{\overline{T}-t_2}(s,w)$$

If $\langle z \rangle^{*,t_2}_{\overline{T}}(s,w) = -\infty$ then we are done. Otherwise, $\tilde{w}_{\overline{T}-t_2} = w + \sum_{k=0}^{\overline{T}-1-t_2} r_k \geq L$ almost surely when using policy $\langle \pi \rangle^{*}_{\overline{T}-t_2}$ from the initial state $(s,w)$. If $s \in G$ we are done because $\langle z \rangle^{*,t_1}_{\overline{T}}(s,w) = \langle z \rangle^{*,t_2}_{\overline{T}}(s,w) = U(w)$. Otherwise if $s$ is not a goal state and since $w \in W^{t_2}$, then it must be $w \leq t_2 * \overline{r}$. If $\tilde{x}_k$ is the state after $k$ steps when using policy $\langle \pi \rangle^{*}_{\overline{T}-t_2}$ from the initial state $(s,w)$, it must be $\tilde{x}_{\overline{T}-t_2} \in G$ almost surely, because otherwise the reward $\tilde{w}_{\overline{T}-t_2}$ would exceed the bound. Then the following policy $\langle \pi \rangle'_{\overline{T}-t_1} = (\langle d_{\overline{T}-t_2} \rangle, \cdots, \langle d_{\overline{T}-1} \rangle, \cdots)$ satisfies

$$\langle z \rangle^{*}_{\overline{T}-t_1}(s,w) \geq \langle z \rangle^{\langle \pi \rangle'_{\overline{T}-t_1}}_{\overline{T}-t_1}(s,w) = \langle z \rangle^{*,t_2}_{\overline{T}}(s,w)$$

because $\tilde{x}_{\overline{T}-t_2} \in G$ almost surely. □

As a corollary, the optimal policy for the augmented MDP is *stationary* and therefore the optimal policy for the original problem is also *stationary*. Intuitively, since we are given an infinite number of steps to reach the goal, the number of steps already taken does not affect the optimal policy.

Notice that under these assumptions (i.e., with a worst case constraints), we can compute the optimal policy for the augmented problem using Algorithm 1, which terminates in a bounded number of steps. In contrast, using the construction in [12], one has to reach a fixed-point using value-iteration procedures in a (countably) infinite state space, which in general requires some form of approximation.

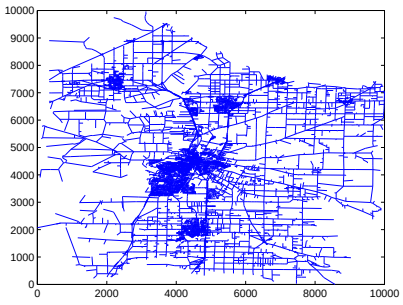



Figure 1: San Joaquin County Road Network

# 6. STOCHASTIC SHORTEST PATHS IN ROAD NETWORKS

In a Stochastic Shortest Path problem, a planner agent is given a graph with vertex set $V$ and edge set $E$, an initial node $s \in V$, and a set of goal nodes $G \subseteq V$. From a node $s \in V \setminus G$, the agent can move to any neighboring node (this is the set of available actions), but unlike standard shortest path problems, the cost of traversing an edge $e \in E$ is stochastic and modeled by a random variable $c_e$ with known probability distribution. The planner stops when a goal node $g \in G$ is reached, and no more costs are incurred. Given an utility function $U$ (see examples below), the goal of the agent is to find a plan that maximizes the expected utility of the total reward, which is defined as minus the total cost. Notice that the problem can be formulated as a finite MDP when the random variables $\{c_e, e \in E\}$ are discrete with finite sample space.

We consider a real-world road network [11] as the underlying graph (see Figure 1 for an example) in our experiments. The edge lengths $\{w_e, e \in E\}$ are also provided in the dataset. The edge costs $\{c_e\}$ model travel times, and are assumed to be discretized Beta-distributed random variables. This is a common modeling assumption for tasks with unknown duration in PERT analysis [6]. In particular, we

assume $c_e = m + (M - m)\mathbf{B}(\alpha, \beta)$ , where $\mathbf{B}$ follows a Beta distribution with shape parameters $\alpha$ and $\beta$. $M$ and $m$ are respectively the upper and lower bound on $c_e$, and are defined as follows:

$$M = (1 + u_1(e)/2)\, w_e$$
$$m = (1 - u_2(e)/2)\, w_e$$

where $u_1(e)$ and $u_2(e)$ are uniformly distributed in $[0, 1]$. The parameters $\alpha$ and $\beta$ are chosen such that the expected edge cost is equal to the edge length for each $e \in E$ (i.e. $E[c_e] = w_e$), with a variance chosen uniformly at random. Note that the rewards $\{r_e\}$ are a *discretized* version of $\{-c_e\}$, so that the finiteness MDP assumption holds.

## 6.1 Utility Functions for SSPs

In our experiments, we consider several types of utility functions. A simple linear utility function $U(x) = x$ leads to the standard maximization of the expected total reward. To maximize the expected total reward (equivalently, minimize expected travel time) with a worst case constraint we use

$$U_L(x) = \begin{cases} -\infty & x \leq L \\ x & x > L \end{cases} \tag{12}$$

In the *Stochastic On Time Arrival* formulations [5], also known as MDPs with Target-Level Utility Functions [12], the utility function has the form $U^K(x) = 1_{[K,+\infty)}(x)$ where $1_{\mathcal{A}}$ is an indicator function for the set $\mathcal{A}$. This corresponds to maximizing the probability of reaching a certain target reward $K$, since it holds that $\mathbb{E}[U^K(w_T)] = \mathbb{E}[1_{[K,+\infty)}(w_T)] = \mathbb{P}[w_T \geq K]$ To maximize the probability of having a total reward at least as large as $K$ with a guaranteed lower bound $L < K$, we can introduce an extended value utility function $U_{K,L}(x)$ that is defined to be $-\infty$ when $x \leq L$, and equal to $U^K(x)$ otherwise. When costs represent travel times, this corresponds to maximizing the probability of reaching the destination by a given deadline, with a worst case constraint. For instance, we might wish to use this criterion in order to maximize the probability of getting to the airport at least 3 hours before our flight departure, but no later than check-in closure time. We can also consider a more general case where the deadline is soft (as in [12]), because partial credit is given for being late, up to some point $D$. We introduce a worst-case constraint $L$ by using the following *extended-value* utility function:

$$U_{K,D,L}(x) = \begin{cases} 1 & K \leq x \\ (x - D)/(K - D) & D \leq x < K \\ 0 & L < x < D \\ -\infty & x \leq L \end{cases}$$

Finally, we consider a worst-case constrained exponential utility function $U_{\gamma,L}(x)$, by introducing a worst-case lower bound $L$ in the standard utility function $U_\gamma(x) = e^{\gamma x}$.

## 6.2 Results

For our experiments we use the San Joaquin County Road Network graph (with 18263 nodes and 23874 edges) represented in Figure 1. Every policy $\pi \in \pi^{HR}$ has an associated probability distribution for the total reward $w_T$. Assuming the costs $\{c_E\}$ represent travel times, this corresponds to a probability distribution for the total travel time $c_T = -w_T$.

For each utility function previously introduced, we compute the corresponding optimal policy with a worst case constraint $L$ using the forward Dynamic Programming method. The optimal max-min values $d_k^*(s)$ are precomputed solving a shortest path problem on the original graph with edge costs given by the worst-case realization. Given a fixed initial position $s \in V$ and destination node $G = \{g\}$, each optimal policy has a different associated probability distribution for the total travel time $c_T$ (notice that they are all optimal, but according to different criteria). In Figure 2, we compare the resulting probability distributions (obtained optimizing different performance metrics), and we also emphasize their worst-case realization (the dashed vertical line on the right). For comparison, we also provide the probability distribution corresponding to the optimal worst-case policy $\pi^{WC}$ (in red). For Markovian policies (such as $\pi^{WC}$), the probability distribution is computed exactly (by evaluating a convolution), while distributions associated with history-dependent policies are obtained by Monte Carlo sampling with 100,000 samples (in green).

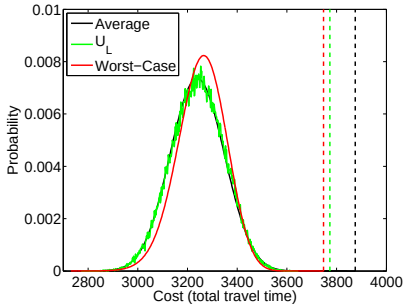
First, we compare the standard linear utility function $U(x) = x$ with its worst-case constrained version $U_L(x)$ defined as in Equation (12). In Figure 2a we see the results for a source-destination pair $s, g$ where we improve the worst-case realization of $w_T$, while at the same time maintaining the same expected value $\mathbb{E}[w_T]$. In other words, the policy for $U_L(x)$ dominates the one for $U(x) = x$ because it achieves the same expected value but it improves the worst-case performance. However, it is not always the case. In Figure 2b, we see that for a different source-destination pair, improving the worst-case realization of $w_T$ leads to a larger expected travel time $\mathbb{E}[c_T] = \mathbb{E}[-w_T]$.

Finally, in Figure 2c and 2d we plot the probability distributions corresponding to $U_{K,D,L}(x)$ (maximizing the probability of reaching the destination by a given soft deadline with a worst-case constraint $L$) and $U_{\gamma,L}(x)$. In both cases, the distributions are significantly different from the one obtained minimizing the expected travel time (in black). Notice that in Figure 2c the probability of reaching the destination by the deadline is significantly improved, and that there is a spike around $c_T = 3800$. This is because according to $U_{K,D,L}(x)$, any realization of $c_T$ larger than $-D$ has the same utility, as long as they satisfy the worst-case requirement. Similarly, the exponential utility function $U_{\gamma,L}(x)$ reflects a strong preference for small realizations of $c_T$. This can be seen in Figure 2d, where for instance the probability $P[c_T < 2900]$ of having a total travel time smaller than 2900 is 4 times larger when optimizing $U_{\gamma,L}(x)$ rather than $U(x) = x$ (area under the green and black curve, respectively). This experiment empirically demonstrates that when the planner cares about different objectives (e.g., a target level criterion), then augmented policies can achieve significant improvements over standard Markovian ones.

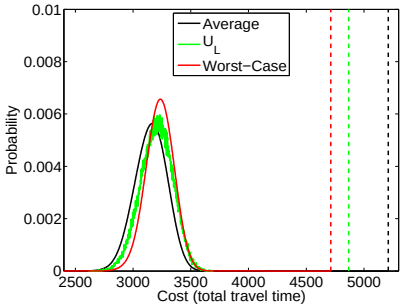
# 7. CONCLUSIONS

In this paper, we combined aspects of the two most widely used frameworks to model risk-sensitive planning, i.e. maximum expected utility and worst-case (*games against nature*) formulations. We introduced a new class of problems, where the goal is to maximize the expected utility of the total reward $w_T$, subject to a linear worst-case constraint on $w_T$. We showed how to encode this constraint using an *extended value utility function* in a maximum expected utility formulation, and we proved several results on the structure of the corresponding optimal policy.

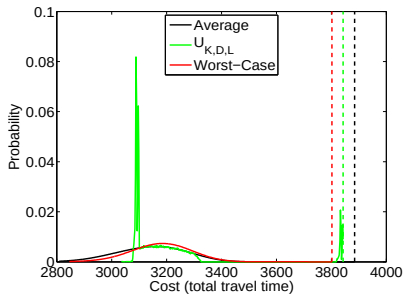
We showed that for finite planning horizons and for a class



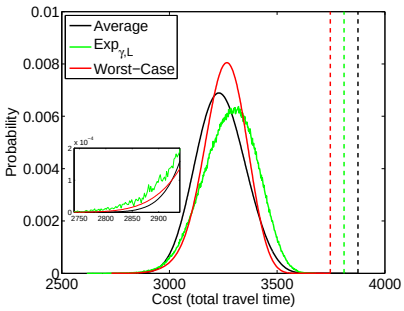

(a) Constrained linear, $L = -3773$




(b) Constrained linear, $L = -4867$




(c) Constrained soft deadline, $D = -3300$, $K = -3100$, $L = -3843$




(d) Constrained exponential, $\gamma = 0.02, L = -3811$

Figure 2: Resulting probability distributions and worst-case bounds (dashed lines). See pdf for colored version.

of infinite horizon problems, the optimal policy is deterministic and although not Markovian, it depends on the history only through the accumulated reward. Therefore, the policy can be represented as a set of functions (one for each state $s \in S$) of the total reward $w$, which, if necessary, can be approximated much more effectively than general history dependent decision rules, i.e. functions defined on the set of all possible histories $H_t$.

Although introducing non-linear utility functions allows the expression of a richer set of planning preferences, it increases the complexity because of the augmentation of the state space. However, adding worst-case constraints does not further increase the complexity, and allows us to speed up the policy search algorithm with additional pruning.

We think this type of formulation can be particularly useful for time-dependent problems where using the augmented space is unavoidable. For instance, in the GreenDriver App [2], they face SSPs where the edge costs probability distributions are dependent on the current time (essentially on $w_t$) because they model traffic lights. Although we used synthetic edge cost probability distributions, we showed our approach scales to large real-world networks, and it leads to significantly different plans with respect to traditional optimization criteria.

# 8. REFERENCES


[1] E. Altman. *Constrained Markov decision processes*. Chapman & Hall, 1999.

[2] J. Apple, P. Chang, A. Clauson, H. Dixon, H. Fakhoury, M. Ginsberg, E. Keenan, A. Leighton, K. Scavezze, and B. Smith. Green driver: AI in a microcosm. In *AAAI*, 2011.

[3] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2005.

[4] S. Ermon, J. Conrad, C. Gomes, and B. Selman. Playing games against nature: optimal policies for renewable resource allocation. *UAI*, 2010.

[5] Y. Fan, R. Kalaba, and J. Moore. Arriving on time. *J. Optimization Theory and Applications*, 127(3).

[6] N. Farnum and L. Stanton. Some results concerning the estimation of beta distribution parameters in PERT. *J. Operational Research Society*, 38(3):pp. 287–290, 1987.

[7] E. Feinberg and A. Shwartz. Constrained discounted dynamic programming. *Mathematics of Operations Research*, 21(4):922–945, 1996.

[8] M. Guay, T. Pham, D. Plotkin, S. Ermon, and A. Vladimirsky. Safer optimal routing in stochastic networks. *Unpublished technical report*, 2010.

[9] M. Heger. Consideration of risk in reinforcement learning. In *ICML*, volume 105, page 111, 1994.

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

[11] F. Li, D. Cheng, M. Hadjieleftheriou, G. Kollios, and S. Teng. On trip planning queries in spatial databases. *Advances in Spatial and Temporal Databases*.

[12] Y. Liu. *Decision-theoretic planning under risk-sensitive planning objectives*. PhD thesis, Georgia Institute of Technology Atlanta, GA, USA, 2005.

[13] Y. Liu. Existence and finiteness conditions for risk-sensitive planning: Results and conjectures. In *UAI*, 2005.

[14] Y. Liu and S. Koenig. Probabilistic planning with nonlinear utility functions. In *ICAPS*, pages 410–413, 2006.

[15] K. Ross and R. Varadarajan. Markov decision processes with sample path constraints: the communicating case. *Operations Research*, pages 780–790, 1989.

[16] J. Von Neumann, O. Morgenstern, A. Rubinstein, and H. Kuhn. *Theory of games and economic behavior*. Princeton Univ Pr, 2007.

[17] D. White. Mean, variance, and probabilistic criteria in finite Markov decision processes: a review. *J. Optimization Theory and Applications*, 56(1):1–29, 1988.

[18] S. Zilberstein, R. Washington, D. Bernstein, and A. Mouaddib. Decision-theoretic control of planetary rovers. *Adv. in Plan-Based Control of Robotic Agents*.