A Temporal Logic for Multi-Agent MDP's

Wojciech Jamroga
Clausthal University of Technology, Germany
wjamroga@in.tu-clausthal.de

ABSTRACT

Typical analysis of Markovian models of processes refers only to the expected utility that can be obtained by the process. On the other hand, modal logic offers a systematic method of characterizing processes by combining various modal operators. A multivalued temporal logic for Markov chains and Markov decision processes has been recently proposed in [8, 9]. Here, we discuss how it can be extended to the multi-agent case.

Keywords

Temporal logic, multi-agent Markov decision processes

1. INTRODUCTION

There are many different models of agents and multi-agent systems; however, most of them follow a similar pattern. First of all, they include information about possible situations (states of the system) that defines relations between states and and their external characteristics (essentially, "facts of life" that are true in these states). Second, they provide information about relationships between states (e.g, possible transitions between states).

Models that share this structure can be, roughly speaking, divided into two classes. *Qualitative models* provide no numerical measures for these relationships. *Quantitative models* assume that relationships are measurable, and provide numerical information about the degrees of relations. In [8, 9], we explored analogies between transition systems and Markovian models in order to provide a more expressive language for reasoning about, and specification of agents in stochastic environments. Here, we tentatively extend the framework to the multi-agent case.

We begin by summarizing our proposal of Markov temporal logic MTL from [8, 9]. Then, we point out that *multi-agent Markov decision processes* (MMDPs) [5] share many features with *concurrent game structures* (which are models of the popular strategic logic ATL [1]). In order to explore the similarity, we first propose a refinement of MMDPs that allows to model both quantitative and qualitative properties of a process. Then, we extend the syntax and semantics of MTL so that strategic abilities of groups of agents can be expressed.

Analysis of quantitative process models is usually based on the notion of expected reward. On the other hand, logical approaches are usually concerned with "limit properties" like the existence of an execution that displays a specific temporal pattern. We believe that both kinds of properties are interesting and worth using to describe processes. For instance, besides the expected value of cumulative future reward, we can ask of the maximal (or minimal) cumulative reward. Or, we might be concerned with the expected value of minimal guaranteed reward etc. A typical analysis of Boutilier's MMDPs is even more constrained, as we assume that all the agents in the system cooperate to achieve a common goal (i.e., maximize their common expected cumulative reward). Our extension allows to study the outcomes that can be obtained by *various* groups of agents.

In the context of multi-robot systems, the framework of Markov temporal logic can be used in several ways:

1. *Models* of MTL can include quantitative as well as qualitative properties. Thus, one can use the models to represent (and study) domains in which both measurable and non-measurable features are important.

2. Given a model of a system, *formulae* of Markov temporal logic can be used for verification of temporal and strategic properties of its components.

3. Perhaps more importantly, formulae of MTL can be used to define the intentional setting of the system. That is, one can use an MTL formula to specify the objective that is supposed to be pursued and the team of agents that is going to pursue it. Additionally, MTL allows to specify the anticipated behavior of the rest of agents.

   Given a model and a specification of the objective, the semantics of MTL provides unambiguous means for evaluation of available policies.

4. Formulae of MTL can be also used for specification of models and their components, since each formula of MTL defines a class of models in which the formula is valid.

In this paper, we focus especially on points 3 (the rescue mission example in Section 4) and 4 (Section 5).

2. MARKOV TEMPORAL LOGIC

In this section we recall Markov Temporal Logic (MTL) from [8, 9]. The logic allows for flexible reasoning about outcomes of agents acting in stochastic environments. The core of the logic is called MTL_0, and addresses outcomes of

AAMAS 2008 Workshop on Formal Models and Methods for Multi-Robot Systems, May 13 2008, Estoril, Portugal.

Markov chains. Intuitively, MTL$_0$ is a quantitative analogue of the branching-time logic CTL* [7]. MTL$_1$ can be used to reason about Markov decision processes, and extends MTL$_0$ with a strategic operator that refers to the outcome that is made by the "best" policy. In consequence, it can be seen as a quantitative analogue of the single-agent fragment of ATL* [1] with nondeterministic models.

2.1 Basic Models: Markov Chains and Markov Decision Processes

Typically, a Markov chain [12, 10] is a directed graph with probabilistic transition relation. In our definition, we include also a device for assigning states with utilities and/or propositional values. This is done through *utility fluents* which generalize atomic propositions in modal logic in the sense that they can take both numerical and qualitative truth values.

DEFINITION 1 (DOMAIN OF TRUTH VALUES). *A* domain $D = \langle U, \top, \bot, {}^- \rangle$ *consists of: (1) a set* $U \subseteq \mathbb{R}$ *of* utility values *(or simply* utilities*); (2) special values* $\top, \bot$ *standing for the logical truth and falsity, respectively;* $\hat{U} = U \cup \{\top, \bot\}$ *will be called the* extended utility set*; and, finally, (3) a complement function* $^- : \hat{U} \to \hat{U}$*. A domain should satisfy the conditions specified in [8, 9], omitted here for lack of space.*

DEFINITION 2 (MARKOV CHAIN). *A* Markov chain *over domain* $D = \langle U, \top, \bot, {}^- \rangle$*, and a set of utility fluents* Π *is a tuple* $M = \langle St, \tau, \pi \rangle$*, where:*

- St *is a set of* states *(we will assume that the set is finite and nonempty throughout the rest of the paper);*
- $\tau : St \times St \to [0, 1]$ *is a* stochastic transition relation *that assigns each pair of states* q_1, q_2 *with a probability* $\tau(q_1, q_2)$ *that, if the system is in* q_1*, it will change its state to* q_2 *in the next moment. For every* $q_1 \in St$*,* $\tau(q_1, \cdot)$ *is assumed to be a probability distribution, i.e.* $\sum_{q \in St} \tau(q_1, q) = 1$*.*
  By abuse of notation, we will sometimes write $\tau(q)$ *to denote the set of states accessible in one step from* q*, i.e.* $\{q' \mid \tau(q, q') > 0\}$*.*
- $\pi : \Pi \times St \to \hat{U}$ *is a* valuation of utility fluents.

A *run* in Markov chain M is an infinite sequence of states $q_0 q_1 \dots$ such that each q_{i+1} can follow q_i with a non-zero probability. The set of runs starting from state q is denoted by $\mathcal{R}_M(q)$.[1] Let $\lambda = q_0 q_1 ...$ be a run and $i \in \mathbb{N}_0$. Then: $\lambda[i] = q_i$ denotes the ith position in λ, and $\lambda[i..\infty] = q_i q_{i+1} \dots$ denotes the infinite subpath of λ from position i on.

Markov decision processes [4, 3] extend Markov chains with an explicit action structure: transitions are now connected to actions that generate them.

DEFINITION 3 (MARKOV DECISION PROCESS). *A* Markov decision process *over domain* $D = \langle U, \top, \bot, {}^- \rangle$*, and a set of utility fluents* Π *is a tuple* $\mathcal{M} = \langle St, Act, \tau, \pi \rangle$*, where:* St, π *are like in a Markov chain,* Act *is a nonempty finite set of actions, and* $\tau : St \times Act \times St \to [0, 1]$ *is a stochastic transition relation;* $\tau(q_1, \alpha, q_2)$ *defines the probability that, if the system is in* q_1 *and the agent executes* α*, the next state will be* q_2*. For every* $q \in St, \alpha \in Act$*, we assume that either (1)* $\tau(q, \alpha, q') = 0$ *for all* q' *(i.e.,* α *is not enabled in* q*), or (2)* $\tau(q, \alpha, \cdot)$ *is a probability distribution.*

Additionally, we define $act(q) = \{\alpha \in Act \mid \exists q'. \tau(q, \alpha, q') > 0\}$ *as the set of enabled actions in* q*.*

[1] If the model is clear from the context, the subscripts will be omitted.

A policy is a conditional plan that specifies future actions of the decision-making agent. Policies can be stochastic as well, thus allowing for randomness in the agent's play.

DEFINITION 4. *A* policy *(or* strategy*) in a Markov decision process* $\mathcal{M} = \langle St, Act, \tau, \pi \rangle$ *is a function* $s : States \times Act \to [0, 1]$ *that assigns each state* q *with a probability distribution over the enabled actions* $act(q)$*. That is,* $s(q, \alpha) \in [0, 1]$ *for all* $q \in St, \alpha \in act(q)$*, and* $\sum_{\alpha \in act(q)} s(q, \alpha) = 1$*. Values of* $s(q, \alpha)$ *for* $\alpha \notin act(q)$ *are irrelevant.*

The set of all policies in a model is denoted by Σ*.*

Note that, if the agent's policy is fixed, a Markov decision process reduces to a Markov chain.

DEFINITION 5. *Policy* $s : States \times Act \to [0, 1]$ instantiates MDP $\mathcal{M} = \langle St, Act, \tau, \pi \rangle$ *to a Markov chain* $\mathcal{M} \dagger s = \langle St', \tau', \pi' \rangle$ *with* $St' = St$*,* $\pi' = \pi$*, and* $\tau'(q, q') = \sum_{\alpha \in act(q)} s(q, \alpha)\ \tau(q, \alpha, q')$*.*

2.2 Logical operators as Minimizers and Maximizers

Note that – when truth values represent utility of an agent – temporal operators "sometime" and "always" have a very natural interpretation. "Sometime p" ($\Diamond p$) can be rephrased as "p is achievable in the future". Thus, under the assumption that agents want to obtain as much utility as possible, it is natural to view the operator as maximizing the utility value along a given temporal path. Similarly, "always p" ($\Box p$) can be rephrased as "p is guaranteed from now on". In other words, $\Box p$ asks for the minimal value of p on the path. On a more general level, every universal quantifier is essentially a minimizer of truth values, while existential quantifiers can be seen as maximizers. Thus, $\mathsf{A}\gamma$ ("for all paths γ") minimizes the utility specified by γ across all paths that can occur, etc. Also, conjunction and disjunction can be seen as a minimizer and a maximizer: $\varphi \vee \psi$ reads easily as "the utility that can be achieved through φ or ψ", while $\varphi \wedge \psi$ reads as "utility guaranteed by both φ and ψ".

2.3 MTL$_0$: A Logic of Markov Chains

Operators of MTL$_0$ include path quantifiers $\mathsf{E}, \mathsf{A}, \mathsf{M}$ for the maximal, minimal, and average outcome of a set of temporal paths, respectively, and temporal operators $\Diamond, \Box, \mathsf{m}$ for the maximal, minimal, and average outcome along a given path.[2] Propositional operators follow the same pattern: $\vee, \wedge, \oplus$ refer to maximization, minimization, and weighted average of outcomes obtained from different utility channels or related to different goals. Finally, we have the "defuzzification" operator $\preccurlyeq$, which provides a two-valued interface to the logic. $\varphi_1 \preccurlyeq \varphi_2$ yields "true" if the outcome of φ_1 is less or equal to φ_2, and "false" otherwise. Among other advantages, it allows to define the classical computational problems of validity, satisfiability and model checking for MTL.

[2] We allow to discount future outcomes with a discount factor c. Also, we introduce the "until" operator $\mathcal{U}$, which is more general than $\Diamond$.

Let $Bool(\omega) = \neg\omega \mid \omega \wedge \omega \mid \omega \oplus_c \omega \mid \omega \preccurlyeq \omega$ denote quasi-Boolean combinations of formulae of type ω. The syntax of MTL$_0$ can be defined by the following production rules:

$$\begin{aligned} \varphi &::= p \mid Bool(\varphi) \mid \mathsf{E}\gamma \mid \mathsf{M}\gamma, \\ \gamma &::= \varphi \mid Bool(\gamma) \mid \bigcirc_c \gamma \mid \Box_c \gamma \mid \gamma \,\mathcal{U}_c\, \gamma \mid \mathsf{m}_c\gamma, \end{aligned}$$

where $p \in \Pi$ is a utility fluent, and $c \in (0, 1]$ is a discount factor. Additionally, we define $\varphi_1 \cong \varphi_2 \equiv (\varphi_1 \preccurlyeq \varphi_2) \wedge (\varphi_2 \preccurlyeq \varphi_1)$. Boolean constants T, F (standing for "true" and "false"), disjunction, and the "sometime" temporal operator $\Diamond$ are defined in the standard way. We may also use the following shorthands for discount-free versions of temporal operators: $\bigcirc \equiv \bigcirc_1, \Diamond \equiv \Diamond_1, \Box \equiv \Box_1, \mathcal{U} \equiv \mathcal{U}_1$.

EXAMPLE 1. *Let* r *be a utility fluent that represents the immediate reward at each state. The following* MTL$_0$ *formulae define some interesting characteristics of a process:* $\mathsf{Mm}_{0.9}\mathsf{r}$ *(expected average reward with time discount* 0.9*),* $\mathsf{Am}_{0.9}\mathsf{r}$ *(guaranteed average reward with the same discount factor),* $\mathsf{M}\Box\mathsf{r}$ *(expected minimal undiscounted reward), and* $\mathsf{A}\Diamond\mathsf{r}$ *(guaranteed maximal reward).*

The main idea behind MTL$_0$ is that formulae can refer to both quantitative utilities and qualitative truth values. Thus, we treat complex formulae as fluents, just like the atomic utility fluents from Π, through a valuation function that assigns formulae with extended utility values from $\hat{U}$. Let $M = \langle St, \tau, \pi\rangle$ be a Markov chain over domain $D = \langle U, \top, \bot, \overline{\ }\rangle$ and a set of utility fluents Π. The valuation function $[\cdot]$ is defined below.

- $[p]_{M,q} = \pi(p, q)$, for $p \in \Pi$;
- $[\neg\varphi]_{M,q} = \overline{[\varphi]_{M,q}}$;
- $[\varphi_1 \wedge \varphi_2]_{M,q} = \min([\varphi_1]_{M,q}, [\varphi_2]_{M,q})$;
- $[\varphi_1 \oplus_c \varphi_2]_{M,q} = (1-c)\cdot[\varphi_1]_{M,q} + c \cdot [\varphi_2]_{M,q}$;
- $[\varphi_1 \preccurlyeq \varphi_2]_{M,q} = \top$ if $[\varphi_1]_{M,q} \le [\varphi_2]_{M,q}$ and $\bot$ else;
- $[\mathsf{E}\gamma]_{M,q} = \sup\{[\gamma]_{M,\lambda} \mid \lambda \in \mathcal{R}(q)\}$;
- The Markovian path quantifier $\mathsf{M}\gamma$ produces the expected truth value γ across all the possible runs, cf. [8, 9] for the formal construction;
- $[\varphi]_{M,\lambda} = [\varphi]_{M,\lambda[0]}$;
- $[\neg\gamma]_{M,\lambda}, [\gamma_1 \wedge \gamma_2]_{M,\lambda}, [\gamma_1 \oplus_c \gamma_2]_{M,\lambda}, [\gamma_1 \preccurlyeq \gamma_2]_{M,\lambda}$: analogous to Boolean combinations of "state formulae" φ;
- $[\bigcirc_c \gamma]_{M,\lambda} = c \cdot [\gamma]_{M,\lambda[1..\infty]}$;
- $[\Box_c\gamma]_{M,\lambda} = \inf_{i=0,1,\ldots}\{c^i[\gamma]_{M,\lambda[i..\infty]}\}$;
- $[\gamma_1 \,\mathcal{U}_c\, \gamma_2]_{M,\lambda} = \sup_{i=0,1,\ldots}$ $\{\min(\ \min_{0\le j<i}\{c^j[\gamma_1]_{M,\lambda[j..\infty]}\},\ c^i[\gamma_2]_{M,\lambda[i..\infty]})\}$;
- The Markovian temporal operator m_c produces the average discounted reward along the given run:

$$[\mathsf{m}_c\gamma]_{M,\lambda} = \begin{cases} (1-c)\sum_{i=0}^{\infty} c^i[\gamma]_{M,\lambda[i\ldots\infty]} & \text{if } c < 1 \\ \lim_{i\to\infty} \frac{1}{i+1}\sum_{j=0}^{i}[\gamma]_{M,\lambda[i\ldots\infty]} & \text{if } c = 1 \end{cases}$$

2.4 MTL$_1$: A Logic of Markov Decision Processes

In order to facilitate strategic reasoning about Markov decision processes, we use a strategic quantifier $\langle\langle a\rangle\rangle$, similar to the *cooperation modality* from alternating-time temporal logic ATL [1]. The intuitive meaning of $\langle\langle a\rangle\rangle\varphi$ is "the most that the decision maker can make out of φ".

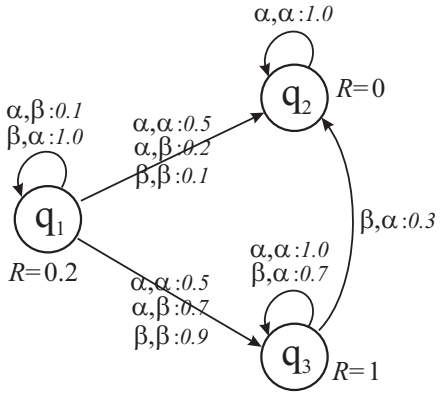

Figure 1: Simple MMDP with two agents $1, 2$

The syntax of MTL$_1$ is given by the following grammar:

$$\begin{aligned} \vartheta &::= p \mid Bool(\vartheta) \mid \langle\langle a\rangle\rangle\varphi, \\ \varphi &::= \vartheta \mid Bool(\varphi) \mid \mathsf{E}\gamma \mid \mathsf{M}\gamma, \\ \gamma &::= \varphi \mid Bool(\gamma) \mid \bigcirc_c \gamma \mid \Box_c\gamma \mid \gamma \,\mathcal{U}_c\, \gamma \mid \mathsf{m}_c\gamma. \end{aligned}$$

An example formula of MTL$_1$ is $\langle\langle a\rangle\rangle\mathsf{Amr}$ which maximizes the guaranteed average reward r with respect to available policies. Note that a is just a fixed symbol and not a parameter of the strategic operator.

Let $\mathcal{M} = \langle St, Act, \tau, \pi\rangle$ be a Markov decision process over domain $D = \langle U, \top, \bot, \overline{\ }\rangle$ and a set of utility fluents Π. The truth value of formulae in M is determined by the valuation function $[\cdot]$ that extends the valuation of MTL$_0$ formulae from Section 2.3 as follows:

- $[p]_{\mathcal{M},q} = \pi(p, q)$, for $p \in \Pi$;
- $[\neg\vartheta]_{\mathcal{M},q}, [\vartheta_1 \wedge \vartheta_2]_{\mathcal{M},q}, [\vartheta_1 \oplus_c \vartheta_2]_{\mathcal{M},q}, [\vartheta_1 \preccurlyeq \vartheta_2]_{\mathcal{M},q}$: analogous as for "state formulae" φ;
- $[\langle\langle a\rangle\rangle\varphi]_{\mathcal{M},q} = \sup\{[\varphi]_{M\dagger s,q} \mid s \in \Sigma\}$;
- $[\vartheta]_{\mathcal{M}\dagger s,q} = [\vartheta]_{\mathcal{M},q}$.

3. BEYOND MDP: THE MULTI-AGENT CASE

In the more general case, a system can include multiple agents/processes, interacting with each other. On the language level, we propose to extend the strategic operator $\langle\langle a\rangle\rangle$ to a family of operators $\langle\langle A\rangle\rangle$, parameterized with groups of agents A. Intuitively, $\langle\langle A\rangle\rangle\varphi$ refers to how much agents A can "make out of" φ by following their best joint policy. This yields a language similar to the alternating-time temporal logic ATL* from [1], albeit with strategic operators separated from path quantifiers.

On the semantic level, we observe the similarity between multi-agent Markov decision processes [5] and concurrent game structures [1] (cf. Figures 1 and 2). As models for our multi-agent MTL, we will therefore use a refinement of MMDPs similar to the versions of MDPs and Markov chains presented in Section 2.1. The semantics of $\langle\langle A\rangle\rangle\varphi$ is based on maximization of the value of φ with respect to A's joint strategies. We assume that the opponents play a strategy that minimizes φ most. This way, operator $\langle\langle A\rangle\rangle$ corresponds to the maxmin of the two-player game where A is the (collective) maximizer, and the rest of agents fills in the

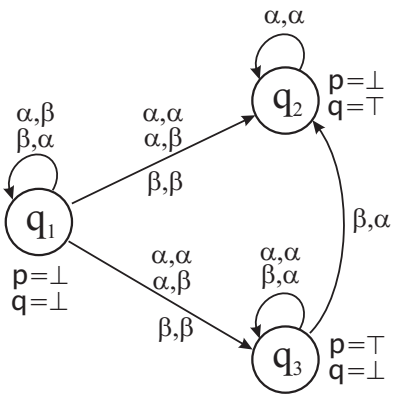

Figure 2: Simple concurrent game structure

role of the (collective) minimizer. Note that such a semantics entails that the opponents of A must also play only *memoryless* (i.e., Markovian) strategies.

3.1 MTL$_2$: Syntax

Let $\mathbb{A}$gt be the set of all agents. The only difference between the syntax of MTL$_2$ and the single-agent MTL$_1$ is that, instead of the single strategic operator $\langle\langle a\rangle\rangle$, we have now a family of operators $\langle\langle A\rangle\rangle$, one for each group of agents $A \subseteq \mathbb{A}$gt. $\langle\langle A\rangle\rangle\varphi$ maximizes the value of φ against the most dangerous response from $\mathbb{A}\text{gt} \setminus A$. In many cases, however, it is not appropriate to assume such a hostile play of $\mathbb{A}\text{gt} \setminus A$. As an alternative, we propose operator $(\mathbf{str}_a\, \xi)$, similar to the "strategic commitment" operator from [13]. $(\mathbf{str}_a\, \xi)\vartheta$ reads as: "suppose that agent a plays strategy ξ, then ϑ holds". Thus, $(\mathbf{str}_a\, \xi)$ assumes a particular strategy on the part of agent a. The strategy can be obtained e.g. by learning, statistical analysis, or game-theoretic rationality assumptions.

3.2 MTL$_2$: Semantics

The semantics of MTL$_2$ is defined for a version of multi-agent Markov decision processes that incorporate qualitative as well as quantitative atomic properties of states.

DEFINITION 6 (MMDP). *A* multi-agent Markov decision process *over domain* $D = \langle U, \top, \bot, \neg\rangle$, *a set of utility fluents* Π, *and a set of strategic terms* Ξ *is a tuple* $\mathcal{M} = \langle \mathbb{A}\text{gt}, St, \{Act_i\}_{i\in\mathbb{A}\text{gt}}, \tau, \pi, [\![\cdot]\!]\rangle$, *where:* $\mathbb{A}\text{gt} = \{1, \ldots, k\}$ *is the set of* agents, Act_i *is the set of* individual actions *of agent* i, *and* $Act = \prod_{i\in\mathbb{A}\text{gt}} Act_i$ *is the space of* joint actions *(*action profiles*). St, τ, π are like in a Markov decision process. The denotation of terms $[\![\cdot]\!]$ will be defined shortly.*

For a joint action α, we define α^i to denote agent i's individual part in α, and we extend the notation to sets of joint actions and agents. Also, let $\mathcal{A}$ be a set of action profiles, and α a collective action of agents A. Then, $\mathcal{A}|\alpha = \{\beta \in \mathcal{A} \mid \beta^A = \alpha\}$ is the set of action profiles that include α.

DEFINITION 7. *An* individual strategy *(*policy*)* $s_i : St \times Act_i \to [0, 1]$ *of agent* i *is defined as in Markov decision processes (only now it refers to* i*'s individual actions* Act_i*). The set of all* i*'s strategies is denoted by* Σ_i. *A* collective strategy s_A *for team* $A \subseteq \mathbb{A}$gt *is simply a tuple of individual strategies, one per agent from* A. *The set of all* A*'s collective strategies is given by* $\Sigma_A = \prod_{i\in A} \Sigma_i$. *The set of all* strategy profiles *in a model is given by* $\Sigma = \Sigma_{\mathbb{A}\text{gt}}$.

Now the denotation of strategic terms can be defined. $[\![\cdot]\!]$ *is a mapping that takes a strategic term* $\xi \in \Xi$ *and an agent* $i \in \mathbb{A}$gt, *and returns a strategy of* i, *that is,* $[\![\xi]\!]_i \in \Sigma_i$.

For a collective strategy s, we define s^i as the i's individual part in s. We also extend the notation to sets of agents.

DEFINITION 8. *Policy* $s \in \Sigma_A$ instantiates MMDP $\mathcal{M} = \langle \mathbb{A}\text{gt}, St, \{Act_i\}_{i\in\mathbb{A}\text{gt}}, \tau, \pi, [\![\cdot]\!]\rangle$ *to a simpler* MMDP $\mathcal{M} \dagger s = \langle \mathbb{A}\text{gt} \setminus A, St, \{Act_i\}_{i\in\mathbb{A}\text{gt}\setminus A}, \tau', \pi, [\![\cdot]\!]\rangle$ *with*

$$\tau'(q, \alpha, q') = \sum_{\alpha' \in (act(q)|\alpha)} \Big(\prod_{i\in A} s^i(q, \alpha')\Big)\ \tau(q, \alpha', q').$$

If $A = \mathbb{A}$gt, *then* s *instantiates* $\mathcal{M}$ *to a Markov chain.*

The semantics of MTL$_2$ formulae extends that of MTL$_1$ with the following clauses:

- $[\langle\langle A\rangle\rangle\varphi]_{\mathcal{M},q} = \sup_{s\in\Sigma_A} \inf_{t\in\Sigma_{\mathbb{A}\text{gt}\setminus A}} \{[\varphi]_{\mathcal{M}\dagger\langle s,t\rangle,q}\}$;
- $[(\mathbf{str}_a\, \xi)\vartheta]_{\mathcal{M},q} = [\vartheta]_{\mathcal{M}\dagger[\![\xi]\!],q}$.

Additionally, we can extend operator $(\mathbf{str}\,)$ to collective strategies:

$$(\mathbf{str}_{\{a_1,\ldots,a_r\}}\, \langle\xi_1, \ldots, \xi_r\rangle) = (\mathbf{str}_{a_1}\, \xi_1) \ldots (\mathbf{str}_{a_r}\, \xi_r).$$

EXAMPLE 2. *Consider the multi-agent Markov decision process from Figure 1. If the agents cooperate, they can maximize the expected achievable reward quite successfully, as* $[\langle\langle 1, 2\rangle\rangle \mathsf{M}\Diamond R]_{q_1} = 0.9$ *(best policy: both agents play* β *in* q_1 *with probability* 1*; the choices at other states are irrelevant). If agent* 1 *is to maximize the expected achievable reward on his own, against adversary behavior of agent* 2*, then he is bound to be less successful:* $[\langle\langle 1\rangle\rangle \mathsf{M}\Diamond R]_{q_1} = 0.6$. *(Note also that in this case* 1 *should employ a different policy, namely play* α *in* q_1 *with probability* 1*.) Finally, assuming random (instead of adversary) behavior of agent* 2 *improves* 1*'s rate of success only slightly:* $[(\mathbf{str}_2\, \xi_u)\langle\langle 1\rangle\rangle \mathsf{M}\Diamond R]_{q_1} = 0.68$*, and the best policy for* 1 *is again to play* α *in* q_1 *(and do anything at the other states).*

4. SPECIFICATION OF TEAMS AND OBJECTIVES

In this paper, we argue that modal logics of strategies and time have much to offer in terms of a specification language for multi-agent Markov processes. In particular, formulae like the ones presented in the previous sections can be used to *specify objectives* behind MMDPs. Note that the evaluation of formula $\langle\langle A\rangle\rangle\varphi$ is in fact underpinned by search for a policy for group A that maximizes the value of φ. Thus, with $\langle\langle A\rangle\rangle\varphi$ we specify both the objective function which is to be maximized (φ), and the team of agents that should perform the task (A).

An MMDP is just a structure of (abstract) agents, states, transitions, and local rewards. It is typically assumed that the global objective is to maximize the expected cumulative (or average) reward, perhaps with a temporal discount. Also, the team is assumed to consist of all the agents in the system. Here, we argue that there are other meaningful objectives for MMDPs, and that it makes sense to consider a subset of agents as the "proponents". We support our argument with a "rescue mission" example.

4.1 Example: Rescue Mission

The scenario is as follows: a group of k robots operates in a burning house in order to save people who are inside. There are n people inside and the house consists of m places. The state of each robot can be characterized by its status (alive or dead), current position, and an indication whether the robot is carrying some person (and, if so, which person). Similarly, a person can be characterized by its current status and position.[3] Each place can be burning, damaged, or still in a good shape. Regarding actions, robots and people that are alive can try to move North, South, East or West. Robots can additionally Pick up a person or Lay it on the ground. Every agent can also decide to do nothing (Nop). Two utility fluents are used: saved represents the percentage of people who are safely outside the building; robs refers to the percentage of robots that are still functioning.

The structure of states, actions, and fluents (including the domain of truth values D) is formally defined below.

$D = \langle [0,1], \top, \bot, ^- \rangle$ `with` $\top = 1$, $\bot = 0$, `and` $\overline{u} = 1 - u$;
$\mathbb{A}\mathrm{gt} = Robots \cup People$, `where`
$Robots = \{1, \ldots, k\}$, $People = \{k+1, \ldots, k+n\}$;
$St = \prod_{i=1}^{k+n} St_i \times PlStat^m$, `where`
$St_i = Places \times Status \times (People \cup \{nobody\})$ `for` $i \in Robots$,
$St_i = Places \times Status$ `for` $i \in People$,
$Places = \{1, \ldots, m, outside\}$, $Status = \{alive, dead\}$,
$PlStat = \{burning, damaged, OK\}$;
$Act = \prod_{i=1}^{k+n} Act_i$, `where`
$Act_i = \{N, S, E, W, Pick, Lay, Wait\}$ `for` $i \in Robots$,
$Act_i = \{N, S, E, W, Wait\}$ `for` $i \in People$;
$\Pi = \{\mathsf{saved}, \mathsf{robs}\}$;
`Let` $q = \langle q_1, \ldots, q_{k+n}, ps_1, \ldots, ps_m \rangle$, `and let` $\#S$ `denote the number of elements of set` S. `Then:`
$\pi(\mathsf{saved}, q) = \frac{\#\{i \in People | q_i = \langle outside, alive \rangle\}}{n}$,
$\pi(\mathsf{robs}, q) = \frac{\#\{i \in Robots | q_i = \langle \_, alive, \_ \rangle\}}{k}$.

The structure of transitions reflects events that can happen during the mission. For instance, a robot's attempt to go North should result with getting to the subsequent place with a high probability if there is a door (or open space) between the places and none of them is burning. In case there is fire in one of the places, the probability should be lower, and the probabilities that the robot becomes dead or staying in the same place should increase etc. We do not give the transitions explicitly here, but an example structure of this kind should be easy to imagine.

At least several different formulae of MTL$_2$ can be used to specify the operating team and its global objective (whose value is to be maximized):

- $\langle\langle Robots \rangle\rangle \mathsf{M}\Diamond\Box\mathsf{saved}$: if the team consists only of the robots (and people inside the house are just objects of the mission), then the robots should seek a policy which maximizes the expected percentage of people who will safely get out of the building (and stay there). Note that, indeed, we should *not* strive to maximize the expected *cumulative* utility (as is usually the case for MDPs): what we are interested in is getting most people out *eventually*, and the intermediary values of [saved] do not matter.
- $\langle\langle Robots \rangle\rangle \mathsf{M}\Diamond\mathsf{saved}$: the above formula can be further simplified if it is enough to get a person out alive (regardless of what happens to him/her afterwards).
- $\langle\langle Robots \rangle\rangle \mathsf{M}\Diamond_{0.95}\mathsf{saved}$: we can use a discount factor to favor more immediate results.
- $\langle\langle Robots \cup P \rangle\rangle \mathsf{M}\Diamond\mathsf{saved}$: the robots can be helped by a subset P of $People$.
- $(\mathbf{str}_{People}\, \xi_u)\langle\langle Robots \rangle\rangle \mathsf{M}\Diamond\mathsf{saved}$: in the previous specifications, we implicitly assumed that the "opponents" will behave in the worst possible way. It is usually more realistic to assume a more balanced pattern of behavior, e.g. the uniform distribution of actions.
- $\langle\langle Robots \rangle\rangle \mathsf{A}\Diamond\mathsf{saved}$: a politician's perspective. Before the elections, it may be a good idea to maximize the *guaranteed* number of the rescued (instead of going for the expected value).
- $(\mathbf{str}_{People}\, \xi_u)\langle\langle Robots \rangle\rangle (\mathsf{M}\Diamond\mathsf{saved} \oplus_{0.1} \mathsf{M}\Box\mathsf{robs})$: finally, keeping the robots themselves from destruction can be taken into account (although with much less importance than rescue of humans).

Note that the usual analysis of MMDPs is just a special case of what we can express with MTL. Namely, it can be specified by the MTL$_2$ formula $\langle\langle \mathbb{A}\mathrm{gt} \rangle\rangle \mathsf{M}\mathsf{m}_c\mathsf{r}$, where r is the fluent representing the local reward at each state, and c is the temporal discount value.

There are two problems with finding optimal strategies in such a setting, as the example clearly demonstrates. First, the complexity of models involving multiple agents is often highly prohibitive. Second, agents do not have perfect information about the current state of the system in most scenarios (i.e., observability is limited). However, both problems are inherent for MMDPs in general, and we do not discuss them further in this paper.

4.2 Some Notes on Specification of Objectives

In general, when team $A \subseteq \mathbb{A}\mathrm{gt}$ is supposed to maximize the objective expressed by an MTL$_0$ formula φ, we can consider 3 types of specifications, depending on what kind of behavior we expect from the rest of agents:

1. *Adversary behavior:* $\langle\langle A \rangle\rangle \varphi$ is used (which refers to the most harmful policy of $\mathbb{A}\mathrm{gt} \setminus A$);
2. *Collaborative behavior:* $\langle\langle \mathbb{A}\mathrm{gt} \rangle\rangle \varphi$ is used, since all the agents are *de facto* members of the team;
3. *Anticipated behavior:* $(\mathbf{str}_{\mathbb{A}\mathrm{gt} \setminus A}\, \xi)\langle\langle \mathbb{A}\mathrm{gt} \rangle\rangle \varphi$ is used, where ξ denotes the behavior of the "opponents" that we anticipate.

5. SPECIFICATION OF MMDP'S WITH MTL$_2$

In Section 4, we showed how various objectives can be specified *for a given multi-agent Markov decision process*. In this section, we take a different perspective, and show how MMDPs themselves can be specified. To this end, we first define what it means for a formula to be valid and/or satisfiable.

In particular, one can specify properties of strategies, thus imposing constraints on the denotation of strategic terms $[\![\cdot]\!]$. We show two important examples of such specifications in Section 5.2.

5.1 Levels of Truth

Since every domain must include a distinguished value for the classical (complete) truth, validity of formulae can be defined in a straightforward way.

[3] If it is being carried by a robot, the information will be included in the robot's state.

DEFINITION 9 (LEVELS OF VALIDITY). *Let $\mathcal{M}$ be a multi-agent Markov decision process, q a state in $\mathcal{M}$, and ϑ a formula of* MTL$_2$. *Then:*

- *ϑ is* true *in $\mathcal{M}, q$ (written $\mathcal{M}, q \models \vartheta$) iff $[\vartheta]_{\mathcal{M},q} = \top$.*
- *ϑ is* valid *in $\mathcal{M}$ (written $\mathcal{M} \models \vartheta$) iff it is true in every state of $\mathcal{M}$.*
- *ϑ is* valid *for multi-agent Markov decision processes (written $\models \vartheta$) iff it is valid in every* MMDP *$\mathcal{M}$.*

Definition 9 allows to define the typical decision problems for MTL$_2$ in a natural way:

- Given a formula ϑ, the *validity problem* asks if $\models \vartheta$;
- Given a formula ϑ, the *satisfiability problem* asks if there are $\mathcal{M}, q$ such that $\mathcal{M}, q \models \vartheta$;
- Given a model $\mathcal{M}$, state q and formula ϑ, the *model checking problem* asks if $\mathcal{M}, q \models \vartheta$.

For example, we can search for a model in which the guaranteed average reward r is at least 0.6 in the long run by solving the satisfiability problem for formula $0.6 \preccurlyeq \mathsf{Amr}$.

5.2 Characterization of Nash Equilibrium

Multi-agent Markov decision processes strictly generalize extensive game forms from game theory (in MMDPs, players act simultaneously, cycles are allowed, and payoffs/utilities can be defined for each state). Thus, extensive games can be seen as special instances of MMDPs.

DEFINITION 10. *By a* game model *we denote an acyclic finite connected* MMDP *$\mathcal{M}$.*

The state with no incoming transitions is called the initial state. *The "sink" states with no outgoing transitions are called* final states.[4] *Let St_F be the set of final states in $\mathcal{M}$. $\mathcal{M}$ includes utility fluents $\mathsf{u}_1, \ldots, \mathsf{u}_k$ such that $\pi(\mathsf{u}_i, q) \in U$ for $q \in St_F$, and $\pi(\mathsf{u}_i, q) = \bot$ otherwise. The fluents represent the payoffs for agents at the end of the game.*

For such models, we can use formulae of MTL$_2$ for a simple characterization of Nash equilibrium and subgame-perfect Nash equilibrium. First, formula $BR_i(\xi, \varphi)$ specifies that the i's strategy within ξ is the best response to the $\mathbb{A}\mathrm{gt} \setminus \{i\}$'s part of ξ if φ specifies the objective (utility) of agent i. Then, $NE(\xi, \varphi_1, \ldots, \varphi_k)$ says that no agent can unilaterally deviate to get a better payoff when i's payoff is defined by formula φ_i.

$$BR_i(\xi, \varphi) \;\equiv\; (\mathbf{str}_{\mathbb{A}\mathrm{gt}\setminus\{i\}}\, \xi[\mathbb{A}\mathrm{gt} \setminus \{i\}])\langle\langle i\rangle\rangle \varphi \;\preccurlyeq\; (\mathbf{str}_{\mathbb{A}\mathrm{gt}}\, \xi)\langle\langle\emptyset\rangle\rangle \varphi,$$

$$NE(\xi, \varphi_1, \ldots, \varphi_k) \;\equiv\; \bigwedge_{a \in \mathbb{A}\mathrm{gt}} BR_i(\xi, \varphi_i).$$

PROPOSITION 1. *Let $\mathcal{M}$ be a game model with initial state q_0. Then $\mathcal{M}, q_0 \models NE(\xi, \mathsf{M}\Diamond\mathsf{u}_1, \ldots, \mathsf{M}\Diamond\mathsf{u}_k)$ iff ξ denotes a Nash equilibrium in $\mathcal{M}$.*

A strategy profile is in subgame-perfect Nash equilibrium iff it is in NE for every subgame of the game:

$$SPN(\xi, \varphi_1, \ldots, \varphi_k) \;\equiv\; \langle\langle\emptyset\rangle\rangle\, \mathsf{A}\Box NE(\xi, \varphi_1, \ldots, \varphi_k).$$

[4]We can add loops at these states to make the model formally consistent with the definition of MMDPs.

PROPOSITION 2. *Let $\mathcal{M}$ be a game model with initial state q_0. Then $\mathcal{M}, q_0 \models SPN(\xi, \mathsf{M}\Diamond\mathsf{u}_1, \ldots, \mathsf{M}\Diamond\mathsf{u}_k)$ iff ξ denotes a subgame-perfect Nash equilibrium in $\mathcal{M}$.*

The above characterizations of Nash equilibrium and subgame-perfect Nash are straightforward adaptations of characterizations from [2, 13]. However, they are much more compact than there because we do not have to enumerate all possible payoff values. Moreover, they work also for games with chance moves and infinite sets of utility values.

6. CONCLUSIONS

We extend the Markov Temporal Logic MTL from [8, 9] to handle Markovian models with multiple agents acting in parallel. We show how formulae of the resulting logic can be used to define global objectives out of local reward values, and we demonstrate the potential on a "rescue mission" example. Finally, we discuss specification of multi-agent Markov decision processes themselves. In particular, we show that the new version of MTL can be used for compact characterizations of Nash equilibria and subgame-perfect Nash equilibria in extensive games.

7. REFERENCES

[1] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time Temporal Logic. *Journal of the ACM*, 49:672–713, 2002.

[2] A. Baltag. A logic for suspicious players. *Bulletin of Economic Research*, 54(1):1–46, 2002.

[3] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[4] R. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6:679–684, 1957.

[5] C. Boutilier. Sequential optimality and coordination in multiagent systems. In *Proceedings of IJCAI*, pages 478–485, 1999.

[6] L. de Alfaro, M. Faella, T. Henzinger, R. Majumdar, and M. Stoelinga. Model checking discounted temporal properties. *Theoretical Computer Science*, 345:139–170, 2005.

[7] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, pages 995–1072. Elsevier Science Publishers, 1990.

[8] W. Jamroga. Markov temporal logic. Technical Report IfI-07-11, Clausthal University of Technology, 2007.

[9] W. Jamroga. A temporal logic for markov chains. In *Proceedings of AAMAS'08*, 2008. To appear.

[10] J. G. Kemeny, L. J. Snell, and A. W. Knapp. *Denumerable Markov Chains*. Van Nostrand, 1966.

[11] A. Lluch-Lafuente and U. Montanari. Quantitative μ-calculus and CTL based on constraint semirings. *Electr. Notes Theor. Comput. Sci.*, 112:37–59, 2005.

[12] A. Markov. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 2(15):135–156, 1906.

[13] W. van der Hoek, W. Jamroga, and M. Wooldridge. A logic for strategic reasoning. In *Proceedings of AAMAS'05*, pages 157–164, 2005.