Game-Theoretic Agent Programming in Golog

Alberto Finzi [1] and **Thomas Lukasiewicz** [1 2]

Abstract. We present the agent programming language GTGolog, which integrates explicit agent programming in Golog with game-theoretic multi-agent planning in Markov games. It is a generalization of DTGolog to a multi-agent setting, where we have two competing single agents or two competing teams of agents. The language allows for specifying a control program for a single agent or a team of agents in a high-level logical language. The control program is then completed by an interpreter in an optimal way against another single agent or another team of agents, by viewing it as a generalization of a Markov game, and computing a Nash strategy. We illustrate the usefulness of this approach along a robotic soccer example.

1 INTRODUCTION

During the recent decades, the development of controllers for autonomous agents has become increasingly important in AI. One way of designing such controllers is the programming approach, where a control program is specified through a language based on high-level actions as primitives. Another way is the planning approach, where goals or reward functions are specified and the agent is given a planning ability to achieve a goal or to maximize a reward function.

Recently, an integration of the programming and the planning approach has been suggested through the language DTGolog [2], which integrates explicit agent programming in Golog [12] with decision-theoretic planning in Markov decision processes (MDPs) [10]. DTGolog allows for partially specifying a control program in a high-level language as well as for optimally filling in missing details through decision-theoretic planning. It can thus be seen as a decision-theoretic extension to Golog, where choices left to the agent are made by maximizing expected utility. From a different perspective, DTGolog can also be seen as a formalism that gives advice to a decision-theoretic planner, since it naturally constrains the search.

The agent programming language DTGolog, however, is designed only for the single-agent framework. That is, the model of the world essentially consists of a single agent that we control by a DTGolog program and the environment that is summarized in "nature". But in realistic applications, we often encounter multiple agents, which may compete against each other, or which may also cooperate with each other. For example, in robotic soccer, we have two competing teams of agents, where each team consists of cooperating agents. Here, the optimal actions of one agent generally depend on the actions of all the other ("enemy" and "friend") agents. In particular, there is a bidirected dependence between the actions of two different agents, which generally makes it inappropriate to model enemies and friends of the agent that we control simply as a part of "nature".

[1] DIS, Università di Roma "La Sapienza", Via Salaria 113, 00198 Rome, Italy; e-mail: {finzi, lukasiewicz}@dis.uniroma1.it.
[2] Institut für Informationssysteme, Technische Universität Wien, Favoritenstraße 9-11, 1040 Vienna, Austria; e-mail: lukasiewicz@kr.tuwien.ac.at.

The field of game theory [14] deals with optimal decision making in the multi-agent framework with competing and cooperating agents. In particular, Markov games [13, 6], also called stochastic games [8], generalize matrix games from game theory, and are multi-agent generalizations of MDPs with competing agents.

In this paper, we present a combination of explicit agent programming in Golog with game-theoretic multi-agent planning in Markov games. The main contributions of this paper are as follows:

- We define the language GTGolog, which integrates explicit agent programming in Golog with game-theoretic multi-agent planning in Markov games. GTGolog is a generalization of DTGolog [2] to a multi-agent setting, which allows for modeling two competing agents as well as two competing teams of cooperative agents.
- The language GTGolog allows for specifying a control program for a single agent or a team of agents, which is then completed by an interpreter in an optimal way against another single agent or another team of agents, by viewing it as a generalization of a Markov game, and computing a Nash equilibrium.
- We show that GTGolog generalizes Markov games. That is, GTGolog programs can represent Markov games, and a GTGolog interpreter can be used to compute their Nash equilibria. We also show that the GTGolog interpreter is optimal in the sense that it computes a Nash equilibrium of GTGolog programs.
- A robotic soccer example gives evidence of the usefulness of our approach in realistic applications. A first prototype implementation (in constraint logic programming) of a simple GTGolog interpreter for two competing teams of agents is given in [3].

The work closest in spirit to this paper is perhaps Poole's one [9], which shows that the independent choice logic can be used as a formalism for logically encoding games in extensive and normal form. Our view in this paper, however, is much different, as we aim at using game theory for optimal agent control in multi-agent systems.

Detailed proofs of all results are given in the extended paper [3].

2 PRELIMINARIES

In this section, we recall the basic concepts of the situation calculus and Golog, of matrix games, and of Markov games.

Situation calculus and Golog. The situation calculus [7, 12] is a first-order language for representing dynamic domains. Its main ingredients are *actions*, *situations*, and *fluents*. A situation is a first-order term encoding a sequence of actions. It is either a constant symbol or of the form $do(a, s)$, where a is an action and s is a situation. The constant symbol S_0 is the *initial situation*, and represents the empty sequence, while $do(a, s)$ encodes the sequence obtained from executing a after the sequence encoded in s. A *fluent* represents a world or agent property that may change when executing an ac-

tion. It is a predicate symbol whose last argument is a situation. E.g., $at(pos, s)$ may express that an agent is at position pos in situation s.

In the situation calculus, a dynamic domain is expressed by a *basic action theory* $BAT = (\Sigma, \mathcal{D}_{S_0}, \mathcal{D}_{ssa}, \mathcal{D}_{una}, \mathcal{D}_{ap})$, where

- Σ is the set of foundational axioms for situations;
- $\mathcal{D}_{una}$ is the set of unique name axioms for actions, which express that different action terms stand for different actions;
- $\mathcal{D}_{S_0}$ is a set of first-order formulas describing the initial state of the domain (represented by S_0);
- $\mathcal{D}_{ssa}$ specifies the *successor state axioms* [11, 12]: for each fluent $F(\vec{x}, s)$ we have an axiom of the form $F(\vec{x}, do(c, s)) \equiv \Phi_F(\vec{x}, c, s)$ providing both action effects and a solution to the frame problem (assuming deterministic actions). For example,

$$at(o, x, y, do(a, s)) \equiv a = moveTo(o, x, y) \vee at(o, x, y, s) \wedge \neg(\exists x', y') a = moveTo(o, x', y'); \quad (1)$$

- $\mathcal{D}_{ap}$ represents the action precondition axioms: for each simple action a, we have an axiom $Poss(a(\vec{x}), s) \equiv \Pi(\vec{x}, s)$.

Golog is an agent programming language that is based on the situation calculus. It allows for constructing complex actions from the primitive actions defined in a basic action theory BAT, where standard (and not so-standard) Algol-like control constructs can be used, in particular, (i) action sequences: $p_1; p_2$, (ii) tests: $\phi?$, (iii) nondeterministic action choices: $p_1|p_2$, (iv) nondeterministic choices of action argument: $(\pi x).p(x)$, and (v) conditionals, while loops, and procedure calls. An example of a Golog program is $\mathbf{while}\ \neg at(\boldsymbol{a}, 1, 2)\ \mathbf{do}\ (\pi x, y) moveTo(\boldsymbol{a}, x, y)$. Intuitively, the nondeterministic choice $(\pi x, y) moveTo(\boldsymbol{a}, x, y)$ is iterated until the agent $\boldsymbol{a}$ is at the position $(1, 2)$.

The semantics of a Golog program δ is specified by a situation-calculus formula $Do(\delta, s, s')$, which encodes that s' is a situation which can be reached from s by executing δ. Thus, Do represents a macro expansion to a situation calculus formula. For example, the action sequence is defined through $Do(p_1; p_2, s, s') = \exists s''(Do(p_1, s, s'') \wedge Do(p_2, s'', s'))$. For more details on the core situation calculus and Golog, we refer the reader to [12].

Matrix games. We briefly recall two-player matrix games from classical game theory [14]. Intuitively, they describe the possible actions of two agents and the rewards that they receive when they simultaneously execute one action each. Formally, a *two-player matrix game* $G = (A, O, R_a, R_o)$ consists of two nonempty finite sets of *actions* A and O for two agents $\boldsymbol{a}$ and $\boldsymbol{o}$, respectively, and two *reward functions* $R_a, R_o \colon A \times O \to \mathbf{R}$ for $\boldsymbol{a}$ and $\boldsymbol{o}$, respectively. The game G is *zero-sum* iff $R_a = -R_o$; we then often omit R_o.

A pure (resp., mixed) strategy specifies which action an agent should execute (resp., which actions an agent should execute with which probability). Formally, a *pure strategy* for agent $\boldsymbol{a}$ (resp., $\boldsymbol{o}$) is any action from A (resp., O). If agents $\boldsymbol{a}$ and $\boldsymbol{o}$ play the pure strategies $a \in A$ and $o \in O$, respectively, then they receive the *rewards* $R_a(a, o)$ and $R_o(a, o)$, respectively. A *mixed strategy* for agent $\boldsymbol{a}$ (resp., $\boldsymbol{o}$) is any probability distribution over A (resp., O). If agents $\boldsymbol{a}$ and $\boldsymbol{o}$ play the mixed strategies π_a and π_o, respectively, then the *expected reward* to agent $k \in \{\boldsymbol{a}, \boldsymbol{o}\}$ is $R_k(\pi_a, \pi_o) = \mathbf{E}[R_k(a, o) | \pi_a, \pi_o] = \sum_{a \in A,\, o \in O} \pi_a(a) \cdot \pi_o(o) \cdot R_k(a, o)$.

We are especially interested in pairs of mixed strategies (π_a, π_o), which are called Nash equilibria, where no agent has the incentive to deviate from its half of the pair, once the other agent plays the other half. Formally, (π_a, π_o) is a *Nash equilibrium* (or *Nash pair*) for G iff (i) for any mixed strategy π'_a, it holds $R_a(\pi'_a, \pi_o) \leq R_a(\pi_a, \pi_o)$, and (ii) for any mixed strategy π'_o, it holds $R_o(\pi_a, \pi'_o) \leq R_o(\pi_a, \pi_o)$. Every two-player matrix game G has at least one Nash pair among its mixed (but not necessarily pure) strategy pairs, and many two-player matrix games have multiple Nash pairs, which can be computed by linear complementary programming and linear programming in the general and the zero-sum case, respectively. A *Nash selection function* f associates with every two-player matrix game G a unique Nash pair $f(G) = (f_a(G), f_o(G))$. The expected reward to agent $k \in \{\boldsymbol{a}, \boldsymbol{o}\}$ under $f(G)$ is denoted by $v_f^k(G)$.

In the zero-sum case, if (π_a, π_o) and (π'_a, π'_o) are Nash pairs, then $R_a(\pi_a, \pi_o) = R_a(\pi'_a, \pi'_o)$, and also (π_a, π'_o) and (π'_a, π_o) are Nash pairs. That is, the expected reward to the agents is the same under any Nash pair, and Nash pairs can be freely "mixed" to form new Nash pairs. Here, agent $\boldsymbol{a}$'s strategies in Nash pairs are given by the optimal solutions of following linear program: $\max v$ subject to (i) $v \leq \sum_{a \in A} \pi(a) \cdot R_a(a, o)$ for all $o \in O$, (ii) $\sum_{a \in A} \pi(a) = 1$, and (iii) $\pi(a) \geq 0$ for all $a \in A$. Moreover, agent $\boldsymbol{a}$'s expected reward under a Nash pair is the optimal value of the above linear program.

Markov games. Markov games [13, 6], or also called stochastic games [8], generalize both matrix games and MDPs.

Roughly, a Markov game consists of a set of states S, a matrix game for every state $s \in S$, and a transition function that associates with every state $s \in S$ and combination of actions of the agents a probability distribution on future states $s' \in S$. We only consider the two-player case here. Formally, a *two-player Markov game* $G = (S, A, O, P, R_a, R_o)$ consists of a finite nonempty set of states S, two finite nonempty sets of actions A and O for two agents $\boldsymbol{a}$ and $\boldsymbol{o}$, respectively, a transition function $P \colon S \times A \times O \to PD(S)$, where $PD(S)$ denotes the set of all probability functions over S, and two *reward functions* $R_a, R_o \colon S \times A \times O \to \mathbf{R}$ for $\boldsymbol{a}$ and $\boldsymbol{o}$, respectively. G is *zero-sum* iff $R_a = -R_o$; we then often omit R_o.

Assuming a finite horizon $H \geq 0$, a pure (resp., mixed) time-dependent policy associates with every state $s \in S$ and number of steps to go $h \in \{0, \ldots, H\}$ a pure (resp., mixed) matrix-game strategy. Formally, a *pure policy* α (resp., ω) for agent $\boldsymbol{a}$ (resp., $\boldsymbol{o}$) assigns to each state $s \in S$ and number of steps to go $h \in \{0, \ldots, H\}$ an action from A (resp., O). The *H-step reward* to agent $k \in \{\boldsymbol{a}, \boldsymbol{o}\}$ under a start state $s \in S$ and the pure policies α and ω, denoted $G_k(H, s, \alpha, \omega)$, is defined as $R_k(s, \alpha(s, 0), \omega(s, 0))$, if $H = 0$, and $R_k(s, \alpha(s, H), \omega(s, H)) + \sum_{s' \in S} P(s' | s, \alpha(s, H), \omega(s, H)) \cdot G_k(H-1, s', \alpha, \omega)$, otherwise. A *mixed policy* π_a (resp., π_o) for $\boldsymbol{a}$ (resp., $\boldsymbol{o}$) assigns to every state $s \in S$ and number of steps to go $h \in \{0, \ldots, H\}$ a probability distribution over A (resp., O). The *expected H-step reward* to agent k under a start state s and the mixed policies π_a and π_o, denoted $G_k(H, s, \pi_a, \pi_o)$, is $\mathbf{E}[R_k(s, a, o) \mid \pi_a(s, 0), \pi_o(s, 0)]$, if $H = 0$, and $\mathbf{E}[R_k(s, a, o) + \sum_{s' \in S} P(s' \mid s, a, o) \cdot G_k(H-1, s', \pi_a, \pi_o) \mid \pi_a(s, H), \pi_o(s, H)]$, otherwise.

The notion of a finite-horizon Nash equilibrium for a Markov game is then defined as follows. A pair of mixed policies (π_a, π_o) is a *Nash equilibrium* (or *Nash pair*) for G iff (i) for any start state s and any π'_a, it holds $G_a(H, s, \pi'_a, \pi_o) \leq G_a(H, s, \pi_a, \pi_o)$, and (ii) for any start state s and any π'_o, it holds $G_o(H, s, \pi_a, \pi'_o) \leq G_o(H, s, \pi_a, \pi_o)$. Every two-player Markov game G has at least one Nash pair among its mixed (but not necessarily pure) policy pairs, and it may have exponentially many Nash pairs.

Nash pairs for G can be computed by finite value iteration from local Nash pairs of two-player matrix games as follows [5]. We assume an arbitrary Nash selection function f for two-player matrix games with the action sets A and O. For every state $s \in S$ and number of steps to go $h \in \{0, \ldots, H\}$, the two-player matrix game $G[s, h] = (A, O, Q_a[s, h], Q_o[s, h])$ is defined by $Q_k[s, 0](a, o) =$

$R_k(s,a,o)$ and $Q_k[s,h](a,o) = R_k(s,a,o) + \sum_{s' \in S} P(s'|s,a,o) \cdot v_f^k(G[s',h-1])$ for all $a \in A$, $o \in O$, and $k \in \{\boldsymbol{a}, \boldsymbol{o}\}$. Let the mixed policy π_k be defined by $\pi_k(s,h) = f_k(G[s,h])$ for all $s \in S$, $h \in \{0,\ldots,H\}$, and $k \in \{\boldsymbol{a}, \boldsymbol{o}\}$. Then, $(\pi_{\boldsymbol{a}}, \pi_{\boldsymbol{o}})$ is a Nash pair of G, and $G_k(H,s,\pi_{\boldsymbol{a}},\pi_{\boldsymbol{o}}) = v_f^k(G(s,H))$ for all $k \in \{\boldsymbol{a}, \boldsymbol{o}\}$ and $s \in S$.

In the case of zero-sum G, by induction on $h \in \{0,\ldots,H\}$, it is easy to see that every $G[s,h]$, $s \in S$ and $h \in \{0,\ldots,H\}$, is also zero-sum. Moreover, all Nash pairs that are computed by the above finite value iteration produce the same expected H-step reward, and they can be freely "mixed" to form new Nash pairs.

3 GAME-THEORETIC GOLOG

In this section, we present the language GTGolog for the case of two competing agents. We first describe the domain theory and the syntax of GTGolog programs. We then define the semantics of GTGolog programs and provide representation and optimality results.

Domain theory. GTGolog programs are interpreted w.r.t. a background action theory AT and a background optimization theory OT.

The background action theory AT is an extension of the basic action theory BAT of Section 2, where we also allow for stochastic actions. We assume two (zero-sum) competing agents $\boldsymbol{a}$ and $\boldsymbol{o}$ (called *agent* and *opponent*, respectively, where the former is under our control, while the latter is not). A *two-player action* is either an action a for agent $\boldsymbol{a}$, or an action b for agent $\boldsymbol{o}$, or two parallel actions $a \| b$, one for each agent. For example, $move(\boldsymbol{a}, M)$, $move(\boldsymbol{o}, O)$, and $move(\boldsymbol{a}, M) \| move(\boldsymbol{o}, O)$ are two-player actions. Note that we introduce parallel actions as primitives to simplify the presentation, more complex parallel actions can be easily treated as well, deploying the concurrent version of the situation calculus [12].

Analogously to [2], we represent stochastic actions by means of a finite set of deterministic actions. When a stochastic action is executed, then "nature" chooses and executes with a certain probability exactly one of its deterministic actions. We use the predicate $stochastic(a,s,n)$ to associate the stochastic action a with the deterministic action n in situation s, and we use the function $prob(a,n,s) = p$ to encode that "nature" chooses n in situation s with probability p. A stochastic action s is indirectly represented by providing a *successor state axiom* for each associated nature choice n. Thus, BAT is extended to a probabilistic setting in a minimal way. For example, consider the stochastic action $moveS(a,x,y)$ such that $stochastic(moveS(a,x,y), s, moveTo(a,x,y')) \equiv 0 \leq y' - y \leq 1$, that is, $moveS(a,x,y)$ can slide to $y+1$. We specify $moveS$ by defining the precondition of $moveTo(a,x,y')$ and assuming the successor state axiom (1). Furthermore, in order to specify the probability distribution over the deterministic components, we define $prob(moveS(a,x,y), moveTo(a,x,y), s) = 0.9$ and $prob(moveS(a,x,y), moveTo(a,x,y+1), s) = 0.1$.

Like [2], we assume that the domain is *fully observable*. To this end, we introduce *observability axioms*, which disambiguate the state of the world after executing a stochastic action. For example, after executing $moveS(a,x,y)$, we test the predicates $at(a,x,y,s)$ and $at(a,x,y+1,s)$ to check which of the deterministic components was executed (that is, $n{=}moveTo(a,x,y)$ or $n{=}moveTo(a,x,y{+}1)$). This condition is denoted by the predicate $condSta(a,n,s)$, e.g., $condSta(moveS(a,x,y), moveTo(a,x,y+1), s) \equiv at(a,x,y+1,s)$. Analogous observability axioms are needed to observe which actions the opponent and the agent have chosen. For this purpose, we introduce the predicate $cond(a,s)$, for example, $cond(moveTo(o,x,y),s) \equiv at(o,x,y,s)$.

The optimization theory OT specifies a reward and a utility function. The former associates with every situation s and two-player action α, a reward to agent $\boldsymbol{a}$, denoted $reward(\alpha, s)$, e.g., $reward(moveTo(a,x,y),s) = y$. As we assume that the rewards to $\boldsymbol{a}$ and $\boldsymbol{o}$ are zero-sum, we need not explicitly specify the reward to $\boldsymbol{o}$. The utility function maps every reward and success probability to a real-valued utility $utility(v,pr)$. We assume that $utility(v,1) = v$ for all v. An example is $utility(v,pr) = v \cdot pr$. The utility function suitably mediates between the agent reward and the failure of actions due to unsatisfied preconditions.

Syntax. Given the two-player actions represented by the domain theory, *programs* p in GTGolog are inductively built using the following constructs (ϕ is a condition, and p_1 and p_2 are programs):

1. *Deterministic or stochastic action:* α (a two-player action).
2. *Nondeterministic action choice:* $\alpha|...|\beta$. Execute α or ... or β, where α,..., β are two-player actions.
3. *Test action*: $\phi?$. Test the truth value of ϕ in the current situation.
4. *Nondeterministic choice of an argument*.
5. *Action sequence*: $p_1; p_2$. Do program p_1 followed by program p_2.
6. *Conditionals*: $\textbf{if } \phi \textbf{ then } p_1 \textbf{ else } p_2$.
7. *While loops*: $\textbf{while } \phi \textbf{ do } p_1$.
8. *Nondeterministic iteration*: ${p_1}^*$. Execute p_1 zero or more times.
9. *Procedures, including recursion.*

To clearly distinguish between the choices of the agent $\boldsymbol{a}$ and the opponent $\boldsymbol{o}$, we use $\textbf{choice}(\boldsymbol{a}: a_1|\ldots|a_n) \,\|\, \textbf{choice}(\boldsymbol{o}: o_1|\ldots|o_m)$ to denote $(\boldsymbol{a}{:}a_1 \| \boldsymbol{o}{:}o_1) | (\boldsymbol{a}{:}a_2 \| \boldsymbol{o}{:}o_1) | \cdots | (\boldsymbol{a}{:}a_n \| \boldsymbol{o}{:}o_m)$.

Semantics. The semantics of a GTGolog program p w.r.t. AT and OT is defined through the predicate $DoG(p,s,h,\pi,v,pr)$. Here, we have given as input the program p, a situation s, and a finite horizon $h \geq 0$. The predicate DoG then determines a strategy π for both agents $\boldsymbol{a}$ and $\boldsymbol{o}$, the reward to agent $\boldsymbol{a}$ under this strategy π, and the success probability $pr \in [0,1]$ of π. Note that due to the finite horizon, if the program p fails to terminate before the horizon h is reached, then it is stopped, and the best partial strategy is returned. Intuitively, our aim is to control agent $\boldsymbol{a}$, which is given the strategy π that DoG computes for program p, and which then executes its part of π. We define $DoG(p,s,h,\pi,v,pr)$ by induction as follows:

1. Zero horizon and null program:

$$DoG(p,s,0,\pi,v,pr) =_{def} \pi = Nil \wedge v = 0 \wedge pr = 1$$
$$DoG(Nil,s,h,\pi,v,pr) =_{def} \pi = Nil \wedge v = 0 \wedge pr = 1$$

   Intuitively, p ends when it is null or the horizon end is reached.

2. Deterministic first program action:

$$\begin{aligned} &DoG(a;p,s,h,\pi,v,pr) =_{def} \\ &\quad \neg Poss(a,s) \wedge \pi = Stop \wedge v = 0 \wedge pr = 0 \vee \exists \pi', v' : \\ &\quad Poss(a,s) \wedge DoG(p, do(a,s), h{-}1, \pi', v', pr) \wedge \\ &\qquad \pi = a;\pi' \wedge v = v' {+} reward(s,a) \end{aligned}$$

   Informally, if a is not executable, then p stops with success probability 0. As in [2], $Stop$ is a fictitious action of zero-cost, which stops the program execution. If a is executable, then the optimal execution of $a;p$ in s depends on that one of p in $do(a,s)$.

3. Stochastic first program action (nature choice):

$$\begin{aligned} &DoG(a;p,s,h,\pi,v,pr) =_{def} \\ &\quad \neg Poss(a,s) \wedge \pi = Stop \wedge v = 0 \wedge pr = 0 \vee \exists \pi_q, v_q, {pr}_q : \\ &\quad Poss(a,s) \wedge \textstyle\bigwedge_{q=1}^{k} DoG(n_q;p,s,h,n_q;\pi_q,v_q,{pr}_q) \wedge \\ &\quad \pi = a; \textbf{if } \phi_1 \textbf{ then } \pi_1 \textbf{ else if } \phi_2 \textbf{ then } \pi_2 \ \ldots \\ &\qquad\qquad \textbf{else if } \phi_k \textbf{ then } \pi_k \wedge \\ &\quad v = \textstyle\sum_{q=1}^{k} v_q \cdot prob(a,n_q,s) \wedge pr = \sum_{q=1}^{k} {pr}_q \cdot prob(a,n_q,s) \end{aligned}$$

where $\{n_q | 1 \leq q \leq k\}$ are the nature choices associated to a (with the same reward as a), and $\phi_1, \ldots, \phi_k$ are the relative conditions (represented by the *observability axioms*). The generated strategy is a conditional plan, that is, a cascade of if-then-else statements, where each possible stochastic action execution is considered.

4. Nondeterministic first program action (choice of agent $\boldsymbol{a}$):

$$DoG(\mathbf{choice}(\boldsymbol{a}\colon a_1|\ldots|a_n); p, s, h, \pi, v, pr) =_{def}$$
$$\exists \pi_i, v_i, pr_i\colon \textstyle\bigwedge_{i=1}^{n} DoG(\boldsymbol{a}{:}a_i; p, s, h, \pi_i, v_i, pr_i) \wedge$$
$$utility(v_k, pr_k) = \max\{utility(v_i, pr_i) \mid i \in \{1, \ldots, n\}\} \wedge$$
$$\pi = \pi_k \wedge v = v_k \wedge pr = pr_k$$

Given several possible actions for agent $\boldsymbol{a}$, the best action is the one where the action execution has the best utility.

5. Nondeterministic first program action (choice of opponent $\boldsymbol{o}$):

$$DoG(\mathbf{choice}(\boldsymbol{o}\colon o_1|\ldots|o_m); p, s, h, \pi, v, pr) =_{def}$$
$$\exists \pi_j, v_j, pr_j\colon \textstyle\bigwedge_{j=1}^{m} DoG(\boldsymbol{o}{:}o_j; p, s, h, \pi_j, v_j, pr_j) \wedge$$
$$\pi = \mathbf{if}\ \psi_1\ \mathbf{then}\ \pi_1\ \mathbf{else\ if}\ \psi_2\ \mathbf{then}\ \pi_2\ \ldots$$
$$\mathbf{else\ if}\ \psi_m\ \mathbf{then}\ \pi_m \wedge$$
$$utility(v_k, pr_k) = \min\{utility(v_j, pr_j) \mid j \in \{1, \ldots, m\}\} \wedge$$
$$v = v_k \wedge pr = pr_k$$

Informally, agent $\boldsymbol{a}$ assumes a rational behavior of $\boldsymbol{o}$, which is connected to its minimal reward (we consider a zero-sum setting). The ψ_i's are the conditions (defined by the *observability axioms*) that agent $\boldsymbol{a}$ has to test to observe the choice of opponent $\boldsymbol{o}$.

6. Nondeterministic first program action (choice of both $\boldsymbol{a}$ and $\boldsymbol{o}$):

$$DoG(\mathbf{choice}(\boldsymbol{a}\colon a_1|\ldots|a_n)\|\mathbf{choice}(\boldsymbol{o}\colon o_1|\ldots|o_m);$$
$$p, s, h, \pi, v, pr) =_{def} \exists \pi_{i,j}, v_{i,j}, pr_{i,j}, \pi_{\boldsymbol{a}}, \pi_{\boldsymbol{o}}\colon$$
$$\textstyle\bigwedge_{i=1}^{n}\bigwedge_{j=1}^{m} DoG(\boldsymbol{a}{:}a_i\|\boldsymbol{o}{:}b_j; p, s, h, \boldsymbol{a}{:}a_i\|\boldsymbol{o}{:}b_j; \pi_{i,j}, v_{i,j}, pr_{i,j}) \wedge$$
$$(\pi_{\boldsymbol{a}}, \pi_{\boldsymbol{o}}) = selectNash(\{r_{i,j} = utility(v_{i,j}, pr_{i,j}) \mid i, j\}) \wedge$$
$$\pi = \pi_{\boldsymbol{a}}\|\pi_{\boldsymbol{o}}; \mathbf{if}\ \phi_1 \wedge \psi_1\ \mathbf{then}\ \pi_{1,1}\ \mathbf{else\ if}\ \phi_2 \wedge \psi_1\ \mathbf{then}\ \pi_{2,1}\ \ldots$$
$$\mathbf{else\ if}\ \phi_n \wedge \psi_m\ \mathbf{then}\ \pi_{n,m} \wedge$$
$$v = \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m} v_{i,j} \cdot \pi_{\boldsymbol{a}}(a_i) \cdot \pi_{\boldsymbol{o}}(o_j) \wedge$$
$$pr = \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m} pr_{i,j} \cdot \pi_{\boldsymbol{a}}(a_i) \cdot \pi_{\boldsymbol{o}}(o_j)$$

Intuitively, we compute a Nash strategy by finite horizon value iteration for Markov games [5]. For each possible pair of action choices, the optimal strategy is calculated. Then, a Nash strategy is locally extracted from a matrix game by using the function $selectNash$. Here, $\pi_{\boldsymbol{a}}$ and $\pi_{\boldsymbol{o}}$ are probability distributions over $\{a_1, \ldots, a_n\}$ and $\{o_1, \ldots, o_m\}$, respectively. Moreover, ψ_i and ϕ_j are the conditions defined by the *observability axioms* to observe what $\boldsymbol{a}$ and $\boldsymbol{o}$, respectively, have actually executed.

7. Test action:

$$DoG(\phi?; p, s, h, \pi, v, pr) =_{def} \phi[s] \wedge DoG(p, s, h, \pi, v, pr) \vee$$
$$\neg\phi[s] \wedge \pi = Stop \wedge v = 0 \wedge pr = 0$$

8. The semantics of nondeterministic iterations, conditionals, while loops, procedures, argument selection, and associate sequential composition is defined in the standard way.

Representation and optimality results. The following result shows that every zero-sum two-player Markov game can be represented in GTGolog, and that DoG computes one of its finite-horizon Nash equilibria and its expected finite-step reward.

Theorem 3.1 *Let $G = (S, A, O, P, R_{\boldsymbol{a}})$ be a zero-sum two-player Markov game, and let $H \geq 0$ be a horizon. Then, there exists an action theory AT, an optimization theory OT, and a GTGolog program p relative to them such that $\vec{\pi} = (\vec{\pi}_{\boldsymbol{a}}, \vec{\pi}_{\boldsymbol{o}})$ is a Nash equilibrium for G, where $\vec{\pi}_k(s, h) = \pi_k$, $k \in \{\boldsymbol{a}, \boldsymbol{o}\}$, is given by $DoG(p, s, h{+}1, \pi_{\boldsymbol{a}}\|\pi_{\boldsymbol{o}}; \pi', v, pr)$ for every $s \in S$ and $h \in \{0, \ldots, H\}$. Furthermore, for every $H \geq 0$ and $s \in S$, it holds that $G(H, s, \vec{\pi}_{\boldsymbol{a}}, \vec{\pi}_{\boldsymbol{o}}) = v$ is given by $DoG(p, s, H{+}1, \pi_{\boldsymbol{a}}\|\pi_{\boldsymbol{o}}; \pi', v, pr)$.*

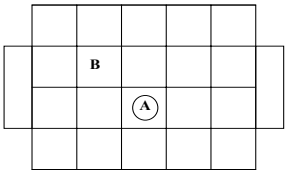

Figure 1. Soccer Example

We next show that DoG produces optimal results. Given a finite horizon $H \geq 0$, a *strategy* π for a GTGolog program p is obtained from the H-horizon part of p by replacing agent and opponent choices by single actions, and choices of both by probability distributions over their actions. The notions of an *expected H-step reward* $G(p, s, H, \pi)$, with a situation s, and of a finite-horizon *Nash equilibrium* can then be defined in a straightforward way as for Markov games. The next theorem shows that DoG is optimal in the sense that it computes a Nash equilibrium and its expected finite-step reward.

Theorem 3.2 *Let AT be an action theory, OT be an optimization theory, and p be a GTGolog program relative to them. Let $DoG(p, s, h{+}1, \pi, v, pr)$ for a situation s and $h \geq 0$. Then, π is a Nash equilibrium, and $utility(v, pr)$ is its expected h-step reward.*

In general, there may be exponentially many Nash equilibria. We assume that the opponent is rational, and thus follows a Nash equilibrium. But we do not know which one it actually uses. The following theorem shows that this is not necessary, as far as the opponent computes its Nash half in the same way as we do. That is, different Nash equilibria computed by DoG can be freely "mixed". The result follows from a similar result for matrix games [14] and Theorem 3.2.

Theorem 3.3 *Let AT be an action theory, OT be an optimization theory, and p be a GTGolog program relative to them. Let π and π' be strategies computed by DoG using different Nash selection functions. Then, π and π' have the same expected finite-step reward, and the strategy obtained by mixing π and π' is also a Nash equilibrium.*

4 EXAMPLE

We consider a slightly modified version of the soccer example by Littman [6] (see Fig. 1): The soccer field is a 4×5 grid. There are two players, A and B, each occupying a square, and each able to do one of the following actions on each turn: N, S, E, W, and stand (move up, move down, move left, move right, and no move, respectively). The ball is represented by a circle and also occupies a square. A player is a *ball owner* iff it occupies the same square as the ball. The ball follows the moves of the ball owner, and we have a goal when the ball owner steps into the adversary goal. When the ball owner goes into the square occupied by the other player, if the other player stands, possession of ball changes. Therefore, a good defensive maneuver is to stand where the other agent wants to go. To axiomatize this domain, we introduce the action $move(\alpha, m)$ (agent α executes $m \in \{N, S, E, W, stand\}$) and the fluents $at(\alpha, x, y, s)$ and $haveBall(\alpha, s)$ defined by the following successor state axioms:

$$at(\alpha, x, y, do(a, s)) \equiv at(\alpha, x, y, s) \wedge a = move(\alpha, stand) \vee$$
$$at(\alpha, x', y', s) \wedge (\exists m).a = move(\alpha, m) \wedge \phi(x, y, x', y', m);$$
$$haveBall(\alpha, do(a, s)) \equiv (\exists \alpha').haveBall(\alpha', s) \wedge$$
$$(\alpha = \alpha' \wedge \neg cngBall(\alpha', a, s) \vee \alpha \neq \alpha' \wedge cngBall(\alpha, a, s)).$$

Here, $\phi(x, y, x', y', m)$ represents the coordinate change due to m, and $cngBall(\alpha, a, s)$ is true iff the ball possession changes after an action a of agent α in situation s. We can now define $goal(\alpha, s) \equiv (\exists x, y)\ haveBall(\alpha, s) \wedge at(\alpha, x, y, s) \wedge goalpos(\alpha, x, y)$. Here, $goalpos(\alpha, x, y)$ represents the coordinates for the α adversary goal. We next define a zero-sum reward function as follows:

$$reward(\alpha, s){=}r \equiv$$
$$(\exists\alpha')goal(\alpha', s) \wedge (\alpha'{=}\alpha \wedge r{=}M \vee \alpha'{\neq}\alpha \wedge r{=}-M) \vee$$
$$(\exists\alpha')\neg goal(\alpha', s) \wedge haveBall(\alpha', s) \wedge at(\alpha', x, y, s) \wedge$$
$$(\exists r')evalPos(x, y, r') \wedge (\alpha{=}\alpha' \wedge r{=}r' \vee \alpha'{\neq}\alpha \wedge r{=}-r').$$

Here, the reward is high (M stands for a "big" integer), if a player scores a goal, and the reward depends on $evalPos(x, y, r)$, that is, the ball-owner position (roughly, r is high if the ball-owner is close to the adversary goal), otherwise. A game session can then be described by the following Golog procedure:

$$\mathbf{proc}(game, \mathbf{while}\neg(\exists x)goal(x)\ \mathbf{do}\ (\mathbf{choice}(\boldsymbol{a}\colon move(\boldsymbol{a}, E)|move(\boldsymbol{a}, S)|move(\boldsymbol{a}, N)|move(\boldsymbol{a}, O)|move(\boldsymbol{a}, stand))\|\mathbf{choice}(\boldsymbol{o}\colon move(\boldsymbol{o}, E)|move(\boldsymbol{o}, S)|move(\boldsymbol{o}, N)|move(\boldsymbol{o}, O)|move(\boldsymbol{o}, stand))\colon nil).$$

Intuitively, while a goal is not reached, the two players (agent $\boldsymbol{a}$ and opponent $\boldsymbol{o}$) can choose a possible move. Consider the situation in Fig. 1 where A is the agent $\boldsymbol{a}$ and B the opponent $\boldsymbol{o}$. The initial situation can be described by $at(a, 3, 2, S_0) \wedge at(o, 2, 3, S_0) \wedge haveBall(a, S_0)$. Assuming the horizon $h{=}3$, a strategy π can be calculated by searching for a constructive proof of $AT \models (\exists v, \pi, pr)DoG(game, S_0, 3, \pi, v, pr)$. Here, the maximal reward is achieved by following the pure strategy π that leads the agent $\boldsymbol{a}$ to score a goal after executing three times $move(\boldsymbol{a}, O)$. Note that if we consider $at(o, 1, 3, S_0)$ as the initial position of $\boldsymbol{o}$, then any pure strategy of $\boldsymbol{a}$ can be blocked by $\boldsymbol{o}$ and the only solution is a randomized strategy. The $game$ procedure introduced above represents a generic soccer game. However, more specialized game playing behavior can also be written. For instance, the agent $\boldsymbol{a}$ could discriminate game situations Φ_i where the game can be simplified (that is, possible agent and/or opponent behaviors are restricted):

$$\mathbf{proc}(game', \mathbf{while}\neg(\exists x)goal(x)\ \mathbf{do}\ (\mathbf{if}\ ?\Phi_1\colon schema_1\ \mathbf{else}\ (\mathbf{if}\ ?\Phi_2\colon schema_2\ \mathbf{else}\ game))\colon nil).$$

For example, consider an attacking ball owner, which is closer to the opponent's goal than the opponent (that is, $\Phi(s) = (\exists x, y, x', y')\ at(a, x, y, s) \wedge at(o, x', y', s) \wedge x' > x$). In this situation, since the opponent is behind, the best agent strategy is to move quickly towards the goal. This strategy can be encoded as a Golog program $schema'$. Note that $game'$ lies between a specification of the game rules and a sketchy denotation of the agent strategy. This is in the same spirit of a standard Golog program, which can balance the tradeoff between a planner and a deterministic program: GTGolog allows for the specification of a partial agent strategy, which can be optimally completed once interpreted.

5 TEAMS

We now generalize to the case where we have two competing teams, rather than only two competing agents. Every team consists of a set of cooperating agents. Hence, all the members of every team have the same reward, while the rewards of two members of different teams are zero-sum. Formally, we assume two teams $\vec{\boldsymbol{a}}$ and $\vec{\boldsymbol{o}}$, where $\vec{\boldsymbol{a}} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ consists of $n \geq 1$ agents $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, while $\vec{\boldsymbol{o}} = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_m)$ consists of $m \geq 1$ opponent agents $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_m$. A *one-team action* is the parallel combination of at most one action for each member of one of the two teams. A *two-team action* is either a one-team action $\vec{a}$ for team $\vec{\boldsymbol{a}}$, or a one-team action $\vec{b}$ for team $\vec{\boldsymbol{o}}$, or two parallel one-team actions $\vec{a} \parallel \vec{b}$, one for each team.

We then easily extend the presented GTGolog for two competing agents to the case of two competing teams of agents, by using two-team actions, rather than two-player actions. Then, a one-agent choice of the form $\mathbf{choice}(\boldsymbol{a}\colon a_1|\ldots|a_k)$ turns into a team choice $\mathbf{choice}(\boldsymbol{a}_i\colon a_{i,1}|\ldots|a_{i,k_i})$ for $i \in \{1, \ldots, n\}$, which is written as $\mathbf{choice}(\vec{\boldsymbol{a}}\colon \vec{a}_1|\ldots|\vec{a}_k)$. Thus, in the one-agent choice, rather than a basic action a_i, we choose a combined action $\vec{a}_i$, which consists of at most one basic action for each member of the team $\vec{\boldsymbol{a}}$, while in the two-agent choice, rather than one probability distribution over the possible actions of each agent, we choose at most one distribution over the possible actions of each member of the two teams.

In general, every team has several options for how to act optimally, and two such options cannot be "freely" mixed for different team members. It is thus necessary that there is some form of coordination to agree on one common optimal strategy inside a team (see e.g. [1] for coordination in multi-agent systems). We assume that the coordination is done by centrally controlling a team. Alternatives are either allowing for local communication between team members, or having a total order on optimal strategies, which allows the team members to independently select a common preferred optimal strategy.

Acknowledgements. This work has been partially supported by the Austrian Science Fund Project Z29-N04 and a Marie Curie Individual Fellowship of the EU under contract number HPMF-CT-2001-001286 (disclaimer: The authors are solely responsible for information communicated and the European Commission is not responsible for any views or results expressed). We thank the reviewers for their constructive comments, which helped to improve our work.

References

[1] C. Boutilier, 'Sequential optimality and coordination in multiagent systems', in *Proceedings IJCAI-1999*, pp. 478–485.
[2] C. Boutilier, R. Reiter, M. Soutchanski, and S. Thrun, 'Decision-theoretic, high-level agent programming in the situation calculus', in *Proceedings AAAI/IAAI-2000*, pp. 355–362.
[3] A. Finzi and T. Lukasiewicz, 'Game-theoretic agent programming in Golog', Technical Report INFSYS RR-1843-04-02, Institut für Informationssysteme, TU Wien, (2004).
[4] A. Finzi and F. Pirri, 'Combining probabilities, failures and safety in robot control', in *Proceedings IJCAI-2001*, pp. 1331–1336.
[5] M. Kearns, Y. Mansour, and S. Singh, 'Fast planning in stochastic games', in *Proceedings UAI-2000*, pp. 309–316.
[6] M. L. Littman, 'Markov games as a framework for multi-agent reinforcement learning', in *Proceedings ICML-1994*, pp. 157–163.
[7] J. McCarthy and P. J. Hayes, 'Some philosophical problems from the standpoint of Artificial Intelligence', in *Machine Intelligence*, volume 4, 463–502, Edinburgh University Press, (1969).
[8] G. Owen, *Game Theory: Second Edition*, Academic Press, 1982.
[9] D. Poole, 'The independent choice logic for modelling multiple agents under uncertainty', *Artif. Intell.*, **94**(1–2), 7–56, (1997).
[10] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, 1994.
[11] R. Reiter, 'The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression', in *Artificial Intelligence and Math. Theory of Computation*, 359–380, (1991).
[12] R. Reiter, *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, MIT Press, 2001.
[13] J. van der Wal, *Stochastic Dynamic Programming*, volume 139 of *Mathematical Centre Tracts*, Morgan Kaufmann, 1981.
[14] J. von Neumann and O. Morgenstern, *The Theory of Games and Economic Behavior*, Princeton University Press, 1947.