Hierarchical Reinforcement Learning in Communication-Mediated Multiagent Coordination

Felix Fischer, Michael Rovatsos, Gerhard Weiss

AI/Cognition Group, Department of Informatics, Technical University of Munich
{fischerf,rovatsos,weissg}@cs.tum.edu

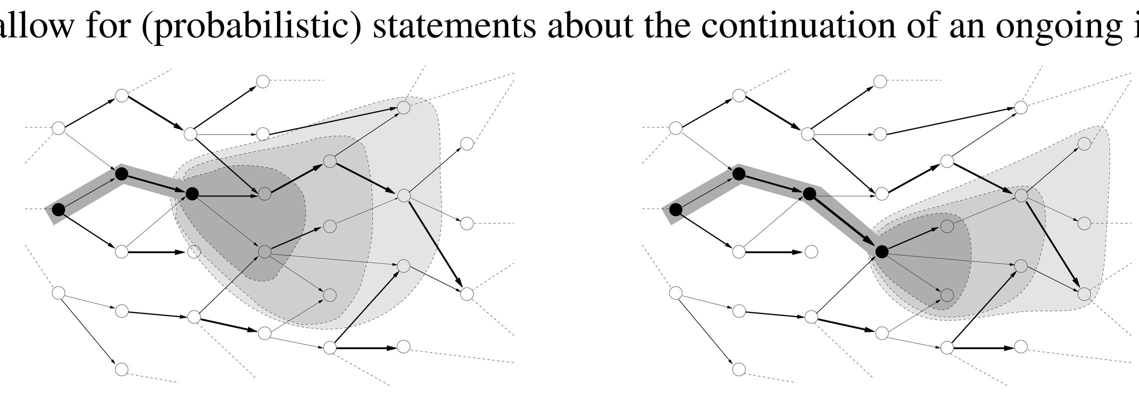
Communication and Openness

- Traditional approach to interaction and communication in a MAS:
  - Agent communication languages with speech-act based semantics (KQML/KIF, FIPA-ACL)
  - Communication protocols (CNP, auctions, . . .)
- Leads to problems in open multiagent systems characterised by
  - Changing populations of autonomous agents (self-interested or even anti-social agents)
  - Heterogenous (and mutually unknown) agent design
- Central question:

  If adherence to communication languages and protocols cannot be taken for granted, how can meaningful and coherent communication be ensured?
- One possible answer: empirical communication semantics [3], i.e. grounding the meaning of communication in its expected consequences

Reasoning about Communication with Empirical Semantics

- Expectation networks [1]: graph-based representation of (statistical) correlations between successive utterances to allow for (probabilistic) statements about the continuation of an ongoing interaction



- Interaction Frames and Framing Architecture InFFrA [2]: a meta-framework for the representation and strategic use of (a simple kind of) expectation networks; no formal semantics, not readily implementable
- Practical problem: utterance-level reasoning about continuation probabilities yields a vast search space
- Proposed solution: view communication as hierachical decision process, use interaction frames as (abstract) communication policies
- m²InFFrA ("Markov-square"): A formalisation of InFFrA for two-party, turn-taking interactions based on this interpretation

A Formalisation of Interaction Frames

- An m²InFFrA *frame* is a tuple $F = (T, \Theta, C, h_T, h_\Theta)$, where
  - $T = \langle p_1, p_2, \ldots, p_n\rangle$, the *trajectory* of the frame, is a sequence of message patterns describing possible instances of F by means of variables in p_i;
  - $\Theta = \langle \vartheta_1, \ldots, \vartheta_m\rangle$ is an ordered list of *variable substitutions*, representing previous enactments;
  - $C = \langle c_1, \ldots, c_m\rangle$ is an ordered list of *condition sets* encoded in a logical language, such that c_j is the condition set relevant under substitution ϑ_j;
  - $h_T \in \mathbb{N}^{|T|}$ is a *trajectory occurrence counter* list counting the occurrence of each prefix of the trajectory T in previous encounters;
  - $h_\Theta \in \mathbb{N}^{|\Theta|}$ is a *substitution occurrence counter* list counting the occurrence of each member of the substitution list Θ in previous encounters.
- Example: Interaction frame for the success path of the FIPA Contract Net Protocol:

$$\begin{aligned}
F_{cn} = \Big\langle \big\langle &\xrightarrow{5} \texttt{cfp}(A_1, A_2, \langle R, P\rangle)) \xrightarrow{3} \texttt{propose}(A_2, A_1, Q) \\
&\xrightarrow{3} \texttt{accept-proposal}(A_1, A_2, Q) \xrightarrow{2} \texttt{do}(A_2, A_1, R)\big\rangle, \\
&\big\langle\{\iota X(P = Q) = Y, \\
&\quad \neg \mathit{Bref}_{A_1}(\mathit{any}\ X\ I_{A_2}\mathit{Done}(R, P)) \wedge \neg B_{A_1} I_{A_2}\mathit{Done}(R) \ @1, \\
&\quad B_{A_2} I_{A_2}\mathit{Done}(R, Q) \ @2, \\
&\quad B_{A_1} I_{A_1}\mathit{Done}(R, Q) \wedge B_{A_1} I_{A_2}\mathit{Done}(R, Q) \ @3, \\
&\quad B_{A_2} Q \ @4\}, \{\}, \{\mathit{existsLink}(A_2, \mathit{agent}_2)\}\big\rangle, \\
&\big\langle \xrightarrow{0} [\,] \\
&\quad \xrightarrow{1} [A_1/\mathit{agent}_1, A_2/\mathit{agent}_2, R/\mathit{addLink}(A_2, A_1, 2), \\
&\qquad P/\mathit{greater}(X, 0), Q/\mathit{equal}(X, 2)], \\
&\quad \xrightarrow{1} [A_1/\mathit{agent}_3, A_2/\mathit{agent}_1, R/\mathit{modifyRating}(A_2, \mathit{agent}_2, -3), \\
&\qquad P/\mathit{greater}(-2, X), Q/\mathit{equal}(X, -3)]\big\rangle\Big\rangle
\end{aligned}$$

- Frame semantics:
  - Given a set $\mathcal{F} = \{F_1, \ldots, F_n\}$ of frames, a conversation prefix w and a knowledge base KB
  - . . . derive a continuation probability

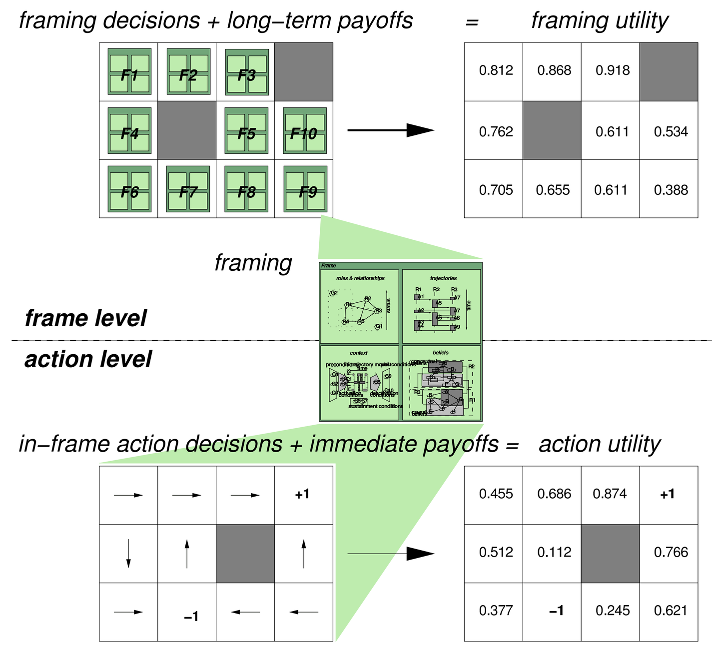
$$P(w'|w) = \sum_{F \in \mathcal{F}} P(w'|F, w) P(F|w) = \sum_{F \in \mathcal{F}, ww' = T(F)\vartheta} P(\vartheta|F, w) P(F|w)$$

  - Probability of ϑ under F is proportional to its *similarity* to F:

$$P(\vartheta|F, w) \propto \sigma(\vartheta, F) = \sum_{i=1}^{|\Theta(F)|} \overbrace{\sigma(T(F)\vartheta, T(F)\Theta(F)[i])}^{\text{similarity}} \overbrace{h_\Theta(F)[i]}^{\text{frequency}} \overbrace{c_i(F, \vartheta, KB)}^{\text{relevance}}$$

Learning and Decision-Making with Frames

- Problem: given a (possibly abstract) state (w, KB) corresponding to a conversation prefix w and a knowledge base KB and a set $\mathcal{F} = \{F_1, \ldots, F_n\}$ of frames, derive the optimal continuation m^*
- Hierarchical approach: select best frame $F^* \in \mathcal{F}$ according to w and KB, then select best continuation m^* according to F^*



- Frame level: use the hierarchical RL framework of options [4] to learn which frame to choose in a given situation
  - $F \in \mathcal{F}$ induces an option $o_F = (\mathcal{I}_F, \pi_F, \beta_F)$
    - Input set $\mathcal{I}_F$ of states $s = (w, KB)$ in which F can be invoked, i.e. where the Prefix of $T(F)$ matches w and the corresponding suffix is executable under KB
    - Policy π_F assigns a probability of 1 to m^* (which is to be determined on the action level)
    - Termination criterion β_F is given by $T(F)$, w and KB (in analogy to $\mathcal{I}_F$) and by a private desirability criterion
  - Update equation for SMDP Q-learning:

$$Q(s, o) \leftarrow (1 - \alpha) Q_k(s, o) + \alpha \left[r + \gamma^\tau \max_{o' \in \mathcal{O}_{s'}} Q_k(s', o') \right]$$

  - Optimal "framing" policy: $\pi^*(s, o) = 1 \Leftrightarrow o = \arg\max_{o'} Q^*(s, o')$
- Action level: expected utility maximisation based on continuation probabilities $P(w'|w, F)$
  - Requires utility estimate $u(w', KB)$
  - Three different kinds of variable substitutions:
    - "fixed" substitution $\vartheta_f(F, w)$: bindings induced by the observation w
    - "own" substitution ϑ_o: binds variables in the frame steps to be executed by the agent
    - "peer" substitution ϑ_p: binds those in the other's steps
  - Expected utility of "own" substitution ϑ_s:

$$E[u(\vartheta_s, F, w, KB)] = \sum_{\vartheta_p} P(\vartheta_p|\vartheta_s, F, w)\ u\big(\mathit{postfix}(T(F), w)\vartheta_f(F, w)\vartheta_s\vartheta_p, KB\big)$$

  - Probability for "peer" substitution estimated from previous instantiations of F:

$$P(\vartheta_p|\vartheta_s, F, w) = \frac{P(\vartheta_p \wedge \vartheta_s|F, w)}{P(\vartheta_s|F, w)} = \frac{P(\vartheta_f(F, w)\vartheta_s\vartheta_p|F, w)}{\sum_\vartheta P(\vartheta_f(F, w)\vartheta_s\vartheta|F, w)} \propto \sigma(\vartheta_f(F, w)\vartheta_s\vartheta_p, F)$$

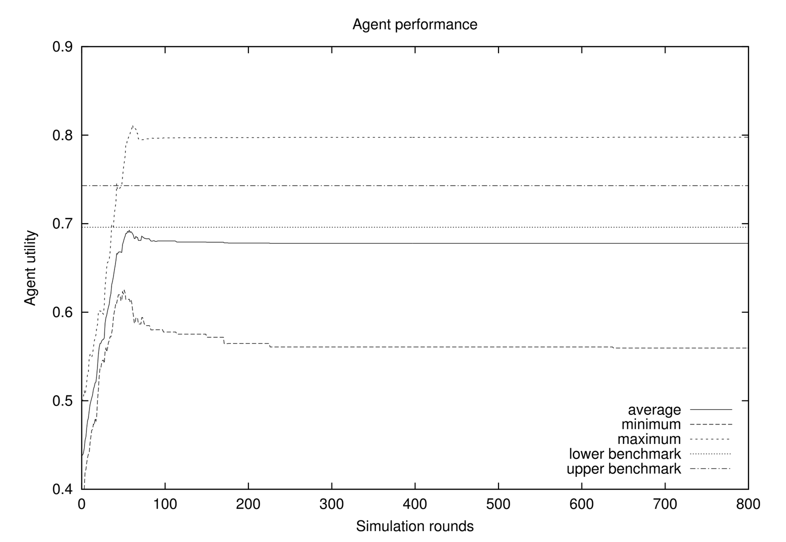
- Selection of optimal substitution and action:

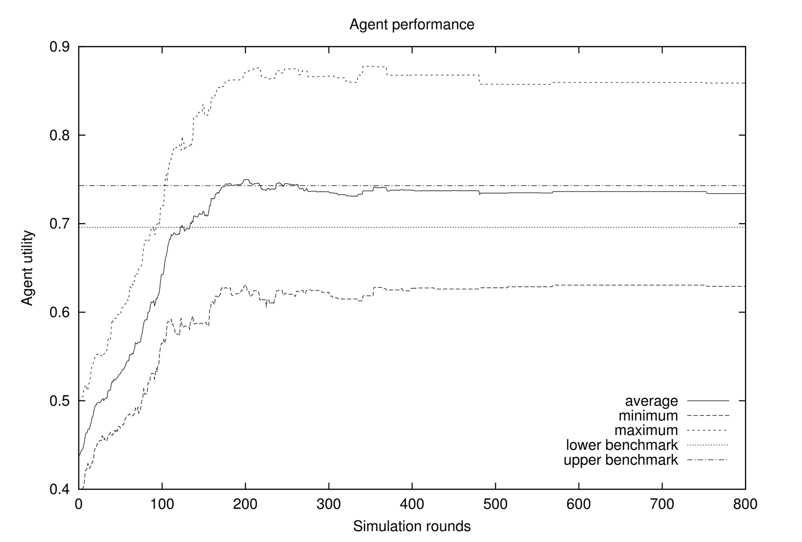
$$\vartheta^*(F, w, KB) = \arg \max_{\vartheta_s \in \Theta_{poss}(F, KB, w)} E[u(\vartheta_s, F, w, KB)]$$

$$m^*(F, w, KB) = T(F)[|w| + 1]\vartheta^*(F, w, KB)$$

Experimental Results

- Scenario: Automated Website Linkage (agents represent Web site owners who negotiate over linkage)
- Agents with Prolog-like inference mechanism and BDI-based goal generation and planning
- Comparison between agents that simply issue requests if they cannot perform an action by themselves and m²InFFrA agents





References

[1] M. Nickles and M. Rovatsos. Communication Systems: A Unified Model of Socially Intelligent Systems. In K. Fischer and M. Florian, editors, *Socionics: Its Contributions to the Scalability of Complex Social Systems*.

[2] M. Rovatsos. Interaction frames for artificial agents. Research Report FKI-244-01, AI/Cognition Group, Department of Informatics, Technical University of Munich, 2001.

[3] M. Rovatsos, M. Nickles, and G. Weiß. Interaction is Meaning: A New Model for Communication in Open Systems. In J. S. Rosenschein, T. Sandholm, M. Wooldridge, and M. Yokoo, editors, *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03)*, Melbourne, Australia, 2003.

[4] R. S. Sutton, D. Precup, and S. P. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.