Autonomous Reinforcement Learning of Multiple Interrelated Tasks

ICDL-EpiRob 2019

Authors

Vieri Giuliano
Santucci

Gianluca
Baldassarre

Emilio
Cartoni

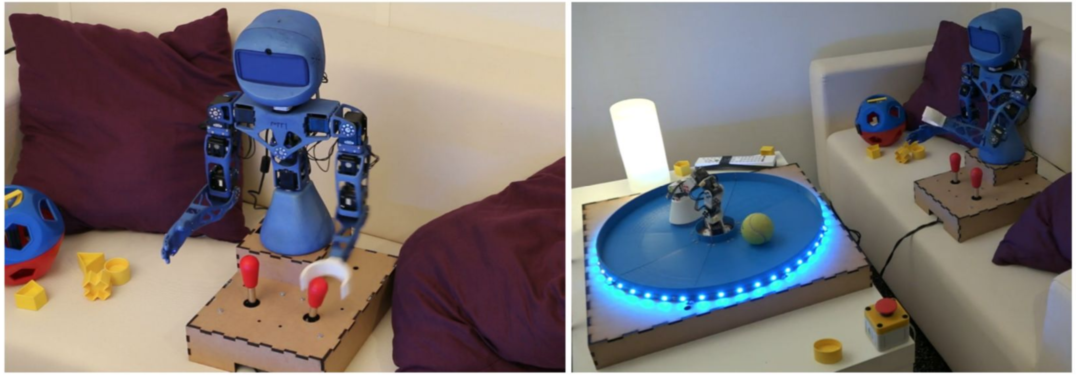
Institute of Cognitive Sciences and Technologies, Rome, Italy
[Laboratory of Computational Embodied Neuroscience]

Motivation of the Study

- Solving interrelated (hierarchical) tasks autonomously
- How this question can be addressed working on the level of task selection?
- Open-ended learning => active task selection

Intrinsic Motivation

“The doing of an activity for its inherent satisfactions rather than for some separable consequence”. (Ryan and Deci, 2000)

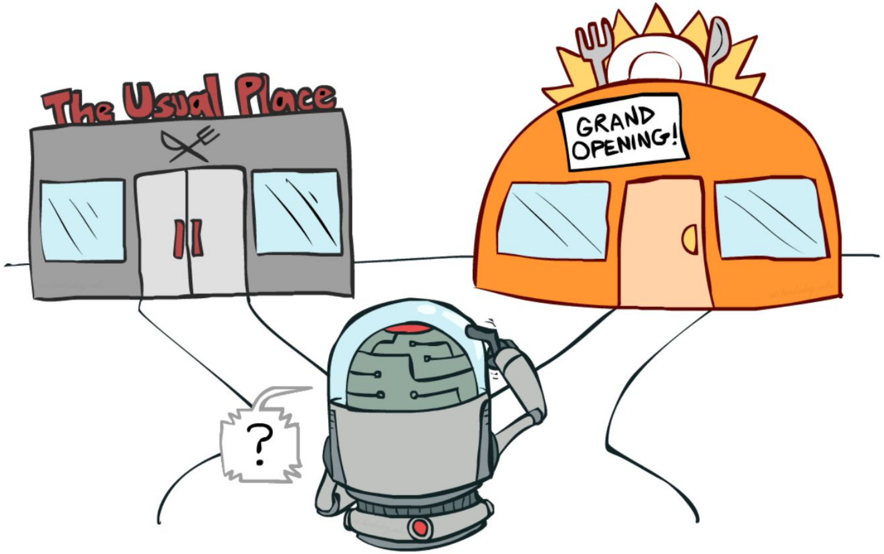


Img. Source: Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning

Preliminaries

N-Armed Bandit

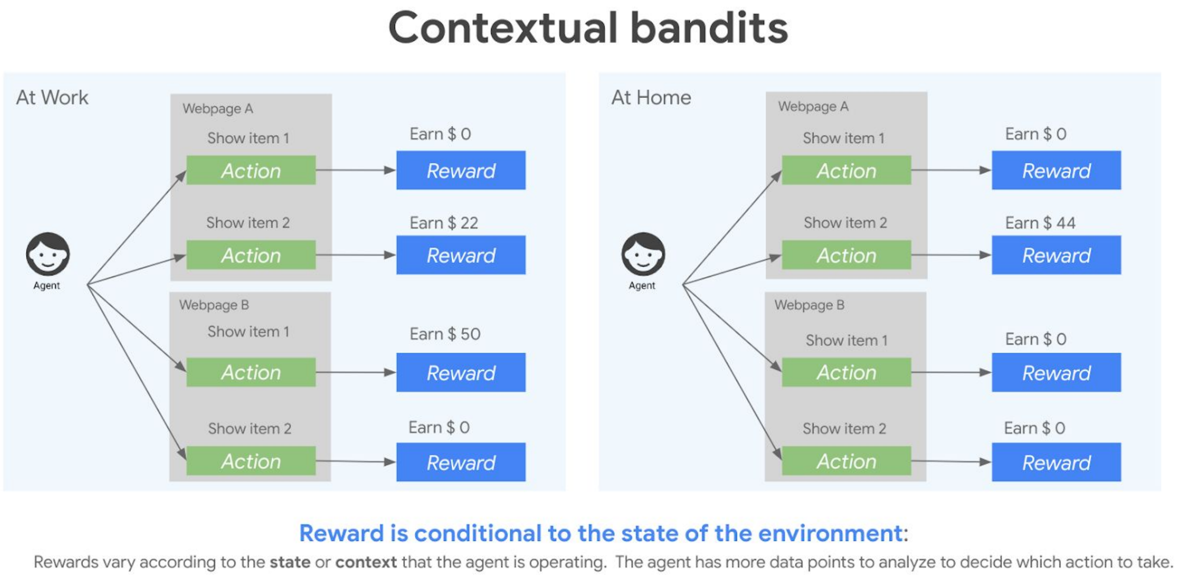
- Repeatedly have to choose among n different actions
- After each action, receive a reward depends on the action you have taken
- **Objective:** Maximize the expected total reward over some time period
- Non-associative



Img. Source: UC Berkeley CS188 Intro to AI - Course Materials

Contextual Bandit

- A.k.a: Associative search tasks
- Intermediate between N-Armed Bandit and RL



Img. Source: Google Cloud Blog - [Better bandit building: Advanced personalization the easy way with AutoML Tables]

Markov Decision Process (MDP)

Markov Decision Process is a 4 tuple (S, A, T, R), where:

- S : Set of states
- A : Set of actions
- T : Transition probability $P(s_{t+1} | s_t, a_t)$
- R : Reward Function

Q-Learning Algorithm

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        Choose A from S using policy derived from Q (e.g., ε-greedy)
        Take action A, observe R, S'
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
        $S \leftarrow S'$
    until S is terminal

Img. Source: Reinforcement Learning: An Introduction, Sutton and Barto

Problem Definition

- Learning of multiple goals => Learning of different policies π_g maximising different reward functions
- For each g the system aims to learn a policy

$$\pi_g^*(a|s) = \underset{\pi}{\operatorname{argmax}}\, R_g(\pi_g) \tag{1}$$

- **Competence:** Ability of the system in achieving a goal
- Maximising **Competence Function (C)** instead of extrinsic reward
- C => sum of the C_g at each goal
- Π_t is the goal selection policy
- Allocate training time to the goals that guarantee the highest competence gain

Proposed solution

- Select the goal with the highest competence improvement
- Agent learns a policy to **select goals**

$$\Pi^* = \underset{\Pi}{\operatorname{argmax}} \; \delta C(\Pi_t) \tag{2}$$

- If depends on environmental conditions

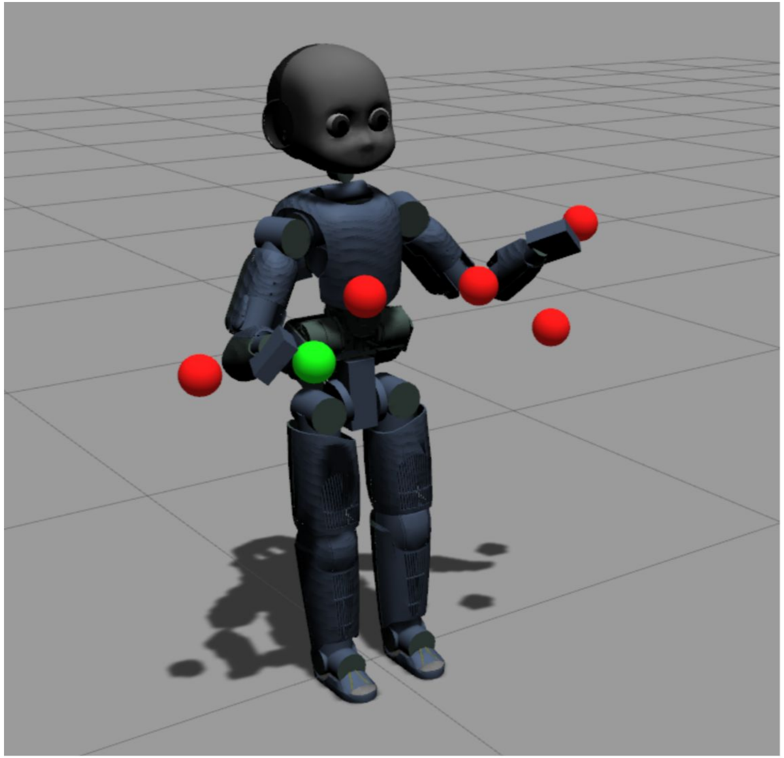
$$\Pi^*(s_t) = \underset{\Pi}{\operatorname{argmax}} \; \delta C(\Pi(s_t)) \tag{3}$$

- If the goals are interrelated => Treat as MDP
    - Transfer IM values between interrelated goals

Experiments and Results

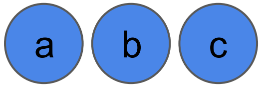
Experiment Setup

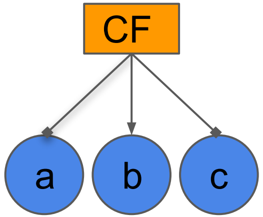
- Simulated iCub robot
- Interrelated tasks of touching-to-activate different spheres
- 2 arms, each 4 DOF
- Wrist joint fixed, end-effectors are scoops
- Sensor at each scoop to determine touch



3 Different Experiment Scenarios

1. No relations / N-armed bandit
2. Environmental dependence / Contextual bandit
3. Multiple Interrelated Tasks / MDP





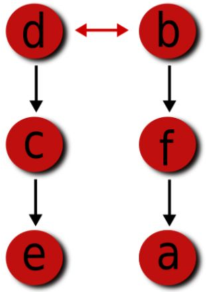


Fig. 6. Structure of the third experimental scenario. Black arrows indicate positive dependencies, and red arrows negative dependencies.

Compared Systems

1. GRAIL (Santucci et al. 2016, TCDS)
2. C-GRAIL
3. M-GRAIL

GRAIL

A four level architecture:

1. Goal formation
2. **Goal Selector**
3. Expert Selector
4. Expert

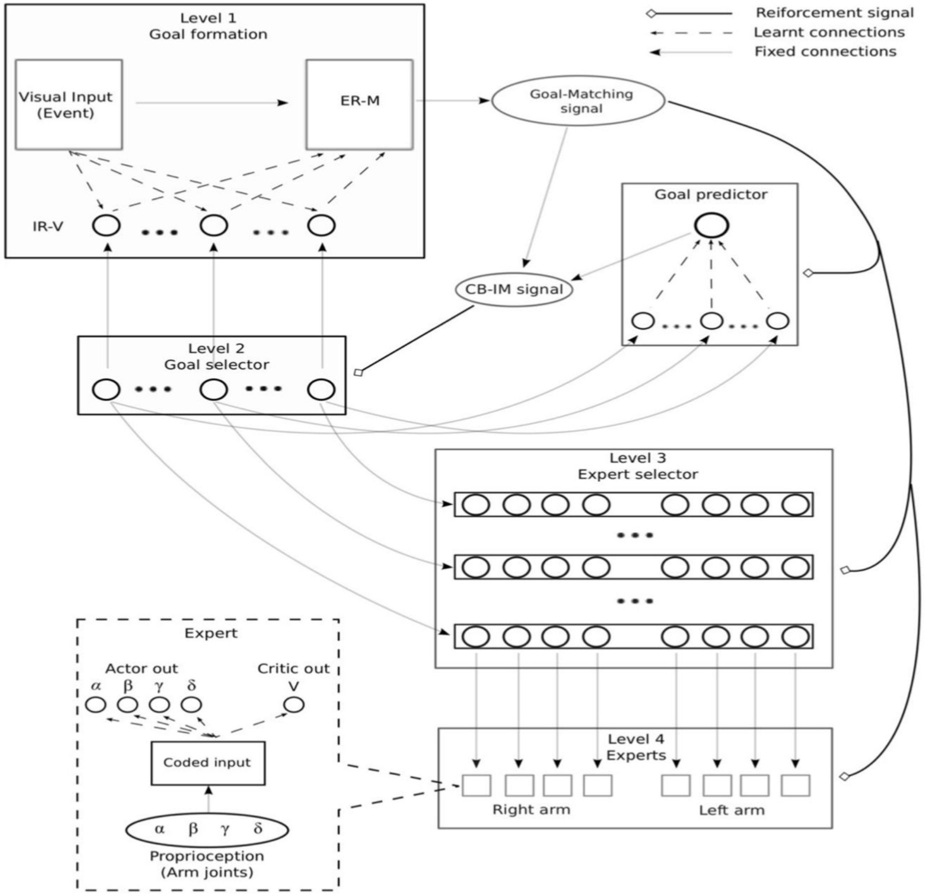


Fig. 3. Four-level hierarchical architecture of GRAIL: 1) goal-formation mechanisms with the IR-V and the ER-M; 2) goal-selector; 3) expert-selector; and 4) experts. The CB-IM signal is also presented in the figure, together with the GM reinforcement signal.

CB-IM Signal

- **Competence:** Ability of the system in achieving a goal
- A predictor that is trying to anticipate the achievement of the desired state
- PEI of the predictor is the IM signal
- **Input:** The selected goal
- **Output:** Predicted probability that the event associated to the selected goal will happen [0,1]

$$PEI_t = \frac{\sum_{i=t-(2T-1)}^{t-T} |PE|_i}{T} - \frac{\sum_{i=t-(T-1)}^{t} |PE|_i}{T} \qquad (8)$$

Selection Rule

Softmax selection rule

$$p_k = \frac{\exp\left(\frac{Q_k}{\tau}\right)}{\sum_{i=0}^{n} \exp\left(\frac{Q_n}{\tau}\right)} \tag{3}$$

Exponential Moving Average (EMA)

$$Q_k^t = Q_k^{t-1} + \alpha\left(\mathrm{ir} - Q_k^{t-1}\right) \tag{4}$$

Equations from: GRAIL: A Goal-Discovering Robotic Architecture for Intrinsically-Motivated Learning

C-GRAIL

- Modify Goal Selector of GRAIL
- Context can be:
  - Standard state features
  - The status of different goals
- Selects goals as in contextual bandit, different EMAs are associated with different contexts
- Goal selection => Softmax Rule

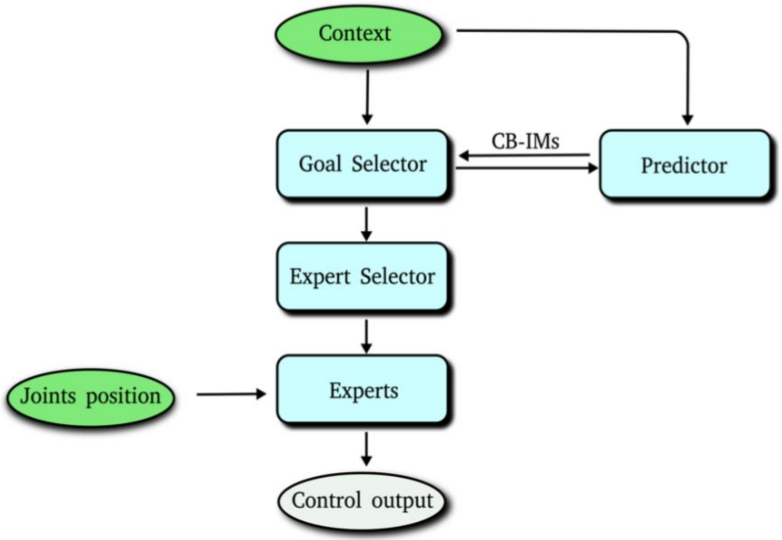


Fig. 2. A schema of the architecture implemented in C-GRAIL and M-GRAIL. Differently from GRAIL, the new architectures use context as input to the goal selector. Note that for all the architectures the expert selector and experts are goal-specific.

M-GRAIL

- Goal Selector takes same input as in C-GRAIL
- Goal selection as an MDP
- Models temporal interdependencies between goals (Q-Learning)
- Goal selection => Softmax Rule

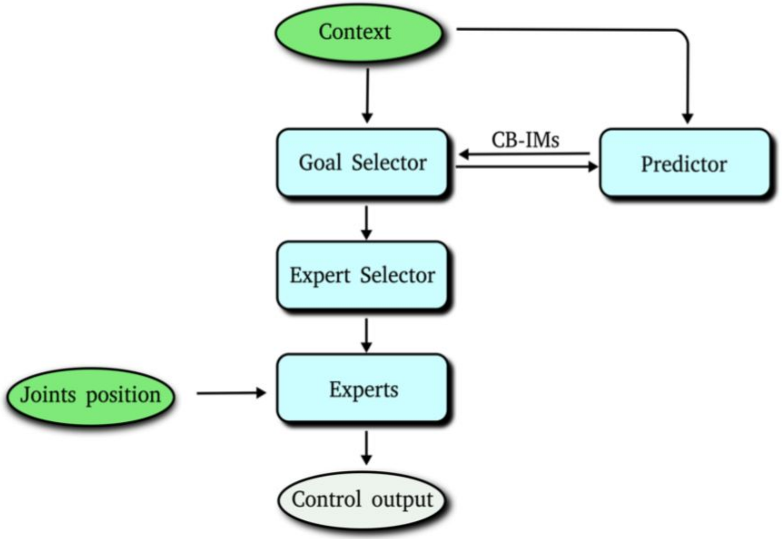
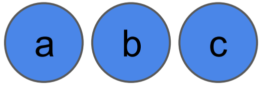


Fig. 2. A schema of the architecture implemented in C-GRAIL and M-GRAIL. Differently from GRAIL, the new architectures use context as input to the goal selector. Note that for all the architectures the expert selector and experts are goal-specific.

First Experiment: No Relations Between Tasks



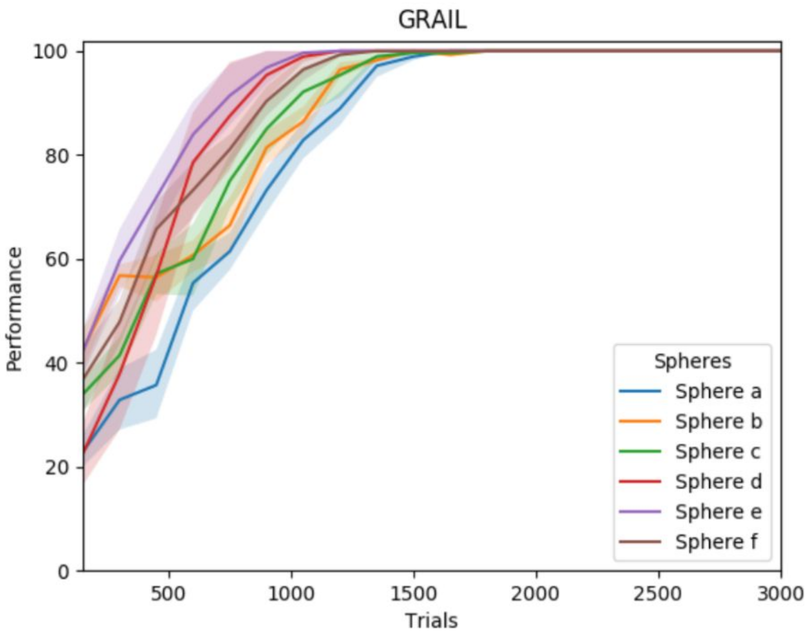
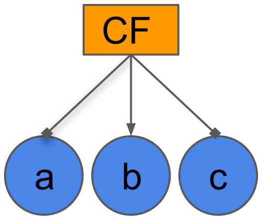
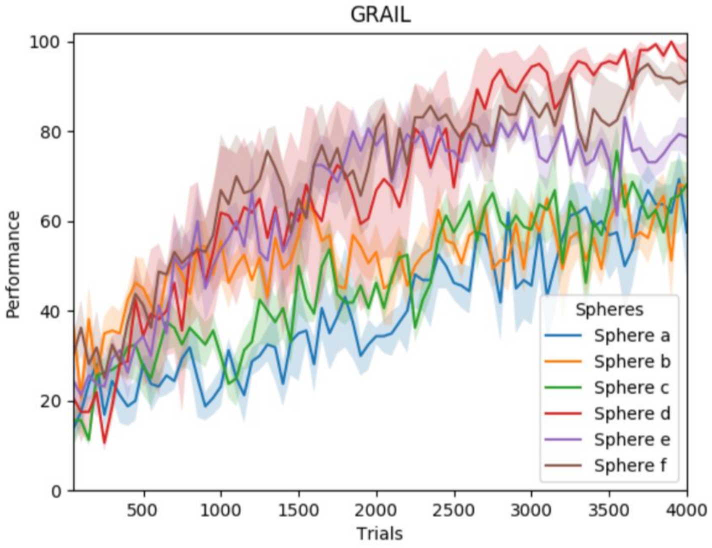
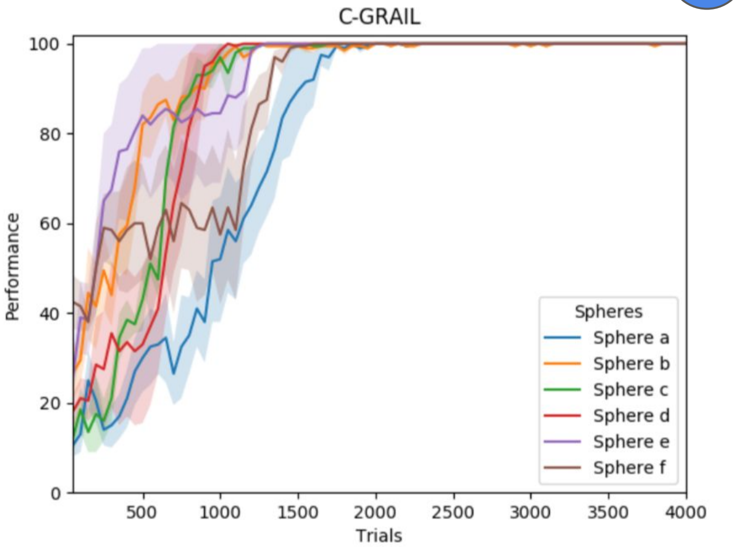


Fig. 3. Performance of GRAIL in the first experiment. Average over 10 replications of the experiment. Shadows show the confidence intervals.

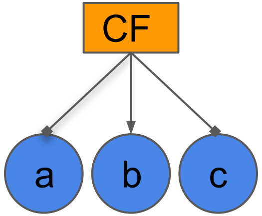
Second Experiment: Environmental Dependence







Second Experiment: Environmental Dependence



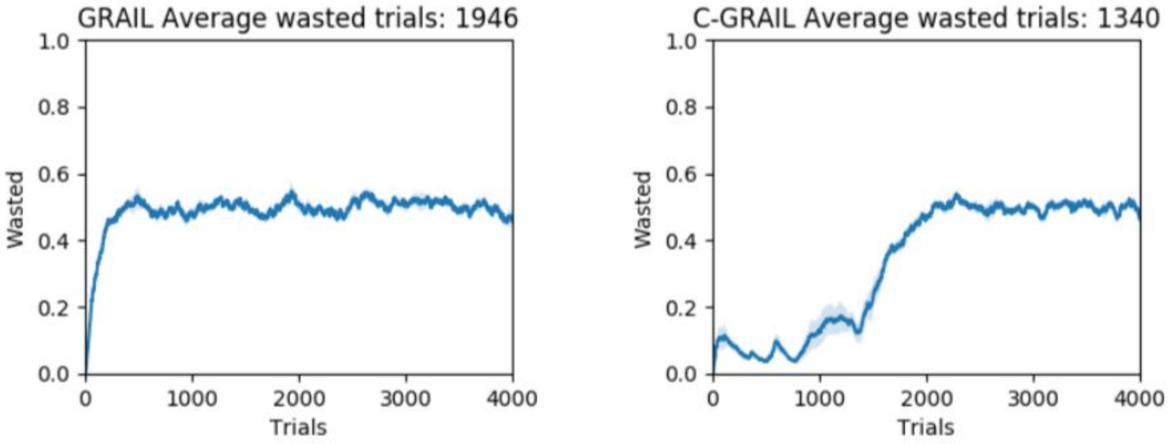
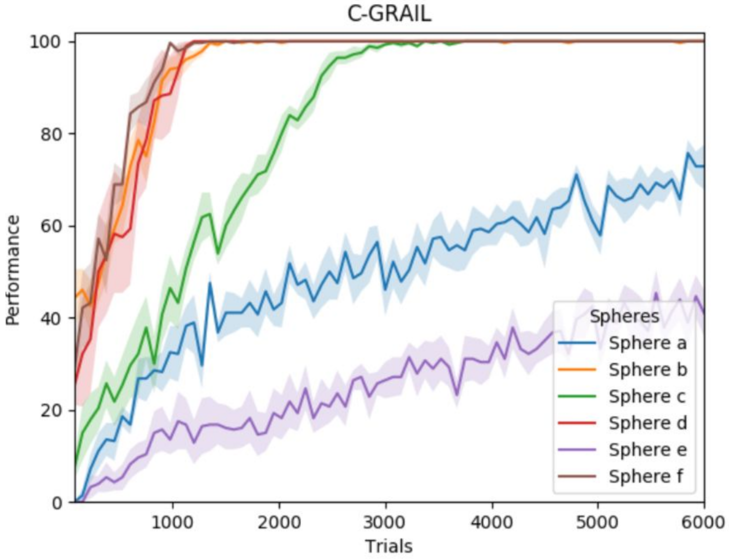
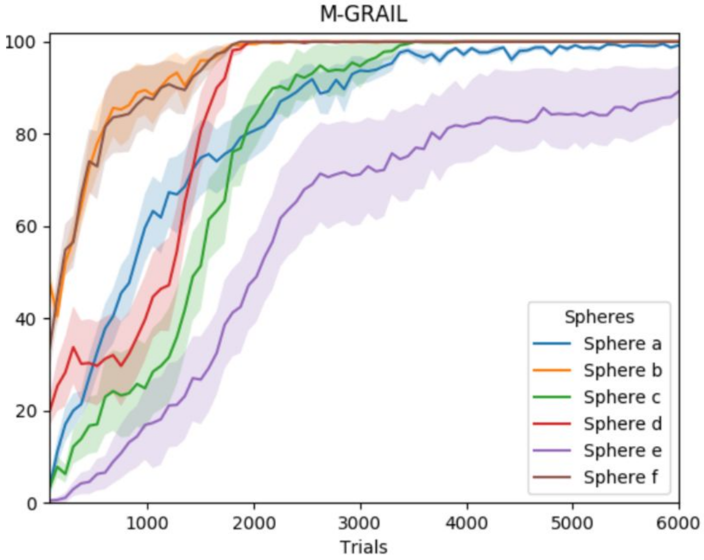
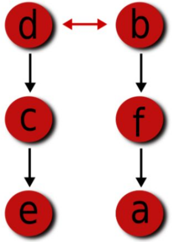


Fig. 5. Trials wasted by GRAIL and C-GRAIL in selecting tasks that cannot be performed.

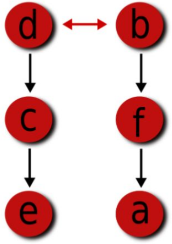
Third Experiment: Multiple Interrelated Tasks







Third Experiment: Multiple Interrelated Tasks



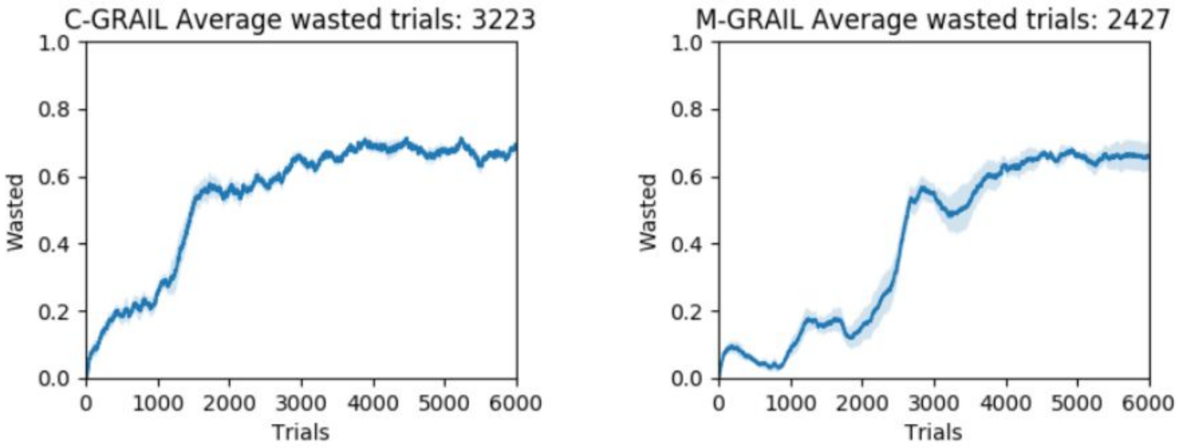


Fig. 8. Trials wasted by C-GRAIL and M-GRAIL in selecting tasks that cannot be performed.