Proceedings of the Post-Graduate Conference on Robotics and Development of Cognition
10-12 September 2012, Lausanne, Switzerland

Whom Will an Intrinsically Motivated Robot Learner Choose to Imitate from?

Sao Mai Nguyen, Pierre-Yves Oudeyer

[1]Flowers Team, INRIA and ENSTA ParisTech, France.
nguyensmai at gmail.com, pierre-yves.oudeyer at inria.fr

Abstract

This paper studies an interactive learning system that couples internally guided learning and social interaction in the case it can interact with several teachers. Socially Guided Intrinsic Motivation with Interactive learning at the Meta level (**SGIM-IM**) is an algorithm for robot learning of motor skills in high-dimensional, continuous and non-preset environments, with two levels of active learning: **SGIM-IM** actively decides at a meta-level when and to whom to ask for help; and an active choice of goals in autonomous exploration. We illustrate through an air hockey game that SGIM-IM efficiently chooses the best strategy.

Index Terms: Active Learning, Intrinsic Motivation, Social Learning, Programming by Demonstration, Imitation.

1. Introduction

In initial work to address multi-task learning, we proposed the Socially Guided Intrinsic Motivation by Demonstration (SGIM-D) algorithm which merges socially guided exploration as defined in[1, 2, 3, 4] and intrinsic motivation [5, 6, 7, 8] based on SAGG-RIAC algorithm [9], to reach goals in a continuous task space, in the case of a complex, high-dimensional and continuous environment [10]. Nevertheless, the SGIM-D learner uses demonstrations given by a teacher at regular frequency. It is passive with respect to the social interaction and the teacher, and does not optimise the timing of the interactions with the teacher, not to mention that it did not consider the everyday situation where it has several human teachers around him, to whom it can ask for help. Some works have considered the choice among the different teachers that are available to be observed [11] where some of them might not even be cooperative [12], but have then overlooked autonomous exploration. Our new **SGIM-IM** (Socially Guided Intrinsic Motivation with Interactive learning at the Meta level) learner is able to choose between active autonomous and social learning strategies, and in the case of social learning, whom to imitate from.

2. General Framework

2.1. Formalisation

In this subsection, we describe the learning problem that we consider. Csibra's theory of human action finds that infants connect actions to both their antecedents and their consequents [13, 14]. Thus, every episode would be described as [context][action][effect].

Let us describe different aspects of the states of a robotic system and its environment by both a state/context space C, and an effect/task space Y (an effect/task can be considered as restricted to the changes caused by the agent's actions). For given contexts $c \in C$, actions $act \in ACT$ allow a transition towards new states $y \in Y$ (fig. 1 and 2). We define the actions act as parameterised dynamic motor primitives, i.e. temporally extended macro-actions controlled by parameters a in the action parameters space A. The association (c, a, y) corresponds to a learning exemplar that will be memorised. Our agent learns a policy through an inverse model $\mathbf{M^{-1}} : (c, y) \mapsto a$ by building local mappings of $\mathbf{M} : (c, a) \mapsto y$, so that from a context c and for any achievable effect y, the robot can produce y with an action a. We can also describe the learning in terms of tasks, and consider y as a desired task or *goal* which the system reaches through the *means* a in a given *context* c. In the following, both descriptions will be used interchangeably.

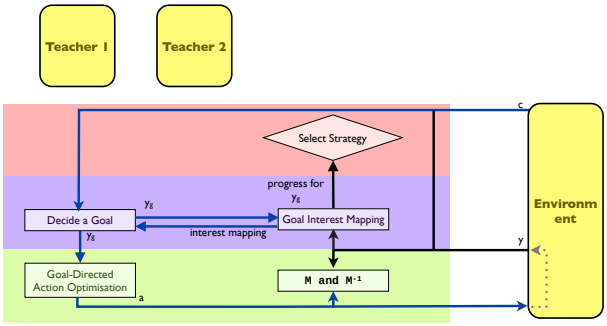

Figure 1: Data Flow under the Intrinsic Motivation strategy

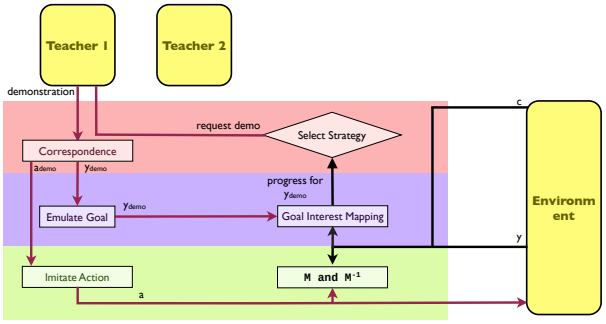

Figure 2: Data Flow under the Social Learning strategy with teacher 1

2.2. SGIM-IM Overview

SGIM-IM learns by episodes during which it chooses actively its learning strategy between intrinsically motivated exploration or social interaction with each of the existing teachers.

In an episode under the intrinsic motivation strategy (fig. 1), it actively generates a goal $y_g \in Y$ of maximal competence improvement, then explores which actions a can achieve the goal y_g in context c, following the SAGG-RIAC algorithm [9]. The exploration of the action space gives a local forward model $\mathbf{M} : (c, a) \mapsto y$ and inverse model $\mathbf{M^{-1}} : (c, y) \mapsto a$, that it can use later on to reach other goals. The SGIM-IM learner explores preferentially goals where it makes progress the fastest. It tries different actions to approach the self-determined goal, re-using and optimising the action repertoire of its past au-

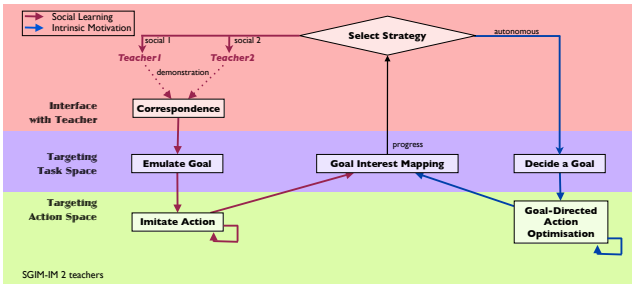

Figure 3: Time flow chart of SGIM-IM

Algorithm 2.1 SGIM-IM

Initialization: $\mathcal{R} \leftarrow$ singleton $C \times Y$
Initialization: $flagInteraction \leftarrow false$
Initialization: $Memo \leftarrow$ empty episodic memory
Initialization: $\Delta_0, ...\Delta_i, ...$: progress values made by strategy i among: autonomous exploration or social learning with either teacher
loop
  $strategy \leftarrow$ Select Strategy$(pref_S, pref_A)$
  if Social Learning Strategy then
    demo $\leftarrow$ ask & perceive demo to the selected teacher i
    $(c_{demo}, a_{demo}, y_{demo}) \leftarrow$ Correspondence (demo)
    Emulate Goal: $y_g \leftarrow y_{demo}$
    $\gamma_s \leftarrow$ Competence for y_g
    Memo $\leftarrow$ Imitate Action(a_{demo}, c)
    $\gamma \leftarrow$ Competence for y_g
    Add $\gamma - \gamma_s$ to stack Δ_i
  else
    Intrinsic Motivation Strategy
    Measure current context c
    $y_g \leftarrow$ Decide a goal$(c, \mathcal{R})$
    $\gamma_s \leftarrow$ Competence for y_g
    repeat
      Memo $\leftarrow$ Goal-Directed Action Optimisation(c, y_g)
    until Terminate reaching of y_g
    $\gamma \leftarrow$ Competence for y_g
    Add $\gamma - \gamma_s$ to stack Δ_0
  end if
  $\mathcal{R} \leftarrow$ Update Goal Interest Mapping$(\mathcal{R}, Memo, c, y_g)$
end loop

tonomous exploration or the actions suggested by the teacher's demonstrations of the social learning strategy. The episode ends after a fixed duration.

In an episode under the social learning strategy with teacher i (fig. 2), our SGIM-IM learner observes the demonstration $[c_{demo}, act_{demo}, y_{demo}]$, memorise this effect y_{demo} as a possible goal, and imitates the demonstrated action act_{demo} for a fixed duration.

The SGIM-IM learner actively decides on a meta level which strategy to choose according to the recent learning progress enabled by each strategy. If it has recently made the most progress in the intrinsic motivation strategy, it prefers exploring autonomously. Conversely, if the demonstrations does not enable him to make progresses higher than by autonomous learning (limited teacher, or inappropriate teacher) it would prefer autonomous exploration.

3. SGIM-IM Architecture

3.1. A Hierarchical Architecture

SGIM-IM (Socially Guided Intrinsic Motivation with Interactive learning at the Meta level) is an algorithm that merges interactive learning as social interaction, with the SAGG-RIAC algorithm of intrinsic motivation [9], to learn local inverse and forward models in complex, redundant, high-dimensional and

Algorithm 3.2 [strategy] = SelectStrategy(Δ_S, Δ_A)

input: $\Delta_0, ...\Delta_i, ...$: progress values made by strategy i among: autonomous exploration or social learning with either teacher
output: $flagInter$: chosen strategy
parameter: $nbMin$: duration of the initiation phase
parameter: ns : window frame for monitoring progress
parameter: $cost_i$: cost of each strategy
Initiation phase
if Social Learning and Intrinsic Motivation Regimes have not been chosen each $nbMin$ times yet **then**
  $p_i \leftarrow 0.5$
else
  Permanent phase
  for all strategies **do**
    $w_i \leftarrow$ average(last ns elements of Δ_i)
  end for
  $p_i \leftarrow min(0.9, max(0.1, \frac{cost_i \times w_i}{\sum cost_j \times w_j}))$
end if
strategy $\leftarrow$ i with probability p_i
return strategy

continuous spaces and with several teachers. Its architecture (alg. 2.1) is separated into three layers (fig. 3) :

- An interface with the teacher, which manages the interaction with the teacher. It decides in an active manner whether to request a demonstration and to whom (*Select Strategy*) and interpreting his actions or his intent and translates the demonstrations into the robot's representation system (*Correspondence*, which is an important issue [15] but will not be addressed in this study.
- The *Task Space Exploration*, a level of active learning which drives the exploration of the task space. With the autonomous learning strategy, it sets goals y_g depending on the interest level of previous goals, by stochastically choosing the ones for which its empirical evaluation of learning progress is maximal (*Decide a Goal*). With the social learning strategy, it retrieves from the teacher information about demonstrated effects y_{demo} (*Emulate a Goal*). Then, it maps $C \times Y$ in terms of interest level (*Goal Interest Mapping*).
- The *Action Space Exploration*, a lower level of learning that explores the action space A to build an action repertoire and local models. With the social learning strategy, it imitates the demonstrated actions act_{demo}, by repeating it with small variations (*Imitate an Action*). During self-exploration, the *Goal-Directed Action Optimisation* function attempts to reach the goals y_g set by the *Task Space Exploration* level, 1) by building local models during exploration that can be re-used for later goals and 2) by optimising actions to reach y_g. Then, the *Action Space Exploration* returns the measure of competence at reaching y_{demo} or y_g.

The active choice of learning strategy will be described hereafter. For the other parts of the architecture, which are common to SGIM-D, please refer to [10] for more details.

3.2. Select Strategy

Based on the recent progress made by each of them, a meta level chooses the best strategy among autonomous exploration and social learning with each of the teachers. For each episode,

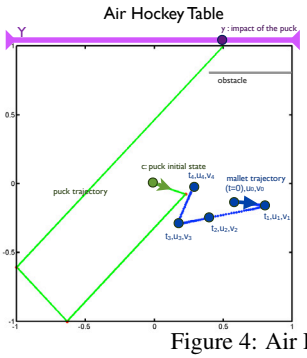

Figure 4: Air Hockey Table

the learner measures its progress as the difference of competence at the beginning and the end of the exploration for the self-determined or the emulated goal, and adds this progress value to stacks Δ_i, where i is the current strategy ($i = 0$ for autonomous exploration, $i = 1$ for social learning with teacher 1, $i = 2$ with teacher 2,...). The preference for each strategy is computed as the average on a window frame of the last ns progress values of Δ_i. Setting the value of n_s does not depend on the complexity of the tasks but more on the size of the task space. It needs to allow appropriate sampling of Y by each method. In our simulations, $n_s = 20$. Besides, to limit the reliance on the teacher and take into account the availability of each teacher, we penalise the preference for social learning with a $cost_i$ factor ($cost_i = 100$). For the autonomous exploration strategy, $cost_0 = 1$. The strategies are selected stochastically with a probability proportional to their preference (alg 3.2).

We applied our hierarchical **SGIM-IM** algorithm with 2 layers of active learning to an illustration experiment.

4. AirHockey Experiment

4.1. Description of the Experimental Setup

Our first experimental setup is a simulated square air hockey table that contains an obstacle (fig. 4). Starting with a fixed position and velocity (1 single context), the puck moves in straight line without friction. The effect is the position of the impact when the puck collides with the top border of the table. Y is thus the top border of the table, mapped into the $[-1, 1]$ segment, which highlights the subregion hidden by the obstacle as difficult to reach.

We control our mallet with a parameterised trajectory determined by 5 key positions $u_0, u_1, u_2, u_3, u_4 \in [-1, 1]^2$ (10 scalar parameters) at times $t_0 = 0 < t_1 < t_2 < t_3 < t_4$ (4 parameters). The trajectory in time is generated by Gaussian distance weighting:

$$u(\mathbf{t}) = \sum_{i=0}^{5} \frac{w_i(\mathbf{t})u_i}{\sum_{j=0}^{5} w_j(\mathbf{t})} \text{ with } w_i(\mathbf{t}) = e^{\sigma * |\mathbf{t}-\mathbf{t_i}|^2}, \sigma > 0 \quad (1)$$

Therefore, A is of dimension 14 and Y of dimension 1. The learner maps which trajectory of the mallet a induces a collision with the top border at position y. This is an inverse model of a highly redundant mapping, which is all the more interesting than the obstacle introduces discontinuities in the model.

4.2. Experimental Protocol

We detail in this subsection the experiments we carry with our air hockey table, how we processed to evaluate SGIM-IM and provide our learner with demonstrations.

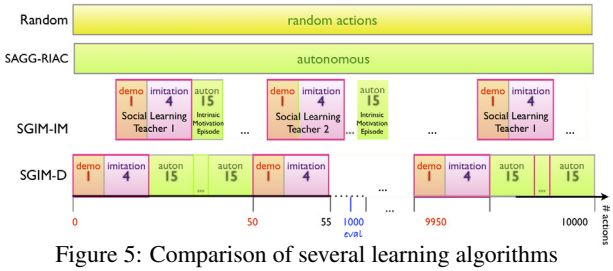

Figure 5: Comparison of several learning algorithms

4.2.1. Comparison of Learning Algorithms

To assess the efficiency of SGIM-IM, we decide to compare the performance of several learning algorithms (fig. 5):

- Random exploration: throughout the experiment, the robot picks actions randomly in the action space A.
- SAGG-RIAC: throughout the experiment, the robot explores autonomously driven by intrinsic motivation. It ignores any demonstration by the teacher.
- SGIM-IM: interactive learning where the robot learns by actively choosing between social learning strategy or intrinsic motivation strategy, and who to imitate from.
- SGIM-D: the robot's behaviour is a mixture between Imitation learning and SAGG-RIAC. When the robot sees a new demonstration, it imitates the action for a short while. Then, it resumes its autonomous exploration, until it sees a new demonstration by the teacher. Demonstrations occur every T actions of the robot.

For each experiment in our air hockey setup, we let the robot perform 10000 actions in total, and evaluate its performance every 1000 actions. For the air hockey experiment, we set the parameters of SGIM-IM to: $cost = 10$ and $ns = 20$, and those of SGIM-D to T= 50.

4.2.2. Demonstrations and Evaluation

We simulate 2 teachers by using the learning exemplars taken from Random and SAGG-RIAC learners. For teacher 1, we choose demonstrations in $[-1, 0.5]$ with each $y_{demo_k} \in [-1 + k\times0.01, -1+(k+1)\times0.01]$. For teacher 2, we likewise choose demonstrations in $[0.5, 1]$, that manage to place the puck behind the obstacle.

We assess the algorithms by measuring how close they can reach a benchmark set distributed over $Y = [-1, 1]$ and placed every 0.05, with the mean error at reaching the benchmark points.

4.3. Results

Fig.6 plots the mean distance error of the attempts to hit the border at the benchmark points, with respect to the number of actions performed by the mallet. It shows that while Random exploration and SAGG-RIAC error decrease, SGIM-IM performs significantly better, and faster. It almost divided by a factor of 10 the final error value compared to SAGG-RIAC. Its error rate is always smaller than for the other algorithms since the very beginning. SGIM-IM has taken advantage of the demonstrations very fast to be able to hit the puck and place it on the top border, instead of making random movements which would have little probability of hitting the puck, let alone placing it at the benchmark position. Its performance is comparable with SGIM-D. This shows that its active choice of strategy was able to choose social learning over autonomous learning to bootstrap its progress, and to vary its choice of teacher to overcome the limited subspaces of the demonstrations of each teacher .

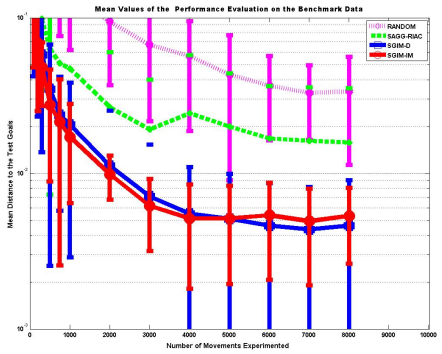

Figure 6: Evaluation of the performance of the robot with respect to the number of actions performed, under different learning algorithms. We plotted the mean distance to the benchmark set with its variance errorbar.

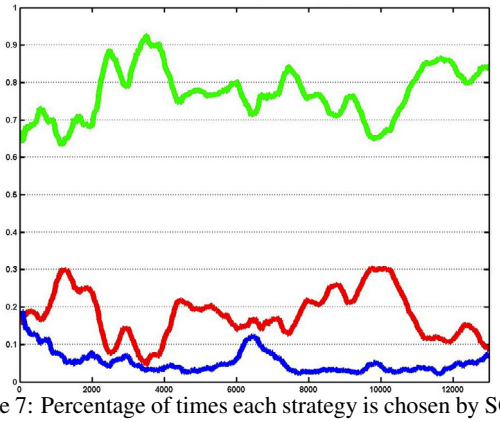

Figure 7: Percentage of times each strategy is chosen by SGIM-IM with respect to the number of actions performed: intrinsic motivation (green), social learning with teacher 1 (red) and with teacher 2 (blue).

4.4. Active Choice of Strategy

As for the strategy adopted, fig.7 shows that total number of demonstration requests increases in the very beginning, as they are most useful in the beginning, as each indicate to the learner which kind of actions can make the mallet hit the puck whereas random movements have low probability of hitting the puck. After this first phase, the learner prefers autonomous learning because of the cost of asking for teachers' help. It then increases again in the second half of the experiment when the progress made by autonomous exploration decreases. Demonstrations then help the learner improve in precision.

Furthermore, requests were asked more often to the teacher 1 as he covers a more important subspace of Y. This indicates that the learner could detect the difference in teaching capabilities of the 2 teachers. We would also like to point out that the number of demonstrations of teacher 2 made a small peak around 6500 when the error curve stops decreasing, showing that his help was most useful once the learner has managed to reach the subspace of Y that is easy to reach before getting interested in the subspace behind the obstacle. This slight peak effect can be more visible with more experiments to improve our statistics, and by complementary figures to analyse this effect.

5. Conclusion

We presented **SGIM-IM** (Socially Guided Intrinsic Motivation with Interactive learning at the Meta level), an algorithm that combines intrinsically motivated exploration and interactive learning with demonstrations. With an architecture organised into 3 layers, it actively decides when and to whom to ask for demonstrations. Through an air hockey experimental setup, we showed that SGIM-IM efficiently learns inverse models in high-dimensional, continuous and non-preset environment despite high redundancy. Its active choice of strategy was able to choose social learning over autonomous learning to bootstrap its progress, and to choose the right teacher to overcome the limited subspaces of the demonstrations of each teacher. It thus offers a framework for more flexible interaction between an autonomous learner and its users.

6. Acknowledgements

This research was partially funded by ERC Grant EXPLORERS 240007 and ANR MACSi.

7. References

[1] A. Whiten, "Primate culture and social learning," *Cognitive Science*, vol. 24, no. 3, pp. 477–508, 2000.

[2] M. Tomasello and M. Carpenter, "Shared intentionality," *Developmental Science*, vol. 10, no. 1, pp. 121–125, 2007.

[3] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, *Handbook of Robotics*. MIT Press, 2007, no. 59, ch. Robot Programming by Demonstration.

[4] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469 – 483, 2009.

[5] E. Deci and R. M. Ryan, *Intrinsic Motivation and self-determination in human behavior*. New York: Plenum Press, 1985.

[6] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, no. 599-600, 2001.

[7] M. Lopes and P.-Y. Oudeyer, "Active learning and intrinsically motivated exploration in robots: Advances and challenges (guest editorial)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 65–69, 2010.

[8] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[9] A. Baranes and P.-Y. Oudeyer, "Intrinsically motivated goal exploration for active motor learning in robots: A case study," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, oct. 2010, pp. 1766 –1773.

[10] S. M. Nguyen, A. Baranes, and P.-Y. Oudeyer, "Bootstrapping intrinsically motivated learning with human demonstrations," in *Proceedings of the IEEE International Conference on Development and Learning*, Frankfurt, Germany, 2011.

[11] B. Price and C. Boutilier, "Accelerating reinforcement learning through implicit imitation," *J. Artificial Intelligence Research*, vol. 19, no. 569-629, 2003.

[12] A. Shon, D. Verma, and R. Rao, "Active imitation learning," in *American Association for Artificial Intelligence*, vol. 22, no. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 756.

[13] G. Csibra, "Teleological and referential understanding of action in infancy," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1431, p. 447, 2003.

[14] G. Csibra and G. Gergely, "Obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans," *Acta Psychologica*, vol. 124, no. 1, pp. 60 – 78, 2007.

[15] C. L. Nehaniv and K. Dautenhahn, *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*. Cambridge: Cambridge Univ. Press, March 2007.