Strategic Robot Learner for Interactive Goal-Babbling

Sao Mai Nguyen *
Flowers Team
INRIA
Bordeaux, France
nguyensmai@gmail.com

Pierre-Yves Oudeyer †
Flowers Team
INRIA
Bordeaux, France pierre-yves.oudeyer@inria.fr

Abstract

The challenges posed by robots operating in human environments on a daily basis and on the long-term point out the importance of adaptivity to changes which can be unforeseen at design time. Therefore, the robot must learn continuously in an open-ended, non-stationary and high dimensional space. It can not possibly explore all its environment to learn about everything within a life-time. We propose to investigate the relationship between two classical learning modes: imitation learning and intrinsically-motivated autonomous exploration. We buildan algorithmic architecture where relationships between the two sampling modes intertwine into a hierarchical structure, called Socially Guided Intrinsic Motivation with Active Choice of Teachers and Strategies (SGIM-ACTS).

Indeed, we have built an intrinsically motivated active learner which learns how its actions can produce varied consequences or outcomes. For instance, the robot learns to throw a ball at different distances, by associating a distance (outcome) to a specific movement (action). It actively learns online by sampling data which it chooses by using several sampling modes. On the meta-level, it actively learns which data collection strategy is most efficient for improving its competence and generalising from its experience to a wide variety of outcomes. The interactive learner thus learns multiple tasks in a structured manner, discovering by itself developmental sequences.

We contribute to different fields of machine learning:

- imitation learning : we propose a unified structure to address simulateneously the fundamental questions of imitation learning: what, how, when and who to imitate. In particular in interactive learning,we identify advantages of combining autonomous exploration and socially guided exploration, and build an agent which decides by itself when to interact with teachers.
- multi-task learning : SGIM-ACTS can discover the structure of its environment by a goal-oriented exploration. We propose a unified architecture to approach goal-oriented imitation learning (to reproduce a demonstrated goal) and goal-directed autonomous exploration (goals guiding policy exploration).
- active learning : we investigate different levels of active learning : the learner can decide which action to take, or which goal to aim, or which sampling mode to use. Its decisions are made online, driven by artificial curiosity based on its monitoring of learning progress.
- hierarchical learning : we propose a hierarchical learning architecture to learn on several levels: policy, outcome, and mode spaces. The learner relies on hierarchical active decisions of what and how to learn driven by empirical evaluation of learning progress for each sampling mode on a meta-level.

Keywords: active learning, interactive learning, imitation learning, goal-oriented exploration, data-collection, exploration, programming by demonstration

Acknowledgements

This work was supported by the French ANR program (ANR 2010 BLAN 0216 01) through Project MACSi, as well by ERC Starting Grant EXPLORERS 240007.

*http://nguyensmai.free.fr
†www.pyoudeyer.com

1 Strategic Active Learning for Life-Long Acquisition of Multiple Skills

Life-long learning by robots to acquire multiple skills in unstructured environments poses challenges of learning in large and high-dimensional sensorimotor spaces, while their life-time allows only limited number of collected data.

1.1 Active Learning for Producing Varied Outcomes with Multiple sampling modes

The choice of a sampling mode can be formalised under the notion of strategic learning [**?**]. One perspective is learning to achieve varied outcomes and aims at selecting which outcome to spend time on. Another perspective is learning how to learn, by making explicit the choice and dependence of the learning performance on the method. However most studies have not addressed the learning of both how to learn and what to learn, to select at the same time which outcome to spend time on, and which learning method to use. Only [**?**] studies the framework of these questions. In initial work to address learning for varied outcomes with multiple methods, we proposed in [**?**] the Socially Guided Intrinsic Motivation by Demonstration (SGIM-D) algorithm which uses both 1) socially guided exploration, especially programming by demonstration [**?**] and 2) intrinsically motivated exploration, which are active learning algorithms based on measures of the evolution of the learning performance [**?**] to reach goals in a continuous outcome space.

In this paper, we extend this work and study how a learning agent can achieve varied outcomes in structured continuous outcome spaces, and how he can learn which sampling mode to choose among 1) active self-exploration, 2) reproduction of the demonstrated outcome or *emulation* of a teacher actively selected among available teachers, 3) reproduction of the demonstrated policy or *mimicry* of an actively selected teacher.

1.2 Actively Learning When, Who and What to Imitate

In this paper, we develop our social guidance into interactive learning. The learner actively requests for the information it needs and when it needs help [**?**]. For the model and experiments presented below, our agent learns to answer the four main questions of imitation learning: "what, how, when and who to imitate" [**?**, **?**] at the same time. We address active learning for varied outcomes with multiple sampling mode, multiple teachers, with a structured continuous outcome space (embedding sub-spaces with different properties). The sampling modes we consider are autonomous self-exploration, emulation and mimicking, by interactive learning with several teachers.

1.3 Our Approach

Let us consider an agent learning motor skills, i.e. how to induce any outcome $\mathcal{A} \in \mathbb{A}$ with motor programs $\pi \in \mathbb{P}$. We parameterise the outcome space with parameters $a \in A$. A policy π_b is described by motor primitives parameterised by $b \in B$.The probability of that the policy parameter b produces the outcome of parameter a is $\tilde{p}(a|b,c)$, where the probability density $\tilde{p}$ represents the physics of the environment which the agent estimates. The association (b,a) corresponds to a learning exemplar that will be memorised.

To solve the problem formalised above, we propose a system, called Socially Guided Intrinsic Motivation with Active Choice of Teacher and Strategy (**SGIM-ACTS**) that allows an online interactive learning of inverse models in continuous high-dimensional robotic sensorimotor spaces with multiple teachers, and sampling mode. SGIM-ACTS learns various outcomes with different types of outcomes, and generalises from sampled data to continuous sets of outcomes.

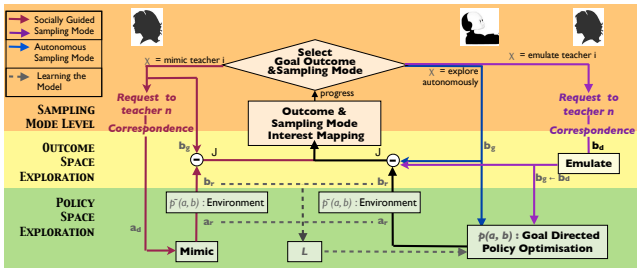
Technically, we adopt a method of generalisation of policies for new outcomes similar to [**?**, **?**], except that instead of using a pool of examples given by the teacher preset from the beginning of the experiment to learn outcomes specified by the engineer of the robot, the SGIM-ACTS algorithm decides by itself which outcomes it needs to learn more to better generalise for the whole outcome space, like in [**?**]. Moreover, SGIM-ACTS actively requests the teacher's demonstrations online, by choosing online a good sampling mode, similarly to [**?**], except that we instead of a discrete, we use a continuous outcome space. SGIM-ACTS also interacts with several teachers and uses several social learning methods.

Our active learning approach is inspired by 1) psychological theories for socially guided learning [**?**], 2) teleological learning [**?**] which considers actions as goal-oriented, and 3) intrinsic motivation in psychology [**?**] which triggers spontaneous exploration and curiosity in humans, which recently led to novel robotic and machine active learning methods which outperform traditional active learning methods [**?**].

After this definition of the problem addressed in this paper, we describe the design of our **SGIM-ACTS** (Socially Guided Intrinsic Motivation with Active Choice of Teacher and Strategy) algorithm. Then we show through an illustration experiment that SGIM-ACTS efficiently learns to realise different types of outcomes in continuous outcome spaces, and it coherently selects the right teacher to learn from.

Algorithm 2.1 SGIM-ACTS

Input: the different modes $\chi_\alpha, ...\chi_\kappa$.
Initialization: partition of outcome space $\mathcal{R} \leftarrow$ singleton A
Initialization: episodic memory (collection of produced outcomes) $\mathcal{H} \leftarrow$ empty memory
Initialization: $e \leftarrow 1$
loop
  $a_i, \chi \leftarrow$ Select Goal Outcome and Strategy$(\mathcal{R})$
  if $\chi =$ Mimic teacher i mode **then**
    $(\zeta_d, a_d) \leftarrow$ ask and observe demonstration to teacher i.
    $\gamma_1 \leftarrow$ Competence for a_g
    $\mathcal{D}_e \leftarrow$ Mimic Action(ζ_d)
    $p_{e+1} \leftarrow \mathcal{L}(p_e, \mathcal{D}_e)$
  else if $\chi =$ Emulate teacher i mode **then**
    $(\zeta_d, a_d) \leftarrow$ ask and observe demonstration to teacher i.
    Emulation: $a_g \leftarrow a_d$
    $\gamma_1 \leftarrow$ Competence for a_g
    $\mathcal{D}_e \leftarrow$ Goal-Directed Policy Optimisation(a_g)
    $p_{e+1} \leftarrow \mathcal{L}(p_e, \mathcal{D}_e)$
  else
    $\chi =$ **Intrinsic Motivation mode**
    $a_g \leftarrow a_i$
    $\gamma_1 \leftarrow$ Competence for a_g
    $\mathcal{D}_e \leftarrow$ Goal-Directed Policy Optimisation(a_g)
    $p_{e+1} \leftarrow \mathcal{L}(p_e, \mathcal{D}_e)$
  end if
  $\gamma_2 \leftarrow$ Competence for a_g
  $nbA \leftarrow$ number of episodes in $\mathcal{D}_e$
  $prog \leftarrow 2(sig(\alpha_p * \frac{\gamma_2 - \gamma_1}{|T_i| \cdot nbA}) - 1)$
  Append $\mathcal{D}_e$ to $\mathcal{H}$
  $\mathcal{R} \leftarrow$ Update Outcome and Strategy Interest Mapping$(\mathcal{R}, \mathcal{H}, a_g, prog, \chi)$
  $e \leftarrow e + 1$
end loop



(a) Time flow chart of SGIM-ACTS, which combines Intrinsic Motivation and Mimicking and Emulation into 3 layers that pertain to the sampling mode, the outcome space and the policy space exploration respectively.

Figure 1: SGIM-ACTS algorithm.

2 Algorithm Description

In this section, we describe the SGIM-ACTS architecture by giving a behavioural outline (Algorithm **??** and fig. **??**).

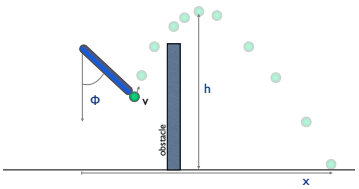
SGIM-ACTS is an architecture that merges intrinsically motivated self-exploration with interactive learning as socially guided exploration. In the latter case, a teacher performs an observed trajectory ζ which achieves an observed outcome b_d (intentional or unintentional). SGIM-ACTS learns by episodes during which it actively chooses simultaneously an outcome $b_g \in T$ to reach and a sampling mode with a specific teacher. Its choice χ is selected among : intrinsically motivated exploration, mimicry from teacher 1, emulation of teacher 1, mimicry from teacher 2, emulation of teacher 2 .... (fig. **??**).

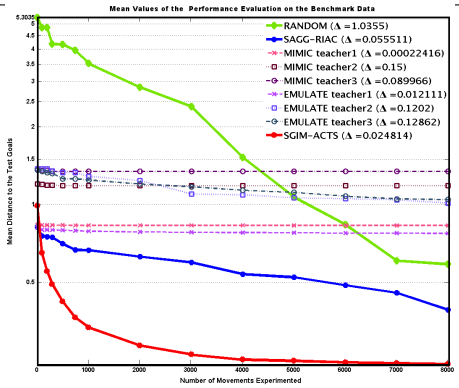
In an episode under a mimicking mode, our SGIM-ACTS learner actively self-generates a goal b_g where its competence improvement is maximal. The SGIM-ACTS learner explores preferentially goal outcomes easy to reach and where it makes progress the fastest. The selected teacher answers its request with a demonstration $[\zeta_d, b_d]$ to produce an outcome b_d that is closest to b_g. The robot mimics the teacher to reproduce ζ_d, for a fixed duration, by performing policies a_θ which are small variations of an approximation of ζ_d. In an episode under an emulation mode, our SGIM-ACTS learner observes from the selected teacher a demonstration $[\zeta_d, b_d]$. It tries different policies using goal-directed optimisation algorithms to approach the observed outcome b_d, without taking into account the demonstrated policy ζ_d. It re-uses and optimises its policy repertoire built through its past autonomous and socially guided explorations. The episode ends after a fixed duration. In an episode under the intrinsic motivation mode, it explores autonomously following the SAGG-RIAC algorithm [**?**]. It actively self-generates a goal b_g where its competence improvement is maximal, as in the mimicking mode. Then, it explores which policy a_θ can achieve b_g best. It tries different policies to approach the self-determined outcome b_g, as in the emulation mode. The episode ends after a fixed duration. The intrinsic motivation and emulation mode differ mainly by the way the goal outcome is chosen. The details of this **3-layered (policy, outcome and mode space explorations) hierarchical architecture** and the different functions and levels can be read in [**?**, **?**].

3 Experiment

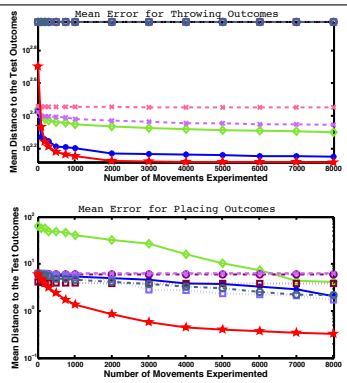
3.1 Experimental setup

Figure 2: An arm, described by its angle ϕ, is controlled by a motor primitive with 14 continuous parameters (taking bounded values) that determine the evolution of its acceleration $\ddot{\phi}$. A ball is held by the arm and then released at the end of the motion. The objective of the robot is to learn the mapping between the parameters of the motor primitive and two types of outcomes he can produce: a ball thrown at distance x and height h, or a ball placed at the arm tip at angle ϕ with velocity smaller than $|v_{max}|$.



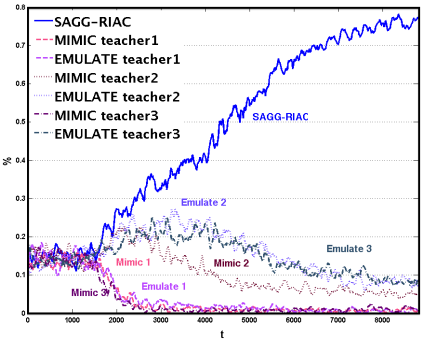


(a) Mean error for the different learning algorithms averaged over the two sub outcome spaces (final variance value Δ is indicated in the legend) .

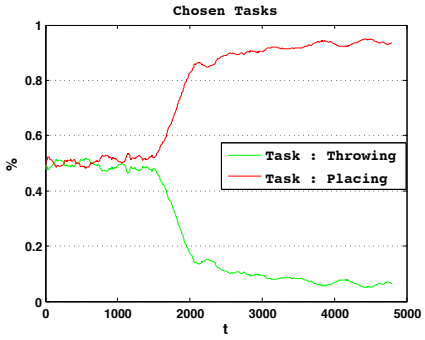


(b) Mean error for the different learning algorithms for each of the throwing outcomes and placing outcomes separately. The legend is the same as in fig. **??**.

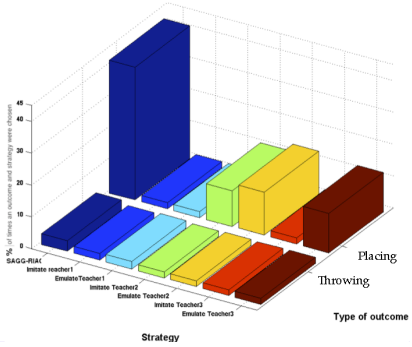
Figure 3: Error plots.



(a) Strategy chosen by SGIM-ACTS through time: percentage of times each sampling mode is chosen for several runs of the experiment.



(b) Outcome chosen by SGIM-ACTS through time: percentage of times each kind of outcome is chosen for several runs of the experiment.



(c) Consistency in the choice of outcome, teacher and mode: percentage of times each sampling mode, teacher and outcome are chosen over all the history of the robot.

Figure 4: sampling modes chosen.

We illustrate in the following section this hierarchical algorithm through a simulation where a robot learns to throw a ball or to place it at different angles (fig. **??**) with 7 sampling modes: intrinsically motivated exploration, mimicry from 3 teachers and emulation from 3 teachers. The 3 teachers considered are respectively an expert in throwing balls, an expert in placing balls, and an expert in placing balls with correspondence problems. We prepared demonstration sets for each teacher, so that the demonstrated outcomes are equally distributed in the reachable space. A demonstration is stored as a pair or policy and outcome parameters. When a teacher is requested a demonstration for emulation, he gives a random demonstration among its demonstration set. The details of the experimental setup can be read in [**?**]. In the next section, we present the results of the experiment.

3.2 Results

We compared SGIM-ACTS with 4 other learning algorithms: random exploration of the policy space, SAGG-RIAC [**?**], mimicry and emulation. Fig. **??** shows that SGIM-ACTS decreases its cumulative error for both placing and throwing. It performs better than autonomous exploration by random search or intrinsic motivation, and better than mimicry or emulation with any teacher. Fig. **??** shows that SGIM-ACTS error rate for both placing and throwing is low. For throwing, SGIM-ACTS performs the best in terms of error rate and speed because it could find the right mode. While mimicking and emulating teacher 1 decreases the error as expected, mimicking and emulating a teacher who is expert in another kind of outcomes and is bad in that outcome leaves a high error rate. For placing, SGIM-ACTS makes less error than all other algorithms. Indeed, as we expected, mimicking the teacher 2, and emulating teachers 2 and 3 enhances low error rates, while mimicking a teacher with correspondence problem (teacher 3) or an expert on another outcome (teacher 1) gives poor result. We also note that for both outcomes, mimicry does not lead to important learning progress, and the error curve is almost flat. This is due to the lack of exploration which leads the learner to ask demonstrations for outcomes only in a small subspace.

Indeed, we see in fig. **??** which illustrates the percentage times each sampling mode is chosen by SGIM-ACTS with respect to time, that mimicry of teacher 3, which lacks efficiency because of the correspondence problem, is seldom chosen by SGIM-ACTS. Mimicry and emulation of teacher 1 is also little used because autonomous learning learns

quickly throwing outcomes. Teachers 2 and 3 are exactly the same with respect to the outcomes they demonstrate, and are emulated in the same proportion. This figure also shows that the more the learner cumulates knowledge, the more autonomous he grows : his percentage of autonomous learning increases steadily.

Not only does he choose the right sampling mode, but also the right outcome to concentrate on. Fig. **??** shows that he concentrates in the end more on placing, which are more difficult.

Finally, fig. **??** shows the percentage of times over all the experiments where he chooses at the same time each outcome type, a sampling mode and a teacher. We can see that for the placing outcomes, he seldom requests help from the teacher 1, as he learns that teacher 1 does not know how to place the ball. Likewise, because of the correspondence problems, he does not mimic teacher 3. But he learns that mimicking teacher 2 and emulating teachers 2 and 3 are useful for placing outcomes. For the throwing outcomes, he uses slightly more the autonomous exploration sampling modes, as he can learn efficiently by himself. The high percentage for the other sampling mode is due to the fact that the throwing outcomes are easy to learn, therefore are learned in the beginning when a lot of sampling of all possible sampling modes is carried out. SGIM-ACTS is therefore consistent in its choice of outcomes , sampling modes and teachers.

4 Conclusion and Discussion

We presented the **SGIM-ACTS** (Socially Guided Intrinsic Motivation with Active Choice of Teacher and Strategy) algorithm that efficiently and actively combines autonomous self-exploration and interactive learning, to address the learning of multiple outcomes, with outcomes of different types, and with different sampling modes. In particular, it learns actively to decide on the fundamental questions of programming by demonstration: *what and how* to learn; but also *what, how, when and who* to imitate. This interactive learner decides efficiently and coherently whether to use social guidance. It learns when to ask for demonstration, what kind of demonstrations (action to mimic or outcome to emulate) and who to ask for demonstrations, among the available teachers. Its hierarchical architecture bears three levels. The lower level explores the policy parameters space to build skills for determined goal outcomes. The upper level explores the outcome space to evaluate for which outcomes he makes the best progress. A meta-level actively chooses the outcome and sampling mode that leads to the best competence progress. We showed that SGIM-ACTS can focus on the outcome where it learns the most, while choosing the most appropriate associated sampling mode. The active learner can explore efficiently a composite and continuous outcome space to be able to generalise for new outcomes of the outcome spaces.

Even in the case of correspondence problems, the system still takes advantage of the demonstrations to bias its exploration of the outcome space, as argued in **[?]**. Future work should test SGIM-ACTS on more complex environments, and with real physical robots and everyday human users. It would also be interesting to compare the outcomes selected by our system to developmental behavioural studies, and highlight developmental trajectories.

References

[1] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *The Journal of Machine Learning Research,*, 5:255–291, 2004.

[2] Adrien Baranes and Pierre-Yves Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

[3] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. *Handbook of Robotics*, chapter Robot Programming by Demonstration. Number 59. MIT Press, 2007.

[4] Cynthia Breazeal and B. Scassellati. Robots that imitate humans. *Trends in Cognitive Sciences*, 6(11):481–487, 2002.

[5] J. Call and M. Carpenter. *Imitation in animals and artifacts*, chapter Three sources of information in social learning, pages 211–228. Cambridge, MA: MIT Press., 2002.

[6] Sonia Chernova and Manuela Veloso. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research*, 34, 2009.

[7] Gergely Csibra. Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):447, 2003.

[8] B.C. da Silva, G. Konidaris, and Andrew G. Barto. Learning parameterized skills. In *29th International Conference on Machine Learning (ICML 2012)*, 2012.

[9] Kerstin Dautenhahn and Chrystopher L. Nehaniv. *Imitation in Animals and Artifacts*. MIT Press, 2002.

[10] Jens Kober, Andreas Wilhelm, Erhan Oztop, and Jan Peters. Reinforcement learning to adjust parametrized motor primitives to new situations. *Autonomous Robots*, pages 1–19, 2012. 10.1007/s10514-012-9290-3.

[11] Manuel Lopes and Pierre-Yves Oudeyer. The Strategic Student Approach for Life-Long Exploration and Learning. In *IEEE Conference on Development and Learning / EpiRob*, San Diego, États-Unis, November 2012.

[12] Sao Mai Nguyen, Serena Ivaldi, Natalia Lyubova, Alain Droniou, Damien Gerardeaux-Viret, David Filliat, Vincent Padois, Olivier Sigaud, and Pierre-Yves Oudeyer. Learning to recognize objects through curiosity-driven manipulation with the icub humanoid robot. In *IEEE International Conference on Development and Learning - Epirob*, 2013.

[13] Sao Mai Nguyen and Pierre-Yves Oudeyer. Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner. *Paladyn Journal of Behavioural Robotics*, 3(3):136–146, 2012.

[14] Sao Mai Nguyen and Pierre-Yves Oudeyer. Properties for efficient demonstrations to a socially guided intrinsically motivated learner. In *21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012.

[15] Sao Mai Nguyen and Pierre-Yves Oudeyer. Socially guided intrinsic motivation for robot learning of motor skills. *Autonomous Robots*, pages 1–22, 2013.

[16] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 2007.

[17] Richard M. Ryan and Edward L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54 – 67, 2000.