Learning how to reach various goals by autonomous interaction with the environment: unification and comparison of exploration strategies

Clément Moulin-Frier, Pierre-Yves Oudeyer

Learning how to reach various goals by autonomous interaction with the environment: unification and comparison of exploration strategies

Clément Moulin-Frier
Flowers team, Inria / ENSTA-Paristech
Bordeaux, France
clement.moulin-frier@inria.fr

Pierre-Yves Oudeyer
Flowers team, Inria / ENSTA-Paristech
Bordeaux, France
pierre-yves.oudeyer@inria.fr

Abstract

In the field of developmental robotics, we are particularly interested in the exploration strategies which can drive an agent to learn how to reach a wide variety of goals. In this paper, we unify and compare such strategies, recently shown to be efficient to learn complex non-linear redundant sensorimotor mappings. They combine two main principles. The first one concerns the space in which the learning agent chooses points to explore (motor space vs. goal space). Previous works (Rolf et al., 2010; Baranes and Oudeyer, 2012) have shown that learning redundant inverse models could be achieved more efficiently if exploration was driven by goal babbling, triggering reaching, rather than direct motor babbling. Goal babbling is especially efficient to learn highly redundant mappings (e.g the inverse kinematics of a arm). At each time step, the agent chooses a goal in a goal space (e.g uniformly), uses the current knowledge of an inverse model to infer a motor command to reach that goal, observes the corresponding consequence and updates its inverse model according to this new experience. This exploration strategy allows the agent to cover the goal space more efficiently, avoiding to waste time in redundant parts of the sensorimotor space (e.g executing many motor commands that actually reach the same goal). The second principle comes from the field of active learning, where exploration strategies are conceived as an optimization process. Samples in the input space (i.e motor space) are collected in order to minimize a given property of the learning process, e.g the uncertainty (Cohn et al., 1996) or the prediction error (Thrun, 1995) of the model. This allows the agent to focus on parts of the sensorimotor space in which exploration is supposed to improve the quality of the model.

This paper shows how an integrating probabilistic framework allows to model several recent algorithmic architectures for exploration based on these two principles, and compare the efficiency of various exploration strategies to learn how to uniformly cover a goal space.

Keywords: Exploration strategies, goal babbling, active learning, developmental robotics.

Acknowledgements

This work was partially financed by ERC Starting Grant EXPLORERS 240 007.

1 Introduction

The learning of sensorimotor tasks, for example reaching objects with the hand or controlling the shape of a vocal tract to produce particular sounds, involves the learning of complex sensorimotor mappings. This latter generally requires to build a model of the relationships between parts of the sensorimotor space. For example, one might need to predict the positions of the hand knowing the joint configurations, or to control the shape the vocal tract to produce the sound of particular words.

Let us introduce the problem more formally. A learning agent interacts with a surrounding environment through motor commands M and sensory perceptions S. We call $f : M \rightarrow S$ the unknown function defining the physical properties of the environment, such that when the agent produces a motor command $m \in M$, it then perceives $s \in S$. Classical robotic problems are e.g. the prediction of the sensory effect of an intended motor command through a forward model $\tilde{f} : M \rightarrow S$, or the control of the motor system to reach sensory goals through an inverse model $\tilde{f^{-1}} : S \rightarrow M$. The agent has to learn such models by collecting (m, s) pairs through its interaction with the environment, i.e. by producing $m \in M$ and observing $s = f(m)$. These learning processes are often difficult for several reasons: 1) the agent has to deal with uncertainties both in the environment and in its own sensorimotor loop, 2) M and S can be highly dimensional, such that random sampling in M to collect (m, s) pairs can be a long and fastidious process, 3) f can be strongly non-linear, such that the learning of $\tilde{f}$ from experience is not trivial, 4) f can be redundant (many M to one S), such that the learning of $\tilde{f^{-1}}$ is an ill-posed problem ($f^{-}1$ does not exist, or cannot be directly recovered from f).

When a learning process faces these issues, random motor exploration (or motor babbling) in M is not a realist exploration strategy to collect (m, s) pairs. Due to high dimensionality, data are precious whereas, due to non-linearity and/or redundancy, data are not equally useful to learn an adequate forward or inverse model.

2 Exploration strategies

Computational studies have shown the importance of developmental mechanisms guiding exploration and learning in high-dimensional M and S spaces and with highly redundant and non-linear f (Oudeyer et al., 2007; Baranes and Oudeyer, 2012). Among these guiding mechanisms, intrinsic motivations, generating spontaneous exploration in humans (Berlyne, 1954; Deci and Ryan, 1985), have been transposed in curiosity-driven learning machines (Schmidhuber, 1991; Barto et al., 2004; Schmidhuber, 2010) and robots (Oudeyer et al., 2007; Baranes and Oudeyer, 2012) and shown to yield highly efficient learning of inverse models in high-dimensional redundant sensorimotor spaces (Baranes and Oudeyer, 2012). Efficient versions of such mechanisms are based on the active choice of learning experiments that maximize learning *progress*, for e.g. improvement of predictions or of competences to reach goals (Schmidhuber, 1991; Oudeyer et al., 2007). This automatically drives the system to explore and learn first easy skills, and then explore skills of progressively increasing complexity.

This led to the implementation of various exploration strategies (Baranes and Oudeyer, 2012), which differ in the way the agent iteratively collects (m, s) pairs to learn forward and/or inverse models (comparing random vs. learning progress based exploration, in either the motor M or the sensory S spaces). These strategies are summarized below (the original name of the corresponding algorithm appears in parenthesis).

- **Random motor exploration (ACTUATOR-RANDOM):** at each time step, the agent randomly chooses an articulatory command $m \in M$, produces it, observes $s = f(m)$ and updates its sensorimotor model according to this new experience (m, s).
- **Random goal exploration (SAGG-RANDOM):** at each time step, the agent randomly chooses a goal $s_g \in S$ and tries to reach it by producing $m \in M$ using an inverse model $\tilde{f^{-1}}$ learned from previous experience. It observes the corresponding sensory consequence $s = f(m)$ and updates its sensorimotor model according to this new experience (m, s).
- **Active motor exploration (ACTUATOR-RIAC):** at each time step, the agent chooses a motor command m by maximizing an interest value in M based on an empirical measure of the learning progress in prediction in its recent experience. The agent uses a forward model $\tilde{f}$ learned from its past experience to make a prediction $s_p \in S$ for the motor command m. It produces m and observe $s = f(m)$. The agent updates its sensorimotor model according to the new experience (m, s). A measure of learning accuracy is computed from the distance between s_p and s, which is used to update the interest model in the neighborhood of m.
- **Active goal exploration (SAGG-RIAC):** at each time step, the agent chooses a goal s_g by maximizing an interest value in S based on an empirical measure of the learning progress in competence to reach goals in its recent experience. It tries to reach s_g by producing $m \in M$ using a learned inverse model $\tilde{f^{-1}}$. It observes the corresponding sensory consequence $s \in S$ and updates its sensorimotor model according to this new experience

(m, s). A measure of learning accuracy is computed from the distance between s_g and s, which is used to update the interest model in the neighborhood of s_g.

In the two active strategies, the measure of interest was obtained by recursively splitting the space (M in ACTUATOR-RIAC, S in SAGG-RIAC) into sub-regions during the agent life. Each region maintains its own empirical measure of learning progress from its learning accuracy history in a relative time window. This accuracy is defined as the opposite of the distance between s_p and s in the active motor strategy, between s_g and s in the active goal one. These active strategies are very related to the field of *active learning*, although this latter often constrains the interest measure to be defined in the input space (M in our formalism).

We have recently suggested to classify these four strategies along two dimensions (Moulin-Frier and Oudeyer, 2013a,b). The first one corresponds to the space X in which the agent drives its exploration, which is here either M (motor strategies) or S (goal strategies). We call it the *choice space*. The second dimension is the kind of interest measure used by this agent at each time step to choose a point in its choice space, either uniform leading to a random sampling in X (random strategies), or based on empirical measurements, here the learning progress in prediction or control (active strategies).

3 Probabilistic modeling

We use a probabilistic framework where the notations are inspired by Jaynes (2003) and Lebeltel et al. (2004). Upper case A denotes a probabilistic variable, defined by its continuous, possibly multidimensional and bounded domain $\mathcal{D}(A)$. The conjunction of two variables $A \wedge B$ can be defined as a new variable C with domain $\mathcal{D}(A) \times \mathcal{D}(B)$. Lower case a will denote a particular value of the domain $\mathcal{D}(A)$. $p(A \mid \omega)$ is the probability distribution over A knowing some preliminary knowledge ω (e.g. the parametric form of the distribution, a learning set ...). Practically, ω will serve as a model identifier, allowing to define different distributions of the same variable, and we will often omit it in the text although it will be useful in the equations. $p(A\ B \mid \omega)$ is the probability distribution over $A \wedge B$. $p(A \mid [B = b]\ \omega)$ is the conditional distribution over A knowing a particular value b of another variable B (also noted $p(A \mid b\ \omega)$ when there is no ambiguity on the variable B). For simplicity, we will often confound a variable and its domain, saying for example *"the probability distribution over the space A"*.

Considering that we know the joint probability distribution over the whole sensorimotor space, $p(M\ S \mid \omega_{SM})$, Bayesian inference provides the way to compute every conditional distribution over $M \wedge S$. In particular, we can compute the conditional distribution over Y knowing a particular value x of X, as long as X and Y correspond to two complementary sub-domains of $M \wedge S$ (i.e. they are disjoint and $X \wedge Y = M \wedge S$). Thus, the prediction of $s_p \in S$ from $m \in M$ in the active motor exploration strategy, or the control of $m \in M$ to reach $s_g \in S$ in the active or random goal exploration strategies, correspond to the probability distributions $p(S \mid M\ \omega_{SM})$ and $P(M \mid S\ \omega_{SM})$, respectively. More generally, whatever the choice and inference spaces X and Y, as long as they are subspaces of $M \wedge S$ and they are disjoint, Bayesian inference allows to compute $p(Y \mid X\ \omega_{SM})$.

Such a probabilistic modeling is also able to express the interest model, that we will call ω_I, such that the agent draws points in the choice space X according to the distribution $p(X \mid \omega_I)$. In the random motor and goal exploration strategies, this distribution is uniform, whereas it is a monotonically increasing function of the empirical interest measure in the case of the active exploration strategies.

Given this probabilistic framework, Algorithm 1 describes our generic exploration algorithm.

Algorithm 1 Generic exploration algorithm

```
 1: set choice space $X$
 2: while true do
 3:     $x \sim p(X \mid \omega_I)$
 4:     $y \sim p(Y \mid x\ \omega_{SM})$
 5:     $m = M\ ((x, y))$
 6:     $s = exec(m)$
 7:     $e = distance(S(x, y), s)$
 8:     $update(\omega_{SM}, (m, s))$
 9:     $update(\omega_I, (x, e))$
10: end while
```

Line 1 defines the choice space of the exploration strategy. For example X is set to M for the motor strategies and to S for the goal strategies described in Section 2, but the formalism can also deal with any part of $M \wedge S$ as the choice space. Line 3, the agent draws a point x in the choice space X according to the current state of its interest model ω_I, through the probability distribution $p(X \mid \omega_I)$ encoding the current interest over X. This distribution is uniform in the case of the random strategies and related to the learning progress in prediction or control in the active strategies of Section 2. Line 4, the agent draws a point y in the inference space Y (remember that Y is such that $X \wedge Y = M \wedge S$) according to the distribution $p(Y \mid x\, \omega_{SM})$, using Bayesian inference on the joint distribution $p(M\ S \mid \omega_{MS})$. If $X = M$, and therefore $Y = S$, this corresponds to a prediction tasks $p(S \mid [M = x])$; if $X = S$, and therefore $Y = M$, this corresponds to a control task $p(M \mid [S = x])$. Line 5, the agent extracts the motor part m of (x, y), noted $M\,((x, y))$, i.e. x if $X = M$, y if $X = S$. Line 6, the agent produces m and observe $s = exec(m)$, i.e. $s = f(m)$ with possible sensorimotor constraints and noises. Line 7 the agent computes a learning error as a distance betwween the sensory part of (x, y), noted $S(x, y)$, i.e. y if $X = M$, x if $X = S$, and the actual sensory consequence s. Line 8 the agent updates its sensorimotor model according to its new experience (m, s). Line 9 the agent updates its interest model according to the choice $x \in X$ it made and the associated learning error e.

In this framework, we are able to more formally express each algorithm presented in Section 2. The random motor strategy (ACTUATOR-RANDOM) is the simpler case where the choice space is $X = M$ and the interest model of line 3 is set to a uniform distribution over X. Inference in line 4 is here useless because motor extraction (line 5) will return the actual choice x and that there is no need to update the interest model in line 9. The active motor strategy (ACTUATOR-RIAC) differs from the previous one by the interest model of line 3 which favors regions of X $(= M)$ maximizing the learning progress in prediction. This latter is computed at the update step of line 9 using the history of previous learning errors computed at line 7, which are here distances between the prediction $y \in Y$ computed on line 4 (with $Y = S$) and the actual realization $s \in S$ of line 6. The random goal strategy (SAGG-RANDOM) is the case where the interest model is uniform and the choice space is S, implying that the inference corresponds to a control task to reach $x \in X$ by producing $y \in Y$ (with $X = S$ and therefore $Y = M$). Finally, the active goal strategy (SAGG-RIAC) differs from the previous one by the interest model which favors regions of X $(= S)$ maximizing the learning progress in control. This latter is computed in the same way that for ACTUATOR-RANDOM, except that the distance is here between the chosen goal $x \in X$ and the actual realization $s \in S$ (with $X = S$).

We do not develop in this abstract how the sensorimotor and the interest distributions can be practically implemented (see e.g. Moulin-Frier and Oudeyer (2013a,b) and further papers of the authors). We therefore directly provide comparative results in the next section, asking the reader to assume that these distributions can be computed in a way or another.

4 Results

In this section, we perform computer simulations with a simulated sensorimotor agent. The motor space M is articulatory (7-dimensional), and the sensory one is auditory (2-dimensional). The unknown function $f : M \rightarrow S$ is provided by the articulatory synthesizer of the DIVA model described in Guenther et al. (2006), a computational model of the human vocal tract. We do not present it here, the only important point being that the articulatory-to-auditory transformation is known to be redundant and non-linear. The agent implements Algorithm 1 with different choice spaces and interest distributions corresponding to the four strategies ACTUATOR-RANDOM, ACTUATOR-RIAC, SAGG-RANDOM and SAGG-RIAC described in Section 2. We evaluate the efficiency of the obtained sensorimotor models to achieve a control task, i.e. to reach a test set of goals uniformly distributed in the reachable auditory space.

Figure 1 shows the performance results of the four exploration strategies on a control task during the life time of learning agents. We observe that the strategies with S as the choice space (random and active goal strategies) are significantly more efficient that those with M (random and active motor strategies), i.e. both convergence speed (say around 100 updates) and generalization at the end of the simulation (500 updates) are better. Moreover, both convergence speed and generalization are better for the active than for the random goal strategy. These results are similar (though less significant) to those obtained in previous experiments (Baranes and Oudeyer, 2012) in other sensorimotor spaces (e.g. a arm reaching points on a plan), and we refer to the corresponding paper for a thorough analysis of these results.

5 Conclusion

We have integrated in this paper two important exploration principles of developmental robotics (exploration in the sensory space and active learning based on an empirical measure of the competence progress) into an integrated probabilistic framework able to express various exploration strategies in a compact and unified manner. This allowed quantitative comparisons of these strategies, showing that an active goal exploration is the most efficient to reach a set of goals uniformly sampled in the reachable part of the sensory space –as already shown in previous works of our team.

Further works should rely the approach to other tentatives of exploration strategy unification (e.g. Lopes and Oudeyer (2012); Oudeyer and Kaplan (2007)). We also want to study the effect of an online adaptation of the choice space, taking advantage of the fact that our formalism does not restrict it to be either M or S. For example, we could study how the agent iteratively adapts which part of the sensorimotor space it is interested in at a given time of its development,

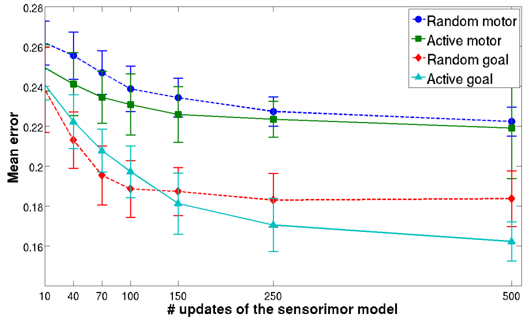

Figure 1: Performance comparison of the four exploration strategies. X-axis: number of update of the sensorimotor model. Y-axis: Mean error on a control task where an agent has to reach 30 test points uniformly distributed in the reachable area of S. For each evaluation point $s_g \in S$, the agent infers 10 motor commands in M from the distribution $p(M \mid s_g \, \omega_{SM})$, where ω_{SM} is the state of the sensorimotor model at the corresponding time step (number of update on the X axis). The error of an agent at a time step is the mean distance between the sensory points actually reached by the 10 motor commands and the evaluation point s_g. Each curve plots the mean and standard deviation of the error for 10 independent simulations with different random seeds, for each of the four exploration strategies described in the previous sections.

favoring exploration in sensorimotor *dimensions* which display higher measures of learning progress. Finally, we are currently extending the implementation to learn how to control sequences of motor commands.

References

Baranes, A. and Oudeyer, P.-Y. (2012). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*.

Barto, A., Singh, S., and Chenatez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proc. 3rd Int. Conf. Dvp. Learn.*, pages 112–119, San Diego, CA.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, 45:180–191.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

Deci, E. and Ryan, R. M. (1985). *Intrinsic Motivation and self-determination in human behavior*. Plenum Press, New York.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, 96(3):280–301.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Lebeltel, O., Bessiere, P., Diard, J., and Mazer, E. (2004). Bayesian robot programming. *Autonomous Robots*, 16:4979.

Lopes, M. and Oudeyer, P.-Y. (2012). The strategic student approach for life-long exploration and learning. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–8. IEEE.

Moulin-Frier, C. and Oudeyer, P.-Y. (2013a). Exploration strategies in developmental robotics: a unified probabilistic framework. In *International Conference on Development and Learning, Epirob, Osaka, Japan*.

Moulin-Frier, C. and Oudeyer, P.-Y. (2013b). The role of intrinsic motivations in learning sensorimotor vocal mappings: a developmental robotics study. In *Proceedings of Interspeech*, page In press, Lyon, France.

Oudeyer, P.-Y. and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286.

Rolf, M., Steil, J., and Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. *IEEE Trans. Autonomous Mental Development*, 2(3):216–229.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In Meyer, J. A. and Wilson, S. W., editors, *Proc. SAB'91*, pages 222–227.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247.

Thrun, S. (1995). Exploration in active learning. *Handbook of Brain Science and Neural Networks*, pages 381–384.