

Maturationally-Constrained Competence-Based Intrinsically Motivated Learning

Adrien Baranes, Pierre-Yves Oudeyer

▶ **To cite this version:**

Adrien Baranes, Pierre-Yves Oudeyer. Maturationally-Constrained Competence-Based Intrinsically Motivated Learning. IEEE International Conference on Development and Learning (ICDL 2010), 2010, Ann Arbor, United States. inria-00541770

HAL Id: inria-00541770
https://inria.hal.science/inria-00541770

Submitted on 1 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Maturationally-Constrained Competence-Based Intrinsically Motivated Learning

Adrien Baranes and Pierre-Yves Oudeyer
INRIA, France

***Abstract*—This paper studies the coupling of intrinsic motivation and physiological maturational constraints, and argues that both mechanisms may have complex bidirectional interactions allowing to actively control the growth of complexity in motor development. First, we introduce the self-adaptive goal generation algorithm (SAGG), instantiating an intrinsically motivated goal exploration mechanism for motor learning of inverse models. Then, we introduce a functional model of maturational constraints inspired by the myelination process in humans, and show how it can be coupled with the SAGG algorithm, forming a new system called McSAGG. We then present experiments to evaluate qualitative properties of these systems when applied to learning a reaching skill with an arm with initially unknown kinematics.**

I. Introduction

In spite of many innate capabilities, human infants are born with a serious lack of knowledge, know-how, and mastery of their own body. Using their aptitude to learn, they increase their capabilities by exploring the world where they evolve and by interacting with other humans, during their entire life span. Intrinsic motivation mechanisms have been shown to be fundamental in their self-exploration behavior [1], [2]. Based on self-determination theories [3], they typically deal with ways to define what is *interesting*.

Various works focused on ways to implement such systems to make a robot learn the relationships existing between parts of its body, or between its body and the outside world (see [4], [5] for an overview). Most of them can be defined as **knowledge based models** according to the terminology introduced in [6]. They study the evolution of knowledge about the world, and typically deal with a notion of interest related to *comparisons between a predicted flow of sensorimotor values, based on an internal forward model, with the actual flow of values*. Efficient when a robot is learning forward models, like consequences of its actions for given contexts, these systems have not been designed for the learning of inverse models of highly-redundant systems. Actually, they do not consider the notion of goal, or task, and only try to improve the quality of forward models with no consideration about how they can be reused for control (this applies for instance to IAC [5] and RIAC [7]). Therefore, they might typically spend large amounts of time exploring variants of actions or sequences of actions that produce the same effect, at the disadvantage of exploring other actions that might produce different outcomes and thus be useful to achieve more tasks. (e.g. learning 10 ways to push a ball forwards, instead of learning to push a ball in 10 different directions).

A way to address this issue is to introduce goals explicitly and drive exploration at the level of these goals, which the system then tries to reach with a lower-level goal-reaching architecture typically based on coupled inverse and forward models, which may include a lower-level goal-directed active exploration mechanism. Such an architecture can be called a **competence-based intrinsic motivation** mechanism, as outlined in [6]; indeed, here they propose a definition of interesting events as related to *comparisons between self-generated goals, which are particular configurations in the sensorimotor space, and the extent to which they are reached in practice, based on an internal inverse model that may be learnt*.

Also conceptualized as active learning heuristics [8], these two kinds of approaches to intrinsic motivation typically have to initially explore the largest part of the whole sensorimotor space, before discriminating which subspaces are the most interesting. This raises the problem of (meta-)exploration in open-ended worlds where typical developmental robots evolve. Actually, in this kind of unprepared spaces where limits are typically unknown, exploring the whole sensorimotor space for extracting "interesting" subspaces is already often impossible in the life-span of an organism.

Biological constraints on the learning process in infants may be a potential solution for this open-ended exploration problem. Actually, the progressive biological maturation of infants' brain, motor and sensor capabilities, introduces numerous important constraints on the learning process [9]. Indeed, at birth, all the sensorimotor apparatus is neither precise enough, nor fast enough, to allow infants to perform complex tasks. The low visual acuity of infants [10], their incapacity to efficiently control distal muscles, and to detect high-frequency sounds, are examples of constraints reducing the complexity and limiting the access to the high-dimensional and open-ended space where they evolve [11].

Some first attempts have been proposed for creating learning systems using these concepts individually. For e.g., [4], [5], [7], [12]–[17], propose different frames for intrinsic motivations systems, typically based on the evolution of the learning process over time. Developmental constraints have also been studied, e.g., Lungarella and Berthouze [18], show experiments about the evolution of locomotion capacities for a walking robot, by releasing progressively degrees of freedoms, and resolving learning problems due to redundancies. As a link between intrinsic motivations, and maturational constraints, Lee et al. [19] introduced the Lift-Constraint, Act, Saturate (LCAS) algorithm, which deals with discrete developmental stages of visual acuity, whose maturation level increases, depending on a notion of saturation linked to the estimation of novelty.

In this paper, we introduce a new exploration algorithm, called **Maturationally-constrained Self-Adaptive Goal Generation (McSAGG)**, also based on the notion of maturation, but focused on linking intrinsic motivations in the competence based framework with the release of different kinds of maturational constraints. Here, we emphasize the idea that *intrinsic motivations do not only have to be seen as*

answering the problem of "what to learn", but can also be considered as providing biological measures, able to make the learning system progressively evolve by itself, by controlling its constraints and its own capabilities.

In the following section, we introduce the **Self-Adaptive Goal-Generation SAGG** algorithm as a new instantiation of the competence based intrinsic motivation framework [6]. Then, we couple this algorithm with a model of the myelination process appearing in the brain, and responsible of different maturational constraints. Finally, we present qualitative results of our architecture with a simulated robot that explores and learn to control its initially unknown arm.

II. Competence Based Intrinsic Motivations: The Self-Adaptive Goal Generation Algorithm

A. Global Architecture

Let us consider the definition of competence based models outlined in [6], and extract from it two different levels for active learning defined at different time scales (Fig. 1):

1) The higher level of active learning (higher time scale) considers the *active self-generation and self-selection of goals*, depending on a feedback defined using the level of achievement of previously generated goals.
2) The lower level of active learning (lower time scale) considers the *goal-directed active choice and active exploration* of lower-level actions to be taken to reach the goals selected at the higher level, and depending on local measures about the evolution of the quality of learnt inverse and/or forward models.

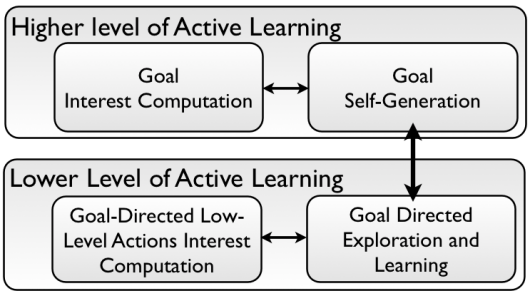


Fig. 1. Global Architecture of the SAGG algorithm. The structure is composed of two parts defining two levels of active learning: a higher which considers the active self-generation and self-selection of goals, and a lower, which considers the goal-directed active choice and active exploration of lower-level actions, to reach the goals selected at the higher level.

B. Model Formalization

Let us consider a robotic system whose configurations are described in both a configuration space S, and an operational/task space S'. For given configurations $(s_1, s'_1) \in S \times S'$, a sequence of actions $a = \{a_1, a_2, ..., a_n\}$ allows a transition toward the new states $(s_2, s'_2) \in S \times S'$ such that $(s_1, s'_1, a) \Rightarrow (s_2, s'_2)$. For instance, in the case of a robotic manipulator, S represents its joint space, and S', the operational space corresponding to the cartesian position of its end-effector.

In the frame of SAGG, we are interested in the reaching of *goals*, from starting states. Also, we formalize starting states as configurations $(s_0, s'_0) \in S \times S'$ and goals, as a desired $s'_g \in S'$. All states are here considered as potential starting states, therefore, once a goal has been generated, the lower level of active learning always try to reach it by starting from the current state (s_c, s'_c) of the system.

When a given goal is set, the low-level process of goal-directed active exploration and learning to reach this goal from the starting state can be seen as exploration and learning of a motor primitive parameterized by both this goal and parameters of already learnt internal forward and inverse models. Also, according to the self-generation and self-selection of goals at the higher level, we deduce that the whole process (higher and lower time scales) developed in SAGG can be defined as an autonomous system that explores and learns *fields of motor primitives*.

We can easily make an analogy of this formalization with the *Semi-Markov Option* framework introduced by Sutton [20]. In the case of SAGG, when considering an option $(\mathcal{I}, \pi, \beta)$, we can firstly define the initiation set $\mathcal{I} : (S, S') \rightarrow [0; 1]$, where $\mathcal{I}$ is true everywhere, because, as presented before, every state can be considered as a starting state. Also, goals are related to the terminal condition β, and the policy π encodes the skill learnt through the process induced by the lower-level of active learning and shall be indiced by the goal $s_{g'}$, i.e. $\pi_{s_{g'}}$. More formally, as induced by the use of semi-markov options, we define policies and termination conditions as dependent on all events between the initiation of the option, and the current instant. This means that the policy π, and β are depending on the *history* $h_{t\tau} = \{s_t, s'_t, a_t, s_{t+1}, s'_{t+1}, a_{t+1}..., s_\tau, s'_\tau\}$ where t is the initiation time of the option, and τ, the time of the latest event. Denoting the set of all histories by Ω, the policy and termination condition become defined by $\pi : \Omega \times \mathcal{A} \rightarrow [0; 1]$ and $\beta : \Omega \rightarrow [0; 1]$.

Moreover, because we have to consider cases where goals are not reachable, we need to define a *timeout* k which allows to stop a goal reaching attempt once a maximal number of actions has been executed. We thus need to consider $h_{t\tau}$, to stop π, (i.e. the low-level active learning process), if $\tau > k$.

Eventually, using the framework of options, we can define the process of goal self-generation, as the self-generation and self-selection of options, a goal reaching attempt corresponding to the learning of a particular option. Therefore, the global SAGG process can be also described as exploring and learning *fields of options*.

C. Measure of Competence

A reaching attempt in direction of a goal is defined as terminated according to two conditions:

1) A timeout related to a maximal number of actions allowed has been exceeded.
2) The goal has effectively been reached.

We then introduce a measure of competence as a measure of the similarity between the state reached when the goal reaching attempt has terminated and the actual goal of this reaching attempt. Before describing the mathematical formulation of competence, let us define what a reached goal is: deciding a goal s'_g as reached lie on the comparison of this precise goal state to the state resulting of a reaching attempt s'_f, using a function D defining a measure of distance. Also, $D(s'_g, s'_f) < \varepsilon_D$ corresponds to a goal reached, with ε_D, a tolerance factor.

In the continuous world where we would like that SAGG behaves, we consider a measure of competence as linked with this criteria of distance: once a reaching attempt has been terminated, we measure competence as the final obtained distance $D(s'_g, s'_f)$, normalized by the original distance

$D(s'_c, s'_g)$, between the starting state s'_c, and the goal. We thus use the following formalization of the competence, described by $\gamma_{s'_g}$:

$$\gamma_{s'_g} = -\frac{D(s'_f, s'_g)}{D(s'_c, s'_g) + 1} \quad (1)$$

Here, $\gamma_{s'_g}$ is equal to the opposite value of the ratio between the obtained distance $D(s'_g, s'_f)$, compared to the original distance $D(s'_c, s'_f)$ with an addition of 1 to avoid a division by zero, such that a high level of competence be represented by the value $\gamma_{s'_g} \approx 0$, and a negative value otherwise.

D. Lower Time Scale: Active Goal Directed Exploration and Learning

The goal directed exploration and learning mechanism can be carried out in numerous ways. Its main idea is to guide the system toward the goal, by executing low-level actions, which allow to progressively explore the world and create a model that may be reused afterwards. Its conception has to respect two imperatives :

1) A model (inverse and/or forward) has to be computed during the exploration, and has to be available for a later reuse, when considering other goals.
2) A learning feedback has to be added, such that the exploration is active, and the selection of new actions depends on local measures about the evolution of the quality of the learnt model.

In the experiment introduced in the following, we will use a method inspired by the SSA algorithm introduced by Schaal & Atkeson [21]. This system is organized around two alternating phases: *reaching* phases, which involve a local controller to drive the system towards the goal, and *local exploration* phases, which allows to learn the inverse model of the system in the close vicinity of the current state, and are triggered when the reliability of the local controller is too low. Other kinds of techniques, such as natural actor-critic architectures in model based reinforcement learning [22], could also be used.

E. Higher Time Scale: Goal Self-Generation and Self-Selection

1) Definition of Local Competence: The active goal self-generation and self-selection lie on a feedback linked with the notion of competence introduced above, and study more precisely the progress of local competences. We firstly define this notion of local competences: let us consider different measures of competence $\gamma_{s'_i}$ computed for reaching attempts to different goals $s'_i \in S', i > 1$. For a subspace called a region $\mathcal{R} \subset S'$, we can compute a measure of competence γ'' that we call a *local measure* such that:

$$\gamma'' = \left(\frac{\sum(\gamma_{s'_j})}{|\mathcal{R}|} | s'_j \in \mathcal{R} \right) \quad (2)$$

with $|\mathcal{R}|$, cardinal of $\mathcal{R}$.

2) Evolution of Competences: Then, let us suppose that each reaching attempt is indexed by the instant t when it has been decided. For a chosen goal $s'_j \in \mathcal{R}$, selected at instant t, we then obtain a measure of competence $\gamma_{s'_j}(t)$. It is important to notice that $\gamma_{s'_j}(t)$ can evolve over time: let us consider a reachable goal s'_j, that needs an important number of attempts to be reached, due to the complexity of the model that has to be explored and learnt to allow its achievement. Selecting a same goal s'_j, n times, and observing the resulting competences $\gamma'_j(1), \gamma'_j(2), ..., \gamma'_j(n)$, would let us observe an increase of $\gamma'_j(t)$ over time.

In the following, we merge the notion of local measure of competence, with the phenomenon of increase of competence for precise goals, to determine a notion of interest in different subspaces of S'.

3) Interest Value from Evolution of Local Competences: Let us consider different regions $\mathcal{R}_i$ of S' such that $\mathcal{R}_i \subset S'$. Each $\mathcal{R}_i$ contains attempted goals $\{s'_{t_1}, s'_{t_2}, ..., s'_{t_k}\}_{\mathcal{R}_i}$, and corresponding competences obtained $\{\gamma_{s'_{t_1}}, \gamma_{s'_{t_2}}, ..., \gamma_{s'_{t_k}}\}_{\mathcal{R}_i}$, indexed by their relative time order $t_1 < t_2 < ... < t_k | t_{n+1} = t_n + 1$ of experimentation inside this precise subspace $\mathcal{R}_i$ (ti are not the absolute time, but integer indices of relative order in the given subspace being considered for goal selection). The interest value, described by equation 3, represents *the derivative of the local competence value inside $\mathcal{R}_i$, over a time window of the ζ more recent goals attempted inside $\mathcal{R}_i$*:

$$interest(\mathcal{R}_i) = \frac{\left(\sum_{j=|\mathcal{R}_i|-\zeta}^{|\mathcal{R}_i|-\frac{\zeta}{2}} \gamma_{s'_j} \right) - \left(\sum_{j=|\mathcal{R}_i|-\frac{\zeta}{2}}^{|\mathcal{R}_i|} \gamma_{s'_j} \right)}{\zeta} \quad (3)$$

This equation shows the utilization of a derivative, to compute the interest. This derivative represents the *competence progress* of the system, computed inside each considered subspace: an increasing competence inside $\mathcal{R}_i$ means that the expected competence gain inside $\mathcal{R}_i$ is important. We deduce that, potentially, selecting new goals in subspaces of high competence progress could bring, on the one hand, a high information gain for the learnt model, and on the other hand, could lead to the reaching of not already reached goals.

4) Goal Self-Generation Using the Measure of Interest: Using the previous description of interest, the goal self-generation and self-selection mechanism has to carry out two different processes:

1) Split of the space S' where goals are chosen, into subspaces, according to heuristics that allows to distinguish approximately maximally areas according to their levels of interest;
2) Select the subspaces where future goals will be chosen;

Such a mechanism has been described in the Robust-Intelligent Adaptive Curiosity (R-IAC) algorithm introduced in [7]. Here, we use the same kind of methods like a recursive split of the space, each split being triggered once a maximal number of goals has been attempted inside. Each split is performed such that is maximizes the difference of the *interest* measure described above, in the two resulting subspaces, this allows to easily separate areas of different interest, and thus, of different reaching difficulty.

Finally, goals are chosen according to the following heuristics which mixes three modes, and once at least two regions exist after an initial random exploration of the whole space:

1. In $A\%$ percent (typically $A = 70\%$) of goal selections, the algorithm chooses a random goal inside a region chosen

proportionally to its interest value:

$$P_n = \frac{|interest_n - \mathbf{min}(interest_i)|}{\sum_{i=1}^{|R_n|} |interest_i - \mathbf{min}(interest_i)|} \quad (4)$$

Where P_n is the probability of selection of the region R_n, and $interest_i$ corresponds to the current $interest$ of regions R_i.
2. In B% of cases (typically $B = 20\%$), the algorithm selects a random goal inside the whole space.
3. In C% (typically $C = 10\%$), it performs a random experiment inside the region where the mean competence level is the lower.

III. Developmental Stages: The Maturationally Constrained Self-Adaptive Goal Generation

A. Maturational Constraints: the Myelination Process

Maturational constraints play an important role in learning, by partially determining a developmental pathway. Numerous biological reasons are part of this process, like the brain maturation, the weakness of infants' muscles, or the development of the physiological sensory system. In the following, we focus on constraints induced by the brain maturation, especially, on the process called **myelination** [23]. Related to the evolution of a substance called myelin, and usually qualified by the term white matter, the main impact of myelination is to help the information transfer in the brain by increasing the speed at which impulses propagate along axons (connections between neurons). Here, we focus on the myelination process for several reasons, this phenomenon being responsible for numerous maturational constraints, effecting the motor development, but also the visual or auditive acuity, by making the number of degrees-of-freedom, and the resolution of sensori-motor channels increase progressively with time.

Actually, infants' brain does not come with an important quantity of white matter, myelination being predominantly a postnatal process, taking place in a large part during the first years of life. Konczak [24] and Berthier [25] studied mechanisms involved in reaching trials in human infants. In their researches, they expose that goal-directed reaching movements are ataxic in a first time, and become more precise with time and training. Also, they show that for all infants, the improving efficiency of control follows a proximo-distal way, which means that infants use in priority their torso and shoulder for reaching movements, and, progressively, their elbow [25]. This evolution of control capabilities comes from the increasing frequency of the muscular impulses, gradually, in shoulders, and elbows. This phenomenon, directly related to the myelination of the motor cortex, then allows muscles to become stronger at the same time, by training, which then increases their possibility to experiment wider sets of positions. Myelin is also responsible for brain responses to high visual and sound frequencies. Therefore, like introduced in [10], children are not able to detect details in images, which is also a reason, of imprecise reaching movements.

In the following, we consider constraints analogous to those induced by the myelination process, in the competence based intrinsic motivation framework introduced above, and present the Maturationally Constrained Self-Adaptive Goal Generation algorithm (McSAGG). To easily illustrate its functioning, we study the system in the precise case of a reaching task using a manipulator, where capabilities of the system are restrained and evolve depending on the learning evolution.

B. Formalization of Constraints

It is important to notice the multi-level aspect of maturational constraints: constraints existing on motor actions, influencing the control, and by analogy in our approach, the efficiency of the low-level active selection of actions performed to reach a goal; and constraints related to sensors, like the capacity to discriminate objects, and so here, to declare a goal as reached. The global idea is to control all of these constraints using an evolving term $\psi(t)$, called **maturational clock**, whose increase, which influence the lifting of constraints, depends on the global learning evolution, and is typically non-linear.

C. Stage Transition: Maturational Evolution and Intrinsic Motivations

Often considered as a process strictly happening in the first years of life, myelin continues to be produced even in adults while learning new complex activities [26]. Also, in a developmental robotics frame, we set the maturational clock $\psi(t)$, which controls the evolution of each release of constraint, as depending on the learning activity, and especially on the progress in learning by itself. Here, the main idea is to increase $\psi(t)$ (lifting constraints), when the system is in a phase of progression, considering its current learnt model. This progression is shown by an increase of the global average competence over time in S', and is analogous to the notion of a positive interest in the whole space S'. Therefore, considering competence values estimated for the ζ last reaching attempts $\{\gamma_{s'_{n-\zeta}}, ..., \gamma_{s'_n}\}_{S'}$, $\psi(t)$ evolves until reaching a threshold ψ_{max} such that:

$$\psi(t+1) = \begin{cases} \psi(t) + \lambda.interest(S') & \text{if } interest(S') > 0 \\ \psi(t) & \text{otherwise} \end{cases}$$

where $0 < \lambda << 1$. As the global interest of the whole space is typically non-stationary, the maturational clock becomes typically non-linear, and stops its progression when the global average of competence decreases, due to the lifting of previous constraints.

D. Constraints Implementation

In our model, we concentrate on three kinds of maturational constraints, directly inspired by consequences of the myelination process, and which are controlled by $\psi(t)$. These constraints are general and can be integrated in numerous kind of robots.

The first considered constraint describes the *limitation of frequency of muscular impulses* allowed for controlling limbs which is responsible of the precision and complexity of control [24]. Also corresponding to the frequency of feedback updating movements to achieve a trajectory, we define the constraint $f(t)$ as increasing with the evolution of the maturational clock:

$$f(t) \quad = \quad \left(-\frac{(p_{max} - p_{min})}{\psi_{max}}.\psi(t) + p_{max} \right)^{-1} \quad (5)$$

Where p_{max} and p_{min} represents maximal and minimal possible time periods between control impulses.

The second studied constraint relies on the sensor abilities. Here, we consider the *capacity to discriminate objects* as evolving over time, which here corresponds to an evolving value of ε_D, the tolerance factor allowing to decide of a goal

as reached. We thus set ε_D as evolving, and more precisely, decreasing over the maturational clock, from $\varepsilon_{D_{max}}$ to $\varepsilon_{D_{min}}$:

$$\varepsilon_D(t) \quad = \quad -\frac{(\varepsilon_{D_{max}} - \varepsilon_{D_{min}})}{\psi_{max}}.\psi(t) + \varepsilon_{D_{max}} \qquad (6)$$

Finally, we set another constraint, analogous to the proximo-distal law described above. Here, we consider the range r_i within which motor commands can be chosen for each joint i of a robotic system, as increasing over maturational time following a proximo-distal way over the structure of the studied embodied system. This typically allows larger movements to become available, and the potential access to the reaching of new goals:

$$r_i(t) \quad = \quad \psi(t).k_i \qquad (7)$$

Where k_i represents an intrinsic value determining the difference of evolution velocities between each joint. Here we fix: $k_1 \geq k_2 \geq ... \geq k_n$, where k_1 is the first proximal joint.

IV. Experiment and Results

A. Robotics Configuration for a Reaching Task

In the following, we consider a n-dimensions manipulator controlled in position and speed (as many today's robots), updated at discrete time values, called *time steps*. The vector $\theta \in \mathbb{R}^n = S$ represents joint angles, and $x \in \mathbb{R}^m = S'$, the position of the manipulator's end-effector in m dimensions, in the euclidian space S' (see Fig. 2 where $n = 3$ and $m = 2$). We introduce the McSAGG algorithm in the case of learning how to reach all reachable points, in the environment S', with this arm's end-effector. To do that, the robot has to learn its inverse kinematics, which answers to the question of what joint movement the robot can do to move in direction of a goal position s'_g fixed in S'. Also, in this precise experiment, where we suppose S' euclidian, and do not consider obstacles, the direction to a goal can always be described as following a straight line between the current end-effector's position and the goal.

Learning the inverse kinematics is an online process that arises each time a movement is executed by the manipulator: by doing movements, the robot stores measures $(\theta, \Delta\theta, \Delta x)$ in its memory; these measures are then reused online to compute the Jacobian $J(\theta) = \Delta x/\Delta\theta$ locally to move the end-effector in a desired direction $\Delta x_{desired}$.

1) Evaluation of Competence: To compute the measure of competence, the function D uses the Euclidian distance (see Fig. 2). Also computing local competence in subspaces typically requires the reaching of numerous goals. Because reaching a goal can necessitate several actions, and thus time, obtaining competence measures can be long. To improve this mechanism, we increase the number of goals artificially, using the fixation of subgoals, allowing the estimation of reaching competences, on the pathway to the generated goal. Considering a current state x'_c in S', and a self-generated goal x'_g, we define the set of l subgoals $\{x'_1, x'_2, ..., x'_l\}$ where $x'_i = (i/l) \times (x'_g - x'_c)$, that have to be reached before attempting to reach the terminal goal x'_g.

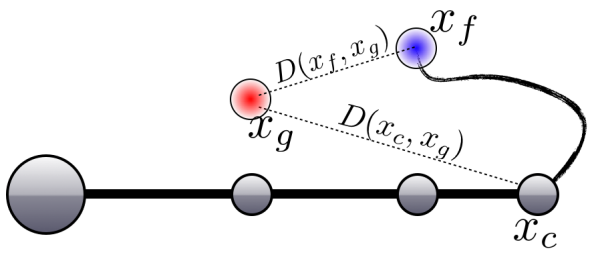


Fig. 2. Values used to compute the competence $\gamma_{s'_g}$, considering a manipulator of 3 degrees-of-freedom, in a 2 dimensions operational space.

B. Exploration and Reaching

In [21], Schaal & Atkeson propose a method called SSA to deal with learning sparse data in high dimensional spaces. Based on the observation that random exploration could be very long, unsafe, or costly, they introduced an exploration algorithm decomposing the problem of motor control, into two separated control tasks: a first one, where it trains a nonlinear regulator, by directing the controlled system to stay close to some chosen setpoints. And a second one where setpoints are shifted to reach a handcrafted goal. This typically allows the system to create a narrow tube of known data, to guide it toward the goal.

Here we propose a method, inspired by the SSA algorithm, to guide the system to learn on the pathway toward the selected goal position s'_g. The system is organized around two alternating phases: *reaching* phases, which involve a local controller to drive the system from s'_c towards the goal, and *local exploration* phases, which allows to learn the inverse model of the system in the close vicinity of the current state, and are triggered when the reliability of the local controller is too low. These mechanisms are stopped once the goal has been reached or a timeout exceeded. Let us here describe the precise functioning of those phases, in our experiment:

1) Reaching Phase: the reaching phase deals with creating a pathway to the goal position x_g. This phase consists of determining, from the current position x_c, an optimal movement to guide the end-effector toward x_g. For this purpose, the system computes the needed end-effector's displacement $\Delta x_{next} = v.\frac{x_c - x_g}{\|x_c - x_g\|}$ (where v is velocity bounded by v_{max}), and performs the action $\Delta\theta_{next} = J^+.\Delta x_{next}$, with J^+, pseudo-inverse of the Jacobian estimated in the close vicinity of θ and given the data collected by the robot so far. After each effected action Δx_{next}, we compute the error $\varepsilon = \|\widetilde{\Delta x}_{next} - \Delta x_{next}\|$, and trigger the exploration phase in cases of a too high value $\varepsilon > \varepsilon_{max} > 0$.

2) Exploration Phase: this phase consists in performing λ small random explorative actions $\Delta\theta_i$, around the current position θ. This allows the learning system to learn the relationship $(\theta, \Delta\theta) \Rightarrow \Delta x$, in the close vicinity of θ, which is needed to compute the inverse kinematics model around θ.

C. Qualitative Results

Let us set the constraint values in the case of a n=3 DOF arm, put in an 2 dimensions environment bounded in intervals $x_g \in [-1; 1] \times [0; 1]$. In this experiment, the considered control problem consists of learning relationships between a 6-dimensional space $(\theta, \Delta\theta)$ and resulting end-effector change Δx, in 2 dimensions such that $(\theta, \Delta\theta) \Rightarrow \Delta x$ (thus the

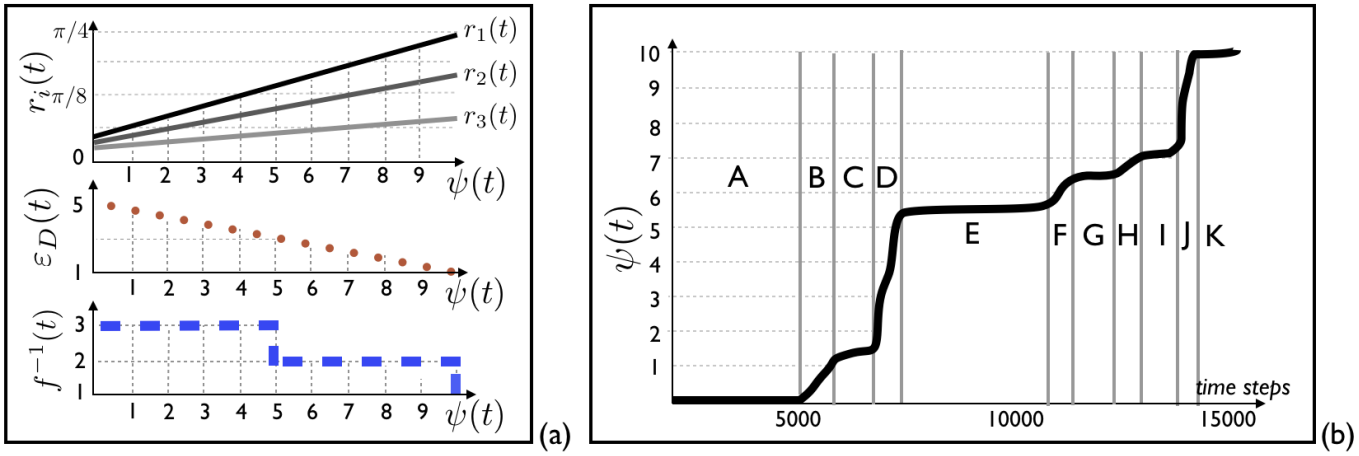


Fig. 3. (a) Exploration of maturational constraints over values taken by the maturational clock $\psi(t)$, for a manipulator of 3-dof. (b) evolution of the maturational clock over time, for a given experiment. Vertical splits are added manually, to let appear what we call *maturational stages*, which are described as periods between important changes of the evolution of $\psi(t)$ (change of the second derivative of $\psi(t)$).

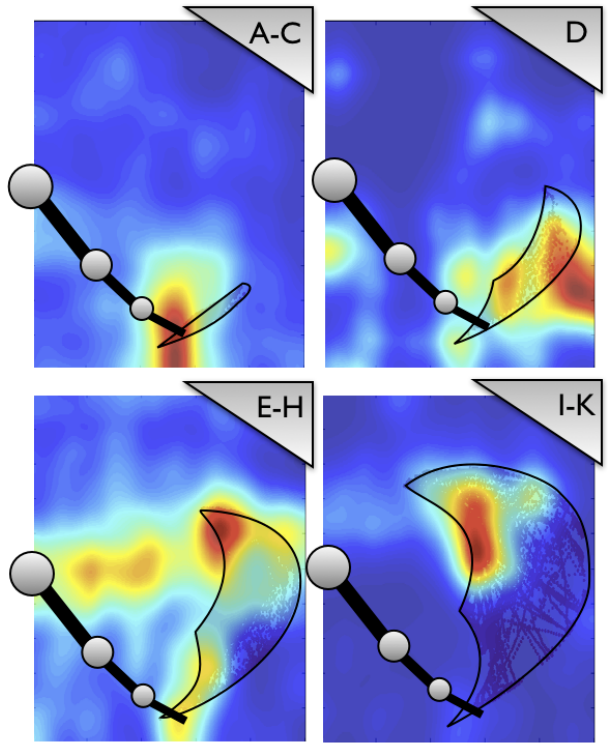


Fig. 4. Histograms of self-selected goals and illustration of reachable surfaces (thin black contours) over maturational stages.

problem space is here 8-dimensional). We set the arm with a global length of 50 units, and fix the proportion of each limb as 3/5, 2/5, and 1/5 of this length and fix $\psi_{max} = 10$.

Fig. 3 (a) shows the different constraints $r_i(t)$, ε_D and $f^{-1}(t)$ over values that take the maturational clock $\psi(t)$.

We can firstly observe increasing ranges $r_i(t)$, defined such that $r_3(t) < r_2(t) < r_1(t)$, which respects the proximo-distal constraint meaning that joints closer to the basis of the arm have a controllable range which increase faster than further joints. Fig. 3 (a) also shows the evolutions of $\varepsilon_D(t)$, from 5 to 1 units over $\psi(t)$, and $f^{-1}(t)$, representative of the time period between the manipulator's update control signals, from 3 to 1 time steps. The evolution of the frequency has been decided as being not continuous, to let us observe the behavior of the algorithm when a sudden change of complexity arises for a constraint. We run an experiment over 15000 time steps, which corresponds to the selection of about 7500 goals. During the exploration, we observe the evolution of the maturational clock $\psi(t)$ over time (black curve in Fig. 3 (b)) which evolves non-linearly, depending on the global progress of competence. Letters from A to K are added from an external point of view, they are described as periods between important changes of the evolution of $\psi(t)$ (evolution of the second derivative of $\psi(t)$) and represent what we call *maturational stages*. We describe two types of stages, *stationary stages* like A, C, E, G, I, K, where the maturational clock evolves slowly, which corresponds to time period (over time steps) where the global competence progress is either stable or negative, and *evolution stages*, like B, D, F, H, J, where the maturational clock is evolving with a high velocity.

We can emphasize two important maturational stages : the first one, A, which corresponds to a non-evolution of $\psi(t)$; this is due to the need of the mechanism to obtain a minimal number of competence measures, before computing the global progress to decide of a release of constraints. Also, the stable stage E, which appears after that $\psi(t)$ reaches the value 5 can be explained by the sudden change of frequency $f(t)$ from $1/3$ to $1/2$ update per time step, that is produced precisely at $\psi(t) = 5$. This is an effective example which clearly shows the adaptiveness of the McSAGG algorithm, which is able to slow down the evolution of the maturational clock in cases of an important change of complexity of the accessible body and world, according to constraints.

We also store all data-points x_i, visited by the end-effector during learning and create histograms of positions of self-generated goals over time windows. Fig. 4 is split in four subfigures, each one representing the behavior of McSAGG over a time window described over maturational stages. In each subfigure, we display all data-points explored by the end-effector in the considered time window, and create a contour (thin black lines) around the surface that is reachable, according to current limited ranges $r_i(t)$. We also superpose

histograms of goals selected over the time-window, red colors being representative of a surface where a high number of goals have been selected. The first global observation that we can deduce from these results is that by focusing in majority in areas which bring the maximum competence progress, the McSAGG algorithm progressively direct its goal self-selection in regions which are accessible by the end-effector: we can globally observe in each subfigure a higher quantity of selected goals inside and/or close to reachable regions (contoured areas). Therefore, using a measure of interest related to the competence progress allows the robot to learn to reach a high number of positions with its end-effector, instead of spending too much exploration time trying to explore non-accessible areas. A more precise study deduced from these results is the progressive focus of McSAGG on areas *which are newly accessible*: we can effectively observe histograms of goals selection progressively shifting on areas that were not accessible for previous maturational stages, for instance, comparing windows of chosen goals over the stage D (upper-right subfigure) and E-H (lower-left subfigure), we observe the change of position of the histogram, which, in E-H, is clearly focusing on areas which were not accessible before (inside the contoured area of the lower-left subfigure, but not the upper-right).

Eventually, we can argue in a qualitative point of view that the bidirectional coupling of maturational constraints and self-adaptive goal generation allows the self-focalization of goals inside maturationally restrained areas, which bring the maximal information needed for constraints to evolve, increasing progressively the complexity of the accessible world, and so on.

V. Conclusion

In this paper, we argued that intrinsic motivations and maturational constraints mechanisms may have complex bidirectional interactions allowing to actively control the growth of complexity in motor development. We proposed an integrated system of these two frameworks which allows a robot to developmentally learn its inverse kinematics progressively and efficiently, and presented qualitative results about the self-adaptive behavior of the algorithm, when considering constrains which evolve with different velocities. An important future direction will consist in a quantitative study of the learnt models, in simulations, but also using real robotic setups, as well as various sensorimotor embeddings to evaluate the scalability of the algorithm.

References

[1] R. White, "Motivation reconsidered: The concept of competence," *Psychol. Rev.*, vol. 66, pp. 297–333, 1959.

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

[3] E. L. Deci, H. Eghrari, B. Patrick, and D. Leone, "Facilitating internalization: The self-determination theory perspective," *Journal of Personality*, vol. 62, no. 1, pp. 119–142, 1994.

[4] A. Barto, S. Singh, and N. Chenatez, "Intrinsically motivated learning of hierarchical collections of skills," in *Proc. 3rd Int. Conf. Development Learn.*, San Diego, CA, 2004, pp. 112–119.

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

[6] P.-Y. Oudeyer and F. Kaplan, "How can we define intrinsic motivations ?" in *Proc. Of the 8th Conf. On Epigenetic Robotics.*, 2008.

[7] A. Baranes and P.-Y. Oudeyer, "Riac: Robust intrinsically motivated exploration and active learning," *IEEE Transation on Autonomous Mental Development*, vol. 1, no. 3, pp. 155–169, December 2009.

[8] V. Fedorov, *Theory of Optimal Experiment*. New York, NY: Academic Press, Inc., 1972.

[9] M. Schlesinger, "Heterochrony: It's (all) about time!" in *Proceedings of the Eighth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, L. U. C. Studies, Ed., Sweden, 2008, pp. 111–117.

[10] G. Turkewitz and P. Kenny, "The role of developmental limitations of sensory input on sensory/perceptual organization." *J Dev Behav. Pediatr.*, vol. 6, no. 5, pp. 302–6, 1985.

[11] D. Bjorklund, "The role of immaturity in human development," *Psychological Bulletin*, vol. 122, no. 2, pp. 153–169, September 1997.

[12] M. Schembri, M. Mirolli, and G. Baldassare, "Evolving internal reinforcers for an intrinsically motivated reinforcement learning robot," in *Proceedings of the 6the IEEE International Conference on Development and Learning (ICDL07)*, Y. Demeris, B. Scasselati, and D. Mareschal, Eds., 2007.

[13] X. Huang and J. Weng, "Novelty and reinforcement learning in the value system of developmental robots," in *Proc 2nd Int. Workshop Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, C. Prince, Y. Demiris, Y. Marom, H. Kozima, and C. Balkenius, Eds., vol. 94. Lund University Cognitive Studies, 2002, pp. 47–55.

[14] J. Schmidhuber, "Curious model-building control systems," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 2, 1991, pp. 1458–1463.

[15] J. Mugan and B. Kuipers, "Towards the application of reinforcement learning to underected developmental learning," in *Proceedings of the 8th International Conference on Epigenetic Robotics*, L. U. C. Studies, Ed., 2008.

[16] K. Merrick and M. L. Maher, "Motivated learning from interesting events: Adaptive, multitask learning agents for complex environments," *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, vol. 17, no. 1, pp. 7–27, 2009.

[17] L. Meeden, D. Blank, D. Kumar, and J. Marshall, "Bringing up robot: fundamental mechanisms for creating a self-motivating self-organizing architecture," *Cybernetics and Systems*, vol. 36, no. 2, 2005.

[18] M. Lungarella and L. Berthouze, "Adaptivity via alternate freeing and freezing of degrees of freedom," in *Proc. of the 9th Intl. Conf. on Neural Information Processing*, 2002.

[19] M. Lee, Q. Meng, and F. Chao, "Staged competence learning in developmental robotics," *Adaptive Behavior*, vol. 15, no. 3, pp. 241–255, 2007.

[20] R. Sutton, D. Precup, and S. Singh, "Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning," *Artificial Intelligence*, vol. 1123, no. 181-211, 1999.

[21] S. Schaal and C. G. Atkeson, "robot juggling: an implementation of memory-based learning," *Control systems magazine*, pp. 57–71, 1994.

[22] J. Peters and S. Schaal, "Natural actor critic," *Neurocomputing*, no. 7-9, pp. 1180–1190, 2008. [Online]. Available: http://www-clmc.usc.edu/publications//P/peters-NC2008.pdf

[23] J. Eyre, *Development and Plasticity of the Corticospinal System in Man*. Hindawi Publishing Corporation, 2003.

[24] J. Konczak, M. Borutta, and J. Dichgans, "The development of goal-directed reaching in infants. learning to produce task-adequate patterns of joint torque," *Experimental Brain Research*, 1997.

[25] N. E. Berthier, R. Clifton, D. McCall, and D. Robin, "Proximodistal structure of early reaching in human infants," *Exp Brain Res*, 1999.

[26] J. Scholz, M. Klein, T. Behrens, and H. Johansen-Berg, "Training induces changes in white-matter architecture," *Nature neuroscience*, vol. 12, no. 11, pp. 1367–1368, November 2009.