Reinforcement Learning Algorithms that Assimilate and Accommodate Skills with Multiple Tasks

Paolo Tommasino, Daniele Caligiore, Marco Mirolli, and Gianluca Baldassarre
Laboratory of Computational Embodied Neuroscience, Istituto di Scienze e
Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche (LOCEN-ISTC-CNR),
Via San Martino della Battaglia 44, I-00185 Roma, Italy,
{paolo.tommasino,daniele.caligiore,marco.mirolli,gianluca.baldassarre}@istc.cnr.it

***Abstract*—Children are capable of acquiring a large repertoire of motor skills and of efficiently adapting them to novel conditions. In previous work [1] we proposed a hierarchical modular reinforcement learning model (RANK) that can learn multiple motor skills in continuous action and state spaces. The model is based on a development of the mixture-of-expert model that has been suitably developed to work with reinforcement learning. In particular, the model uses a high-level gating network for assigning responsibilities for acting and for learning to a set of low-level expert networks. The model was also developed with the goal of exploiting the Piagetian mechanisms of assimilation and accommodation to support learning of multiple tasks. This paper proposes a new model (TERL - Transfer expert reinforcement learning) that substantially improves RANK. The key difference with respect to the previous model is the decoupling of the mechanisms that generate the responsibility signals of experts for learning and for control. This led made possible to satisfy different constraints for functioning and for learning. We test both the TERL and the RANK models with a two-DOFs dynamic arm engaged in solving multiple reaching tasks, and compare the two with a simple, flat reinforcing learning model. The results show that both models are capable of exploiting assimilation and accommodation processes in order to transfer knowledge between similar tasks, and at the same time to avoid catastrophic interference. Furthermore, the TERL model is shown to significantly outperform also the RANK model thanks to its faster and more stable specialization of experts.**

I. Introduction

One fascinating and still unexplained aspect regarding animals, and especially primates, is their capability to acquire a large repertoire of skills by autonomously interacting with the environment. In comparison, artificial agents and machine learning algorithms are often very effective when solving single tasks, but are affected by poor generalization capabilities and catastrophic interference when they face multiple tasks.

Caligiore et al. [1] have proposed a hierarchical modular reinforcement learning algorithm, here called "RANK", for learning multiple tasks. The model (derived from previous work, [2], [3]) developed the mixture-of-expert neural network model (ME) [4], designed for supervised learning problems, so to address reinforcement learning problems. As the ME, RANK used a high-level gating neural network to assign responsibilities to low-level expert networks that solved the tasks at hand. Although RANK was shown to be capable of learning different tasks, the results of its tests highlighted that further research was needed to better understand its functioning and to make its learning more robust [1].

This paper proposes a new model, called TERL (Transfer Expert Reinforcement Learning), which has one key difference and various minor improvements with respect to RANK. The key modification, a departure from the philosophy of ME, is based on the decoupling of the responsibility signals that establish the contribution of experts to the generation of actions with respect to the signals that establish the entity of their learning. As we shall see, this modification allowed us to make the processes of functioning and learning of TERL significantly more efficient.

The RANK model was also proposed as a tool to investigate the Piagetian concepts of assimilation and accommodation [5]. However, in previous work only preliminary evidence was shown on the fact that the model could actually capture these processes. Here we present evidence that both RANK and TERL can indeed exploit such processes to generalise when learning similar tasks and at the same time avoid the problem of catastrophic interference [6] when learning different tasks. In doing so, we will also show how the models allow us to provide an operational definition of assimilation and accommodation .

The rest of the paper is organised as follows. Sec. I-A reviews previous relevant, while Sec. I-B introduces some issues related to assimilation and accommodation. Sec. II presents the simulated robot and tasks used to test the models. Sec. III presents TERL and highlights its differences with RANK. Sec. IV shows the results of the tests both in terms of performance and in terms of the capacity to assimilate and accommodate. Finally, Sec. V draws the conclusions.

A. Related models

In the supervised learning literature the *mixture of experts* model (ME) has been proposed as a means to avoid catastrophic interference and enhance generalization [4]. The ME has a hierarchical and modular architecture formed by a number of *experts* modules, which compete to produce the answer of the system, and a *gating network*, which learns to assign responsibilities to experts. A key idea of ME is that the gating network uses a Bayesian accumulation of evidence on the capacity of experts to give a proper answer to the current input. This idea, first adapted to a RL context in [2], [3], is also at the core of both RANK and TERL.

In the last decade, Hierarchical RL systems (HRL) have been proposed as a preferential route to speed up the convergence of RL. These systems are used to either perform task-decomposition or, as here, to learn multiple tasks. However, the majority of these systems work with discrete states and action spaces and have not been used with continuous actions and states (e.g., [7]; see [8] for a review). Indeed, few RL models have been developed that are capable to cope with continuous actions and states or have been shown to work within robotic setups (see [9]–[11] for notable examples; see also [12] for a model that shares some features with TERL)

Although very interesting, these systems do not directly face the problem tackled here, that is the problem of deciding if storing different skills in the same or different experts. This type of problem has recently received attention within the RL community under the research agenda called *transfer reinforcement learning* (TRL). Within this context, the problem consists in identifying possible "source tasks", among those previously learned, on the basis of which to learn a new "target task" so as to maximise the transfer of knowledge and decrease the learning time and the steady-state performance. A recent important survey of TRL [13] highlights the fact that we still lack systems that can solve this problem in principled ways. TERL contributes to face this problem by proposing mechanism for resources allocation that is based, as in ME, on a Bayesian accumulation of evidence regarding which are the experts that are most suitable to solve a given task.

B. Assimilation and accommodation

Piaget held a *constructivist* approach according to which knowledge has *form* and *content*. Form is the innate organizational structure (schemas) that allows humans to process and categorise knowledge. Content is the representation of the world acquired with experience. According to Piaget [5], cognition develops on the basis of two complementary phenomena, assimilation and accommodation. Assimilation incorporates new environmental information in pre-existing schemas without modifying them. Accommodation, instead, modifies pre-existing schemas to fit new information. This idea has been operationalised with neural networks capable of self-changing not only the connection weights (content) but also their architecture (form) [14]. According to another interpretation of assimilation and accommodation [15], neural networks assimilate when they treat new inputs with their existing internal structure (generalisation) whereas they accommodate when this internal structure is updated to store new information (learning).

With respect to the model presented in this paper, the hard-wired and fixed architecture based on critic and actor experts can be considered as "innate form" whereas the knowledge it acquires through learning is the "content" (i.e. which skill for a given task). Hence, the learning processes taking place within the model presented here allows us to assign a novel meaning to assimilation and accommodation. *Assimilation* can be considered the process through which an expert trained for solving an already learned task is used, *as it is*, for solving a novel task that requires the *same sensorimotor mappings*. *Accommodation* occurs when the model recruits a copy of the expert developed for solving a given task and *suitably modifies it* for solving another task that requires *similar sensorimotor mappings*. The model also exhibits a third process, here called *generation*, used to face novel tasks that require *very different sensorimotor mappings*, and for which it is convenient to recruit non-trained novel experts.

II. The simulated robot and task

Fig. 1 shows the simulated dynamic planar arm and its work space with four different "objects" representing possible *goals* for reaching. Note that reaching objects A and B requires completely different sensorimotor mappings, A-C similar mappings, and B-D the same mapping: this is important for studying the assimilation/accommodation capabilities of the models.

The arm is formed by two links measuring respectively 25 cm (upper arm) and 35 cm (forearm). The arm has two actuated DOFs, one for the shoulder joint (θ_s) and one for the elbow joint (θ_e). The movement ranges were set to [-30°;+100°] for the shoulder and [0°;+160°] for the elbow. The equations describing the dynamics of the arm are as follows:

$$\begin{aligned} u_s &= (I_s + I_e + 2M_eL_sS_e\cos\theta_e + M_eL_s{}^2)\ddot{\theta_s} \\ &\quad +(I_e + M_eL_sS_e\cos\theta_e)\ddot{\theta_e} - M_eL_sS_e(2\dot{\theta_s} \\ &\quad +\dot{\theta_e})\dot{\theta_e}\sin\theta_e + B_s\dot{\theta_s} \\ u_e &= (I_e + M_eL_sS_e\cos\theta_e)\ddot{\theta_s} + I_e\ddot{\theta_e} + \\ &\quad M_eL_sS_e\dot{\theta_s}^2\sin\theta_e + B_e\dot{\theta_e}. \end{aligned} \tag{1}$$

where u is the actuated torque of a joint and the parameters M, L, S, I, and B are respectively the mass, the length, the distance from the centre of mass to joint, the rotational inertia of links, and the coefficient of viscosity (the parameters were set to $\{0.9, 0.25, 0.11, 0.065, 0.08\}$ for the shoulder joint and to $\{1.1, 0.35, 0.15, 0.1, 0.08\}$ for the elbow joint). The equations were integrated with a 4-th order Runge-Kutta method using a time step of 0.01 s.

A *proportional derivative* controller (PD) was used to supply the torque to each arm joint. A PD produces a torque proportional to the difference between the desired joint angle set by the model and the actual joint angle, and a damping proportional to the rate of change (time derivative) of the joint angle: $u = K_p(\theta - \theta_{des}) - K_d \cdot \dot{\theta}$. In this formula K_p and K_d are respectively the proportional gains and damping gains ($K_p = 25$ and $K_d = 4$ for both joints).

The environment is a working plane with four object goals having a radius equal to 3 cm. The object define four different reaching tasks. Each task requires that the arm learns to touch one of the four objects starting from the position showed in Figure 1 (simulations show that once the system has been trained it is capable of reaching it from any position). The system gets a reward of one when the hand touches an object, zero otherwise.

Note that the low complexity of the tasks and the set-up was very important for developing the algorithm and for

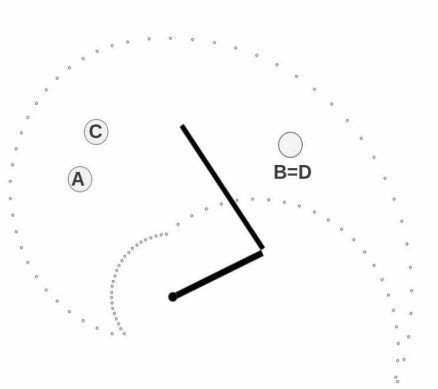

Fig. 1: The planar arm and the four objects A,B,C and D. Dots represent the borders of the work space due to the length of the arm link and the range of joints.

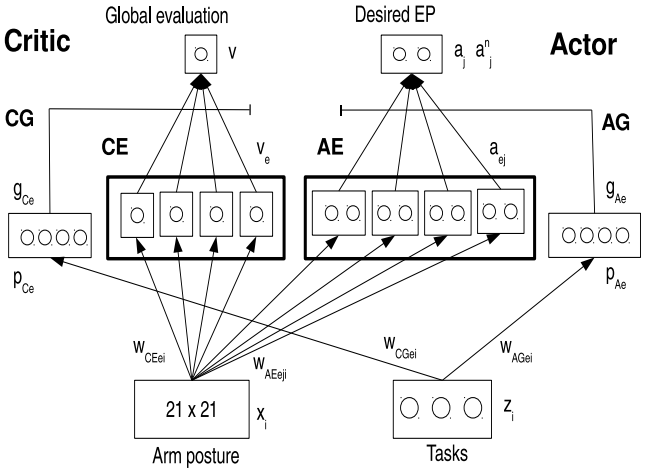

Fig. 2: The TERL Hierarchical Architecture.

understanding its functioning in depth. However, preliminary experiments not reported here indicate that the model can scale up to a robotic arm acting in 3D with a 4-DOF redundant arm.

III. Architectures and Algorithms

The TERL system (Fig. 2) is formed by two components: an *actor* that controls actions and a *critic* that evaluates states. Both these components have a hierarchical architecture formed by a *gating network* and a number of *experts*, as in ME [4]. We now explain the functioning of TERL, then its learning, and then its differences with RANK.

A. Functioning of TERL

1) Input: The system gets two types of inputs: (a) the gating networks get as input the current task, or *goal*, encoded with a different binary vector for different objects: A=[1,0,0,0], B=[0,1,0,0], C=[0,0,1,0] and D=[0,0,0,1]; (b) the experts get as input the arm postures ($\theta_s(t)$, $\theta_e(t)$) encoded in a neural map (with *population coding*, cf. [16]) formed by 21 x 21 normalised Gaussian radial basis function units x_i (as in in [9]).

The difference in the input between the gating and the experts networks reflects what is done in the TRL literature where the systems is typically informed about the task it is facing. The different task in most cases have to be accomplished in the same environment (as here), and the input (here the arm proprioception) sent to the part of the system that have to solve the task (here the experts) does not change (but there are other possibilities, see [13]). We cannot expand this issue here, but this arrangement seems also to reflect the organization of the striato-cortical loops in real brains, the core structures that underpin trial-and-error learning in organisms (e.g., see [17], [18]).

2) Actor gating network: The actor gating network (AG) has ten output units (indexed with e) which receive the task input z_i via connections with weights w_{AGei}. The activation potential, p_{Ae}, of output unit e is filtered with a soft-max function, and the resultant activation, g_{Ae}, represents the expert *responsibility* (Bayesian *prior*):

$$g_{Ae} = \frac{e^{(p_{Ae}/T)}}{\sum_{e=1}^{10} e^{(p_{Ae}/T)}} \tag{2}$$

the T temperature parameter, set to 0.1, allows to enhance slight differences between priors and therefore promotes a fast specialization of the experts.

3) Actor experts: Each actor expert (AE_e) has two output units with sigmoidal activation a_{ej} which encode the control signals to the arm (the two desired joint angles). These output units receive input from the arm-posture map units x_i via connections with weights w_{AEeji} and a bias weight (input constantly set to one). The global action a_j (desired arm angles) of the actor is computed on the basis of the *priors*:

$$a_j = \sum_e g_{Ae} \cdot a_{ej} \tag{3}$$

To foster exploration, the executed action, a^n_{jt}, includes noise, as explained in sec. III-D.

4) Critic gating network: The critic gating network (CG) works analogously to the AG on the basis of the connection weights, w_{CGei}, the unit activation potentials, p_{Ce}, and the *priors* of the critic experts g_{Ce}.

5) Critic experts: Each critic expert (CE) has a linear output unit v_e encoding the evaluation of the current state and receives input from the arm-posture map units x_i via connections with weights w_{CEei}. The global evaluation v of the critic is computed on the basis of the *priors*:

$$v = \sum_e g_{Ce} \cdot v_e \tag{4}$$

B. Learning signals

1) Global TD-error: Couples of successive global evaluations, together with the reward r_t, are used to compute the global TD-error, δ_t, as in standard reinforcement learning [19]:

$$\delta_t = \begin{cases} r_t - v_{t-1} & if\ end\ of\ trial \\ (r_t + \gamma v_t) - v_{t-1} & if\ during\ trial \\ 0 & if\ start\ of\ trial \end{cases} \quad (5)$$

where γ is a discount factor (γ = 0.99).

2) Critic Experts TD-error: The expert TD-error signals are calculated as follows:

$$\delta_{et} = \begin{cases} r_t - v_{et-1} & if\ end\ of\ trial \\ (r_t + \gamma v_{et}) - v_{et-1} & if\ during\ trial \\ 0 & if\ start\ of\ trial \end{cases} \quad (6)$$

3) Actor experts posterior responsibilities: To train the actor experts and gating network the algorithm computes the adjusted responsibilities (Bayesian *posteriors*, [4]) of the experts as follows:

$$h_{Ae} = \frac{c_{Ae} \cdot g_{Ae}}{\sum_e [c_{Ae} \cdot g_{Ae}]} \quad (7)$$

where c_{Ae} is a measure of the *likelihood* that the actor expert, e, chose the global action, $\mathbf{a}^n_{t-1}$:

$$c_{Ae} = e^{-0.5 \frac{\left(D\left[\mathbf{a}^n_{t-1}, \mathbf{a}_{et-1}\right]\right)^2}{\sigma^2}} \quad (8)$$

where $D\left[\mathbf{a}^n_{t-1}, \mathbf{a}_{et-1}\right]$ is the Euclidean distance between the two vectors encoding respectively the global action $\mathbf{a}^n_{t-1}$ and the action $\mathbf{a}_{et-1}$, computed by expert e. The width of the Gaussian (σ) is kept constant at 0.5.

4) Critic experts posterior responsibilities: The *posteriors* of the critic experts are computed as follows:

$$h_{Ce} = \frac{c_{Ce} \cdot g_{Ce}}{\sum_e [c_{Ce} \cdot g_{Ce}]} \quad (9)$$

where c_{Ce} is a measure of the *likelihood* that the critic expert, e, produced an accurate evaluation producing a zero TD-error.

$$c_{Ce} = e^{-0.5(\delta_{et})^2} \quad (10)$$

C. Learning

1) Actor gating network learning: The learning of the AG has been developed in analogy with ME. Intuitively, the learning rule tends to increase the responsibility of an expert if its likelihood (i.e., the similarity of its action with the executed action) is higher than average and if it has produced a positive surprise; otherwise it is decreased. Formally:

$$\Delta w_{AGei} = \eta_{AG} \cdot \delta_t \cdot (h_{Ae} - g_{Ae}) \cdot z_{it-1} \quad (11)$$

where η_{AG} is the learning rate (here set to 3.0).

2) Actor experts learning: Filtering the gating outputs with the soft-max favors the quick specialization of the experts. This means that the prior of the best expert will be close to one and those of other experts will be close to zero. In this case the Bayes rule returns a posterior close to one for the best expert and posteriors close to zero for the remaining experts. Therefore if posteriors are used to modulate the experts' learning rates, as in ME (and as in RANK), it is not possible to create multiple copies of the behavior of the best experts. To solve this issue TERL uses a different learning rule. The soft-max priors g_{Ae} are ranked and the ranks are used to calculate a *learning rate modulation parameter*, l_{Ae}:

$$l_{Ae} = b^{-k_e} / \sum_{e=1}^{N} b^{-k_e} \quad (12)$$

where $b = 6$, $k_e = [0, 1, 2, 3, ..., 10]$. The resulting l_{Ae} are [0.834,0.139, 0.023,0.004,0,0,0,0,0,0]. Note that here we use the same function as in RANK (cf Sec. III-E) to keep the two models comparable, but in the case of TERL ranks do not determine the priors for actions and therefore they do not need to sum up to one as in RANK. This means that the rank-based mechanism used for regulating learning is *decoupled* from the priors used to act: this gives much flexibility to TERL because allows the user to establish the number of copies the algorithm develops and the rate with which those copies are trained.

The TD(0) learning rule adapted to TERL is:

$$\begin{aligned} e_{AEejit} &= (a^n_{jt} - a_{ejt}) \cdot (a_{ejt} \cdot (1 - a_{ejt})) \cdot x_{it} \\ w_{AEejit} &= w_{AEejit-1} + \eta_{AE} \cdot l_{Ae} \cdot \delta_t \cdot e_{AEejit-1} \end{aligned} \quad (13)$$

where η_{AE} is a learning rate ($\eta_{AE} = 1.2$), and $(a_{ejt} \cdot (1 - a_{ejt}))$ is the derivative of the sigmoid function.

3) Critic gating network learning: Even this rule has been developed in analogy with ME: the responsibility of an expert is increased if the expert likelihood was higher (i.e., its reward prediction error was smaller) than average, and decreased otherwise (but differently from AG, δ_t is not needed as the likelihood is already informative of the expert's output quality). Formally:

$$\Delta w_{CGei} = \eta_{CG} \cdot (h_{Ce} - g_{Ce}) \cdot z_{it-1} \quad (14)$$

where η_{CG} is a learning rate ($\eta_{CG} = 1$).

4) Critic experts learning: As for the actor we rank the critic priors and obtain the coefficient l_{Ce} to modulate learning rates. The learning rule becomes:

$$w_{CEeit} = w_{CEeit-1} + \eta_{CE} \cdot l_{Ce} \cdot \delta_{et} \cdot x_{it} \quad (15)$$

where η_{CE} is the learning rate (here $\eta_{CE} = 0.01$). Note that here the expert TD error δ_{et} is used to update the critics experts instead of the global TD error δ_t.

D. Exploratory behavior

One important challenge in RL is the regulation of exploratory noise. Different solutions have been proposed for discrete action/state stationary environments(e.g. [20], [21]),

but solutions for continuous action/state environments are still preliminary (e.g., see [9]).

Here we use a noise regulation that exploits the fact that TRL involves episodic RL problems [13]. In particular, each trial is divided in two phases: a first exploitation phase, with low noise, and a second exploration phase, with high noise. The exploration phase starts when a close to optimal system is expected to be able to solve the task, i.e. after 1.5 sec.

Formally, an *exploratory module* produces stochastic actions obtained by filtering a uniform random noise:

$$a_j^{EM}t = (1 - \frac{1}{\tau}) \cdot a_{jt-1}^{EM} + \frac{1}{\tau} \cdot n_t \tag{16}$$

where $1/\tau = 0.01$ is the filter time constant and n_t is a random variable uniformly distributed in $[-20, +20]$. The result of the integration is cut in $[0; 1]$.

This stochastic action is then mixed via a coefficient c_t with the global action a_j to obtain the executed action a_{jt}^n:

$$a_{jt}^n = c_t \cdot a_j + (1 - c_t) \cdot a_j^{EM}t \tag{17}$$

The key point is that c_t is modulated during two phases of each trial so as to suitably regulate noise. In particular:

$$c_t = \begin{cases} c_0 & if\ t \leq t_e \\ \beta \cdot c_{t-1} & if\ t_e < t \leq t_T) \end{cases} \tag{18}$$

where t_T ($t_T = 10$ s) is the trial duration, t_e ($t_e = 1.5$ s) is the exploitation time during which $c_t = c_0$ ($c_0 = 0.99$), β ($\beta = 0.996$) is a decay coefficient progressively decreasing c during the exploration phase. The small noise during the exploitation phase ($c_0 = 0.99$) allows the system to slowly refine the policy even during this phase. Actions range in $[0; 1]$ and desired angles a_{jt}^n are mapped onto the joint ranges before being sent to the arm.

E. Functioning of the RANK system

The main differences of RANK with respect to TERL are: (a) Functioning: at each step, RANK ranks the activation potential (p_{Ae} and p_{Ce}) of the gating networks based on Eq. 12 and uses the ranks *for deciding the responsibilities of experts*; TERL, instead, uses the soft-max responsibilities to act; (b) Learning: RANK first computes the ranks and then transforms them into posteriors h with the Bayes rule (Eqs. 7 and 9) and uses these posteriors to modulate experts learning; TERL, instead, applies the ranking function to the priors g and uses the ranked priors for learning (therefore responsibility signals commanding actions are *decoupled* from the coefficients regulating experts learning). The mechanism (a) is not efficient as constrains RANK to use experts other than the best one to act: indeed, the ultimate reason for introducing ranking in RANK was to regulate learning and to obtain copies in background, but there are no good reasons for using it also for regulating expert responsibilities for functioning. The mechanism (b), directly derived from ME, has the problem that in RL the likelihood with which posteriors are computed are very unstable due to exploratory and environmental noise; moreover, once expert learning has been decoupled from functioning to be fully controllable, the motivation for using the Bayesian posterior to modulate it (as in ME) is no more theoretically founded.

F. The SINGLE model

The performances of TERL and RANK are compared with a third baseline RL model (SINGLE) formed by a single expert for both the critic and the actor and no gating networks.

IV. Results

Task A and B require very *different* sensorimotor mappings and so allow testing the capacity for generation (see Sec. I-B) of the models. Task C is close to A and so allows us to measure the *accommodation* capability of the models as in this case the models can transfer knowledge from A to C. Finally, task D is the same as task B, but the gating networks are informed that a different task is being solved, so to allow us to test the assimilation capability of the models.

Training was carried out with a simulation lasting 3000 trials in total and involving two phases. In the first phase, lasting 1000 trials, in each trial the task was switched between task A and B. In the second phase, lasting other 2000 trials, all four tasks were trained.

A. Learning Performance

Fig. 3 shows the average performance of TERL, RANK and SINGLE over the first and second phase of the simulation. For each trial, the figure reports the reaching time of the models (10s if the object was not touched) averaged over 10 replications of the simulations. A first result, which confirms what found in [1], is that SINGLE does not find a suitable solution to the problem as catastrophic interference and the limited amount of computational resources prevent it from learning even two tasks.

Regarding the comparison between TERL and RANK for tasks A-B, Fig. 3 shows that TERL is much faster then RANK. Introducing new tasks (from trial 1000 on) compromises performance only very briefly, indicating that the new tasks, being similar or equal to previous ones, are solved very rapidly. Furthermore, TERL has a better performance also when tasks C and D are introduced.

Fig 4 shows the performance of TERL and RANK for each single task aligned to the time when the task is introduced (A and B from the beginning, C and D from trial 1000). TERL learns task A and B approximately 10 times faster than RANK. Also for task C and D TERL largely outperform RANK. This higher performance is in part due to the fact that TERL can fully exploit the ability of the best expert once discovered, while RANK mixes the actions of the best expert with those of the experts with a non-zero rank-based responsibility.

Importantly, Fig 4 also shows that for task C, similar to the previously experienced task A, both models are capable of transferring knowledge, as they learn the new task much faster than task A itself. A similar result is achieved for task D, equal to the previously learned task B: also in this case

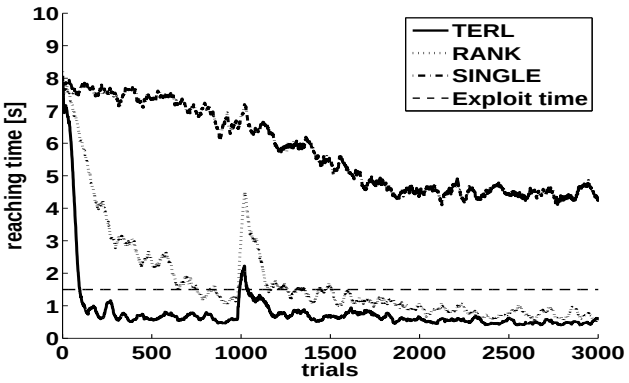

Fig. 3: Average performance (y-axis) of TERL, RANK, and SINGLE during the simulation (x-axis). Each curve represents an average over 10 replications of the simulation, and has been smoothed with a moving average of 30 trials. Performance is averaged over tasks A and B for the first 1000 trials and over tasks A, B, C, and D for the last 2000 trials.

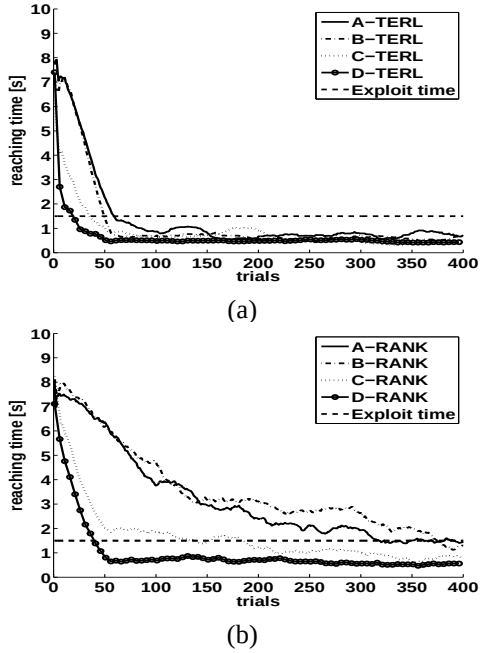

Fig. 4: (a) Learning curves of TERL in each of the four task aligned with the time of their introduction; (b) Same data for RANK.

both models learn very fast the new task as they realize they can exploit previously acquired experts. We now analyze in detail the processes underlying these results.

B. Assimilation, accommodation, and generation

To understand how TERL and RANK behave when learning different, similar, and same tasks, we investigated the dynamics of the value of the responsibility priors of the actor and critic gating networks during the simulation. This values establish the responsibility of experts in action and contribute to the entity of their learning (filtered by the ranks in TERL, and multiplied by the likelyhood in RANK). Thus the priors give a good indication of: (a) which expert has the

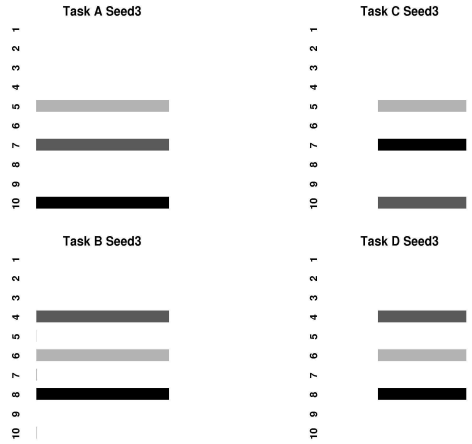

Fig. 5: Use of actor experts by TERL during one simulation. The four graphs refer to tasks A, C, B and D. Each graph reports the priors of the 10 experts during trials. For each trial of the simulation the highest, second highest, and third highest priors are respectively marked with black, dark gray, and light gray, while all other priors are not marked (white). Recall that learning on tasks C and D begins after the 1000th trial, so the priors for them are not shown before such trial.

main responsibility for the selection of actions and which are the other experts that contribute to it; (b) which experts are learning a task "in background" (i.e., with a smaller intensity) with respect to the main expert, and so become "copies" of the skill available for future exploitation; (c) which expert is used when a new task is introduced (e.g., the expert most used for a previously learned task, a "copy" of it, or a completely new expert).

Fig. 5 shows the evolution of the prior responsibilities of actor experts of TERL recorded at the end of each trial of a representative simulation. Importantly, Fig. 5 shows that when task C (similar to A) is learned, a copy formed during the learning of A is recruited as expert with highest prior (expert 7). Using the definitions proposed in Sec. I-B, this represents a case of *accommodation*: a (copy of a) skill previously used to accomplish task A is now recruited for the similar task C and suitably and efficiently modified. Notably, catastrophic forgetting is avoided thanks to the fact that for solving task C the systems does not use the best expert used for task A, but another expert that has learned the same skill.

Fig. 5 also shows that when task D (which is identical to task B) is learned, the best expert used for learning B is recruited as the expert with the highest prior (expert 8). According to the definitions proposed in Sec. I-B, this represents a case of *assimilation*: the skill developed for B is now recruited for the identical task D without any modification.

Finally, Fig. 5 also shows that the experts with the three highest priors for tasks A and C, on one side, and those for tasks B and D, on the other side, differ: the system has "understood" that the tasks are different and so has recruited different experts (a case of "generation", see Sec. I-B).

Fig. 6 shows that similar results are obtained for RANK,

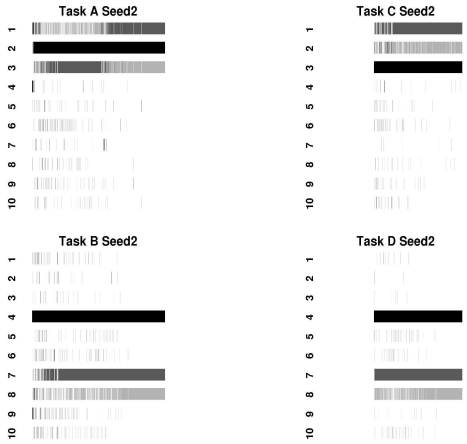

Fig. 6: Same data as in Fig. 5 but for the RANK model.

but for a crucial difference: RANK takes a lot of time before specializing the second and the third best experts, and initially oscillates between several different experts. This unstable selection of experts produces unstable learning signals and hence slows down convergence. This instability is due the fact that in RANK ranks are not directly used to train the experts but are filtered with the likelihood, which is rather unstable. Instead, TERL uses ranked priors for regulating learning, and priors are rather stable as they are based on the Bayesian accumulation of evidence collected by the gating networks.

Table I summarizes the behavior of all the 10 repetitions of the simulations with TERL and RANK. Overall, both models present four type of behaviors: (a) Assimilation: the expert with the highest prior is used for solving another task; (b) Assimilation with copies: a new task is solved on the basis of a copy of the skill developed for an identical task rather than with its best expert; (c) Accommodation: a new task is solved on the basis of a copy of a similar task, suitably modified; (d) Generation: a new task is solved with a completely new expert. The table shows that TERL and RANK have similar behaviors in terms of these different processes. Moreover, it also shows that in some cases task D, identical to B, is solved through a copy, rather than through the best expert, of B: as the copies have become as good as the best expert, either can be used for the new task.

Both TERL and RANK sometimes (once and twice, respectively) sub-optimally use the same expert for the similar tasks A and C. Furthermore, for the critic experts both models tend to use assimilation not only for tasks B and D (identical) but also for tasks A and C (similar). The reason might be that the evaluation gradient of A and C is very similar (roughly, a hill centered on the target) and that the copies for the two tasks are similar but not equal (data not shown), so maybe the systems use slightly different mixtures to produce slightly different evaluations. Further investigations are needed to explain this behavior.

TABLE I: Assimilation (Assi), assimilation with a copy (Assi$_c$), accommodation (Acco), and generation (Gene) for the actor (Act) and Critic (Cri) for the four tasks and 10 seeds.

MODELS		*AB* Act	dif. Cri	*AC* Act	sim. Cri	*BD* Act	same Cri
TERL	Assi$_c$	0	0	0	0	6	10
	Assi	0	0	1	10	4	0
	Acco	0	0	9	0	0	0
	Gene	10	10	0	0	0	0
RANK	Assi$_c$	0	0	0	0	6	10
	Assi	0	0	2	10	4	0
	Acco	0	0	8	0	0	0
	Gene	10	10	0	0	0	0

V. Conclusions

This article has presented TERL, a model capable of learning multiple tasks while exploiting their similarities and avoiding catastrophic interference. The model represents a substantial improvement of a previous similar model that adapted the key ideas of the mixture of expert network, developed for supervised learning, for working with reinforcement learning problems that have continuous states and actions spaces. The key innovation of the new model is the decoupling between the responsibility signals used to exploit and to train the experts. This decoupling allows the model to refine the Bayesian mechanism through which TERL collects evidence on which expert is best suited to face the current task and to form background copies of skills that can be exploited for new tasks.

The model has been shown to be able to nicely adapt to the requests of new encountered tasks. In particular, the model is able to: (a) decide that a novel non-trained expert has to be used if the new task is substantially different from previously learned ones, thus preventing catastrophic interference; (b) exploit a copy of the skill already developed for a task if the new task is sufficiently similar to the previous one, so that knowledge can be transferred between tasks; (c) exploit the same skill used for a previously acquired task also for the new task if the sensorimotor mapping required for the former are the same as those required for the latter.

These processes also lead to an operational definition of the concepts of accommodation and assimilation introduced by Piaget. In particular, the model implements assimilation when it uses experts previously used to solve very similar/same tasks, and implements accommodation when it modifies copies of experts previously used to solve similar tasks so to adapt to the new conditions.

These results show that the principles behind TERL have a high potential to allow the construction of autonomous robots capable of learning multiple skills while exploiting their similarities and avoiding catastrophic interference. At the same time, they can be suitably used to investigate the processes underlying development, for example assimilation and accommodation processes. Although there is not space to expand this issue here, we also think that the mechanisms underlying the functioning of TERL are also suitable to

investigate various aspects of brain organization and plasticity, in particular those related to the hierarchical organization of behavior in cortico-basal ganglia loops [17], [18].

ACKNOWLEDGEMENTS

This was funded by the European Commission 7th Framework Programme (FP7/2007-2013), *Challenge 2 - Cognitive Systems, Interaction, Robotics*, Grant Agreement No. ICT-IP-231722, Project *IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots*.

REFERENCES

[1] D. Caligiore, M. Mirolli, D. Parisi, and G. Baldassarre, "A bioinspired hierarchical reinforcement learning architecture for modeling learning of multiple skills with continuous states and actions," in *Proceedings of the Tenth International Conference on Epigenetic Robotics (EpiRob2010)*, Lund University, Sweden, 2010.

[2] G. Baldassarre, "A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours," *Journal of Cognitive Systems Research*, vol. 3, pp. 5–13, 2002.

[3] ——, "Planning with neural networks and reinforcement learning," PhD Thesis, Computer Science Department, University of Essex, Colchester, UK, 2002.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, pp. 79–87, 1991.

[5] J. Piaget, *The Origins of Intelligence in Children*. London: Routledge and Kegan Paul, 1953.

[6] M. McCloskey and N. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *The psychology of learning and motivation*, G. H. Bower, Ed. San Diego, CA: Academic Press, 1989, vol. 24, pp. 109–165.

[7] S. P. Singh, "Transfer of learning by composing solutions of elemental sequential tasks," *Machine Learning*, vol. 8, pp. 323–339, 1992.

[8] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dynamic Systems*, vol. 13, no. 4, pp. 341–379, 2003.

[9] K. Doya, "Reinforcement learning in continuous time and space," *Neural Comput*, vol. 12, no. 1, pp. 219–245, Jan 2000.

[10] J. Peters and S. Schaal, "Natural actor-critic," *Neurocomputing*, vol. 71, pp. 1180–1190, 2008.

[11] J. Mugan and B. Kuipers, "Autonomous exploration and the qualitative learner of action and perception, qlap," *IEEE Transactions on Autonomous Mental Development*, inpr.

[12] M. Ring, T. Schaul, and J. Schmidhuber, "The two-dimensional organization of behavior," in *IEEE International Conference on Development and Learning (ICDL-2011)*, vol. 2. IEEE, 2011, pp. 1–8.

[13] M. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *The Journal of Machine Learning Research*, vol. 10, pp. 1633–1685, 2009.

[14] D. Mareschal and T. R. Shultz, "Generative connectionist networks and constructivist cognitive development," *Cognitive Development*, vol. 11, no. 4, pp. 571–603, 1996. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0885201496900180

[15] D. Parisi and M. Schlesinger, "Artificial life and piaget," *Cognitive Development*, vol. 17, no. 3-4, pp. 1301–1321, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12691760

[16] A. Pouget and P. E. Latham, "Population codes," in *The Handbook of Brain Theory and Neural Networks*, 2nd ed., M. A. Arbib, Ed. Cambridge, MA, USA: The MIT Press, 2003.

[17] J. C. Houk, J. Davis, and D. Beiser, Eds., *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: The MIT Press, 1995.

[18] M. M. Botvinick, Y. Niv, and A. Barto, "Hierarchically organized behavior and its neural foundations: A reinforcement-learning perspective," *Cognition*, 2008.

[19] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge MA, USA: The MIT Press, 1998.

[20] J. C. Gittins and D. M. Jones, "A dynamic allocation index for the discounted multiarmed bandit problem," *Biometrika*, vol. 66, no. 3, pp. 561–565, 1979. [Online]. Available: http://biomet.oxfordjournals.org/content/66/3/561.short

[21] S. B. Thrun, "Efficient exploration in reinforcement learning," Tech. Rep., 1992.