Simultaneous Acquisition of Task and Feedback Models

Manuel Lopes, Thomas Cederbourg and Pierre-Yves Oudeyer
INRIA, Bordeaux, France

***Abstract*— We present a system to learn task representations from ambiguous feedback. We consider an inverse reinforcement learner that receives feedback from a user with an unknown and noisy protocol. The system needs to estimate simultaneously what the task is, and how the user is providing the feedback. We further explore the problem of ambiguous protocols by considering that the words used by the teacher have an unknown relation with the action and meaning expected by the robot. This allows the system to start with a set of known symbols and learn the meaning of new ones. We present computational results that show that it is possible to learn the task under a noisy and ambiguous feedback. Using an active learning approach, the system is able to reduce the length of the training period.**

I. Introduction

Learning from demonstration has provided several examples of efficient learning in robotic systems. A feature of most of those systems is that the data is provided in a batch perspective where data acquisition is done before the learning phase. Recently it has been suggested that *interactive learning* [1] might be a new perspective of robot learning that combines the ideas of learning by demonstration, learning by exploration and tutor guidance. Under this approach the user interacts with the robot and provides extra feedback. Approaches have considered extra reinforcement signals [2], action requests' [3], [4], disambiguation among actions [5] or preferences among states [6]. In [7] the authors compare the results when the robot has the option to ask or not the user for feedback.

Several studies discuss the different behaviors that naive users use when instructing robots [2], [8]. An important aspect is that, many times, the feedback is ambiguous and deviates from the mathematical interpretation of a reward or a sample from a policy. For instance, in the work of [2] the users frequently gave a reward to exploratory actions even if the signal was used as a reward of a performed state-action and not just for getting closer to the goal. Also, in some problems we can define an optimal teaching sequence but humans do not behave according to those strategies [8].

In this work we consider a setting where the robot must learn a task description from interacting with a user that provides feedback signals such as the name of the correct action to be used or by explicitly saying if an action is correct or wrong. We extend previous approaches by learning simultaneously how the feedback is being provided and what is the meaning of the user's feedback utterances. Note that we will call what the user says/writes `feedback utterances` and the meaning of the feedback `feedback meaning`. In a human-robot interaction setting we consider the case where the robot tries an action and then receives a feedback signal from the teacher. Such feedback is not restricted to a pre-defined protocol, with a pre-defined set of symbols or words, but should allow for new interaction types and instruction commands. For instance, a user might tell a robot if an action was right or wrong while another user might instruct the robot by saying the name of the correct action. The users will also utter different words not expected by the robot. A simple case is when the user gives synonyms of feedback utterances.

The authors are with the Flowers Team at INRIA, France. Contact email: `manuel.lopes@inria.fr`. Work (partially) supported by INRIA, Conseil Régional d'Aquitaine and the ERC grant EXPLORERS 24007.

II. Inverse Reinforcement Learning with Ambiguous Feedback

We consider a standard *markov-decision process* (MDP) and follow the notation of [9] . In our case we are not interested in learning a task by self-exploration but will use data from a user. In this situation we do not have a reward function from which we can sample but will have instead samples from the policy. This formalism is called the *inverse reinforcement learning* (IRL) problem [10]. The goal is to find the reward function that the demonstrator is trying to maximize and later on use it to select the best actions. Using a Bayesian perspective, we follow the *Bayesian IRL* approach (BIRL)[11]. In that setting we consider that, if the demonstrator is performing the task described by the reward function r, the samples of the demonstration are generated by $p(x,a|r) = \frac{e^{\eta Q(x,a)}}{\sum_b e^{\eta Q(x,b)}}$, where η is a confidence parameter where high values will correspond to the optimal policy and lower values will allow samples of non-optimal actions. We assume a uniform state sampling. To learn the task we compute the posterior distribution of the reward function after observing a given data vector $D_t = \{A_{0:t}, X_{0:t}\}$:

$$p(R_{t+1}|A_t, X_t) \quad \propto \quad p(A_t|R_t, X_t)p(R_t) \tag{1}$$

for a suitable choice of prior distribution on R, see [11].

A. Feedback Model

We will change the standard setting and, for a given state action pair (x,a), consider the probability of receiving a given feedback meaning f. Table I shows all the possible feedback protocols that can range from a pure learning from demonstration behavior (protocol 1) to a pure binary

reinforcement one (protocol 8). Each protocol is defined with the feedback that the teacher provides the learner when it does the correct action and when it does the wrong action. The teacher might choose to say the correct action (A), say nothing ($\emptyset$), give a confirmation (O) or inform the robot that the selected action is wrong (W). This protocol is ambiguous and the same feedback ($\emptyset$) can either mean correct or incorrect. If more than one correct action is available in a state then the teacher provides, randomly, one of them. To model perceptual errors there is a probability of "listening" the wrong feedback and "hearing" a random symbol instead.

TABLE I

FEEDBACK PROFILES. POSSIBLE FEEDBACK INSTRUCTIONS GIVEN BY THE USER WHEN THE ROBOT DOES THE **CORRECT** OR **WRONG** ACTION ARE: THE ACTION NAME (A), NOTHING ($\emptyset$), CORRECT (O) OR WRONG (W). EIGHT FEEDBACK PROFILES WERE CONSIDERED.

Action \ Feedback	1	2	3	4	5	6	7	8
Correct	A	A	A	$\emptyset$	$\emptyset$	O	O	O
Wrong	A	$\emptyset$	W	A	W	A	$\emptyset$	W

Each different teacher will be modeled as a convex combination of these profiles. For the teacher model we will consider a set of parameters M that describe the mixture of profiles in Table I.

We have to extend the model in Eq. 1 to include the ambiguous feedback. The goal is now to learn simultaneously the task R and the feedback model M, based on the pairs of executed actions A and the feedback meaning received F. Our posterior now depends not only on the demonstration but also on the feedback model. By independence, and removing the state dependency to simplify notation, we get the following factored model:

$$\begin{aligned} & p(R_{t+1}, M_{t+1} | A_{0:t}, F_{0:t}) \\ & \quad \propto \quad p(F_t | A_t, R_t, M_t) p(A_t | R_t) p(R_t, M_t) \end{aligned} \tag{2}$$

B. Utterance-Meaning Correspondences

Another aspect of human-robot interaction systems is that the feedback is often given using a natural interface such as gestures or speech. Most of the times there is an implicit assumption that the vocal symbols are assumed to have a known semantics for the robot. Now, we will relax this assumption and allow the user to provide instructions to the robot that are unknown. We will define the feedback meaning as the instruction the user wants to provide to the robot, as defined in Table I, and the feedback utterances as the words actually provided by the user. In this way it is possible for the robot to accept new words and learn their meanings. As an example, the user might say "good", or "ok", or "correct" and the robot should always understand it as a confirmation, i.e. the different utterances all correspond to the same feedback, as in Figure 1.

We have to extend the previous model, in Equation 2, to include the uncertainty in the symbols received. We will consider a new relation that gives the probability of having an utterance g when the user wants to provide a given feedback f, $p(g|f,.)$. As the feedback is no longer observed, we have to integrate it out from the observation of the utterance. Finally, we get the following expression:

	Utterances	Feedback Meanings
Known	up	$\uparrow$
	down	$\downarrow$
	left	$\leftarrow$
	right	$\rightarrow$
	$\emptyset$	CORRECT/WRONG
	ok	CORRECT
	error	WRONG
Unknown	good	?
	b	?
	$\vdots$	?

Fig. 1. Relation between uttered feedback and its intended meaning. There are only $Na + 3$ feedback meanings, one corresponding to each available action and the meanings of CORRECT and WRONG. They are fixed and known from the beginning. We assume that there is at least one utterance with a known correspondence to a feedback signal, there is the possibility of unknown feedback symbols to exist and their relation to the feedback must be learned. For instance the user might say good instead of ok.

$$p(G_{t+1}|D_t) = \sum_g p(G_t|F_t) p(g|D_t) \tag{3}$$

This posterior distribution on the utterance-meaning model can also be implemented as a particle filter.

C. Algorithm

The algorithm involves the estimation of three entities from data: the reward, the feedback model and the meanings of the feedback symbols. We will use a particle filter to estimate all the variables of interest. To reduce the number of particles we will not represent the full joint distribution but only an approximate of each marginal. We update the weight of each particle taking into account the maximum likelihood estimate of the other variables. Table II, summarizes the algorithm.

We can follow the active learning extension for IRL as presented in [4] to allow the learner to request the most informative samples. In that approach the policy distribution is inferred from the distribution on the rewards. Then, for each state, a measure of the uncertainty is made to select the state where the policy posterior has higher variance.

III. RESULTS

We now consider an environment where the robot can navigate and where there is a probability of finding three different objects. The robot has to learn which objects it should collect, or not, and for each of the object classes learn where they must be delivered. The number of actions is 7, the 5 navigation ones plus collect and release. The number of feedback symbols is 10, again we assume that we have an initial known set of symbols and the user will provide 10 new synonyms. The robot executes an action and then it receives the feedback.

TABLE II

ALGORITHM FOR THE JOINT ESTIMATION OF THE TASK, FEEDBACK AND UTTERANCE-MEANING MODELS. IT COMBINES THREE PARTICLE FILTERS TO APPROXIMATE THE POSTERIOR DISTRIBUTION OF THE THREE VARIABLES.

- Select number of samples n_r, n_g and n_m
- Sample n_r reward vectors
- Sample n_g utterance-meaning parameters
- Sample n_m meanings tables

1) Sample state x
2) Choose and execute action a
3) Observe utterance g
4) Sample feedback from f_t $p(f|g_t)$
5) Find best feedback parameters $M = argmax_i w_f^{(i)}$
6) $w_r^{(i)} \leftarrow p(f_t|A_t, R_t^i, M)p(A_t|R_t)w_r^{(i)}$
7) Resample reward particles
8) Find best reward parameters $r^* = argmax_i w_r^{(i)}$
9) $w_f^{(i)} \leftarrow p(f_t|A_t, r^*, M_t)p(A_t|r^*)w_f^{(i)}$
10) Resample feedback model
11) $w_g^{(i)} \leftarrow \sum_i p(g_t|f_t)w_g^{(i)}$
12) Resample utterance-meaning model
13) goto 1

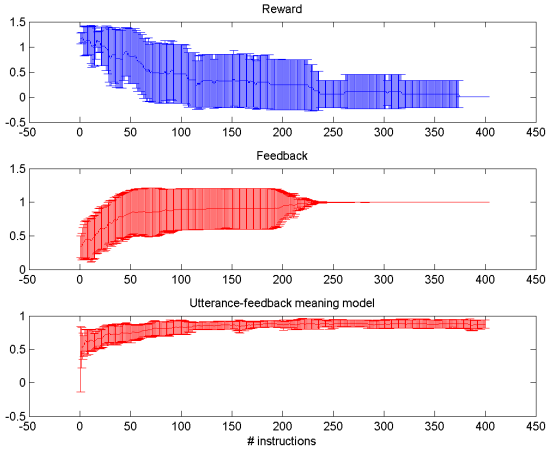

Fig. 2. Mean and variance for the active learning method in the "Object Collecting" Task. The system is able to learn the task, the feedback system and new feedback symbols. Top - policy loss; Middle - likelihood of correct feedback model; Bottom - number of correctly assigned symbols.

Figures 2 and 3 give the results for a problem with three objects and 64 possible locations. In each execution of the problem the system randomly selects the objects that should be collected and their delivery locations. Results show that the system can learn the task, the feedback model and (part of) the novel feedback symbols.

IV. CONCLUSIONS

Computational approaches in learning by demonstration have evolved a lot in recent years. These methods can now be applied in realistic human-robot interaction settings to effectively provide an intuitive way for untrained users to program robots. Under this setting most algorithms have to be adapted to the noise and ambiguity usually present in human dialog. In this work we showed how a robot can learn

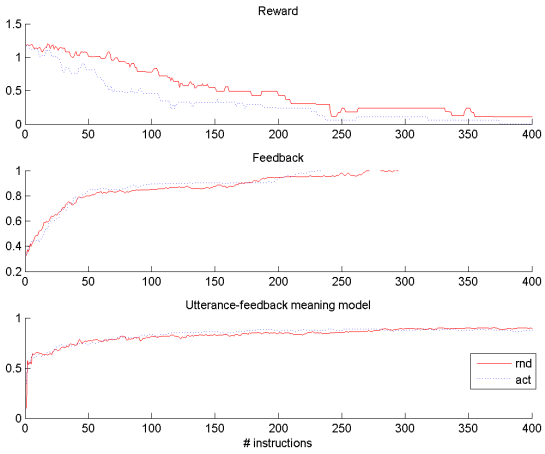

Fig. 3. Comparison between active and randomly sampling in the "Object Collecting" Task. The system is able to learn the task, the feedback system and new feeback symbols. Top - policy loss; Middle - likelihood of correct feedback model; Bottom - number of correctly assigned symbols.

a task description when the feedback it gets from the user does not follow a *rigid protocol* and is *very noisy* (10% error in correctly recognizing the feedback symbols). We showed that a learning system can simultaneously estimate the feedback protocol and the task representation in a reasonable amount of time and computational complexity. We took a further challenge and only assumed partial knowledge of the guidance symbols. By bootstrapping the systems with some known guidance-feedback correspondences, the system could successfully estimate the correspondences of new guidance symbols.

REFERENCES

[1] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A. L. Thomaz, and D. Mulanda., "Tutelage and collaboration for humanoid robots," *International Journal of Humanoid Robotics*, vol. 1, no. 2, 2004.

[2] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence Journal*, vol. 172, pp. 716–737, 2008.

[3] D. Grollman and O. Jenkins, "Dogged learning for robots," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 2483–2488.

[4] M. Lopes, F. S. Melo, and L. Montesano, "Active learning for reward estimation in inverse reinforcement learning," in *European Conference on Machine Learning (ECML/PKDD)*, Bled, Slovenia, 2009.

[5] S. Chernova and M. Veloso, "Interactive policy learning through confidence-based autonomy," *J. Artificial Intelligence Research*, vol. 34, pp. 1–25, 2009.

[6] M. Mason and M. Lopes, "Robot self-initiative and personalization by learning through repeated interactions," in *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*, 2011.

[7] M. Cakmak, C. Chao, and A. Thomaz, "Designing interactions for robot active learners," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 108–118, 2010.

[8] M. Cakmak and A. Thomaz, "Optimality of human teachers for robot learners," in *Proceedings of the International Conference on Development and Learning (ICDL)*, 2010.

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

[10] A. Y. Ng and S. J. Russel, "Algorithms for inverse reinforcement learning," in *Proc. 17th Int. Conf. Machine Learning*, USA, 2000.

[11] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *20th Int. Joint Conf. Artificial Intelligence*, India, 2007.