2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Reinforcement Learning of Variable Admittance Control for Human-Robot Co-manipulation

Fotios Dimeas and Nikos Aspragathos

Abstract— **In this paper, a variable admittance controller based on reinforcement learning is proposed for human-robot co-manipulation tasks. Setting as the goal of the reinforcement learning algorithm the minimisation of the jerk throughout a point-to-point movement, the proposed controller can learn the appropriate damping for effective cooperation without any prior knowledge of the target position or other task characteristics. The performance of the proposed variable admittance controller is investigated on a co-manipulation task with a number of subjects using a KUKA LWR robot, demonstrating considerable reduction both in the effort required by the operator and in the completion time of the task.**

I. Introduction

Physical human-robot cooperation is a rapidly emerging field that aims to combine the complementary skills between a robotic manipulator and a human in tasks such as effortless carrying of long or heavy objects, precise co-manipulation for assembly or robotic surgery, programming by demonstration etc. Admittance control is a widely used technique that allows the compliant behaviour of a robot by imposing a relationship of a mass-spring-damper system between the external forces and the motion of the manipulator [1]. The selection of the admittance parameters, namely the virtual inertia, spring constant, and damping, determines the form and the effectiveness of the interaction.

The most dominant parameter, that greatly affects the physical cooperation, is the virtual damping [2]. Although low damping values facilitate low-effort co-manipulation, the accuracy in fine positioning tasks is reduced because the robot becomes over-responsive. Variable admittance control can overcome this problem by regulating appropriately the admittance parameters during the cooperation, e.g. apply low damping for accelerating into fast motion and higher damping in fine positioning. The main objective in variable admittance is to find the most appropriate way to regulate the damping. A number of techniques have been proposed that regulate the admittance parameters based on the velocity of the robot [3], the rate of change of the force applied by the operator [4] or a combination of them [5]. Although the results demonstrate an improved performance over constant admittance, the algorithms that tune the parameters are heavily based on the designer's intuition and do not guarantee any optimality.

Systematic approaches have also been proposed to tune the admittance parameters that consider human arm impedance characteristics [6], or the minimum jerk trajectory model [7]. However, most of these methods require an *a priori* knowledge of the movement to be conducted for the training. Within this scope, a Fuzzy Model Reference Learning Controller (FMRLC) was presented in [8], that combines human domain knowledge and a supervised learning algorithm in order to regulate the virtual damping during a predefined movement towards the minimum jerk trajectory. This method utilised measurements of the applied force and the velocity of the robot to create a mapping among the provided input-output data pairs, assuming that there is no dependency among them. This approach turns out to be quite restrictive in human-robot co-manipulation, as it requires knowledge of the movement characteristics and it eliminates the consideration that the current state of the robot is affected by past actions.

The regulation of the admittance control parameters can be considered as an optimal control problem, where it is very difficult to acquire perfect knowledge of the system mainly because of the uncertainties of a human model. Reinforcement Learning (RL) is an unsupervised learning method that discovers an optimal solution through interaction with the environment [9] and can overcome the limitations of conventional optimal control algorithms imposed by insufficient environment modelling. In RL, an agent performs actions to the environment and measures their effect by the received rewards or punishments. The key difference with the supervised learning, is that the training feedback in RL is evaluative rather instructive, meaning that the correct action is not provided. Moreover, RL has a mechanism that can associate the dependency among the provided training data [10]. Reinforcement learning methods have been applied in robotic applications for variable impedance/admittance control of manipulators [11], [12] and human-robot coordinated motion [13], [14]. In [13], user intervention is required to speed up learning, while in [14] the controller does not allow compliance to external forces, confining the effectiveness of the physical interaction. To the authors' knowledge, a systematic approach to optimise the admittance gains online during cooperation without prior knowledge of the task characteristics and no user intervention, has not been investigated.

In this work, a variable admittance controller is proposed for optimal human-robot co-manipulation based on reinforcement learning. The Fuzzy Q-Learning algorithm is selected for the training of the RL agent, which is a widely used value function approximation technique to deal with continuous state, real-world problems. The RL agent regulates the virtual

Authors are with the Robotics Group, Dept. of Mechanical Engineering & Aeronautics, University of Patras, 26500 Patra, Greece. Fotios Dimeas is funded by "IKY fellowships of excellence for postgraduate studies in Greece - Siemens program". {dimeasf,asprag}@mech.upatras.gr

damping towards minimisation of the jerk of the movement, without any prior knowledge of the initial and target position or the duration of the movement. Within a small number of iterations and without the need to embed expert knowledge to speed up the learning process, the RL agent can converge to an appropriate mapping between the human-robot state and the virtual damping. The trained variable admittance controller is able to adjust the damping appropriately by maximising the effectiveness of the co-manipulation task in terms of effort and time to complete. The proposed system is tested on an experimental setup with a number of subjects using a KUKA LWR robot, indicating promising results for efficient physical human-robot cooperation.

II. Reinforcement Learning of Variable Admittance Control

In this section, the proposed variable admittance controller is presented in detail, consisting of an admittance control scheme that allows compliant behaviour of the manipulator and an agent that determines the appropriate damping for the admittance controller. The ability of the proposed method to optimise the effectiveness in the co-manipulation, lies in the training of the agent using the RL-based Fuzzy Q-Learning (FQL) algorithm.

A. The Admittance Controller

In the considered co-manipulation task, the human is the leader of the motion and the robot is the follower. Since the interaction point lies at the robot's end-effector (Fig. 1), the admittance controller accepts as an input the externally applied force by the operator and outputs a desired motion for the end-effector. The admittance controller is expressed relative to the Cartesian frame attached to the end-effector as:

$$\mathbf{M_d}\dot{\mathbf{V}}_{\mathbf{ref}} + \mathbf{C_d}\mathbf{V_{ref}} = \mathbf{F_h} \tag{1}$$

where $\mathbf{V_{ref}} \in \mathbb{R}^6$ is the reference Cartesian velocity of the end-effector and $\mathbf{F_h} \in \mathbb{R}^6$ is the measured human force/torque vector. The admittance controller gains $\mathbf{M_d}$, $\mathbf{C_d} \in \mathbb{R}^{6\times 6}$ are positive definite diagonal matrices representing the desired inertia and the damping of the second order-relationship that the controller imposes to the manipulator. The virtual stiffness is omitted because no restoring force is desired in the investigated tasks, where only free-space cooperation is conducted.

Among the gains $\mathbf{M_d}$ and $\mathbf{C_d}$ we focus on the damping parameter because it has greater effect at the effectiveness of the cooperation than the inertia term [2]. The lower the damping of the admittance, the more responsive the manipulator is to external forces. In a point-to-point movement, the motion can be divided into two major phases [15]: a high velocity motion with low accuracy to approach the target, that benefits from low damping and a lower velocity motion for accurate positioning, that is facilitated with an increased damping.

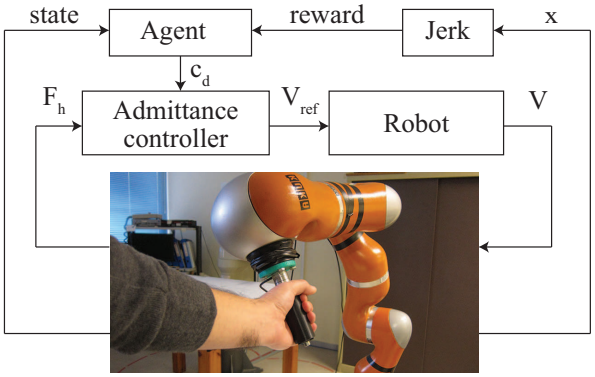


Fig. 1: Reinforcement learning of variable admittance control for human-robot co-manipulation.

B. Reinforcement Learning in co-manipulation

The training of the agent to appropriately regulate the desired damping is based in reinforcement learning (RL). RL is formalised as a Markov Decision Process (MDP), a discrete time stochastic control process. At each time step t, an agent observes the current state s_t of the human-robot system and performs an action U_t by adjusting the desired damping of the admittance controller, selected from a set of discrete actions A. After applying this action, the system arrives to a new state s_{t+1} and a reward r_{t+1} is obtained. The action performed at each state is determined by the policy π of the agent. Each policy is rated by a value function according to the received rewards. When there are no prior policies or reward models of the system available, such as those involving human factors, Q-Learning is a well-known model-free method that can estimate the value functions $Q^{\pi}(S, A)$. A point-to-point movement in a certain direction of the end-effector starts and ends with zero velocity and can be treated as an *episodic* task. The goal of RL is to maximise the total reward received during that episode.

Human-robot cooperation involves a bilateral adaptation between the RL agent and the human operator. It is necessary that both the agent and the human have the same goal [16]. Since the goal of the human is given by the central nervous system to move between two points (in the investigated task) with the least change in acceleration [7], the reward provided to the agent should be selected similarly. Equally important is the observability of both the robot's and the operator's state, a precondition for accurate approximation of the MDP. However, it has been shown in practice that RL can work very well without fulfilment of the last requirement [17]. In the case of the Cartesian admittance controller, the partial representation of the system's state is selected for the benefit of generalisation and simplicity. The relative or absolute position of the robot's end-effector is omitted from the state representation, since it is an *ad hoc* variable that limits the generalisation to arbitrary movements. For cooperation along a single direction of the end-effector, the state provided to the agent is given by the actual Cartesian velocity V in the direction of the motion, the external force F_h applied by the operator and their first derivatives $\dot{V}$, $\dot{F}_h$.

1) Fuzzy Q-Learning: Fuzzy Q-Learning (FQL) is a method that can overcome the curse of dimensionality when representing continuous states and actions of real world problems [18], [19]. Instead of using discrete sets that would yield the problem intractable, FQL realises state representation using fuzzy states, constituted by the fuzzy set S_t. The agent can visit a state partially, in the sense that the real-valued input variables $X = \{V,\ F_h,\ \dot{V},\ \dot{F}_h\}$ may belong up to a degree to the membership functions of the fuzzy sets. All combinations of the input membership functions form the rule base, but unlike the standard fuzzy systems, the actions are selected and not combined. In FQL, the conclusion of each rule $R_i,\ i = 1, ..., n$, where n is the number of rules, is a crisp action a'_i selected from the set of discrete actions A, according to the policy π. The action set A consists of a discrete number of crisp damping values. The selected action by each rule R_i contributes to a continuous global action U_t, which is the damping value C_d provided to the admittance controller, according to the premise strength ϕ_i of that rule. The Fuzzy Inference System (FIS) is implicitly used as a function approximation of the Q value functions. A rule R_i of the FQL agent has the following form:

$$R_i := IF\ X\ is\ S_i\ with\ \phi_i\ THEN\ a_1\ with\ q(S_i, a_1)$$
$$...$$
$$OR\ a_m\ with\ q(S_i, a_m)$$

where S_i is the fuzzy state of the rule R_i composed by a vector of fuzzy sets, ϕ_i is the firing strength of the rule R_i, $A = \{a_1, ..., a_m\}$ are the m possible discrete actions of the rule and $q(S_i, a_j)$ is the q-value that determines the probability of choosing the action j $(j = 1, ..., m)$ of the rule R_i. In order to explore all possible actions and rate them according to the received reward, the policy π selects the action a'_i according to an exploration-exploitation strategy:

$$a'_i = \underset{a_j \in A}{\operatorname{argmax}}(q(S_i, a_j) + \eta(S_i, a_j) + \rho(S_i, a_j)) \quad (2)$$

where $\eta(S_i, a_j)$ and $\rho(S_i, a_j)$ indicate the undirected (random) and directed exploration of actions respectively. The term $\eta(S_i, a_j)$ provides random values from an exponential distribution [19], while $\rho(S_i, a_j)$ forces the selection of actions that have rarely been elected [18]. The selected damping is the global action U_t given by the aggregation of all n rules:

$$U_t(X_t) = \sum_{i=1}^{n} a'_i \phi_i \quad (3)$$

A Q-function quantifies the quality of a given action with respect to the current state and is given by:

$$Q_t(X_t, U_t) = \sum_{i=1}^{n} q_t(S_i, a_j)\phi_i \quad (4)$$

The optimal action for a rule is given by the Q*-function:

$$Q^*_t(X_t) = \sum_{i=1}^{n} (\max_{a_j \in A} q_t(S_i, a_j))\phi_i \quad (5)$$

The q-values are updated at each iteration of the algorithm according to:

$$q_{t+1}(S_i, a_j) = q_t(S_i, a_j) + \beta\tilde{\epsilon}_{t+1}e_t(S_i, a_j) \quad i = 1, 2, ..., n,\ j = 1, 2, ..., m \quad (6)$$

where β is the learning rate, e_t the eligibility trace of the past visited rules (described in II-B.2) and $\tilde{\epsilon}_{t+1}$ is the Temporal Difference (TD) error given by:

$$\tilde{\epsilon}_{t+1} = r_{t+1} + \gamma Q^*_t(X_{t+1}) - Q_t(X_t, U_t) \quad (7)$$

The term r_{t+1} is the reward received at time $t + 1$ and γ is a discount factor that weights the effect of the future rewards. γ must be selected high enough so that the agent will try to collect long term rewards during an episode. The TD error determines whether the quality of an action has to be increased/decreased by comparing the action currently applied with the previous optimal action Q^*_t.

2) Effect of past actions in the current state: A very powerful mechanism of TD learning is the use of eligibility traces. This technique considers the effect of past actions, weighted by $\lambda \in [0, 1]$, and enables their adaptation at the current time step t, in proportion to their proximity to the current state. For each action a_j of rule R_i the trace $e_t(S_i, a_j)$ is calculated as:

$$e_t(S_i, a_j) = \begin{cases} \gamma\lambda e_{t-1}(S_i, a_j) + \phi_i & if\ a_j = a'_i \\ \gamma\lambda e_{t-1}(S_i, a_j) & else \end{cases} \quad (8)$$

Unlike the supervised learning methods [8] that only train according to the current state-action pair, the eligibility traces of FQL memorise past state-action pairs and reduce significantly the training time. Furthermore, reinforcement learning with eligibility traces does not treat the training data as a sequence of independent states, but considers the effect of past decisions to the current state.

3) Reward: The reward r_{t+1} provided to the agent must be specified according to the desired behaviour. In a human-robot co-manipulation task, the reward can be constructed so as to minimise the effort of the operator or reduce the completion time of the task. Alternatively, the minimum jerk trajectory model can be used, which describes the trajectory of the human arm in a reaching task and is characterised with a minimum change in the acceleration throughout the movement as:

$$N^* = min\{N\} = min\left\{\sum_{\tau=0}^{\tau_f} ||\dddot{x}_\tau||^2\right\} \quad (9)$$

where $N^*,\ N$ are non-negative terms and τ_f is the duration of the motion in discrete steps, which is unknown beforehand.

Using as a reference the minimum jerk model with a supervised learning algorithm in a previous work [8], the results demonstrated effective co-manipulation with a decrease in the operator's effort and the completion time, however prior knowledge of the target point and the duration τ_f of movement were required. In the proposed reinforcement learning training procedure, no such prior knowledge is

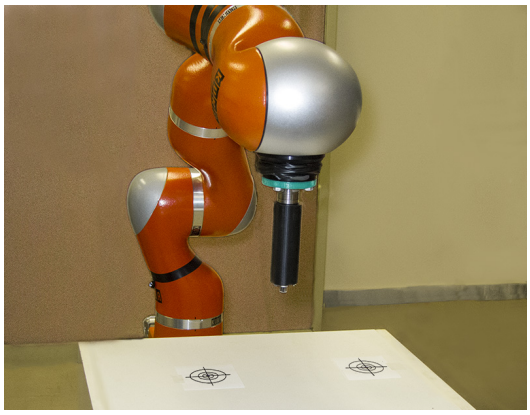


Fig. 2: Experimental setup with a KUKA LWR. The marked targets are unknown to the proposed algorithm.

necessary since the goal of the agent can be the maximisation of the reward provided at the end of the episode. However, this approach requires a significant number of episodes to be conducted to discover an optimal policy, which is not very practical. To overcome this limitation and speed up the learning of the optimal policy, the reward r_{t+1} is provided at intermediate steps to the FQL agent:

$$r_{t+1} = -\sum_{\tau=\tau_t}^{\tau_{t+1}} ||\dddot{x}_\tau||^2 \tag{10}$$

The terms $\tau_t,\ \tau_{t+1} \in \mathbb{N}$ of Eq. (10) represent the discrete steps for the sampling rate of the admittance controller and correspond to the discrete steps of the FQL agent $t,\ t+1 \in \mathbb{N}$. A correlation between τ_t and t using $\tau_t = \kappa t$ is necessary because the FQL agent operates at a rate lower than the sampling loop of the admittance controller. By using an asynchronous agent operating κ times slower than the rate of the controller the training becomes more robust, since the human-robot system is given a sufficient period of time to respond to the action U_t of the previous step. The accumulated reward r_{tot} received until the end of an episode of duration τ_f, equals to the sum of the individual rewards acquired from the first step t_1 until the last step t_f:

$$r_{tot} = \sum_{i=1}^{t_f} r_i = -\sum_{\tau=\tau_0}^{\tau_f} ||\dddot{x}_\tau||^2 = -N \tag{11}$$

The goal of the FQL agent is to regulate the damping accordingly, so as to maximise the reward r_{tot}, which is the opposite of the non-negative jerk N throughout the movement.

III. Experimental evaluation

The evaluation of the proposed variable admittance control scheme is conducted experimentally using a manipulator in cooperation with a human for a translational reaching task. The purpose of the experiment is twofold. First, it is investigated the convergence of the reinforcement learning algorithm within a practically small number of steps. Secondly, the effectiveness of the overall variable admittance controller is evaluated with respect to the required effort and completion time of the task.

The experimental setup consists of a 7DOF LWR robot (Fig. 2). A handle is attached to the end-effector that allows the operator to cooperate with the robot. The external force $\mathbf{F_h}$ applied by the operator is measured using a 6DOF force/torque sensor mounted between the end-effector and the handle. The admittance controller of Eq. (1) is implemented for a single Cartesian direction of the frame attached to the end-effector, allowing compliant behaviour only to the task direction.

To acquire the continuous state input vector $\{V, F_h, \dot{V}, \dot{F_h}\}$, each variable is partitioned with five fuzzy sets using triangular membership functions that are uniformly distributed to the universe of discourse. A complete rule base is formed with a total of $n = 625, (5^4)$ rules. The action set A is selected the same for all rules and consists of three ($m = 3$) damping values ranging from $5Ns/m$ to $50Ns/m$. These values have been experimentally found to allow fast cooperation and accurate positioning respectively [8]:

$$A = \{5, \quad 22.5, \quad 50\}\, Ns/m \tag{12}$$

Selecting among this relatively small set, the FQL agent requires less steps to converge and generates a continuous action $C_d = U_t \in [5\ 50] Ns/m$ provided to Eq. (1) at a rate equal to the admittance controller that operates at 1kHz. The reward to the agent is provided at a slower rate of 100Hz using $\kappa = 10$. This frequency is experimentally found to allow an accumulated reward signal that is more robust to the noise generated by the numeric differentiation for $\dddot{x}$. The learning rate and the discount factor of the agent in (6), (7), (8) are set $\beta = 0.05$, $\gamma = 0.95$, and $\lambda = 0.9$ respectively. The virtual inertia M_d during the experiments is constant and equal to half of the manipulator's effective inertia at the direction of the movement, maintaining the passivity of the system for stability issues.

The desired Cartesian velocities $\mathbf{V_r}$ from Eq. (1) are transformed into robot joint velocities $\dot{\mathbf{q}}_{\mathbf{ref}} \in \mathbb{R}^6$ using the pseudo-inverse Jacobian, and are provided to the position control system via incremental joint position commands. The actual velocity $\mathbf{V}$ of the end-effector is realised by the inner joint control of the robot that operates at higher sampling frequency than the admittance control loop of Eq. (1). This control scheme is also known as position based admittance control and can be implemented in the majority of motion controlled manipulators. Alternatively, the desired dynamic behaviour of Eq. (1) can also be achieved using the impedance control scheme, but only on torque controlled manipulators.

A. Learning variable admittance

In the experimental evaluation participated 7 subjects, aged from 23 to 33 years old, 5 of them are male and all right handed. Each subject is placed in front of the robot, where the initial and target position of the task are visually marked on a white surface (Fig. 2). A red laser pointer

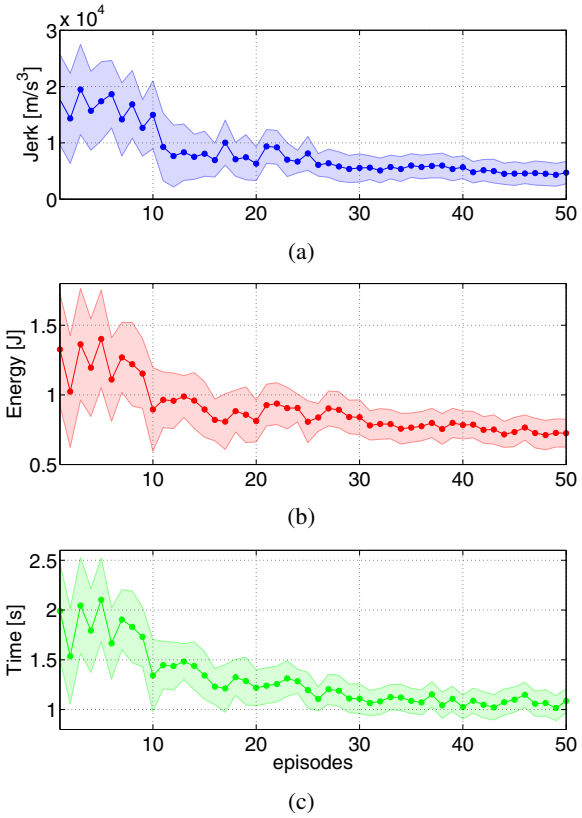


Fig. 3: Performance measures of the cooperation and of the FQL agent using the mean and standard deviation from all subjects for the jerk, energy and completion time.

attached to the end-effector projects the actual position of the robot to that surface for visual assistance. The targets are unknown to the learning algorithm. The subject is asked to cooperate with the robot and move the red pointer from one target to the other. An episode is defined by the time when the motion starts until the end-effector reaches the target with zero velocity. Two consecutive sets of 50 episodes are recorded for each subject, from which only the second one is taken into consideration in order to minimise the learning effects of the subject. Before the experiment, the subjects are informed that the robot will adjust its behaviour and might feel "heavier" or "lighter". After the experiments, a questionnaire is given to the subjects which are asked to characterise the assistance provided by the robot as "helpful", "unhelpful" or "indifferent".

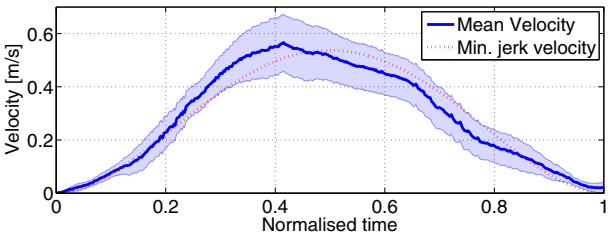
To evaluate both the reinforcement learning algorithm and the effectiveness of the cooperation, the mean value and standard deviation of all subjects for each episode are plotted in Fig. 3. The accumulated jerk of the movement that is calculated from Eq. (11), is the objective function N to be minimised. The energy spent by the operator is calculated by integrating the external force over the distance travelled $\int |f_h| dx$. The goal of the FQL agent is to regulate the damping appropriately in order to maximise the total reward r_{tot} over an episode. A small damping during the motion increases the instantaneous jerk because the robot becomes over-responsive, but it accumulates higher reward over time because of shorter τ_f. On the other hand, with very high damping the instantaneous jerk is decreased but more time and energy are required to reach the target, resulting in lower reward. In practice, the FQL tries to balance between the two contradictory damping values to maximise its reward. At the first 10 episodes of each movement the actions of the agent are exploratory, producing high jerk motion with large deviation among the subjects (Fig. 3a). However, between episodes 11 and 25, the agent has acquired a rough knowledge of the appropriate actions and produces lower jerk, while performing the exploitation strategy. Exploitation actions tend to improve the rough policy until the agent converges to an optimum solution, which happens averagely after the first 25 episodes. Although this solution is not guaranteed to be globally optimum, it converges to a policy that produces approximately the same damping for all subjects.

Having acquired a policy without prior knowledge of the motion to be conducted, the agent has implicitly managed to extract the movement characteristics from the state input vector $\{V, F_h, \dot{V}, \dot{F}_h\}$ and create a mapping to the appropriate damping to assist the minimum jerk trajectory. The effectiveness of such a mapping to the cooperation task is evaluated by the energy transferred from the operator to the robot (Fig. 3b) and the required time to reach the target (Fig. 3c). Indeed, the mean energy in the 5 final episodes from all subjects is reduced by 42% relative to the initial 5 episodes, while the corresponding time to complete is reduced by 44%. Moreover, all subjects (100%) characterised the overall assistance provided by the robot as "helpful". These results suggest that with the proposed method the agent can learn a policy to regulate the damping gain C_d, even without prior knowledge of the movement.

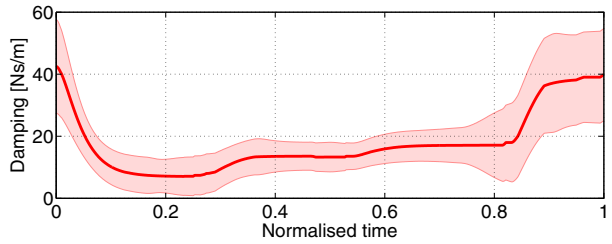
Useful information can be extracted by investigating the progression of the velocity and the damping during the motion in Fig. 4. The mean values and standard deviation are illustrated respectively from all subjects for the last 5 episodes, when the proposed method has converged. The mean velocity profile as a function of normalised time (Fig. 4a) approaches at a great extent the theoretical minimum jerk trajectory given by Eq. (9). The damping provided from the developed FQL agent to the admittance is illustrated in Fig. 4b. The lowest damping value appears approximately between 0.1 and 0.3 (normalised time) in order to assist the operator in the initial acceleration phase and overcome the virtual inertia. From 0.3 until 0.8 the agent selects a slightly increased damping, which can be attributed to the desired reduction in the acceleration towards minimisation of the jerk. Interestingly, the ability of the trained FQL agent to detect the motion features appears with a significant increase in the damping towards the end of the movement, when the target is approached and a deceleration is desired.

IV. Discussion

The proposed method is an initial work that indicates promising results towards effective and effortless physical cooperation between a robot and a human in everyday tasks.



(a)



(b)

Fig. 4: Mean and standard deviation values for the velocity and damping normalised over time, for the last 5 episodes from all subjects.

The training procedure is conducted on a single DOF of the end-effector, where the agent learns an appropriate mapping from the movement characteristics to the virtual damping, facilitating cooperation in a point-to-point movement.

Concerning the reinforcement learning method that was selected, alternative training methods with the ability to search in the continuous space may present similar performance, although their testing is not within the scope of this paper. The selection of the FQL algorithm proves to be sufficient for demonstrating the learning capability for variable damping without requiring knowledge about the movement characteristics. Fuzzy representation is an effective mechanism to deal with the curse of dimensionality of continuous space searching but requires a number of parameters to be manually tuned. Experimental testing has shown that the most sensitive parameter is the eligibility trace weight λ, followed by the rate of the reward provisioning κ.

V. Conclusion

A novel method is presented in this paper based on reinforcement learning for online training of a variable admittance controller in human-robot co-manipulation. The main contribution of this work is the decoupling of the movement characteristics such as the goal position from the learning of the variable admittance controller. The learning is conducted under the following assumption; if the damping is regulated so that the robot trajectory approaches the minimum jerk, the cooperation becomes more effective. Indeed, the agent learns to adjust the damping of the admittance appropriately and reduces of the operator's effort and the time required to complete the task. Experiments with a number of subjects indicated a considerable decrease of those subjective measures by half of their initial values.

Under investigation is the encoding of the acquired mapping between the state variables and the appropriate damping of the admittance controller, into an inference system that can operate to arbitrary movements, involving motion in all Cartesian directions of the end-effector and without any additional training. Since the desired behaviour of Eq. (1) describes a decoupled controller, the trained agent can be implemented independently in each direction of the end-effector.

References

[1] N. Hogan, "Impedance Control: An Approach to Manipulation," in *1984 American Control Conference*, vol. 107 of *Proceedings of the 1984 American Control Conference.*, pp. 304–313, IEEE, 1984.

[2] R. Ikeura, H. Monden, and H. Inooka, "Cooperative motion control of a robot and a human," in *IEEE International Workshop on Robot and Human Communication*, pp. 112–117, IEEE, 1994.

[3] M. S. Erden and B. Marić, "Assisting manual welding with robot," *Robotics and Computer-Integrated Manufacturing*, vol. 27, pp. 818–828, Aug. 2011.

[4] V. Duchaine and C. M. Gosselin, "General Model of Human-Robot Cooperation Using a Novel Velocity Based Variable Impedance Control," in *Second Joint EuroHaptics Conference WHC07*, pp. 446–451, IEEE, 2007.

[5] A. Lecours, B. Mayer-St-Onge, and C. Gosselin, "Variable admittance control of a four-degree-of-freedom intelligent assist device," in *IEEE International Conference on Robotics and Automation*, no. 2, pp. 3903–3908, Ieee, May 2012.

[6] T. Tsumugiwa, R. Yokogawa, and K. Hara, "Variable impedance control based on estimation of human arm stiffness for human-robot cooperative calligraphic task," in *IEEE International Conference on Robotics and Automation*, vol. 1, pp. 644–650, IEEE, 2002.

[7] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *The Journal of Neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[8] F. Dimeas and N. Aspragathos, "Fuzzy Learning Variable Admittance Control for Human-Robot Cooperation," in *IEEE International Conference on Intelligent Robots and Systems*, (September 13-18, Chicago, IL, USA), pp. 4770–4775, 2014.

[9] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction.," *MIT Press*, 1998.

[10] M. Wiering and M. van Otterlo, eds., *Reinforcement Learning, State-of-the-Art*. Springer Berlin Heidelberg, 2012.

[11] S. M. Prabhu and D. P. Garg, "Fuzzy-logic-based Reinforcement Learning of Admittance Control for Automated Robotic Manufacturing," *Engineering Applications of Artificial Intelligence*, vol. 11, pp. 7–23, Feb. 1998.

[12] J. Buchli, F. Stulp, E. Theodorou, and S. Schaal, "Learning variable impedance control," *The International Journal of Robotics Research*, vol. 30, pp. 820–833, Apr. 2011.

[13] U. Kartoun, H. Stern, and Y. Edan, "A Human-Robot Collaborative Reinforcement Learning Algorithm," *Journal of Intelligent & Robotic Systems*, vol. 60, pp. 217–239, May 2010.

[14] A. Thobbi, Y. Gu, and W. Sheng, "Using Human Motion Estimation for Human-Robot Cooperative Manipulation," *IEEE International Conference on Intelligent Robots and Systems*, 2011.

[15] E. Burdet and T. E. Milner, "Quantization of human motions and learning of accurate movements.," *Biological cybernetics*, vol. 78, pp. 307–318, Apr. 1998.

[16] T. Tamei and T. Shibata, "Policy gradient learning of cooperative interaction with a robot using user's biological signals," *Advances in Neuro-Information Processing*, no. 20300071, pp. 1029–1037, 2009.

[17] J. Kober, J. a. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, pp. 1238–1274, Aug. 2013.

[18] L. Jouffe, "Fuzzy inference system learning by reinforcement methods," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 28, no. 3, pp. 338–355, 1998.

[19] M. Er and C. Deng, "Online Tuning of Fuzzy Inference Systems Using Dynamic Fuzzy Q-Learning," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 34, pp. 1478–1489, June 2004.