2016 IEEE International Conference on Robotics and Automation (ICRA)
Stockholm, Sweden, May 16-21, 2016

Learning Assistive Strategies from a Few User-Robot Interactions: Model-based Reinforcement Learning Approach

Masashi Hamaya[1,2], Takamitsu Matsubara[1,3], Tomoyuki Noda[1], Tatsuya Teramae[1] and Jun Morimoto[1]

Abstract— **Designing an assistive strategy for exoskeletons is a key ingredient in movement assistance and rehabilitation. While several approaches have been explored, most studies are based on mechanical models of the human user, i.e., rigid-body dynamics or Center of Mass (CoM)-Zero Moment Point (ZMP) inverted pendulum model, or only focus on periodic movements with using oscillator models. On the other hand, the interactions between the user and the robot are often not considered explicitly because of its difficulty in modeling. In this paper, we propose to learn the assistive strategies directly from interactions between the user and the robot. We formulate the learning problem of assistive strategies as a policy search problem. To alleviate heavy burdens to the user for data acquisition, we exploit a data-efficient model-based reinforcement learning framework. To validate the effectiveness of our approach, an experimental platform composed of a real subject, an electromyography (EMG)-measurement system, and a simulated robot arm is developed. Then, a learning experiment with the assistive control task of the robot arm is conducted. As a result, proper assistive strategies that can achieve the robot control task and reduce EMG signals of the user are acquired only by 30 seconds interactions.**

I. Introduction

There has been a growing demand for developing exoskeleton robots and active orthoses with several structures and actuators, that can perform motion assistance to augment able-bodied persons or the assisted rehabilitation of physically challenged persons [1]–[10]. Compared to the notable recent advances in terms of robot hardware, assistive control strategies in software are still in an exploration stage. Due to bidirectional physical interactions between the robot and the user, the advanced robot control strategies established for manipulation and biped locomotion are less useful.

Several assistive strategies have been explored so far. A typical strategy in power assistance is based on gravity compensation control [3], [4], [11], [12]. It is effective for supporting the load taken by users wearing the exoskeleton robot in a static posture. Another popular strategy is the electromyography (EMG)-based method [13], [14]. With the EMG-to-force model, it predicts the torques required for the motion to be compensated based on the EMG signals.

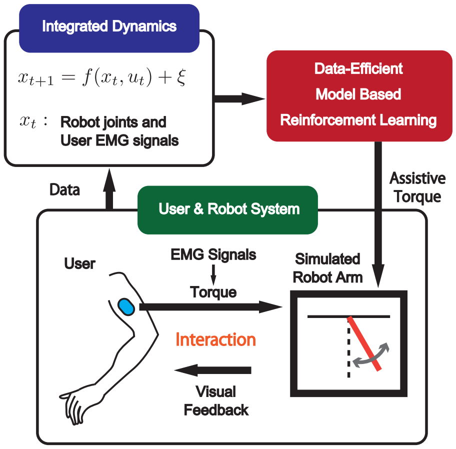


Fig. 1. Schematic diagram of our approach. We formulate the learning problem of assistive strategy from direct interaction. We consider the human and the robot integration system and apply data-efficient model based reinforcement learning.

For walking and balancing assistance, an inverted pendulum model with Center of Mass (CoM) and Zero Moment Point (ZMP) can be used to derive stable gait patterns [9], [15]. The adaptive oscillator-based strategy has also received much attention because of its simplicity. This approach uses an oscillator model to generate coordinated periodic trajectories with user intentions as references to the robot controller [2], [16]–[18]. A recent study combined it with a state-machine based controller [19].

As described above, most previous studies are based on mechanical models of the human user, i.e., rigid-body dynamics or CoM-ZMP inverted model, or only focus on periodic movements with using oscialltor models. On the other hand, the interactions between the user and the robot are often not considered explicitly because of its difficulty in modeling. Therefore, if we are able to construct the model to represent the interactions, assistive control performance can be significantly improved.

To cope with this interaction modeling problem, in this paper, we propose to learn the assistive strategies directly from interactions between the user and the robot. We formulate the learning problem of assistive strategies as a policy search problem. In a practical sense, a very long interactions between the user and the robot should not be required for learning. Therefore, we exploit a data-efficient model-based reinforcement learning framework [20] to alleviate burdens to the user for learning. When we learn a complex model from a small data set, we cannot expect that a good determin-

[1] All authors are with the Department of Brain Robot Interface, ATR-CNS, Kyoto, Japan

[2] MH is with Graduate School of Frontier Bioscience, Osaka University, Osaka, Japan

[3] TM is with Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan

This study is supported by SRPBS from AMED, by MIC-SCOPE, by MEXT KAKENHI 23120004, by NEDO, by "Development of Medical Devices and Systems for Advanced Medical Services" from AMED, and by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan).

978-1-4673-8026-3/16/$31.00 ©2016 IEEE

istic model can be obtained. In the framework, a *probabilistic* model is used because it can express the uncertainty of the model learning by a probability distribution. Moreover, the uncertainty can be taken into account to derive a practical and effective controller.

The schematic diagram of our proposed method is depicted in Fig. 1. Our approach considers the human-robot integrated dynamics explicity, since such a view point makes the formulation of learning problem of assistive strategy tractable. Several reinforcement learning methods have been applied in several human-in-the-loop scenarios, for example, human-robot collaborative lifting [21], myoelectric prosthesis control [22], walking-aid robot control [23], robotic training for a dart-throwing task [24], and telemanipulation of non-rigid objects [25]. However, model-based reinforcement learning for the human-in-the-loop human-robot integrated system has not been explored. Again, using a model based approach is critical to efficiently learn control policies from limited amount of data.

As the proof of concept study of our approach, an experimental platform composed of a real subject, a EMG-measurement system, and a simulated robot arm is developed. Then, a learning experiment with the assistive control task of the robot arm is conducted. As a result, proper assistive strategies that can achieve the robot control task and reduce EMG signals of the user are acquired only by 30 seconds interactions. We believe that our approach is readily applicable for a wide range of assistive robots.

The remainder of this paper is structured as follows. Section II presents our formulation of learning problem of assistive strategy. Section III briefly summarizes the data-efficient model-based reinforcement learning framework. Then, Section IV and Section V show simulation and experimental studies, repectively. Finally, we conclude this paper and state some future work in Section VI.

II. Problem Formulation

The aim of this section is to formulate the learning problem of assistive strategy from direct interactions as shown in Fig. 1. We assume that the robot is tightly coupled to the user such as in typical exoskeleton movement assistance scenarios.

The future state of the robot can depend on the current state of the robot, the robot action, and the user action. Therefore, the robot dynamics can be written as follows:

$$\boldsymbol{s}_{t+1} = \boldsymbol{g}(\boldsymbol{s}_t, \boldsymbol{u}_t, \boldsymbol{v}_t) + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \Sigma_\nu) \tag{1}$$

where $\boldsymbol{s}_t$ is the state of the robot (e.g., joint angles and velocities), and $\boldsymbol{u}_t$ is the robot action (e.g., joint torques generated by the actuators). $\boldsymbol{v}_t$ is the user's action (e.g., joint torques generated or muscle activations). ν_t is an additive Gaussian noise.

One hand, the user's action is decided by the user's control policy that may be based on the robot state, the robot action and the previous user's action. Thus, the user's control policy can be modeled by:

$$\boldsymbol{v}_{t+1} = \boldsymbol{h}(\boldsymbol{s}_t, \boldsymbol{u}_t, \boldsymbol{v}_t) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \Sigma_\eta) \tag{2}$$

where η_t is an additive Gaussian noise.

By integrating them into one equation, the human-robot integrated dynamics can be represented as follows:

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t) + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \Sigma_\xi) \tag{3}$$

where

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{v} \end{bmatrix}, \quad \Sigma_\xi = \begin{bmatrix} \Sigma_\nu & 0 \\ 0 & \Sigma_\eta \end{bmatrix}. \tag{4}$$

Based on the above system, we formulate our learning problem of assistive strategies. Our objective is to find the robot control policy (assistive strategy) $\pi : \pi(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{u}$ which minimizes long-term cost

$$J^\pi(\boldsymbol{\theta}) = \Sigma_{t=0}^{T} \mathbb{E}_{\boldsymbol{x}_t}[c(\boldsymbol{x}_t)], \ \boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{5}$$

where J^π evaluates the cost for T steps, $\boldsymbol{\theta}$ is policy parameter and $c(\boldsymbol{x}_t)$ is the cost in state $\boldsymbol{x}$ at time t. This formulation can cover a wide range of applications in robotic assistance. For example, if the cost is composed of the error between the robot state and the time-dependent target, and the norm of user action, it can be applied for assistance of trajectory tracking.

The difficulty to solve the above problem comes from the modeling of such dynamics. Due to the inclusion of the user action policy and the interaction effects between the user and the robot, it does not follow the standard physics like rigid body dynamics anymore. Identifying it from scratch would require unrealistic amount of data that is impossible to collect from the human-in-the-loop system.

III. Model-Based Reinforcement Learning Approach using PILCO

Deisenroth et al. proposed Probabilistic Inference for Learning Control (PILCO), a model based policy search method [20]. PILCO uses probabilistic models, non-parametric Gaussian processes to consider the uncertainty of the models. Since PILCO computes long-term predictions, policy evaluation and policy improvement analytically, it can perform data-efficient learning. In this section, we briefly summarize PILCO. More details can be found in [20].

PILCO considers the dynamical system in Eq. (3) and the long-term cost in Eq. (5). The cost $c(\boldsymbol{x}_t)$ is typically given as

$$c(\boldsymbol{x}_t) = 1 - \exp\left(-\frac{1}{2\sigma_c^2}(\boldsymbol{x}_t - \boldsymbol{x}_t^d)^\top \boldsymbol{T}^{-1}(\boldsymbol{x}_t - \boldsymbol{x}_t^d)\right) \tag{6}$$

where $\boldsymbol{x}_t^d$ is desired trajectory, σ_c is the width of the cost function, $\boldsymbol{T}$ is a diagonal matrix which expresses the weight of each element in the state for the cost.

A. Model Learning

To create the model, PILCO uses the Gaussian process regression [26] where $(\boldsymbol{x}_t, \boldsymbol{u}_t) \in \mathbb{R}^{D+F}$ is training input, $\boldsymbol{\Delta}_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \in \mathbb{R}^D$ is training output. It typically uses the following kernel function:

$$\begin{aligned} k(\tilde{\boldsymbol{x}}_p, \tilde{\boldsymbol{x}}_q) = \ & \sigma_f^2 \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{x}}_p - \tilde{\boldsymbol{x}}_q)^\top \boldsymbol{\Lambda}^{-1}(\tilde{\boldsymbol{x}}_p - \tilde{\boldsymbol{x}}_q)\right) \\ & + \delta_{pq}\sigma_\xi^2 \end{aligned} \tag{7}$$

with $\tilde{\boldsymbol{x}} := [\boldsymbol{x}^\top, \boldsymbol{u}^\top]$, $\boldsymbol{\Lambda}$ is a diagonal matrix which expresses characteristic length and σ_f is the bandwidth parameter. These parameters are learned with n training inputs $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_n]$ and targets $\boldsymbol{y} = [\boldsymbol{\Delta}_1, ..., \boldsymbol{\Delta}_n]$.

The predictive distribution of $\boldsymbol{x}_{t+1}$ is analytically given as follows:

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{u}_t) = \mathcal{N}(\boldsymbol{x}_{t+1}|\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}), \tag{8}$$

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{x}_t + \mathbb{E}_f[\boldsymbol{\Delta}_t], \quad \boldsymbol{\Sigma}_{t+1} = \mathrm{Var}_f[\boldsymbol{\Delta}_t], \tag{9}$$

where

$$\mathbb{E}_f[\boldsymbol{\Delta}_t] = m_f(\tilde{\boldsymbol{x}}_t) = \boldsymbol{k}_*^\top(\boldsymbol{K} + \sigma_\xi^2\boldsymbol{I})^{-1}\boldsymbol{y} = \boldsymbol{k}_*^\top\beta \tag{10}$$

$$\mathrm{Var}_f[\boldsymbol{\Delta}_t] = k_{**} - \boldsymbol{k}_*^\top(\boldsymbol{K} + \sigma_\xi^2\boldsymbol{I})^{-1}\boldsymbol{k}_* \tag{11}$$

$\boldsymbol{k}_* := k(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{x}}_t)$, $k_{**} := k(\tilde{\boldsymbol{x}}_t)$, and $\beta := (\boldsymbol{K} + \sigma_\xi^2\boldsymbol{I})^{-1}\boldsymbol{y}$, where $\boldsymbol{K}$ is the kernel matrix each of which element follows $K_{ij} = k(\tilde{x}_i, \tilde{x}_j)$.

B. Control Policy

The following control policy is employed:

$$\pi(\boldsymbol{x}_*) = \sum_{i=1}^{N} k(\boldsymbol{m}_i, \boldsymbol{x}_*)(\boldsymbol{K} + \sigma_\pi^2\boldsymbol{I})^{-1}\boldsymbol{t} = k(\boldsymbol{M}, \boldsymbol{x}_*)^\top\boldsymbol{\alpha} \tag{12}$$

$\boldsymbol{\alpha} = (\boldsymbol{K} + \sigma_\pi^2\boldsymbol{I})^{-1}\boldsymbol{t}$, where $\boldsymbol{x}_*$ is a test input, $\boldsymbol{t}$ is a training target, $\boldsymbol{M} = [\boldsymbol{m}_1, ..., \boldsymbol{m}_N]$ are the centers of the Gaussian basis functions, σ_π^2 is noise variance and k is kernel function.

C. Policy Evaluation

To evaluate the control policy, we need to compute the long-term cost J^π. While it cannot be obtained analytically due to the complexity of the Guassian process model, PILCO employs a reasonable approximation scheme with an analytic moment matching technique.

To predict $\boldsymbol{x}_{t+1}$, PILCO assumes the distribution $p(\tilde{\boldsymbol{x}}_t) = p(\boldsymbol{x}_t, \boldsymbol{u}_t)$ as the Gaussian and calculates $p(\boldsymbol{\Delta}_t)$ as follows:

$$p(\boldsymbol{\Delta}_t) = \iint p(f(\tilde{\boldsymbol{x}}_t)|\tilde{\boldsymbol{x}}_t)p(\tilde{\boldsymbol{x}}_t)\mathrm{d}f\mathrm{d}\tilde{\boldsymbol{x}}_t. \tag{13}$$

Eq. (13) is calculated analytically. PILCO also assumes the mean $\boldsymbol{\mu}_{\boldsymbol{\Delta}}$ and the covariance $\boldsymbol{\Sigma}_\Delta$ of the distribution $p(\boldsymbol{\Delta}_t)$ are known. Then, the mean and covariance of $p(\boldsymbol{x}_{t+1}) = \mathcal{N}(\boldsymbol{x}_{t+1}|\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1})$ are obtained by

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \boldsymbol{\mu}_{\boldsymbol{\Delta}}, \tag{14}$$

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}_{\boldsymbol{\Delta}} + \mathrm{cov}[\boldsymbol{x}_t, \boldsymbol{\Delta}_t] + \mathrm{cov}[\boldsymbol{\Delta}_t, \boldsymbol{x}_t]. \tag{15}$$

Based on this prediction distribution, expected value $\mathbb{E}_{\boldsymbol{x}_t}[c(\boldsymbol{x}_t)]$ can be computed analytically:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_t}[c(\boldsymbol{x}_t)] &= \int c(\boldsymbol{x}_t)\mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\mathrm{d}\boldsymbol{x}_t && (16)\\
&= 1 - |\boldsymbol{I} + \Sigma_t\boldsymbol{T}^{-1}|^{-1/2}\\
&\times \exp(-\frac{1}{2}(\boldsymbol{\mu}_t - \boldsymbol{x}_t^d)^\top\tilde{\boldsymbol{S}}^{-1}(\boldsymbol{\mu}_t - \boldsymbol{x}_t^d), && (17)\\
\tilde{\boldsymbol{S}} &:= (\boldsymbol{I} + \Sigma_t\boldsymbol{T}^{-1})^{-1}. && (18)
\end{aligned}$$

By using the above equations, we can compute the approximation of J^π analytically.

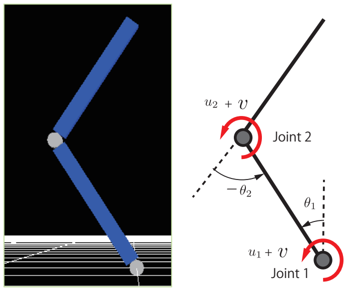


Fig. 2. The two-links rigid robot arm. Each joints generates torque which PILCO learns u_1, u_2 and assumed human state v

D. Policy Improvement with Analytic Gradient

The policy improvement is to find $\boldsymbol{\theta}$ which minimizes $J^\pi(\boldsymbol{\theta})$. The gradient $\partial J^\pi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be computed analytically by using the chain-rule because of the analytic expression of the policy evaluation. The gradient is expressed with $\varepsilon_t := \mathbb{E}_{\boldsymbol{x}_t}[c(\boldsymbol{x}_t)]$

$$\begin{aligned}
\frac{\mathrm{d}J^\pi(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} &= \sum_{t=1}^{T}\frac{\mathrm{d}\varepsilon_t}{\mathrm{d}\boldsymbol{\theta}},\\
\frac{\mathrm{d}\varepsilon_t}{\mathrm{d}\boldsymbol{\theta}} &= \frac{\mathrm{d}\varepsilon_t}{\mathrm{d}p(\boldsymbol{x}_t)}\frac{\mathrm{d}p(\boldsymbol{x}_t)}{\mathrm{d}\boldsymbol{\theta}} := \frac{\partial\varepsilon_t}{\mathrm{d}\boldsymbol{\mu}_t}\frac{\mathrm{d}\boldsymbol{\mu}_t}{\partial\boldsymbol{\theta}} + \frac{\partial\varepsilon_t}{\mathrm{d}\boldsymbol{\Sigma}_t}\frac{\mathrm{d}\boldsymbol{\Sigma}_t}{\partial\boldsymbol{\theta}}.
\end{aligned} \tag{19}$$

Therefore, standard gradient-based non-convex optimization methods, such CG or BFGS, can be applied to find an locally optimal parameter $\boldsymbol{\theta}$.

IV. Simulation

The aim of this simulation is to investigate whether a proper control policy can be learned for a human-robot integrated system composed of the robot and the human models with their interactions even when there is large uncertainty in the human movement model.

We consider a squat-trajectory tracking task to evaluate our proposed approach.

A. Human-Robot Integrated System

We considered a planar two-link rigid body model in Fig. 2. The weight, inertia and length of both links were 1.0 kg, 1.0 kgm^2 and 1.0 m, respectively. The state of the robot $\boldsymbol{s}$ is composed of joint angles and angular velocities: $\mathrm{s} = [\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2]^\top$. In this simulation, dynamics of the human state v was defined as:

$$v_{t+1} = v_t + 0.1\Delta t + \eta_t, \tag{20}$$

where the time step was $\Delta t = 0.004$ s and the system noise which represents uncertainty of the human movement dynamics was given as $\eta_t \sim \mathcal{N}(0, 0.25)$. With considering the robot dynamics, state of the human-robot integraded system can be written as $\boldsymbol{x} = [\boldsymbol{s}^\top, v]^\top = [\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2, v]^\top$.

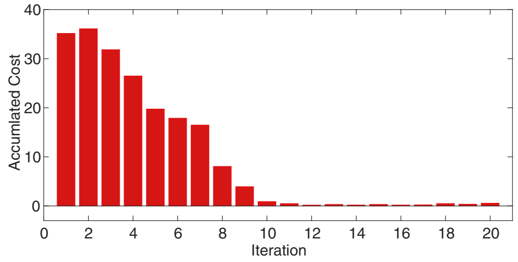


Fig. 3. The accumulated cost on each iterations with the strong observation noise. The cost decreases as the number of the iterations increases.

B. Task Setup

Desired squat trajectories were defined as:

$$\begin{cases}\theta^d_{1\,t} = a_1 \sin(\omega t) + b_1 \\ \theta^d_{2\,t} = -a_2 \sin(\omega t) + b_2,\end{cases} \tag{21}$$

where angular frequency was $\omega = 0.25\pi$, amplitudes were $a_1 = 0.1$ and $a_2 = 0.2$, and nominal postures were defined with $b_1 = 0.6$ and $b_2 = -1.2$.

The weight matrix of the cost function in Eq. (6) was given as

$$\boldsymbol{T} \;=\; \begin{pmatrix} \boldsymbol{T}_s & 0 \\ 0 & T_v \end{pmatrix}, \tag{22}$$

where $\boldsymbol{T}_s = \mathrm{diag}(1.0, 1.0, 1.0, 1.0)$ and $T_v = 0$. The initial state mean was $\boldsymbol{\mu}_0 = [0.6,\ -1.2,\ 0.0,\ 0.0,\ 0.0]^\top$. Maximum and minimum torques both for u_1 and u_2 were $[-3.0, 3.0]$ Nm. The prediction horizon for PILCO algorithm was 4.0 s. The controller period was 0.1 s. We utilized a PILCO open source code [27]. OpenHRP3 was used as a dynamic simulator [28]. The robot dynamics was calculated at each 0.004 s and PILCO updated the control output at each 0.1 s.

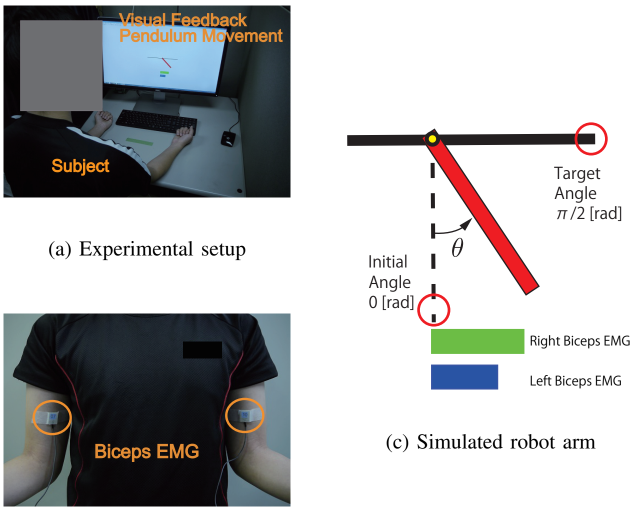
C. Simulation Result

After the initial exploration trial for system identification phase of PILCO algorithm, we conducted 20 learning iterations. The total learning time in the simulated environment for the 20 learning iterations was 80 s, while additional 4 s was required for the exploration phase. PILCO parameterized the policy in Eq. (12) with 100 basis functions resulting in 816 parameters. As shown in Fig. 3, the accumulated cost was decreased as the number of iterations increased even with considering the human movement dynamics.

These simulation results suggest that our approach can be effective for learning control policies of the human-robot integrated system.

V. Experiment

To evaluate our approach, a learning experiment of assistive strategies was conducted. The reaching task was selected for the evalution: the subject tries to move a simulated robot arm to a desired angle by using the subject's EMG signals. The goal of our approach is to learn a proper assistive controller which can assist the user's task.



(a) Experimental setup

(b) EMG

(c) Simulated robot arm

Fig. 4. Experimental setup. The subject tries the reaching task, moves the simulated robot arm (a) to the target angle by using his biceps EMG signals (b). He obtains visual feedback of the moving arm from PC monitor (c).

A. Experimental Setup

Our experimental platform was composed of a real human subject, EMG-measurement system and the simulated robot arm in Fig. 4. For simplicity, we used an one-link rigid robot arm, where the weight, inertia, length and damping of the robot were 1.0 kg, 1.0 kgm^2, 1.0 m and 0.5 Nms/rad, respectively. Muscle activities of the subject's biceps were measured and the measured EMGs were used to control the simulated robot arm. Visual feedback of the robot movement were provided to the subject. Then the subject adjusted his muscle activities to reach to a target angle. PILCO learns the controller to reduce the subject's EMG signals.

1) Human-Robot Integrated System: In Fig. 5, we show the block diagram of the learning system. The robot state $\boldsymbol{s}$ in Eq. (1) is represented by the joint angle θ and the angular velocity $\dot{\theta}$. The human state $\boldsymbol{v}$ in Eq. (2) is defined with the subject's right and left biceps EMG siganls, E_r and E_l. The system can be written as $\boldsymbol{x} = [\boldsymbol{s}^\top, \boldsymbol{v}^\top]^\top = [\theta, \dot{\theta}, E_r, E_l]^\top$.The input torque to the robot model is computed as follows:

$$\tau_{in} = u + K_r E_r - K_l E_l \tag{23}$$

where u is the torque generated by the learned policy of PILCO, K_r and K_l are coefficients and they are set by $K_r = 0.5$ Nm/mV and $K_l = 0.5$ Nm/mV.

2) The Reaching Task: The subject tried to move the simulated robot to reach the target angle, $\theta = \pi/2$ rad. The torques generated by the policy of PILCO were limited within $[-10, 10]$ Nm. The initial state mean was $\boldsymbol{\mu}_0 = [0.0,\ \ 0.0,\ \ 1.5,\ \ 1.5]^\top$. The control period was 0.1 s. The prediction horizon was 3.0 s. The weight matrix for the cost

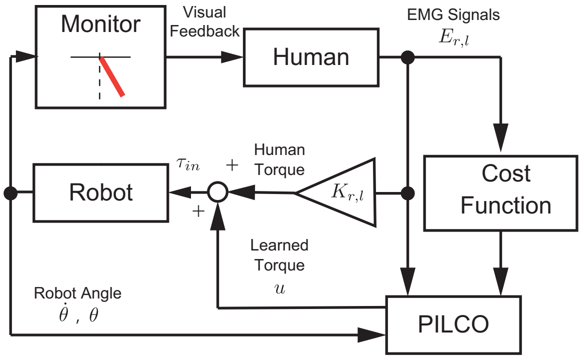


Fig. 5. Block diagram of our proposed system

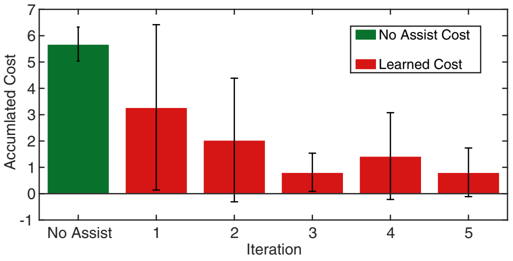


Fig. 6. Results of experiment. The subject tried the task 7 sessions. The red bar shows the accumulated cost on each iterations. The green bar is the cost without assistance by the learned controller.

function was given as:

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_s & 0 \\ 0 & \boldsymbol{T}_v \end{pmatrix}, \quad (24)$$

where $\boldsymbol{T}_s = \boldsymbol{0}$ and $\boldsymbol{T}_v = \mathrm{diag}([0.0005, \ 0.0001])$. The desired trajectory was given as: $E^d_{r\,t} = 0, \ E^d_{l\,t} = 0$, so that the robot tried to assist human movements. To create initial models, PILCO conducted 15 s interaction with the subject by applying random Gaussian torque $u_{int} \sim \mathcal{N}(5.0, 3.0^2)$. The sampling time of the EMG signal was $\Delta t = 0.004$ s, PILCO control period was 0.1 s, therefore the state E was calculated at each 0.1 s as:

$$E_{r,l} = \sum_{k=t-25}^{t} e_k, \quad (25)$$

where e_k is a full-wave rectified EMG signal.

B. Experimental Result

A subject conducted the task seven sessions. One session was composed of five learning iterations with 15 s initial interaction. The total simulated interaction time was 30 s. To compare the learned controller, the subject tried it without robot assistive torque five times. PILCO parameterized the policy in Eq. (12) with 20 basis functions resulting in 125 parameters. Fig. 6 shows the accumulated cost on each iterations. The mean of the accumulated costs was much decreased as the number of the iterations increased. Furthermore, the accumulated cost was much lower than that of without robot assistive control inputs.

To confirm the details of the tracking performance and muscle activities during the task, Fig. 7 shows control performances in different learning stages. The muscle activities were evaluated by percentages of maximum voluntary contraction (%MVC). It used the maximum value of EMG signals ($E_{\max}$): %MVC $= E/E_{\max}$. At the 0, 1st and 2nd iterations, the measured joint angles were apart from the target position, and the subject needed to generate large %MVC. On the other hand, at the 5th iteration, the measured joint angle was reached to target position, and the subject's muscle activities were close to zero.

These experimental results verified the effectiveness of our method for learning proper assistive strategies.

VI. Conclusion and Future Work

In this paper, we proposed to learn the assistive strategies directly from interactions between the user and the robot. We formulated the learning problem of assistive strategies. To reduce the interactions between the user and robot, we applied the data-efficient model-based reinforcement learning framework. As the first step, we conducted the experiment for the proof of concept. The system was composed of a subject, EMG-measurement and a simulated robot arm. As a result, the assistive strategies which reduce the subject's EMG signals were learned for 30 seconds interactions.

As the future work, we will compare our propose method with other reinforcement learning methods [21] [22] [23] [25] whether the assistive strategy can be learned fewer interactions. In addition, we will recruit more subjects to conduct statistical evaluation of our proposed method. Finally, we will apply our approach to a lower exoskeleton robot control for walking assistance. [7].

References

[1] K. Suzuki, G. Mito, H. Kawamoto, Y. Hasegawa, and Y. Sankai, "Intention-based walking support for paraplegia patient with robot suithal," *Advanced Robotics*, vol. 21, no. 12, pp. 1441–1469, 2007.

[2] X. Zhang and M. Hashimoto, "SBC for motion assist using neural oscillator," in *Proc. IEEE Int'l Conf. on Robotics and Automation*, 2009, pp. 659–664.

[3] S. Jacobsen, "On the development of XOS, a powerful exoskeletal robot," in *Plenary Talk IEEE/RSJ Int'l Conf. on Intelligent Robots and System*, 2007.

[4] K. Kazerooni, A. Chu, and R. Steger, "That which does not stabilize, will only make us stronger," *The Int'l J. of Robotics Research*, vol. 26, no. 1, pp. 75–89, 2007.

[5] H. Kwa, J. Noorden, M. Missel, T. Craig, J. Pratt, and P. Neuhans, "Development of the ihmc mobility assist exoskeleton," in *Proc. IEEE Int'l Conf. on Robotics and Automation*, 2009, pp. 2556–2562.

[6] K. Yamamoto, K. Hyodo, M. Ishii, and T. Matsuo, "Development of power assisting suit for assisting nurse labor," *JSME Int'l J. Series C*, vol. 45, no. 3, pp. 703–711, 2002.

[7] S. Hyon, J. Morimoto, T. Matsubara, T. Noda, and M. Kawato, "XoR: Hybrid Drive Exoskeleton Robot That Can Balance," in *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 2011, pp. 2715–2722.

[8] A. Asbeck, S. De Rossi, I. Galiana, Y. Ding, and C. Walsh, "Stronger, smarter, softer: Next-generation wearable robots," *IEEE Robotics Automation Magazine*, vol. 21, no. 4, pp. 22–33, Dec 2014.

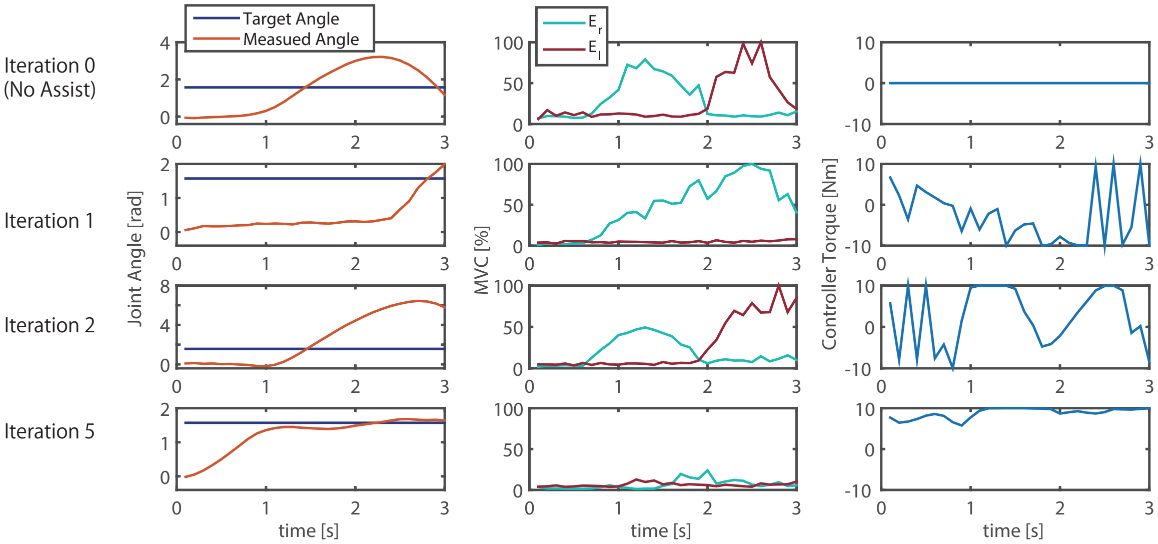


Fig. 7. The tracking performance and muscle activities with and without learned controller.

[9] S. Wang, L. Wang, C. Meijneke, E. van Asseldonk, T. Hoellinger, G. Cheron, Y. Ivanenko, V. La Scaleia, F. Sylos-Labini, M. Molinari, F. Tamburella, I. Pisotta, F. Thorsteinsson, M. Ilzkovitz, J. Gancet, Y. Nevatia, R. Hauffe, F. Zanow, and H. van der Kooij, "Design and control of the mindwalker exoskeleton," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 2, pp. 277–286, March 2015.

[10] R. Farris, H. Quintero, and M. Goldfarb, "Preliminary evaluation of a powered lower limb orthosis to aid walking in paraplegic individuals," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 6, pp. 652–659, Dec 2011.

[11] S. Banala, S. Agrawal, A. Fattah, V. Krishnamoorthy, W.-L. Hsu, J. Scholz, and K. Rudolph, "Gravity-balancing leg orthosis and its performance evaluation," *IEEE Trans. on Robotics*, vol. 22, no. 6, pp. 1228–1239, 2006.

[12] C. J. Walsh, K. Endo, and H. Herr, "A quasi-passive leg exoskeleton for load-carrying augmentation," *Int'l J. of Humanoid Robotics*, vol. 4, no. 3, pp. 487–506, 2007.

[13] H. Kawamoto, S. Kanbe, and Y. Sankai, "Power assist method for HAL-3 using EMG-based feedback controller," in *Proc. Int'l Conf. on Systems, Man and Cybernetics*, 2003, pp. 1648–1653.

[14] C. Fleischer, C. Reinicke, and G. Hommel, "Predicting the intended motion with EMG signals for an exoskeleton orthosis controller," in *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 2005, pp. 2029–2034.

[15] T. Kagawa and Y. Uno, "Gait pattern generation for a power-assist device of paraplegic gait," in *IEEE Int'l Symposium on Robot and Human Interactive Communication*, 2009, pp. 633–638.

[16] R. Ronsse, N. Vitiello, T. Lenzi, J. van den Kieboom, M. Carrozza, and A. Ijspeert, "Adaptive oscillators with human-in-the-loop: Proof of concept for assistance and rehabilitation," in *Proc. IEEE/RAS-EMBS Int'l Conf. on Biomedical Robotics and Biomechatronics*, 2010, pp. 668–674.

[17] T. Matsubara, A. Uchikata, and J. Morimoto, "Full-body exoskeleton robot control for walking assistance by style-phase adaptive pattern generation," in *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 2012, pp. 3914–3920.

[18] T. Matsubara, D. Uto, T. Noda, T. Teramae, and J. Morimoto, "Style-phase adaptation of human and humanoid biped walking patterns in real systems," in *Proc. IEEE/RAS Int'l Conf. on Humanoid Robots*, 2014, pp. 128–133.

[19] T. Yan, M. Cempini, C. M. Oddo, and N. Vitiello, "Review of assistive strategies in powered lower-limb orthoses and exoskeletons," *Robotics and Autonomous Systems*, vol. 64, pp. 120 – 136, 2015.

[20] M. P. Deisenroth, D. Fox, and C. E. Rasmussen, "Gaussian processes for data-efficient learning in robotics and control," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol. 37, no. 2, pp. 408–423, 2015.

[21] T. Tamei and T. Shibata, "Fast Reinforcement Learning for Three-Dimensional Kinetic Human-Robot Cooperation with an EMG-to-Activation Model," *Advanced Robotics*, vol. 25, no. 5, pp. 563–580, 2011.

[22] P. Pilarski, M. Dawson, T. Degris, F. Fahimi, J. Carey, and R. Sutton, "Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning," in *Proc. IEEE/RAS-EMBS Int'l Conf. on Rehabilitation Robotics*, 2011, pp. 1–7.

[23] W. Xu, J. Huang, Y. Wang, and H. Cai, "Study of reinforcement learning based shared control of walking-aid robot," in *Proc. IEEE/SICE Int'l Symposium on System Integration*, 2013, pp. 282–287.

[24] C. Obayashi, T. Tamei, and T. Shibata, "Assist-as-needed robotic trainer based on reinforcement learning and its application to dart-throwing," *Neural Networks*, vol. 53, pp. 52–60, 2014.

[25] T. Matsubara, T. Hasegawa, and K. Sugimoto, "Reinforcement Learning of Shared Control for Dexterous Telemanipulation: Application to a Page Turning Skill," in *Proc. IEEE Int'l Symposium on Robot and Human Interactive Communication*, 2015, pp. 343–348.

[26] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Springer, 2006.

[27] "Pilco web site." [Online]. Available: http://mlg.eng.cam.ac.uk/pilco/

[28] "Openhrp." [Online]. Available: http://fkanehiro.github.io/openhrp3-doc/