Toward a mechanistic account for imitation learning: an analysis of pendulum swing-up

Takuma Torii[1] and Shohei Hidaka[1]

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
{tak.torii,shhidaka}@jaist.ac.jp

Abstract. Learning an action from others require to infer their underlying goals, and recent psychological studies have reported behavioral evidences that young children do infer others' underlying goals by observing their actions. The goal of the present study is to propose a mechanistic account for how this goal inference is possible by observing others' actions. For this purpose, we performed a series of simulations in which two agents control pendulums toward different goals, and analyzed with which types of features it is possible to infer their different latent goals and control schemes. Our analysis showed that pointwise dimension, a type of fractal dimension, of the pendulum movements is sufficiently informative to classify the types of agents. With respect to its invariant nature, this result suggests that the fine-grained movement patterns such as the fractal dimension reflect the structure of the underlying control schemes and goals.

Keywords: imitation learning, goal inference, dynamical systems, pendulum swing-up

1 Introduction

It is crucial for human being as a social being to learn actions by observing others' actions. We refer to a sequence of movements with a certain goal or plan behind it as *action* [2]. Thus, learning an action includes inference of the underlying goal or plan as well as production of bodily movements to achieve the goal inferred. In this study, we exclusively use the term *action* in the sense of Bernstein [2], which is a movement controled to solve a certain problem. In this sense, the actions should be classified by their goals, although movements in general can be classified by their physical appearence. Beyond mere replication of movements or mimicking on the basis of apparent similarity, learners need to infer goals behind actions. Thus, our main question here is how this is possible with less knowledge on the goals or the actions.

Importantly, inference of the goal of an action does not require complete knowledge of bodily movements. According to accumulating pieces of empirical evidence, children as young as 18 month old can infer the goal without observing the intended outcome of action, with which the goal becomes evident [13, 14]. In

their experiments, the psychologists presented to each infant an incomplete accidental and/or intentional action performed by an adult and observed response of the infant. In an experimental condition, children observe a person trying to get an out-of-reach object he *accidentally* dropped on the floor – that is an act repeating the unfulfilled actions. On the other hand, in the control condition, the person *intentionally* threw the object on the floor. For their purpose, the experimenters acted carefully so as not to make apparent difference between the movements (i.e., "drop" or "throw") for 18 month old. The psychologists found that only in the former condition, infants show helping behavior to complete the unfulfilled action of the person (i.e., the children handed him the object). This suggests that children can predict what movements follow the adult's action and thus discriminate the difference in intentions behind seemingly similar movements.

This goal inference underlying actions is a key step toward social learning from others' behavior. In this study, we seek for a mechanistic account for goal or plan inference by observation of unfulfilled actions. There are, however, two major problems to address the mechanism of the goal inference of actions. First, the learner may have a different body from the instructor, and thus observation is never complete and the learner needs to specify missing variables. In theoretical studies, this problem is often formulated as an *inverse problem* [9, 8]. In this formulation, the unobservables are inferred using a generative model (or forward model) and an assumed optimality principle, and this inferrence goes backward to the generating process.

Second, a sequence of bodily movements is typically just one of possible means to achieve a given goal, and thus seemingly different movements may be similar in terms of a certain type of goals, and vice versa – two very similar movements may be performed for two different goals. This problem requires to solve another type of inverse problem, in which the one needs to identify and differentiate bodily movements according to the goals.

To solve the inverse problem, it is often assumed that the learner knows an appropriate class of forward models (a generator of bodily movements), which defines the generative process, from the goal to movements, and allows to estimate a likely model for given observations [9, 8]. Thus, this approach is limited in learning actions, especially to goal inference from incomplete actions, as in such cases no or little observations is available for the learner to successfully infer the hidden goal (e.g., [11, 10, 3]).

To seek for a new theoretical account for goal inference on unfulfilled actions, as instantiated in the empirical studies [13, 14], we analyze seemingly similar movements generated to follow different intentions to accomplish a task. In a prior stage of learning actions, without knowing the goals, learners have to classify given movements by supposed intentions. Here by the term "intention" we refer to as a motor control scheme that outputs the motor action for a given bodily state as the input, by which a sequence of bodily movements is generated for a given initial state. The motor control scheme is constructed by near-optimally fulfilling a given goal for the actor.

As a first step toward understanding the mechanism how children recognize the intention behind the other's actions, we ask the two basic questions:

1. How can we identify multiple seemingly different actions generated by the same motor control scheme?
2. How can we differentiate multiple seemingly similar actions generated by two different motor control schemes?

To address these questions, we resorted a computer simulations of the simplest possible physical body — the classical pendulum swing-up task of one degree-of-freedom, that can produce seemingly similar movements with different intentions. By identifying the motor control scheme underlying the unfulfilled actions in a simple physical model, we address the goal inference found in the psychological studies [13, 14].

2 Simulation

2.1 Rationale

In this study, we idealize and simplify the psychological experimental paradigm of goal inference reported by [13, 14]. Their goal inference experiments can be summarized by the two conditions, goal-achieved and/or goal-failed demonstration. In the goal-achieved demonstration (i.e., intentionally "throw" on the floor), the demonstration is successful: the demonstrator's behavior does meet his/her goal. In the goal-failed demonstration (i.e., accidentally "drop" on the floor), the demonstration is unsuccessful: the demonstrator's behavior does not meet his/her goal, but the resulting movements of the unsuccessful demonstration (accidentally "drop") is quite similar to those of the successful demonstration (intentionally "throw"). At a coarse-grain observation, the movements of the goal-achieved and goal-failed demonstration look similar (due to the experimenters' careful acts in front of children [13, 14]), but the intended goals are actually different.

To capture difference in intended goals and similarity in movements of physical body, our simulation has two agents, goal-achieved (GA) and goal-failed (GF) agent; each agent goes through two phases, a learning and demonstration phase. In the learning phase, each agent forms the motor control scheme by learning to meet a goal given its physical body structure. In the demonstration phase, each agent shows movements according to the learned motor control scheme. In our simulation, the types of agents, goal-achieved and goal-failed, are defined by the consistency of the goals between the learning and demonstration phase. We set two different goals in the learning phase. The GA agent demonstrates movements under the same condition as the agent learned his/her motor control. Thus, the GA agent's movements in the demonstration phase are supposed to be the best or closely-optimized to be intended for the given goal in the learning phase. In contrast, the GF agent demonstrates movements under a different condition from the one the agent learned his/her motor control. Thus, the GF

agent's movements in the demonstration phase are supposed to be not the best or sub-optimal that count as failure or unintended movements. Due to the same constraint for both GA and GF agents, they show apparently similar movements with different intentions in the demonstration phase.

How can we discriminate the GA and GF agents, who both show similar movements, by only observing their movements in the demonstration phase? Specifically, we choose the pendulum swing-up task, which is one of the simplest motor control task and has been analyzed in the past studies. The pendulum model has the control of one degree-of-freedom and two dimensional state space, that is a minimally sufficient setting we can use to answer our questions. In the following two sections, we briefly introduce the basic physical model of the pendulum, and next describe the learning/demonstration phases and the GA/GF agents in this framework.

2.2 Pendulum swing-up task

Figure 1 (left) depicts a mathematical model of simple pendulum. A simple pendulum is composed of the rod of length $l = 1.0$ and the ball of mass $m = 1.0$. A state of this pendulum at an instant of time is determined by the rod angle θ and velocity $\dot{\theta}$. The equation of motion [5] is given by

$$ml^2\ddot{\theta} - mgl\sin\theta = -b\dot{\theta} + u \tag{1}$$

where $g = 9.8$ is the gravity constant, $b = 0.01$ is dumping force (torque), and $u \in [-5, 5]$ is control input (torque) from a control scheme.

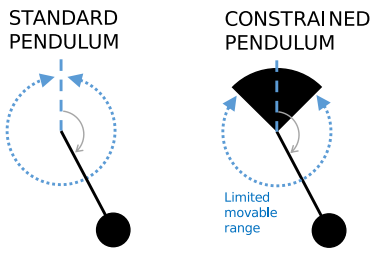

Fig. 1. The standard (left) and constrained pendulum (right).

The pendulum swing-up task [11, 5] is originally to design a control scheme that can swing the pendulum up and hold it at the inverted position given by $\theta = 0$. We will introduce a modified version of this task in the next section.

In the learning phase, the learner (the demonstrator) is required to learn a control scheme from experience, without knowing the mechanics of the pendulum. The control scheme is learned using a reinforcement learning technique, which is described in-depth in Appendix A. A control scheme for this task is defined by the function $u = g(\theta, \dot{\theta})$, which outputs torque u for a given state

$(\theta, \dot{\theta})$. The task for the agent in the learning is to construct the function g, which maximize the reward $\sum_t \cos(\theta_t)$ under the condition that the control input is restricted by $u \in \pm 5$, which prevents the agent from getting the maximum reward easily. For each trial in the learning phase, the initial position of the pendulum is set to randomly within $\theta \in \pm[\pi/8, \pi)$ and $\dot{\theta} = 0$.

A characteristic of the simple pendulum is the mechanistic energy that is the sum of the kinetic energy and the potential energy:

$$E(\theta, \dot{\theta}) = \frac{1}{2} ml^2 \dot{\theta}^2 + mgl(\cos\theta - 1) \tag{2}$$

For the pendulum swing-up task, the goal state, holding about the inverted position given by $\theta = 0$ and $\dot{\theta} = 0$, is characterized by $E(\theta, \dot{\theta}) = E(0, 0) = 0$. Thus, if the learner knows the pendulum model of his/her task, the control problem is reduced to adjust the energy E at the moment to be closer to zero for any $(\theta, \dot{\theta})$ [1]. In the simulated learning, instead of applying this knowledge directly, the agent interatively update their control scheme g by maximizing the reward $\sum_t \cos(\theta_t)$, which is specifically implemented by the reinforcement learning framework.

2.3 Goal-achieved and Goal-failed agent

To construct the goal-achieved and goal-failed agent, we consider two different conditions for the learning phase and one condition for the demonstration phase, which are summarized in Table 1 and depicted in Figure 2. Besides the standard pendulum introduced in the previous section, we additionally introduce the *constrained* pendulum, Figure 1 (right). The constrained pendulum is the exactly same as the standard pendulum except for the "wall" (see Figure 1 (right)), that is, its limitation in the movable angular range ($\theta \in \pm\pi/8$) to which the agent cannot move the pendulum. In addition, the constrained pendulum is limited to have $u = 0$ in the range $\theta \in \pm\pi/6$.

Table 1. Correspondence to the goal-inference experiment in psychology

	Goal-achieved	Goal-failed
Reward	$\sum_t \cos\theta_t$	$\sum_t \cos\theta_t$
Pendulum in Learning	Constrained	Standard
Pendulum in Demonstration	Constrained	Constrained
Human experiment	Intended action	Unfulfilled action

By introducing the constrained pendulum, we made up a clear dissociation between the learning and demonstration phase of the GF agent, and design both GA and GF agents move similarly with different their own control schemes.

Specifically, the GA agent learns the motor control scheme with the constrained pendulum, and demonstrates movements with it (the left panel of Figure 2). The GF agent learns the motor control scheme with the standard pendulum, but demonstrates movements with the constrained pendulum (the right

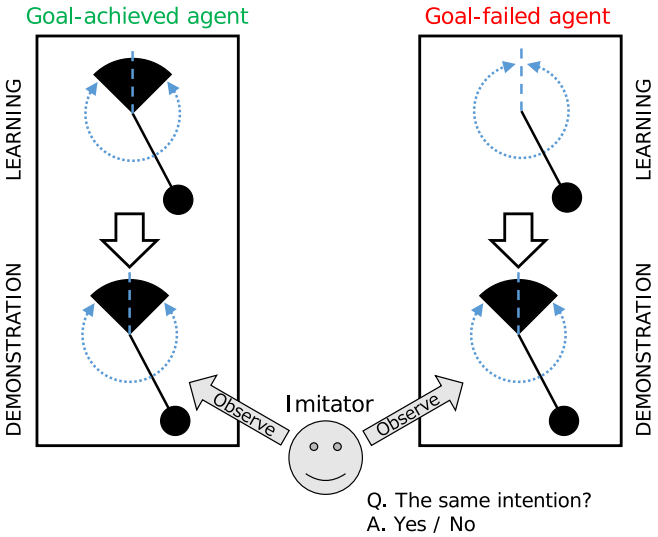

Fig. 2. Simulation design: the goal-achieved (GA) agent learns and demonstrates with the constrained pendulum (left), the goal-failed (GF) agent learns with the standard pendulum but demonstrates with the constrained one (right).

panel of Figure 2). Both GA and GF agent learn to maximize the same goal indicated by the reward $\sum_t \cos\theta_t$, but they control the different pendulums. Thus, their motor control schemes are different – the GF agent tries to swing the standard pendulum up to the top most position, which is allowed in his/her learning phase, but the GA agent tries to swing the constrained pendulum up to the top-most of the feasible region (Figure 1). In the demonstration with the constrained pendulum, the GF agent cannot reach the top-most position which was reachable in only his/her learning phase, while the GA agent can reach the top most position reachable as well in his/her learning phase. We treat the consistency in the GA agent and inconsistency in the GF agent in their learned motor control schemes and physical body in the demonstration phase as "intended" and "unfulfilled" action in the human experiment reported by [13, 14] (i.e., intentionally "throw" and accidentally "drop").

2.4 Reinforcement learning

To construct motor control scheme g for each agent, we used a reinforcement learning framework. Reinforcement learning [12] is a framework rooted in behavioral psychology and control theory. The key idea is that in the task environment in state s, the learner takes an action a, and next, the learner observes the environment in a new state s' and receives a reward r from the environment. The learner continues the task from the new state s'. The goal of learning is to acquire a control scheme $g(a|s)$ that maximizes the cumulative reward. See Appendix A for details.

For this pendulum swing-up task, the state transition of the environment is given by the motion of the pendulum. The learner can partially modify the future movements of the pendulum by supplying control inputs (i.e., actions taken by the learner). The reward function for this task must characterize the angular error from the inverted position. From [5], we adopted $\cos(\theta)$, which gives the highest reward $\cos(\theta) = 1$ at the inverted position $\theta = 0$ and the lowest reward $\cos(\theta) = -1$ at the hanging-down position $\theta = \pm\pi$. The pendulum swing-up task is solved successfully after 5000 trials, each consists of 10000 time steps (about 100 seconds).

3 Typical pendulum movements

We generated datasets for the imitator's task (the demonstration phase in Figure 2) using motor controls learned in the way of Section 2.4 (the learning phase in Figure 2). We systematically sampled thirty different initial conditions from the ranges $\theta = \pm[\pi/8, \pi)$ with the same interval. Below we first show the typical movements of the standard pendulum and constrained pendulum controlled by the GA and GF agents. In the next section, we analyze the movements (i.e., datasets for the imitator) in place of the imitator to answer the questions.

3.1 Standard pendulum

A typical successful movement of the standard pendulum controlled was started from a random initial position, and the position of the pendulum was eventually reached to the inverted position $\theta = 0$ (or $\theta = 2\pi$) by gradually increasing the amplitude. The mechanical energy of the pendulum increases together with its amplitude, until it got charged a sufficient energy to swing up to the inverted position. Since no constraints here, the agent learned the control scheme with the standard pendulum can swing up the standard pendulum.

3.2 Constrained pendulums

Figure 3 show typical movements of the constrained pendulums by the goal-achieved (GA) and goal-failed (GF) agent, respectively. The initial condition is the same for both the cases in the figure. Unlike the standard pendulum, the constrained pendulum cannot be swung up to the inverted position, and the mechanical energy cannot be exactly zero. For the reason of the constraints added to the pendulum, the mechanical energy reduces near the limits of the movable range. Recall that the task of the imitator is to decide whether or not the two movements exposed were generated by the same control scheme. From Figure 3, it seems that it is uneasy to differentiate the GF agent from the GA agent behind the movements observable.

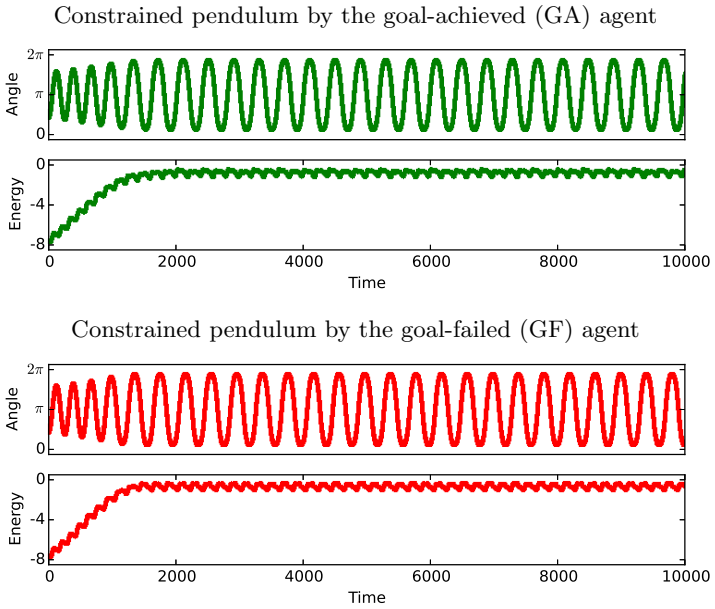

Fig. 3. Constrained pendulum swing-up movements from the same initial condition. Each figure contains the pendulum angle $\theta \in [0, 2\pi)$ and the mechanical energy. Note that $\theta = 0$ or $\theta = 2\pi$ is the inverted position.

4 Analysis: Discrimination of motor control schemes

Given the simulated movements of the two different agents, here we analyze them from two perspectives. One is an observer with little knowledge on the internal parameters of the pendulum, and the other is that with full knowledge on it. The former ignorant observer only accesses the angle and angular velocity $(\theta, \dot{\theta})$, but the latter knowledgeable observer knows the other physical parameters, such as mass, length, and friction of the pendulum, as well as energy of the system. A typical imitator, with respect to young children in the goal inference experiment [13, 14], is supposed an ignorant observer. Thus, to understand the mechanism to recognize intended or unfulfilled action, the GA and GF agents should be discriminable from the perspective of the ignorant observer with a minimal possible access to the actor-specific parameters.

To test whether the minimal accessible feature, the angle and angular velocity $(\theta, \dot{\theta})$, is sufficient to discriminate the type of agents, GA or GF, we performed classification analysis of the agent types using the features of the demonstrated movements. Here we consider angle, angular velocity, mechanical energy, frequency spectrum of the angle, and a type of fractal dimensions of the pendulum position as candidate features for the classification analysis. The mechanical energy requires knowledge of the physical model of the pendulum, and thus it is accessible only by the knowledgeable observer.

4.1 Hypothesis

Our working hypothesis is that intention behind the movements would result in the structural complexity of the motor-controlling system as a whole. In our simulation, the task in the learning phase is defined by the reward and the physical constraints on the pendulum. The pair of the reward (objective function) and the physical constraints (the domain or state space to find the maximum of the objective function) together forms the motor control scheme through the reinforcement learning. Thus, the motor control formed as the result of reinforcement learning is supposed to reduce unnecessary movements according to the task. In the other words, the motor control scheme formed for the standard pendulum is generally not the best choice for the constrained pendulum to maximize the reward. This sub-optimality in the motor control scheme is expected to increase unnecessary movements which does not directly gain the reward.

Specifically, let us consider the GA agent in our simulation. The GA agent is readily given the wall (inadmissible region in the state space) in the learning phase, and thus it should avoid hitting the wall but try to stay longer in time near the wall, at which the agent gains the largest reward value. Namely, the ideal learning leads a pendulum movement free from the wall (as it never hits the wall). Thus, we expect that the dynamical system of the GA agent can be analyzed by the pendulum movements without consideration to the wall. On the other hand, the GF agent learns the control scheme with the standard pendulum. As the GF agent finds the best rewarding region of the standard pendulum in the inadmissible region of the constrained pendulum, its ideally learned motor control swings up through the potential wall in the constrained pendulum. In its demonstrating with the constrained pendulum, it is unavoidable for the GF agent to hit the wall (unintentionally), and the wall comes as one of key factor in the dynamical system of the GF constrained pendulum. Therefore, this observation naturally leads the idea that the dynamical system of the GA agent has lower dimension than that of the GF agent, as the GF dynamical system has the additional factor, the unexpected wall in the way swinging up, playing a substantial role forming the dynamics.

4.2 Pointwise dimension

To characterize this type of complexity in the pendulum movements, we analyze the attractor dimension of the movements treating it as dynamical systems. Specifically, we exploited a sort of fractal dimension called *pointwise dimension* for the classification analysis.

Here, we briefly introduce the basic nature of pointwise dimension. The pointwise dimension is a type of dimension, which is defined for a small open set or measure on it including a point in a given set (for the formal definition, see [15, 4]). It is invariant under arbitrary smooth transformation. As it is associated for each point, we can analyze distribution of pointwise dimension across points in a set by estimating the pointwise dimension for a set of data points. Informally speaking, pointwise dimension of a point characterize “degree of freedom”

around the point. With this nature of the pointwise dimension, we expect it to capture the differences in the motor control scheme reflected by some local differences in the pendulum movements. We have developed a statistical technique to estimate the pointwise dimension for a set of data points [7]. Applying this technique, each point in the dataset is assigned with a positive value of pointwise dimension, with which the topological nature around each point is characterized.

Figure 3 shows a representative time series of pendulum angle θ and the mechanical energy E for the GA and GF agent in the demonstration. Around 2000 time step, the movements of the pendulums reached to the limit of angular range $\theta \in \pm\pi/8$ with the maximum amplitude, and the mechanical energy reaches at its maximum. We call the time interval after the mechanical energy reaches at its maximum *stationary period*, and the time interval before it *transient period*.

Figure 4 shows the time series of pointwise dimension for the same dataset above. The pointwise dimension is estimated for the time-delay coordinate

$$(x_t, x_{t+1}, \ldots, x_{t+9}, y_t, y_{t+1}, \ldots, y_{t+9})$$

corresponding time series of the positions $(x_t, y_t) = (\sin\theta_t, \cos\theta_t)$.

In the transient period, the pointwise dimension of both GA and GF agent stay nearly constant (Figure 4). In the stationary period, the GF agent tended to show larger pointwise dimensions than those in the transient period. In contrast, the GA agent tended to show smaller pointwise dimensions that those in the transient period. As theoretically predicted in the previous section, this trends suggest that the GF pendulum would be described as a dynamical system with higher dimension than the GA one.

For more detailed inspection of the pointwise dimension, Figure 5 shows a representative set of the pointwise dimensions as a function of the vertical positions, $\cos\theta - 1$, for each of the GA and GF pendulum in the stationary period. Figure 5 shows that the GA and GF agent can be discriminable by the pointwise dimensions, especially in the time interval with the higher position. As the high-position period corresponds with the impact with the wall in the constrained pendulum, this result suggests that the differences between the GA and GF agent in the pointwise dimension reflect the dynamical irregularity involving with the wall.

Hitting the wall generally decreases the mechanical energy, and the agent needs to regain the energy for the next swing. Thus, this energy loss and regain is expected to form a periodic dynamics. To visualize this energy oscillation, we plot the mechanical energy $E = U + V$, which is the sum of the kinetic U and the potential energy V, difference between the kinetic and potential energy $\Delta E = U - V$, and the pointwise dimension for each of the GA and GF pendulum (Figure 6). This figure shows the limit cycle on the $(E, \Delta E)$ plane, and the peak of the pointwise dimension in this oscillation corresponds to the period of the higher mechanical energy E (hitting the wall). We found large changes in mechanical energy E when ΔE is close to zero and the pendulum 's hitting the wall, and these changes are coincident with the higher pointwise dimensions. This result means that the patterns in pointwise dimension capture the critical point in the energy dynamics of the pendulums of the GA and GF agent.

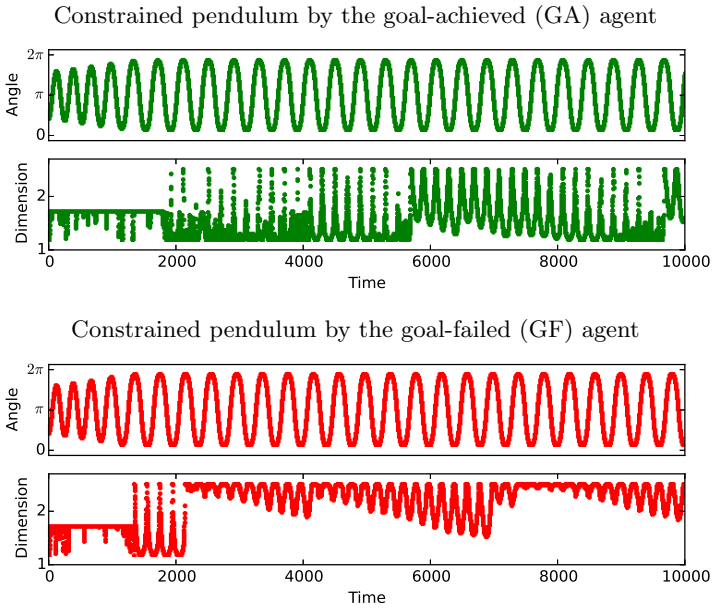

Fig. 4. Time series of pointwise dimensions in the demonstration phase.

4.3 Fourier analysis

Next, we analyzed the power spectrum of the angle time series as an alternative measure compared with the pointwise dimension. Figure 7 shows the power spectrums of the GA (green) and GF agent (red) corresponding to the angle time series shown in Figure 3. This figure shows that both have the peaks at the nearly same frequency (Hz) with similar magnitude. We also analyzed the other 29 pairs of GA and GF samples with different initial positions, and confirmed the similar trend with the trend in Figure 7. This result suggests that the types of agents, the GA or GF, would be difficult to discriminate with the power spectrum of the angle time series. In the next section, we tested this suggestion quantitatively (see the next Section 4.4).

4.4 Classification of agent types

Lastly, we performed classification analyses of agent types based on each of features, which can be derived from the pendulum movements. The performance in this classification analysis is considered as an indicator how informative for the imitator to identify the seemingly different movements with different initial positions and discriminate the GF agent from the GA agent, which can produce seemingly similar movements with the same initial positions.

Specifically, we performed a two-class classification of the types of agents, the GA or GF, for each data point using one of features, angle, angle velocity,

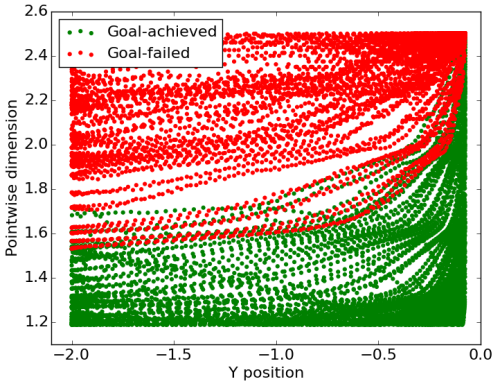

Fig. 5. Pointwise dimension as a function of the vertical position, $\cos\theta - 1$

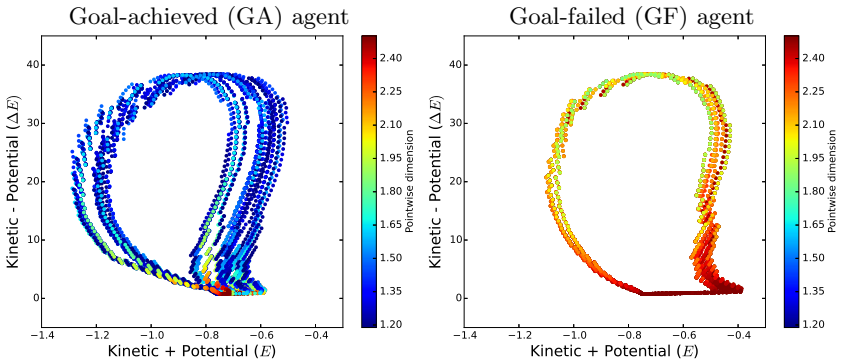

Fig. 6. Pointwise dimension (color map) as a function of $E = U + V$ and $\Delta E = U - V$, where U is the kinetic and V is the potential energy. The large change in mechanical energy E near $\Delta E = 0$ indicates hitting the wall.

mechanical energy, the maximum power of the short interval of time series, and pointwise dimension. Suppose that the imitator is given a training dataset, consisting of a subset of all 30 time series of one out of these features, in which each data point is labeled which type of agents, the GA or GF. As our goal is to numerically test whether the pointwise dimension has significantly high information to classify the two types of agents, we chose one of well known classifiers, the Gaussian mixture model. This choice was motivated by its simplicity in computation, rather than its classification performance.

Given the training data points, the imitator constructs sample probability functions of a variable x for the GA agent $p_{GA}(x)$ and for the GF agent $p_{GF}(x)$. Using the sample probability functions, the imitator asserts that a given new sample x is of the GF agent, if $p_{GF}(x) > p_{GA}(x)$, otherwise the imitator asserts that is of the GA agent in the test. The performance is evaluated by the correct

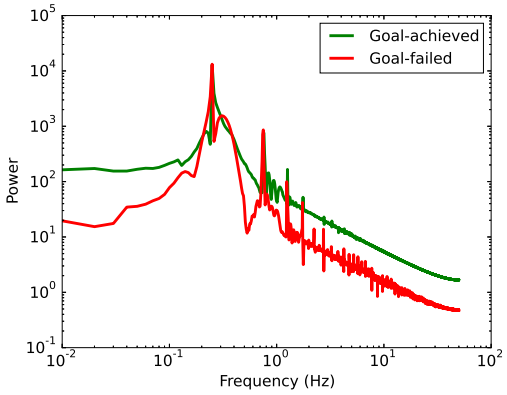

Fig. 7. Power spectrum of the angular time series in the GA and GF pendulum.

ratio of the imitator's prediction. Each of probability distributions $P_{GA}(x)$ and $P_{GF}(x)$ is represented by a Gaussian mixture model with the minimum Bayesian information criterion for one to 30 components. Gaussian mixture models represent a probability distribution of data points as a mixture of multiple Gaussian (normal) distributions. In this analysis, we found the optimal Gaussian mixture models each consists of 5 to 18 components for angle, 8 to 23 components for velocity, 5 to 19 components for power, 4 to 22 components for energy, and 4 to 11 components for pointwise dimension.

We include either or both transient or/and stationary phases in our training and test dataset, as the patterns in these phases are different qualitatively. We divided each time series, using the mechanical energy as an indicator of convergence, into the transient part (before convergence) $E(\theta, \dot{\theta}) < -1.3$ and stationary part (after convergence) $E(\theta, \dot{\theta}) \geq -1.3$.

The results of the classification analysis are summarized in Figure 8. As both training and test dataset consists of equally balanced number of sampled labeled GA and GF, the chance level of this binary classification was 50%. The performance of the classification with the angle, angle velocity, and power spectrum were at or barely above the chance level. The correct ratio of angle was 51.0% for the transient dataset and 51.2% for the stationary one. Similarly, that of angular velocity and the maximum power was at the chance level: about 50% for both datasets.

The classification of the mechanical energy was significantly higher than the chance level: 70.9% for the stationary data and 52.0% for the transient data. Note that, however, the mechanical energy is not accessible by the ignorant imitator, as calculating the mechanical energy requires the full knowledge of the physical model such as gravity constant or friction coefficient. Moreover, the mechanical energy is the direct measure of the task, as both the GA and GF agent aim to maximize the mechanical energy, which is identical to maximization of the reward in the learning phase. Thus, the classification performance of the

mechanical energy is supposed to give an upper bound, if the imitator have the complete knowledge on the pendulum model.

Lastly, the performances with the pointwise dimension are as good as these bounds of the mechanical energy : 63.9% for the stationary and 50.4% for the transient data. Note that this classification accesses externally observeable measures, as the pointwise dimension can be derived by only angle or position time series. Thus, an ignorant imitator can reach this level of performance, if he or she is asked to classify the types of agents using the pointwise dimension or other comparable measures. This implies that pointwise dimension could be a potential characteristics to identify and discriminate the motor control scheme underlying the observed movements, with little prior knowledge on the system of interest.

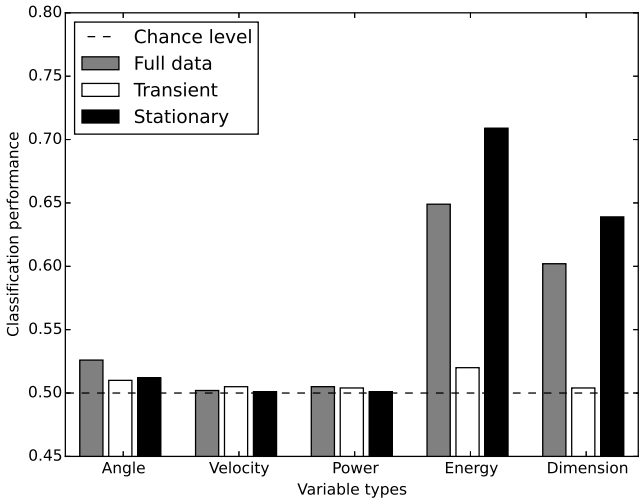

Fig. 8. Classification performance for each features

5 Discussion

In the present study, we explore a mechanistic account for intention inference from other's actions, which is crucial step toward understanding the goal-level imitation. In order to tackle this problem specifically, we consider a classical pendulum swing-up task with constraints as a idealization of the intention inference experimental paradigm by [13, 14]. By defining intention as motor control scheme in our computational framework, intention inference can be viewed as identification of the dynamical system with latent variables by its observeable variables. Our analysis on the two types of agents demonstrated that the latent motor control scheme can be identified with the observeable movements, if the

imitator computes pointwise dimension of the trajectories in state space of the pendulum.

In past literature, identification of the motor control scheme is supposed to require the prior knowledge on the class of systems to be estimated [9, 8]. As expected by this theory and confirmed in our analysis, the model-specific measure such as the mechanical energy has sufficient information on the motor control system. This approach, however, needs explicit modeling of the demonstrator, which is hard to be accessible by an ignorant imitator. In contrast, our approach based on pointwise dimension requires little prior knowledge on the system, but yet it performed as good as the alternative in identification of it. Thus it gives a possible mechanistic account for ignorant imitator to infer the imitation underlying other's actions.

Our next step on this research program is to build an autonomous imitator which can generates actions to meet the hidden goal identified by observing the other's actions. Empirical evidence for the use of pointwise dimension by human subjects is another line of our future research.

A Reinforcement learning

Reinforcement learning [12] is a framework rooted in behavioral psychology and control theory. In the task environment in state s, the learner takes an action a and receives a reward r from the environment in response to the action. Next, the learner faces with the environment in a new state $s' = Q(s'|s)$, where Q is a transition function. The goal of learning is to acquire a control scheme $g(a|s)$ that maximizes the cumulative reward.

The pendulum swing-up task is a classic control problem with continuous space and time [11, 5]. There are many researches to solve this task (e.g., [6] for recent updates). The simple and basic algorithm for this task is so-called actor-critic architecture [12, 6]. It is composed of two, the actor and critic components. The actor represents the control scheme $g(a|s)$. On the other hand, the critic represents the value function $V(s)$, that tells the learner the discounted expected reward of state s.

Since the task is in continuous space and time, it involves several engineering problems. The typical approach is discretization of the continuous space and time. For continuous time, we used discretized time steps for Eular integration (step size $\mathrm{d}t = 0.01$) and we sampled per 3 time steps. For continuous state space, we adopted a discretized representation (tile coding [12]) in which the continuous state space $(\theta, \dot{\theta}) \in [-\pi, \pi] \times [-2\pi, 2\pi]$ is equally divided into the grid of size 40×40 (one of them is called here state s).

Learning proceeds with estimation of state values, and that shapes $g(a|s)$ to navigate to more rewarding states. Suppose the pendulum is now in state s (one of the grid) at time t. The learner supplies a control input u sampled from $g(u|s) = N(\mu_s, \sigma_s)$, a normal distribution of mean μ_s and variance σ_s^2. In response, the learner observes the pendulum in a new state s' and a reward r.

Then the learner increments his value function $V(s)$ for every s by

$$\Delta V(s) = \alpha_c \left[r + \gamma_c V(s') - V(s) \right] E_c(s) \tag{3}$$

where $\alpha_c = 0.1$ is a learning rate, $\gamma_c = 0.97$ is a discount rate, and E_c is an eligibility trace with exponential decay (given by $\lambda_c = 0.65$) that is a device for continuous tasks that assigns higher weights for recently visited states.

For continuous control inputs, the control scheme $g(\cdot)$ is expressed by a collection of normal distributions for each state s. So the learner has to determine the mean μ and variance σ^2 for each state s. Formally, the learner modifies his control scheme μ_s and σ_s for every s by

$$\Delta \mu_s = \alpha_a \left[r + \gamma_a V(s') - V(s) \right] \frac{\partial N(\mu_s, \sigma_s)}{\partial \mu_s}(u) \, E_a(s) \tag{4}$$

$$\Delta \sigma_s = \alpha_a \left[r + \gamma_a V(s') - V(s) \right] \frac{\partial N(\mu_s, \sigma_s)}{\partial \sigma_s}(u) \, E_a(s) \tag{5}$$

where $\alpha_a = 0.001$ is a learning rate, $\gamma_a = 0.65$ is a discount rate, and E_a is an eligibility trace with decay (given by $\lambda_a = 0.0$).

The reward function $r = f(s, a)$ must be designed carefully. From [5], we set the reward function $r = \cos(\theta)$ for this task. It only depends on angle θ. Remark that $\cos(\theta)$ characterizes the goal of this task, because the inverted position $\theta = 0$ gives the highest reward $\cos(0) = 1$, and the hanging-down position $\theta = \pm\pi$ gives the lowest reward $\cos(\pi) = -1$.

Acknowledgment

This study is supported by the JSPS KAKENHI Grant-in-Aid for Young Scientists JP 16H05860.

References

1. Astrom, K.J., Furuta, K.: Swinging up a pendulum by energy control. Automatica 36(2), 287–295 (2000)
2. Bernstein, N.A.: Dexterity and Its Development. Psychology Press (1996)
3. Breazeal, C., Scassellati, B.: Robots that imitate humans. TRENDS in Cognitive Sciences 6(11), 481–487 (2002)
4. Cutler, C.D.: A review of the theory and estimation of fractal dimension, vol. 1, pp. 1–107. World Scientific (1993)
5. Doya, K.: Reinforcement learning in continuous time and space. Neural Computation 12, 243–269 (1999)
6. Grondman, I., Vaandrager, M., Busoniu, L., Babuska, R., Schuitema, E.: Efficient model learning methods for actor-critic control. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42(3) (2012)
7. Hidaka, S., Kashyap, N.: On the estimation of pointwise dimension. ArXiv:1312.2298 (https://arxiv.org/abs/1312.2298) (2013)
8. Kawato, M.: Computational theory of brain (in Japanese). Sangyo Tosho (1996)

9. Marr, D.: Vision. Cambridge: MIT Press (1982)
10. Ng, A., Russell, S.J.: Algorithms for inverse reinforcement learning. In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML 00). pp. 663–670 (2000)
11. Schaal, S.: Learning from demonstration. In: Mozer, M., Jordan, M., Petsche, T. (eds.) Advances in Neural Information Processing Systems 9, pp. 1040–1046. Cambridge: MIT Press (1997)
12. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press (1998)
13. Warneken, F., Tomasello, M.: Altruistic helping in human infants and young chimpanzees. Science 311, 1301–1303 (2006)
14. Warneken, F., Tomasello, M.: The roots of human altruism. British Journal of Psychology 100, 455–471 (2009)
15. Young, L.S.: Dimension, entropy, and lyapunov exponents. Ergodic Theory and Dynamical Systems 2(1), 109–124 (1982)