# Efficient Reuse of Previous Experiences to Improve Policies in Real Environment

Norikazu Sugimoto[1,3], Voot Tangkaratt[2], Thijs Wensveen[4], Tingting Zhao[2], Masashi Sugiyama[2], and Jun Morimoto*[1]

[1]Dept. of Brain Robot Interface, ATR Computational Neuroscience Labs
[2]Tokyo Institute of Technology
[3]National Institute of Information and Communications Techonology
[4]Delft University of Technology

September 1, 2018

**Abstract**

In this study, we show that a movement policy can be improved efficiently using the previous experiences of a real robot. Reinforcement Learning (RL) is becoming a popular approach to acquire a nonlinear optimal policy through trial and error. However, it is considered very difficult to apply RL to real robot control since it usually requires many learning trials. Such trials cannot be executed in real environments because unrealistic time is necessary and the real system's durability is limited. Therefore, in this study, instead of executing many learning trials, we propose to use a recently developed RL algorithm, importance-weighted PGPE, by which the robot can efficiently reuse previously sampled data to improve it's policy parameters. We apply importance-weighted PGPE to CB-i, our real humanoid robot, and show that it can learn a target reaching movement and a cart-pole swing up movement in a real environment without using any prior knowledge of the task or any carefully designed initial trajectory.

## 1 INTRODUCTION

Reinforcement Learning (RL) is becoming a popular approach to acquire a nonlinear optimal policy through trial and error. However, it is considered very difficult to apply RL to real robot control since it usually requires many learning trials. Such learning trials cannot be executed in real environments because unrealistic time is necessary and the real system's durability is limited. Therefore, previous studies used prior knowledge or properly designed initial trajectories to apply RL to a real robot so that the parameters of a robot controller can be improved within a realistic amount of time [1, 6, 7, 10].

However, since prior knowledge may not be always available, it is desirable to use a learning method that can improve policies only from limited experiences. In this study, we consider a recently developed RL algorithm: importanceweighted policy gradients with parameter based

*xmorimo@atr,jp


Figure 1: Humanoid robot CB-i [2] grabbing Wii controller.

exploration (IW-PGPE) [17]. With the IW-PGPE algorithm, we can efficiently use previously sampled data to improve policies. In other words, a robot can use its previous experiences to improve the current policy parameters. The usefulness of this approach has been thoroughly evaluated by comparing with previously proposed RL methods [18, 19, 20, 21] in numerical simulations. In this study, we evaluate how this RL approach using prevoius experiences can work efficiently in the real system.

We apply IW-PGPE to our real humanoid robot called CB-i [2] (see Fig. 1) and show that it can learn a target-reaching movement and a cart-pole swing-up movement in a real environment within 1.5 hours without using any prior knowledge of the task or any initial trajectory. For a target-reaching task, we used five degrees of freedom (DOF) composed of one-DOF torso joint, three-DOF shoulder joints, and one-DOF elbow joint of the humanoid robot. To use these five DOFs, the policy needs to be improved in ten-dimensional state space. In this moderately high-dimensional state space, it is usually considered that RL cannot be directly applied to real systems. However, we show that a target-reaching policy can be improved within a realistic amount of time by using IW-PGPE.

The rest of this paper is organized as follows. Policy update methods by using the standard policy gradient method and by using PGPE are explained in Section 2. The extension of PGPE to efficiently use previous experiences is introduced in Section 3. In Section 4, we evaluate our approach to improve policies in a real environment. First, we show the performances of policy updates by the IW-PGPE method with different parameters in a simulation environment. Then, we apply the learning method to our humanoid robot CB-i to improve a target-reaching policy. In section 5, we introduce a newly developed experimental setup in which our real humanoid robot can interact with a virtual environment through a Wii controller. In this

setup, a cart-pole swing-up policy is learned. Finally, Section 6 concludes the paper with discussion and following work in the future.

## 2 Policy Gradient Methods

In this section, we first review a standard formulation of policy gradient methods [15, 5, 7]. Then we show an alternative formulation adopted in the PGPE (policy gradients with parameter based exploration) method [9].

### 2.1 Standard Policy Update

We assume that the underlying control problem is a discrete-time MDP. At each discrete time step $t$, the agent observes a state $\boldsymbol{x}_t \in \mathcal{X}$, selects an action $u_t \in \mathcal{U}$, and then receives an immediate reward $r_t$ resulting from a state transition in the environment. The dynamics of the environment are characterized by $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, u_t)$, which represents the transition probability density from the current state $\boldsymbol{x}_t$ to the next state $\boldsymbol{x}_{t+1}$ when action $u_t$ is taken, and $p(\boldsymbol{x}_1)$ is the probability density of initial states. The immediate reward $r_t$ is given according to the reward function $r(\boldsymbol{x}_t, u_t, \boldsymbol{x}_{t+1})$.

The robot's decision making procedure at each time step $t$ is characterized by a parameterized policy $p(u_t|\boldsymbol{x}_t, \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$, which represents the conditional probability density of taking action $u_t$ in state $\boldsymbol{x}_t$. We assume that the policy is continuously differentiable with respect to its parameter $\boldsymbol{\theta}$.

A sequence of states and actions forms a *trajectory* denoted by

$$h := [\boldsymbol{x}_1, u_1, \ldots, \boldsymbol{x}_T, u_T],$$

where $T$ denotes the number of steps called horizon length. In this paper, we assume that $T$ is a fixed deterministic number. Note that the action $u_t$ is chosen independently of the trajectory given $\boldsymbol{x}_t$ and $\boldsymbol{\theta}$. Then the discounted cumulative reward along $h$, called the *return*, is given by

$$R(h) := \sum_{t=1}^{T} \gamma^{t-1} r(\boldsymbol{x}_t, u_t, \boldsymbol{x}_{t+1}),$$

where $\gamma \in [0, 1)$ is the discount factor for future rewards.

The goal is to optimize the policy parameter $\boldsymbol{\theta}$ so that the *expected return* is maximized. The expected return for policy parameter $\boldsymbol{\theta}$ is defined by

$$J(\boldsymbol{\theta}) := \int p(h|\boldsymbol{\theta}) R(h) \mathrm{d}h,$$

where

$$p(h|\boldsymbol{\theta}) = p(\boldsymbol{x}_1) \prod_{t=1}^{T} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, u_t) p(u_t|\boldsymbol{x}_t, \boldsymbol{\theta}).$$

The most straightforward way to update the policy parameter is to follow the gradient in policy parameter space using gradient ascent:

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \varepsilon \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}),$$

where $\varepsilon$ is a small positive constant, called the learning rate.

This is a standard formulation of policy gradient methods [15, 5, 7]. The central problem is to estimate the policy gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ accurately from trajectory samples.

## 2.2 PGPE Policy Update

However, standard policy gradient methods were shown to suffer from high variance in the gradient estimation due to randomness introduced by the stochastic policy model $p(a|\boldsymbol{s}, \boldsymbol{\theta})$ [16]. To cope with this problem, an alternative method called *policy gradients with parameter based exploration* (PGPE) was proposed recently [9]. The basic idea of PGPE is to use a deterministic policy and introduce stochasticity by drawing parameters from a prior distribution. More specifically, parameters are sampled from the prior distribution at the start of each trajectory, and thereafter the controller is deterministic[1]. Thanks to this per-trajectory formulation, the variance of gradient estimates in PGPE does not increase with respect to trajectory length $T$. Below, we review PGPE.

PGPE uses a deterministic policy with typically a linear architecture:

$$p(u|\boldsymbol{x}, \boldsymbol{\theta}) = \delta(u = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})), \tag{1}$$

where $\delta(\cdot)$ is the *Dirac delta function*, $\boldsymbol{\phi}(\boldsymbol{s})$ is an $\ell$-dimensional basis function vector, and $^\top$ denotes the transpose. The policy parameter $\boldsymbol{\theta}$ is drawn from a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\rho})$ with hyper-parameter $\boldsymbol{\rho}$.

The expected return in the PGPE formulation is defined in terms of expectations over both $h$ and $\boldsymbol{\theta}$ as a function of hyper-parameter $\boldsymbol{\rho}$:

$$\mathcal{J}(\boldsymbol{\rho}) := \iint p(h|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\rho}) R(h) \mathrm{d}h \mathrm{d}\boldsymbol{\theta}.$$

In PGPE, the hyper-parameter $\boldsymbol{\rho}$ is optimized so as to maximize $\mathcal{J}(\boldsymbol{\rho})$, i.e., the optimal hyper-parameter $\boldsymbol{\rho}^*$ is given by

$$\boldsymbol{\rho}^* := \arg\max_{\boldsymbol{\rho}} \mathcal{J}(\boldsymbol{\rho}).$$

In practice, a gradient method is used to find $\boldsymbol{\rho}^*$:

$$\boldsymbol{\rho} \longleftarrow \boldsymbol{\rho} + \varepsilon \nabla_{\boldsymbol{\rho}} \mathcal{J}(\boldsymbol{\rho}),$$

where $\nabla_{\boldsymbol{\rho}} \mathcal{J}(\boldsymbol{\rho})$ is the derivative of $\mathcal{J}$ with respect to $\boldsymbol{\rho}$:

$$\nabla_{\boldsymbol{\rho}} \mathcal{J}(\boldsymbol{\rho}) = \iint p(h|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\rho}) \nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho}) R(h) \mathrm{d}h \mathrm{d}\boldsymbol{\theta}. \tag{2}$$

Note that, in the derivation of the gradient, the logarithmic derivative,

$$\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho}) = \frac{\nabla_{\boldsymbol{\rho}} p(\boldsymbol{\theta}|\boldsymbol{\rho})}{p(\boldsymbol{\theta}|\boldsymbol{\rho})},$$

[1] Note that transitions are stochastic, and thus trajectories are also stochastic even though the policy is deterministic.

was used. The expectations over $h$ and $\boldsymbol{\theta}$ are approximated by the empirical averages:

$$\nabla_{\boldsymbol{\rho}} \widehat{\mathcal{J}}(\boldsymbol{\rho}) = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}_n | \boldsymbol{\rho}) R(h_n), \tag{3}$$

where each trajectory sample $h_n$ is drawn independently from $p(h|\boldsymbol{\theta}_n)$ and parameter $\boldsymbol{\theta}_n$ is drawn from $p(\boldsymbol{\theta}_n|\boldsymbol{\rho})$. We denote samples collected at the current iteration as

$$D = \{(\boldsymbol{\theta}_n, h_n)\}_{n=1}^{N}.$$

Following [9], in this paper we employ a Gaussian distribution as the distribution of the policy parameter $\boldsymbol{\theta}$ with the hyper-parameter $\boldsymbol{\rho}$. When assuming a Gaussian distribution, the hyper-parameter $\boldsymbol{\rho}$ consists of a set of means $\{\eta_i\}$ and standard deviations $\{\tau_i\}$, which determine the prior distribution for each element $\theta_i$ in $\boldsymbol{\theta}$ of the form

$$p(\theta_i|\rho_i) = \mathcal{N}(\theta_i|\eta_i, \tau_i^2),$$

where $\mathcal{N}(\theta_i|\eta_i, \tau_i^2)$ denotes the normal distribution with mean $\eta_i$ and variance $\tau_i^2$. Then the derivative of $\log p(\boldsymbol{\theta}|\boldsymbol{\rho})$ with respect to $\eta_i$ and $\tau_i$ are given as

$$\begin{aligned}
\nabla_{\eta_i} \log p(\boldsymbol{\theta}|\boldsymbol{\rho}) =& \frac{\theta_i - \eta_i}{\tau_i^2}, \\
\nabla_{\tau_i} \log p(\boldsymbol{\theta}|\boldsymbol{\rho}) =& \frac{(\theta_i - \eta_i)^2 - \tau_i^2}{\tau_i^3},
\end{aligned}$$

which can be substituted into Eq.(3) to approximate the gradients with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$. These gradients give the PGPE update rules.

An advantage of PGPE is its low variance of gradient estimates: Compared with a standard policy gradient method REINFORCE [15], PGPE was empirically demonstrated to be better in some settings [9, 16]. The variance of gradient estimates in PGPE can be further reduced by subtracting an optimal baseline.

# 3 Efficient Reuse of Previous Experiences

The original PGPE is categorized as an *on-policy* algorithm [12], where data drawn from the current target policy is used to estimate policy gradients. On the other hand, *off-policy* algorithms are more flexible in the sense that a data-collecting policy and the current target policy can be different. In this section, we introduce an *off-policy* algorithm for PGPE proposed by [17]. In this algorithm, importance-weighting is used so that we can reuse previously collected data (experience) in a consistent manner.

## 3.1 Importance-Weighted PGPE

Let us consider an off-policy scenario where a data-collecting policy and the current target policy are different in general. In the context of PGPE, we consider two hyper-parameters, $\boldsymbol{\rho}$ for the target policy to learn and $\boldsymbol{\rho}'$ for data collection. Let us denote data samples collected with hyper-parameter $\boldsymbol{\rho}'$ by $D'$:

$$D' = \{(\boldsymbol{\theta}'_n, h'_n)\}_{n=1}^{N'} \overset{i.i.d}{\sim} p(h, \boldsymbol{\theta}|\boldsymbol{\rho}') = p(h|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\rho}').$$

If we naively use data $D'$ to estimate policy gradients by Eq.(3), we have an inconsistency problem:

$$\frac{1}{N'}\sum_{n=1}^{N'} \nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}'_n|\boldsymbol{\rho}) R(h'_n) \overset{N'\to\infty}{\nrightarrow} \nabla_{\boldsymbol{\rho}} \mathcal{J}(\boldsymbol{\rho}).$$

*Importance sampling* [3] is a technique to systematically resolve this distribution mismatch problem. The basic idea of importance sampling is to weight samples drawn from a sampling distribution to match the target distribution, which gives a consistent gradient estimator:

$$\begin{aligned}\nabla_{\boldsymbol{\rho}} \widehat{\mathcal{J}}_{\mathrm{IW}}(\boldsymbol{\rho}) := \tfrac{1}{N'} \textstyle\sum_{n=1}^{N'} w(\boldsymbol{\theta}'_n) \nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}'_n|\boldsymbol{\rho}) R(h'_n) \\ \overset{N'\to\infty}{\longrightarrow} \nabla_{\boldsymbol{\rho}} \mathcal{J}(\boldsymbol{\rho}),\end{aligned} \tag{4}$$

where

$$w(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{\rho})}{p(\boldsymbol{\theta}|\boldsymbol{\rho}')}$$

is called the *importance weight*.

An intuition behind importance sampling is that if we know how "important" a sample drawn from the sampling distribution is in the target distribution, we can make adjustment by importance weighting. This extended method is called *importance-weighted PGPE* (IW-PGPE) [17].

### 3.2 Variance Reduction by Baseline Subtraction for IW-PGPE

To further reduce the variance of gradient estimates in IW-PGPE, we use variance reduction technique which uses a constant *baseline* [11, 14, 4, 13] as suggested in [17].

A policy gradient estimator with a baseline $b \in \mathbb{R}$ is defined as

$$\nabla_{\boldsymbol{\rho}} \widehat{\mathcal{J}}^b_{\mathrm{IW}}(\boldsymbol{\rho}) := \frac{1}{N'}\sum_{n=1}^{N'} (R(h'_n) - b) w(\boldsymbol{\theta}'_n) \nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}'_n|\boldsymbol{\rho}).$$

It is well known that $\nabla_{\boldsymbol{\rho}} \widehat{\mathcal{J}}^b_{\mathrm{IW}}(\boldsymbol{\rho})$ is still a consistent estimator of the true gradient for any constant $b$ [4]. Here, the constant baseline $b$ is determined so that the variance is minimized. Let $b^*$ be the optimal constant baseline for IW-PGPE that minimizes the variance:

$$b^* := \arg\min_b \mathbf{Var}[\nabla_{\boldsymbol{\rho}} \widehat{\mathcal{J}}^b_{\mathrm{IW}}(\boldsymbol{\rho})].$$

The optimal constant baseline for IW-PGPE is derived as [17]:

$$b^* = \frac{\mathbb{E}_{p(h,\boldsymbol{\theta}|\boldsymbol{\rho}')}[R(h) w^2(\boldsymbol{\theta}) \|\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho})\|^2]}{\mathbb{E}_{p(h,\boldsymbol{\theta}|\boldsymbol{\rho}')}[w^2(\boldsymbol{\theta}) \|\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho})\|^2]}.$$

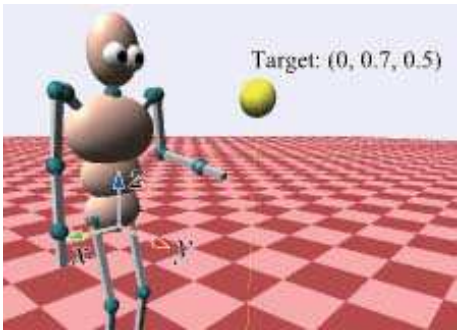


Figure 2: Humanoid robot simulator "SL" [8].

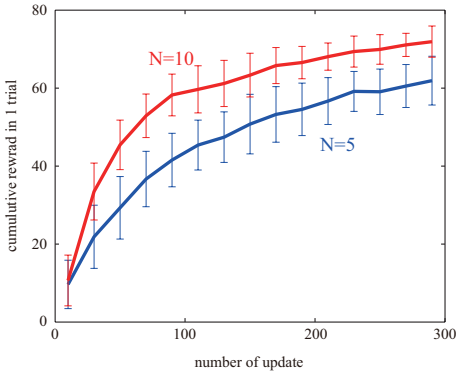


Figure 3: Learning curve of cumulative reward. The horizontal and vertical axes are number of update and cumulative reward. Blue and red lines represent the case that parameters were updated every five trials ($N = 5$) and ten trials ($N = 10$), respectively.

## 4 Experimental Results: Hand reaching task

We applied our proposed approach to a target-reaching task in simulated and real environments. Fig. 2 shows a setup of the target-reaching task. The end-effector is left-hand of the humanoid robot and the target is placed in front of the robot.

In this task, the robot controls 5 joints of the upper body, yaw joint of torso, three joints of left-shoulder, and joint of left-elbow, and learns policy to reach the target. The length of one trial was 1 sec.

The controller outputs a desired velocities of each joint:

$$\dot{\psi}_i^{\mathrm{des}}(t) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}(t)), \tag{5}$$

where $\psi_i^{\mathrm{des}}(i = \{1, 2, 3, 4, 5\})$ are the desired trajectory of controlled joint. In this paper, we employed linear basis function:

$$\boldsymbol{\phi}(\boldsymbol{x}(t)) = \begin{bmatrix} \psi_1(t) & \cdots & \psi_5(t) & \dot{\psi}_1(t) & \cdots & \dot{\psi}_5(t) & 1 \end{bmatrix}^\top, \tag{6}$$

where $\psi_i(i = \{1, 2, 3, 4, 5\})$ are the actual joint angle. The PD controller outputs the torque

command $\tau_i$ for each joint to track the desired trajectories,

$$\tau_i = -K_{\mathrm{P}}(\psi_i - \psi_i^{\mathrm{des}}) - K_{\mathrm{D}}(\dot{\psi}_i - \dot{\psi}_i^{\mathrm{des}}), \tag{7}$$

where $K_{\mathrm{P}}$ and $K_{\mathrm{D}}$ are positive constant.

We defined the objective function as sum of a state dependent reward $q(\boldsymbol{x}(t))$ and a cost of an action $c(\boldsymbol{x}(t), \boldsymbol{u}(t))$:

$$r(\boldsymbol{x}(t), \boldsymbol{u}(t)) = q(\boldsymbol{x}(t)) - c(\boldsymbol{x}(t), \boldsymbol{u}(t)). \tag{8}$$

A state dependent reward is given based on the error between the end-effector and targets:

$$q(\boldsymbol{x}(t)) = \exp\big[-\alpha \|\boldsymbol{p}_{\mathbf{E}}(t) - \boldsymbol{p}_{\mathbf{T}}\|^2\big], \tag{9}$$

where $\boldsymbol{p}_{\mathbf{E}}(t)$ and $\boldsymbol{p}_{\mathbf{T}} = (0.5, 0.7, 0)$ are the end-effector position and the target position. The origin is the center of torso joints. The parameter $\alpha(= 10)$ is constant. The cost of control is given based on the difference between actual angle and desired angle,

$$c(\boldsymbol{x}(t), \boldsymbol{u}(t)) = \beta \sum_{i=1}^{5} (\psi_i(t) - \psi_i^{\mathrm{des}}(t))^2, \tag{10}$$

where $\beta(= 0.0005)$ is constant. To maximize the future cumulative reward, we updated the parameter $\boldsymbol{\theta}$ with the discount factor $\gamma = 0.999$ and the learning rate $\varepsilon = 0.1$.

### 4.1 Simulation

We first apply the proposed approach to the simulated environment. We evaluate the learning performance of the proposed approach with the different settings of parameters of the learning algorithm, $N$ and $N'$, We tested proposed method using humanoid robot simulator "SL" [8] (See Fig.2). To evaluate average learning performances of the proposed approach with different parameter settings, we simulated five runs with different random seeds.

Figure 3 shows the result of learning. The horizontal and vertical axes are number of update and cumulative reward. Two lines (blue and red) represent the performance for $N = 5$ and $N' = 10$.

### 4.2 Real humanoid robot experiment

Finally, we implement proposed approach to real humanoid robot "CB-i" (See Fig. 4).

In this experiment, we updated the parameters $\boldsymbol{\theta}$ every five trial and used 10 trials for policy update ($N = 5$ and $N' = 10$). The experiment's conditions are same as simulation except for the hyper parameters $N$ and $N'$.

Figure 5 shows the learning curve of cumulative reward. The horizontal and vertical axes are number of update and cumulative reward. After 120 iterations, the end-effector reached target perfectly. Fig. 6 shows acquired behavior at 120th iteration.

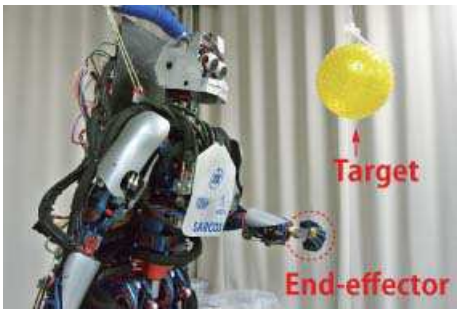


Figure 4: Humanoid robot "CB-i". The end-effector of reaching is right hand. The ball represent the target of reaching.

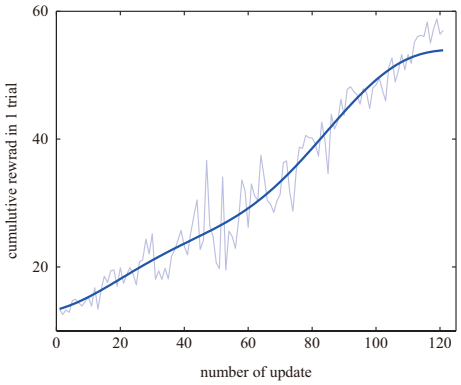


Figure 5: Learning curve of cumulative reward in the real humanoid robot experiment. The horizontal and vertical axes are number of update and cumulative reward.

## 5 Experimental Result: Cart-pole swing-up task

Next, we tested more challenging motor-control task using a virtual environment. As see Fig.7, we developed a virtual dynamics simulator and the robot interact with the virtual environment using a Wii controller. In this task, the robot controls a cart-pole to swing-up from hanging position by swinging a Wii controller. The dynamics of cart-pole is under actuated: the robot can apply a force to the cart only.

In this task, the robot also controls five joints of the upper body. The cart-pole has two joints: horizontal position of the cart and angle of the pole. The length of one trial was 1 sec.

We also employed linear basis function:

$$\boldsymbol{\phi}(\boldsymbol{x}(t)) = \left[\boldsymbol{s}(t)^\top \quad \dot{\boldsymbol{s}}(t)^\top \quad 1\right]^\top, \tag{11}$$

$$\boldsymbol{s} = [\psi_1(t) \quad \cdots \quad \psi_5(t) \quad z(t) \quad \theta(t)]^\top, \tag{12}$$

where $z$ and $\theta$ are horizontal position of the cart and angle of the pole.

The controlling force for the cart is generated based on an angular velocity of wii-controller:

$$F(t) = -k(\dot{z}(t) - \alpha\omega), \tag{13}$$

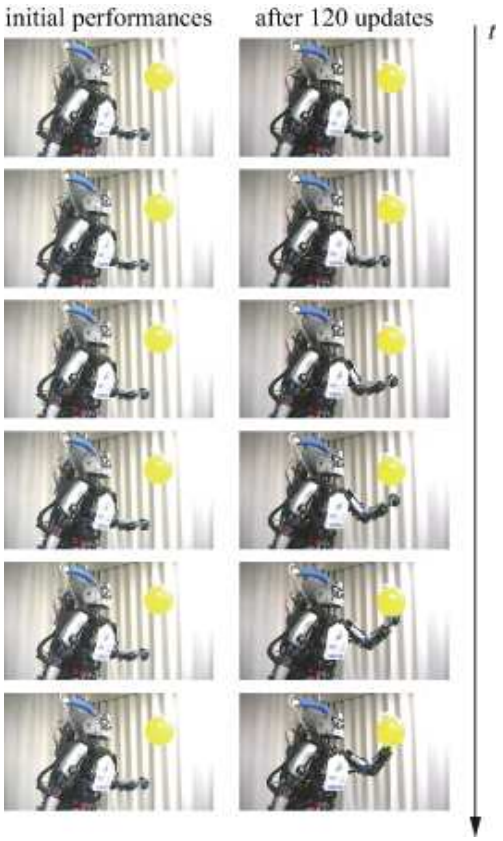


Figure 6: Acquired behavior of humanoid robot.

where $k = 100$ and $\alpha = 5$ are constant parameters.

We also defined the objective function as sum of a state dependent reward $q(\boldsymbol{x}(t))$ and a cost of an action $c(\boldsymbol{x}(t), \boldsymbol{u}(t))$. A state dependent reward is given based on the angle of the pole:

$$q(\boldsymbol{x}(t)) = \exp\big[-\alpha(z(t)^2 + \theta(t)^2)\big], \tag{14}$$

where the parameter $\alpha = 1$ is a constant. When the position of the cart is center and the angle of pole is upright position ($z = 0$, $\theta = 0$), the state dependent reward becomes maximum value. The cost of control is given based on the difference between actual angle and desired angle,

$$c(\boldsymbol{x}(t), \boldsymbol{u}(t)) = \beta \sum_{i=1}^{5} (\psi_i(t) - \psi_i^{\mathrm{des}}(t))^2, \tag{15}$$

where $\beta(= 0.0005)$ is a constant. To maximize the future cumulative reward, we updated the parameter $\boldsymbol{\theta}$ with the discount factor $\gamma = 0.999$ and the learning rate $\varepsilon = 0.1$. The parameters of policy were updated every five trials ($N = 5$) with sample size $N' = 10$. Note that we only consider swing-up movement and do not consider stabilizing the pole at upright position because a Wii controller is not accurate enough to do the pole stabilizing.

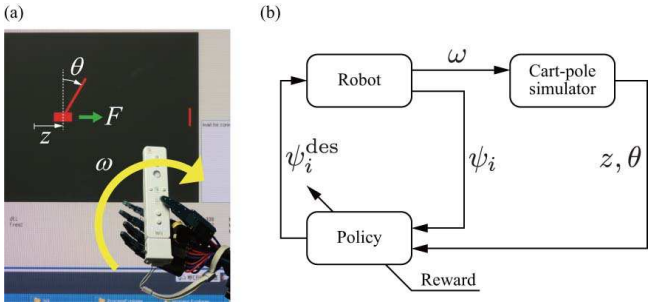


Figure 7: (a) Robot controls a cart-pole which is simulated in the virtual environment to swing-up from hanging position. The angular velocity $\omega$ is converted to the driving force of the cart. (b) A diagram of experimental system. The desired trajectory of controlled joint ($\psi_i^{\mathrm{des}}$) is given to the robot, then the angular velocity of wii-controller is given to the cart-pole simulator. The actual trajectory of the robot and the cart-pole ($\psi_i$, $z$, $\theta$) is feed-backed to the policy. The policy is optimized to maximize the cumulative reward.

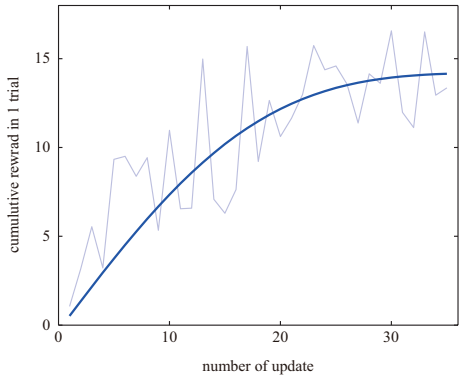


Figure 8: Learning curve of cumulative reward in the cart-pole swing-up task using the real humanoid robot.

### 5.1 Real humanoid robot experiment

Figure 8 shows the learning result of the experiment. The horizontal and vertical axes are the number of iteration and the mean cumulative reward in each iteration respectively. The performance was reached maximum value around $20^{\mathrm{th}}$ iteration. The learning speed was faster than the case of reaching task. It was because that the parameter of reward function ($\alpha$) was different. Fig. 9 shows the initial and acquired swing-up movements of the cart-pole at 30th iteration.

## 6 CONCLUSIONS

In this study, we show that the target-reaching policies can be efficiently acquired by using the previous experiences of the robot in the real environment. To improve the target-reaching policy, we used recently proposed IW-PGPE algorithm [17]. We also evaluated the learning

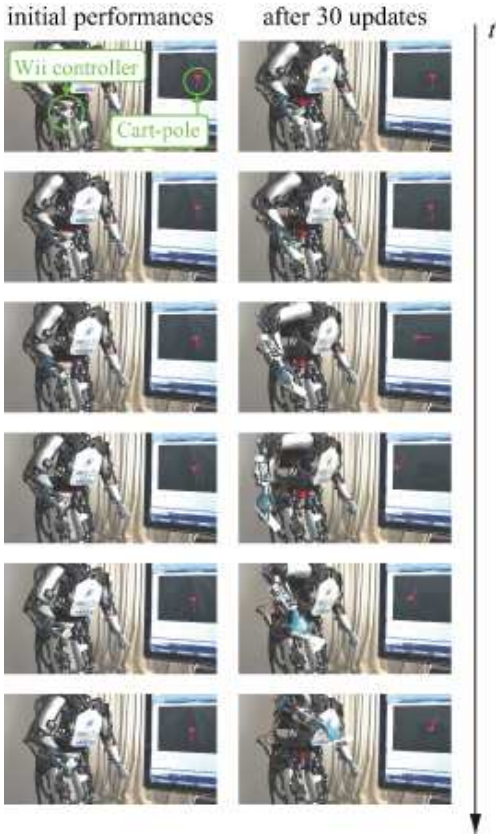


Figure 9: Acquired behavior of humanoid robot in the cart-pole swing-up task.

performance with different parameter settings. Moreover, we introduced newly developed real-virtual hybrid experimental setup and showed that the real humanoid were able to swing up the virtual pole by using Wii controller. As a future study, we will consider to develop a method to find the appropriate number of previously collected data that is used to improve the current policy.

## ACKNOWLEDGMENT

This research was supported by MEXT KAKENHI 23120004.

## References

[1] C. G. Atkeson and S. Schaal. How Can A Robot Learn From Watching A Human? In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 12–20. Morgan Kaufmann, San Francisco, CA, 1997.

[2] G. Cheng, S. Hyon, J. Morimoto, A. Ude, G.H. Joshua, Glenn Colvin, Wayco Scroggin, and C. J. Stephen. Cb: A humanoid research platform for exploring neuroscience. *Advanced Robotics*, 21(10):1097–1114, 2007.

[3] G. S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag, Berlin, Germany, 1996.

[4] E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

[5] S. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems 14*, pages 1531–1538, 2002.

[6] T. Matsubara, J. Morimoto, J. Nakanishi, M. Sato, and K. Doya. Learning CPG-based Biped Locomotion with a Policy Gradient Method. *Robotics and Autonomous Systems*, 54(11):911–920, 2006.

[7] J. Peters and S. Schaal. Policy gradient methods for robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2219–2225, 2006.

[8] Stefan Schaal. The SL simulation and real-time control software package. Technical report, 2009.

[9] F. Sehnke, C. Osendorfer, T. Rückstiess, A. Graves, J. Peters, and J. Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.

[10] N Sugimoto and J Morimoto. Phase-dependent trajectory optimization for cpg-based biped walking using path integral reinforcement learning. In *Proceedings of the IEEE/RAS International Conference on Humaniod Robos*, pages 255–260, 2011.

[11] R. Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts, 1984.

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

[13] L. Weaver and N. Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Processings of the Seventeeth Conference on Uncertainty in Artificial Intelligence*, pages 538–545, 2001.

[14] R. J. Williams. Toward a theory of reinforcement-learning connectionist systems. Technical Report NU-CCS-88-3, College of Computer Science, Northeastern University, Boston, MA, 1988.

[15] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[16] T. Zhao, H. Hachiya, G. Niu, and M. Sugiyama. Analysis and improvement of policy gradient estimation. *Neural Networks*, 26:118–129, 2012.

[17] T. Zhao, H. Hachiya, V. Tangkaratt, J. Morimoto, and M. Sugiyama. Efficient sample reuse in policy gradients with parameter-based exploration. *Neural Computation*, 25:1512–1547, 2013.

[18] J. Peters, S. Schaal. Natural actor-critic. *Neurocomputing*, 71(7 9):1180–1190, 2008.

[19] H. Hachiya, J. Peters, M. Sugiyama. Reward weight regression with sample reuse for direct policy search in reinforcement learning. *Neural Computation*, 23(11):2798–2832, 2011.

[20] R.J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[21] F. Sehnke, C. Osendorfer, T. R̈uckstiess, A. Graves, J. Peters, J. Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.