Reinforcement Learning with Adaptive Curriculum Dynamics Randomization for Fault-Tolerant Robot Control

Wataru Okamoto
Graduate School of Science and Engineering,
Chiba University
Chiba, Japan

Hiroshi Kera
Graduate School of Engineering,
Chiba University
Chiba, Japan

e-mail: kawa@faculty.chiba-u.jp

Kazuhiko Kawamoto
Graduate School of Engineering,
Chiba University
Chiba, Japan

Abstract—This study is aimed at addressing the problem of fault tolerance of quadruped robots to actuator failure, which is critical for robots operating in remote or extreme environments. In particular, an adaptive curriculum reinforcement learning algorithm with dynamics randomization (ACDR) is established. The ACDR algorithm can adaptively train a quadruped robot in random actuator failure conditions and formulate a single robust policy for fault-tolerant robot control. It is noted that the hard2easy curriculum is more effective than the easy2hard curriculum for quadruped robot locomotion. The ACDR algorithm can be used to build a robot system that does not require additional modules for detecting actuator failures and switching policies. Experimental results show that the ACDR algorithm outperforms conventional algorithms in terms of the average reward and walking distance.

I. INTRODUCTION

Robots are being increasingly applied in a wide range of fields such as factory automation and disaster assessment and rescue. In the existing approaches, the control law for robots are manually defined. However, in the presence of significant uncertainties in operating environments, the control law and states to be considered may not be easily identifiable, rendering such manual description challenging. To address these issues, reinforcement learning has attracted attention as a promising method that can automatically establish the control law without a clear description of how to realize a task of interest. A reinforcement learning framework learns, by trial and error, a policy that maximizes the reward, which is a measure of the goodness of the behavior in a certain environment. To realize reinforcement learning, it is necessary to only specify a task to be performed and design an appropriate reward function for the task.

Notably, to implement reinforcement learning in the real world, trial and error processes must be performed in real time. However, this approach may pose safety risks for the robot and surrounding environment. To ensure safety and cost effectiveness, robots are often trained in simulation environments before being applied in real world (Sim2Real). Reinforcement learning in simulation environments has achieved human-level performance in terms of the Atari benchmark [1] and robot control [2, 3]. However, Sim2Real is often ineffective owing to the gap between the simulation and real world, named reality gap, which can be attributed to the differences in physical quantities such as friction and inaccurate physical modeling. A straightforward approach to solve this problem is system identification, in which mathematical models or computer simulators of dynamical systems are established via measured data. However, the development of a realistic simulator is extremely costly. Even if system identification can be realized, many real-world phenomena associated with failure, wear, and fluid dynamics cannot be reproduced by such simulators. A typical approach to bridge the reality gap is domain randomization [4]. Domain randomization randomizes the parameters of the simulators to expose a model of interest to various environments. The model is expected to learn a policy that is effective in various environments, thereby narrowing the reality gap. Domain randomization has been used to successfully apply Sim2Real for the visual servo control of robot arms [4, 5] and locomotion control of quadruped robots [6].

Notably, reinforcement learning involves certain limitations. Figure 1 shows that even though a quadruped robot can walk owing to reinforcement learning (top row), the robot may turn over owing to the actuator failure of one leg (bottom row). In this context, fault-tolerant control is critical for robots working in remote or extreme environments such as disaster sites [7] because such robots can often not be repaired onsite. Several researchers have focused on fault-tolerant robot control based on reinforcement learning [8, 9]. However, these studies assume that the robots can detect internal failures and switch their policies according to the failures. Such robots require an additional module for diagnosing internal failures along with a module for control. In this context, robot systems designed for detecting all possible failures may be prohibitively complex [10]. To address this aspect, in this study, we establish a reinforcement learning algorithm for fault-tolerant control of quadruped robots against actuator failure. In this framework, the control law can ensure the locomotion of quadruped robots even if several actuators in the robot legs fail. Furthermore, we
develop a fault-tolerant control system that does not include any self-diagnostic modules for actuator failures. To this end, an adaptive curriculum dynamics randomization (ACDR) algorithm is proposed that can learn a single robust control policy against actuator failures by randomizing the dynamics of quadruped robots with actuator failures. In addition, we develop a curriculum learning algorithm that adaptively trains the quadruped robots to achieve an enhanced walking ability. It is noted that the hard2easy curriculum, in which the training is initiated in difficult conditions that progressively becomes easier, is more effective than an easy2hard curriculum, which is commonly used. Experiments are conducted to demonstrate the effectiveness of the ACDR algorithm in comparison with conventional algorithms.

The key contributions of this study can be summarized as follows:

- We establish a reinforcement learning algorithm to address the problem of fault tolerance of quadruped robots to actuator failures and realize fault-tolerant control.
- The ACDR algorithm to proposed effectively learn a single robust control policy against actuator failures. The ACDR algorithm does not require any self-diagnosis modules for detecting internal failures, which must be implemented in conventional systems.
- We demonstrate that the ACDR algorithms outperform conventional algorithms in terms of the average reward and walking distance.
- For fault-tolerance control of quadruped robots, we find that the hard2easy curriculum is effective than the easy2hard curriculum.

II. RELATED WORK

Domain randomization [4], [11] is aimed at enhancing generalization performance of the policy identified in simulations. Domain randomization requires a prescribed set of $N$ dimensional simulation parameters, $\xi \in \Xi \subset \mathbb{R}^N$, to randomize and the corresponding sampling interval. The $i$-th element of $\xi$ is sampled from the uniform distribution $\text{Unif}(L, U)$ on the closed interval $[L, U]$. Using a set of the parameters $\{\xi\}$, we can realize robot learning in various simulation environments. This randomization is expected to bridge the reality gap between the simulation and real world. Tobin et al. [4] successfully applied Sim2Real for visual robot control by randomizing the visual rendering of the environments. Moreover, Sim2Real has been successfully applied to control a robot arm [5] and quadruped robot [6] by randomizing physical parameters such as the robot mass and floor friction coefficient. In particular, we refer to the physical parameter randomization [5], [6] as dynamics randomization, thereby distinguishing it from visual domain randomization [4]. To implement dynamics randomization, the interval $[L, U]$ must be carefully determined, as an inappropriate interval is likely to result in failure to learn, a policy with considerable variance [13], or a conservative policy [14].

For reinforcement learning, it is effective to start with an easy task in order to succeed in a difficult task. Curriculum learning [15] is a learning process in which the difficulty of a task is gradually increased during training. Curriculum learning is mainly applied in navigation [16] and pickup tasks involving robot arms [17]. Luo et al. [18] proposed a precision-based continuous curriculum learning (PCCL) strategy that adjusts the required accuracy in a sequential manner during learning for a robot arm and can accelerate learning and increase the task accuracy. In general, it is difficult to determine the appropriate level of difficulty of a task in curriculum learning. Consequently, approaches that adaptively adjust the difficulty level for the agent through the learning process are attracting attention. Wang et al. [19] proposed a paired open-ended trailblazer (POET) strategy, which ensures an appropriate level of difficulty by co-evolving the environment and policies, and enables walking in difficult terrain. Moreover, curriculum learning in which the randomization interval is adaptively varied can help realize domain randomization [13], [20]. Xie et al. [20] considered a stepping-stone walking task with constrained scaffolding. The authors estimated the agent performance on each grid of the random parameters and adaptively varied the difficulty by decreasing and increasing the probability of occurrence of grids with high and low performance values, respectively. This algorithm can enhance the performance and robustness of the model. In general, the design of the environment and curriculum considerably influence the performance of learning to work [21]. To realize
natural walking behavior, it is necessary to limit the torque. To this
trend, a curriculum that starts with a higher value, such as
1.6×, in the early stages of learning and gradually reduces the
torque to the target value is considered to be effective [14].

To realize fault-tolerant robot control, Yang et al. [8]
considered the joint damage problem in manipulation and
walking tasks. The authors established a method to model
several possible fault states in advance and switch among
them during training. When the system encountered a failure,
the method increased the robustness against joint damage by
switching the policy. Kume et al. [9] proposed a method for
storing failure policies in a multidimensional discrete map in
advance. When the robot recognized a failure, it retrieved and
applied these policies, thereby exhibiting a certain adaptability
to robot failure. Notably, the existing failure-aware learning
methods assumed that the robot can recognize its damage state.
It remains challenging to develop a system that can achieve
high performance with a single policy regardless of whether
the robot is in a normal or failed state, even if the robot does
not recognize the exact damage state.

III. METHOD

To realize fault-tolerant robot control without additional
modules, we propose an ACDR algorithm for reinforcement
learning. The objective is to formulate a single robust policy
that can enable a robot to perform a walking task even if
actuator failure occurs. We assume that the quadruped robot
can adaptively change the interval of the random dynamics
parameter to efficiently train the robot.

A. Reinforcement Learning

We focus on the standard reinforcement learning prob-
lem [22], in which an agent interacts with its environment
according to a policy to maximize a reward. The state space
and action space are expressed as S and A, respectively. For
state \( s_t \in S \) and action \( a_t \in A \) at timestep \( t \), the reward
function \( r: S \times A \rightarrow \mathbb{R} \) provides a real number that represents
the desirability of performing a certain action in a certain state.
The goal of the agent is to maximize the multistep return
\( R_t = \sum_{t=0}^{T-1} \gamma^{t-1}r(s_{t},a_{t}) \) after \( T \) steps, where \( \gamma \in [0,1] \)
is the discount coefficient, which indicates the importance of
the future reward relative to the current reward. The agent decides
its action based on the policy \( \pi_{\theta}: S \times A \rightarrow [0,1] \). The policy
is usually modeled using a parameterized function with respect
to \( \theta \). Thus, the objective of learning is to identify the optimal
\( \theta^* \) as

\[
\theta^* = \arg \max_{\theta} J(\theta)
\]

where \( J(\theta) \) is the expected return defined as

\[
J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[ \sum_{t=0}^{T-1} r(s_{t},a_{t}) \right]
\]

where

\[
p_{\theta}(\tau) = p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_{t},a_{t})p_{\theta}(s_{t},a_{t})
\]

Algorithm 1: Adaptive curriculum dynamics randomization

Require: \( L, U \)

Require: \( D \): Performance data buffer

Require: \( m, g_{th} \): Threshold

Require: \( \Delta L, \Delta U \): Update step size

\( \theta \leftarrow \text{random weights} \)

\( D \leftarrow \phi \)

while not done do

\( k \sim \text{Uni}(L,U) \)

\( l \sim \text{Uni}\{0,1,2,3\} \) choose leg

\( o_{l,t} \leftarrow k_{o_{l,t-1}} \) update \( l \)-th actuator by \( k \)

Run policy \( \pi_{\theta} \) in environment with \( k \)

Optimize \( \theta \) with PPO

Generate rollout \( \tau = (s_0,a_0,\ldots,s_T) \)

\( g \leftarrow 0 \)

for \( s_t,a_t \) in \( \tau \) do

\( g \leftarrow g + r(s_t,a_t) \)

end for

\( D \leftarrow D \cup \{g\} \) \{Add performance to buffer\}

if \( |D| \geq m \) then

\( \bar{g} \leftarrow \text{Average of } D \)

if \( \bar{g} \geq g_{th} \) then

if curriculum is easy2hard then

\( U \leftarrow U - \Delta U \)

\( L \leftarrow L - \Delta L \)

else if curriculum is hard2easy then

\( U \leftarrow U + \Delta U \)

\( L \leftarrow L + \Delta L \)

end if

\( g_{th} \leftarrow \bar{g} \)

end if

\( D \leftarrow \phi \)

end if

end while

represents the joint probability distribution on the series of
states and actions \( \tau = (s_0,a_0,s_1,\ldots,a_{T-1},s_T) \) under policy
\( \pi_{\theta} \).

B. Definition of Actuator Failure

The actuator failure is represented through the variation in
the torque characteristics of the joint actuators. By denoting
\( o_{t} \in \mathbb{R} \) as the actuator output at timestep \( t \), we simply represent
the output value of the failure state, \( o'_{t} \), as follows:

\[
o'_{t} = k o_{t}.
\]

where \( k \in [0,1] \) is a failure coefficient. Failure coefficients
of \( k = 1.0 \) and \( k \neq 1.0 \) indicate that the actuator is in
the normal or failed state, respectively. A larger deviation of
\( k \) from 1.0 corresponds to a larger degree of robot failure,
and \( k = 0.0 \) means that the actuator cannot be controlled.
Figure 1 illustrates an example of failure of an 8-DOF robot
that had learned to walk. The top row in Fig. 1 shows a normal

Figure 1: Examples of failures of an 8-DOF robot.
robot with \( k = 1.0 \), and the bottom row in Fig. 1 shows a failed robot (red leg damaged) with \( k = 0.0 \). Because the conventional reinforcement learning does not take into account environmental changes owing to robot failures, the robot falls in the middle of the task and cannot continue walking, as shown in the bottom row in Fig. 1.

C. Adaptive Curriculum Dynamics Randomization

To implement fault-tolerant robot control the ACDR algorithm employs dynamics randomization by randomly sampling the failure coefficient \( k \) and curriculum learning by adaptively changing the interval of \( k \). The process flow of the ACDR algorithm is presented as Algorithm 1 and can be described as follows.

For dynamics randomization, one leg \( l \) of the quadruped robot is randomly chosen as

\[
l \sim \text{Uni}\{0, 1, 2, 3\}
\]

where \( \text{Uni}\{0, 1, 2, 3\} \) is the discrete uniform distribution on the possible leg index set \( \{0, 1, 2, 3\} \) of the quadruped robot. The leg \( l \) is broken at the beginning of each training episode, and the failure coefficient \( k \) is randomly chosen as

\[
k \sim \text{Uni}(L, U)
\]

where \( \text{Uni}(L, U) \) is the continuous uniform distribution on the interval \([L, U]\). According to Eq. 4, the broken actuator of the leg \( l \) outputs \( \tilde{a}_{l,t} = k o_{l,t} \) in each training episode. As the actuator output changes, the dynamics model \( p(s_{t+1}|s_t, a_t) \) also changes during training. This dynamics randomization means that a policy can be learned under varying dynamics models instead of being learned for a specific dynamics model. Under dynamics randomization, the expected return is determined by taking the expectation over \( k \), as indicated in Eq. 2, as

\[
\mathbb{E}_{k \sim \text{Uni}(L, U)} \left[ \mathbb{E}_{\tau \sim \text{p}(\tau|k)} \left[ \sum_{t=0}^{T-1} r(s_t, a_t) \right] \right] \rightarrow \text{max}\,.
\]

By maximizing this expected return, the agent can learn a policy to adapt to the dynamics changes owing to robot failures.

For curriculum learning, the interval \([L, U]\) in Eq. 6 is adaptively updated according to the return earned by the robot in training. As mentioned previously, a large interval \([L, U]\) is likely to result in a policy with large variance \([13]\) or a conservative policy \([14]\). Thus, a small interval \([L, U]\) is assigned to Algorithm 1 as the initial value. The interval \([L, U]\) is updated as follows: First, a trajectory of states and actions \( \tau = (s_0, a_0, \ldots, s_T) \) is generated based on a policy \( \pi_0 \), where \( \theta \) is the policy parameter optimized using a proximal policy optimization (PPO) algorithm \([23]\). Subsequently, the return \( g \) of the trajectory \( \tau \) is calculated by accumulating the rewards as

\[
g = \sum_{t=0}^{T-1} r(s_t, a_t).
\]

By repeating the parameter optimization and return calculation \( m \) times, the average value of the returns \( \bar{g} \) can be determined. This value is used to evaluate the performance of the robot and determine whether the interval \([L, U]\) must be updated. If the average \( \bar{g} \) exceeds a threshold \( g_{th} \), the interval \([L, U]\) is updated, i.e., the robot is trained with the updated failure coefficients.

In this study, we consider two methods to update the interval \([L, U]\). The manner in which the interval is updated during curriculum learning is illustrated in Fig. 2. In the first method, the upper and lower bounds of the interval \([L, U]\) gradually decrease as

\[
\begin{align*}
U & \leftarrow U - \Delta_U \\
V & \leftarrow V - \Delta_V
\end{align*}
\]

given the initial interval \([1.5, 1.5]\). In all experiments, we set \( \Delta_U = \Delta_V = 0.01 \). This decrease rule represents an easy2hard curriculum because smaller values of \( L \) and \( U \) correspond to a higher walking difficulty. For example, when \( k = 0 \), the leg does not move. The second method pertains to a hard2easy curriculum, i.e.,

\[
\begin{align*}
U & \leftarrow U + \Delta_U \\
V & \leftarrow V + \Delta_V
\end{align*}
\]

We set the initial interval as \([0.0, 0.0]\). In this state, the robot cannot move the broken leg, and its ability to address this failure is gradually enhanced. Notably, although the easy2hard curriculum is commonly used in curriculum learning, the hard2easy curriculum is more effective than the easy2hard configuration in accomplishing the task. The difference between the two frameworks is discussed in Section IV-D.

![Fig. 2. Time update of intervals \([L, U]\) of the failure coefficient \( k \) in curriculum learning. The intervals are represented by the colored regions. ACDR_e2h and ACDR_h2e represent easy2hard and hard2easy curricula for ACDR, respectively.](Image)

![Image 20x300 to 300x741]
D. Reward function

In the analysis, we introduce a slightly modified reward function instead of using the commonly applied function described in [24], because the existing reward function [24] is likely to result in a conservative policy in which the robot does not walk but remains in place. The existing reward function [24] can be defined by

\[ r = v_{\text{fwd}} - 10^{-6} \| u \|^2 - 10^{-3} \| f_{\text{impact}} \|^2 + s, \]

where \( s = \begin{cases} 0 & \text{if the robot is falling.} \\ 1 & \text{otherwise} \end{cases} \) Preliminary experiments demonstrate that the modified reward function in Eq. (12) can ensure more active walking of the robots.

IV. Experiments

We conduct experiments in the Ant-v2 environment provided by OpenAI Gym [25]. This environment runs on the simulation software MuJoCo [26], and the task is aimed at realizing rapid advancement of the quadruped robot, as shown in Fig. 1. To realize reinforcement learning, we use the PPO algorithm [23], which is suitable for high-dimensional continuous action environments.

A. Experiment setup

For reinforcement learning, we employ the actor and critic network consisting of two hidden layers of 64 and 64 neurons with tanh activation. All networks use the Adam optimizer [27]. The related hyperparameters are listed in Table I.

| Parameter          | Value     |
|--------------------|-----------|
| Learning rate      | 0.00022   |
| Horizon            | 128       |
| Minibatch size     | 4         |
| epochs             | 4         |
| Clipping parameter | 0.2       |
| Discount \( \gamma \) | 0.99     |
| GAE parameter \( \lambda \) | 0.95 |
| VF coefficient     | 0.5       |
| Entropy coefficient| 0.01      |

The ACDR algorithm is compared to the following three algorithms. Baseline: The baseline is a basic reinforcement learning algorithm for normal robots that implements neither dynamics randomization nor curriculum learning. Uniform Dynamics Randomization (UDR): UDR is a dynamics randomization algorithm that is based on the uniform distribution \( \text{Uni}(0.0, 1.5) \) and does not implement curriculum learning. Linear Curriculum Dynamics Randomization (LCDR): In the LCDR algorithm, the interval \([L, U]\) is linearly updated. The progress of the two curricula, the legends of which are "LCDR\_e2h" and "LCDR\_h2e", is shown in Fig. 2.

We evaluate the algorithms in terms of the average reward and average walking distance in two conditions: plain and broken. In the plain, the robots function normally, i.e., the failure coefficient is fixed at \( k = 1.0 \). In the broken, the robots are damaged, and \( k \) randomly takes a value in \([0.0, 0.5]\). We implement each algorithm on the two conditions, with five
random seeds for each framework. The average reward and average walking distance are calculated as the average of 100 trials for each seed. The average walking distance is one that is suitable for evaluating the walking ability of the robot.

B. Comparative evaluation with conventional algorithms

The average reward and average walking distance for all algorithms are shown in Fig. [3] and Fig. [4] respectively. In both the plain and broken conditions, UDR is inferior to the Baseline, which does not consider robot failures. This result indicates that the simple UDR cannot adapt to robot failures. The hard2easy curriculum of LCDR, denoted by LCDR_h2e, outperforms the Baseline in terms of the average reward, as shown in Fig. [3]. However, LCDR_h2e corresponds to an inferior walking ability compared to that of the Baseline for both plain and broken conditions, as shown in Fig. [4]. This result indicates that LCDR_h2e implements a conservative policy that does not promote walking as active as that generated by the Baseline. For both LCDR and ACDR, the hard2easy curriculum outperforms the easy2hard curriculum, as discussed in Section [IV-D]. Moreover, ACDR_h2e achieves the highest performance in all cases. This finding indicates that ACDR_h2e can avoid the formulation of a conservative policy by providing tasks with difficulty levels that match the progress of the robot.

Figure [3] shows the average reward for each failure coefficient \( k \) in the interval \([0, 1]\). ACDR_h2e earns the highest average reward for any \( k \), and ACDR_h2e achieves the highest performance in any failure state. Moreover, the performance of UDR, LCDR_e2h, and ACDR_e2h, deteriorate as \( k \) increases. This result indicates that the robots implementing these strategies cannot learn to walk. Therefore, even when \( k \) increases and it is easy to move the leg, the robots fall and do not earn rewards.

C. Comparative evaluation with non-curriculum learning

The experiment results show that the hard2easy curriculum is effective against robot failures. To evaluate whether the hard2easy curriculum contributes to an enhanced performance, we compare this framework with a non-curriculum algorithm in which the robot is trained with a fixed failure coefficient \( k \). The values of \( k \) are varied as \([0.0, 0.2, 0.4, 0.6, 0.8, 1.2, 1.4]\). For each \( k \), Fig. [6] shows the average reward earned by the robot on interval \([0, 1]\) of the failure coefficient. ACDR_h2e still achieves the highest performance over the interval \([0, 1]\). Therefore, the hard2easy curriculum can enhance the robustness against robot failures.

D. Comparative evaluation between easy2hard and hard2easy curricula

The hard2easy curriculum gradually decreases the degree of robot failure and makes it easier for the robot to walk. In contrast, the easy2hard curriculum gradually increases the degree of difficulty, and eventually, the leg stops moving. Figures [3] and [4] show that the hard2easy curriculum is more effective than the easy2hard curriculum. We discuss the reason below.

Figure [6] shows that the average reward tends to increase as \( k \) increases and deteriorates as \( k \) decreases. For example, \( k = 1.4 \) corresponds to the highest average reward over the interval \([0, 1]\), whereas \( k = 0.0 \) or \( k = 0.2 \) yield low rewards. This result suggests that it is preferable to avoid training at \( k = 0.0 \) to allow the robot to walk even under failure. Hence, we can hypothesize that the easy2hard curriculum, in which training is performed at \( k = 0.0 \) at the end of training, deteriorates the robot’s ability to walk. To verify this hypothesis, we perform the same comparison as described in Section [IV-B] for robots trained on the interval \([k, 1.5]\); in particular, in this framework, the robots are not trained at \( k = 0.0 \). Figures [7] and [8] show the average reward and average walking distance, respectively, and Fig. [9] shows the average reward for each \( k \).
Fig. 7. Average reward for the plain (blue) and broken (green) quadruped tasks. Error bars indicate standard error. The comparison algorithm is trained in the interval \( k \in [0.5, 1.5] \).

Fig. 8. Average progress for the plain (blue) and broken (green) quadruped tasks. Error bars indicate the standard error. The comparison algorithm is trained in the interval \( k \in [0.5, 1.5] \).

Fig. 9. Average reward of all algorithms across various failure coefficients \( k \in [0.5, 1.5] \).

V. CONCLUSION

This study was aimed at realizing fault-tolerant control of quadruped robots against actuator failures. To solve the problem, we propose a reinforcement learning algorithm with adaptive curriculum dynamics randomization, abbreviated as ACDR. We apply domain randomization to fault-tolerant robot control toward Sim2Real in robotics. Furthermore, we developed an adaptive curriculum learning framework to enhance the effectiveness of domain randomization. The ACDR algorithm could formulate a single robust policy to realize fault-tolerant robot control. Notably, the ACDR algorithm can facilitate the development of a robot system that does not require any self-diagnostic modules for detecting actuator failures.

Experiment results demonstrate that the combination of curriculum learning and dynamics randomization is more effective than non-curriculum learning or non-randomization of dynamics. In addition, the hard2easy curriculum is noted to be more effective than the easy2hard curriculum. This finding suggests that it is preferable to train quadruped robots in the order of difficulty opposite to that implemented in standard curriculum learning. In future work, there are several possible research directions of improving the ACDR algorithm. A combination of automatic curriculum design (e.g. [28]) and domain randomization can be a promising direction.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP19K12039.

REFERENCES

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. nature, 518(7540):529–533, 2015.

[2] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning, pages 1889–1897. PMLR, 2015.
Linear curriculum dynamics randomization

The LCDR updates the interval \([L, U]\) at the discrete timestep \(t = \Delta_t, 2\Delta_t, \ldots, (N-1)\Delta_t\), where \(\Delta_t\) is defined as

\[
\Delta_t = \lfloor T / N \rfloor,
\]

where \(T\) is the predetermined total learning time, and \(\lfloor \cdot \rfloor\) is the floor function. We set \(N = 11\) in the experiments. Similar to the ACDR, the LCDR implements two curricula: easy2hard and hard2easy. Given the initial interval \([1.5, 1.5]\), the easy2hard curriculum updates the interval as

\[
\begin{cases}
    L &\leftarrow L - \Delta t \\
    U &\leftarrow U - \Delta t
\end{cases}
\]

(15)

where the update amount \(\Delta\) is defined as

\[
\Delta = \frac{k_{\text{max}}}{N - 1},
\]

(16)

where \(k_{\text{max}} = 1.5\) is the maximum value of the failure coefficient. In addition, the hard2easy curriculum updates the interval as

\[
\begin{cases}
    L &\leftarrow L + \Delta t \\
    U &\leftarrow U + \Delta t
\end{cases}
\]

(17)

given the initial interval \([0.0, 0.0]\)