Intelligent Trainer for Model-Based Reinforcement Learning

Yuanlong Li, Member, IEEE, Linsen Dong, Student Member, IEEE, Yonggang Wen, Senior Member, IEEE, and Kyle Guan, Member, IEEE

Abstract—Model-based deep reinforcement learning (DRL) algorithm utilizes the cyber data sampled from a cyber mirror of the target physical system to accelerate the training process and reduce real sampling cost. As a potential solution to the high sampling cost problem caused by the large amount of data required by DRL, it is yet not applicable in practice due to issues such as the accuracy of the approximate model maybe not sufficient, the tuning process of sampling and training strategy from the real and cyber environment may incur high sampling cost. To address these issues, we propose an intelligent trainer framework to properly utilize the approximate model and do online tuning of the control parameters in the sampling and training procedure in model-based DRL. More specifically, we package the training process of a model-based DRL as a standard RL environment to decouple the onward optimization from the training algorithm of the target controller. Three control actions for this training process environment (TPE): the first action determines how many cyber data should be sampled and used to train the target controller; the second and third actions control where to sample in the cyber and real environment respectively. On top of the designed TPE, we develop various RL trainers to optimize the training inside TPE in an online manner, and develop an ensemble trainer that can train multiple target controllers without incurring additional sampling cost. The proposed framework is evaluated on five different tasks of OpenAI gym. Results show that the proposed trainer framework we can achieve better performance than a fixed parameter baseline algorithm in most cases. The ensemble trainer can perform almost as good as a manually optimized algorithm without incurring additional sampling cost. The proposed trainer framework can be extended to more control actions with more sophisticated trainer design.

Index Terms—Model-based reinforcement learning, AutoML, intelligent trainer, ensemble algorithm.

I. INTRODUCTION

Deep reinforcement learning (DRL), which combines reinforcement learning (RL) and deep neural networks (DNN), has demonstrated its prowess in solving for complex decision-making problem like Go [1]. A series of recent breakthroughs have also shown that DRL algorithm like Deep Deterministic Policy Gradients (DDPG) and Trust Region Policy Optimization (TRPO) can also work equally well with continuous control problems [2] [3]. While DRL has shown its capability and promise, the high training cost associated with DRL presents great challenges in practice. For example, in [4], the authors show that a DRL agent can score a goal with high probability after about one million trials during training. For practical control that relies on data directly sampled from real physical systems, this is hardly viable as the time and resource costs incurred during training can be prohibitively high, as noticed in [5]. “Huge amounts of noisy data, slow training and testing on real robots, the reality gap between simulators and the real world”.

How to bridge this gap remains one of the key challenges in applying DRL to real-world applications. Existing approaches to tackle this problem include the model-based reinforcement learning approach [6]. Model-based RL is meant to utilize the data sampled from the real-world system to train a system dynamics model, which in turn generates synthesized data for training controller (value/policy function) and search for good actions. Such approach has been applied to robot arm training [6], online tree search based planning [7] [8] [9]. These approaches can reduce the training cost because training in the real physical environment is usually much more expensive than that with a cyber environment that can emulate the real system.

Though model-based RL provides a viable approach to address the sampling cost problem associated with DRL, the following crucial issues have not been sufficiently investigated in current research and development efforts: 1) The effectiveness of the model-based approach depends on whether the model of underlying system dynamics can be learned faster than the corresponding value/policy function. As such, the benefit of cyber emulator could vary for different underlying physical systems. 2) In the existing model-based RL approaches, random sampling strategy with manually tuned hyperparameters is used, which makes it difficult for the algorithm to work in practice. Moreover, though these algorithms are claimed to be able to reduce the training cost, the tuning costs are often unaccounted for in the overall training process.

To solve these problems and develop a practically useful model-based DRL algorithm, we propose a cost-sensitive intelligent trainer enhanced model-based DRL training framework, which can learn the optimal model-related parameters, training and sampling settings in an online manner. Different to the existing approach that directly modifying the training algorithm of the target controller [9], we propose to package the standard model-based DRL training process as a standard RL environment termed as Training Process Environment (TPE) and implement an RL agent as the intelligent trainer to optimize the TPE system. We rely on the trainer to take three...
actions to optimize the training process inside TPE. These actions are designed to regulate how sampling and training are carried out in the real and cyber environment. In particular, the first action determines how many data we should sample from the cyber environment and the probability to train with these cyber data. The second and third actions control how we start a new episode when sampling in the cyber and real environment. For example, in the cyber environment the trainer may choose to start a new episode from a state which has been sampled from the real environment to further the exploitation of this state, or it can choose to start from a random state in favor of exploration. This action affects the balance between exploration and exploitation. In a general model-based RL approach, these actions are usually manually tuned; in our approach, we utilize the trainer to learn to select the proper action on the fly. Since the online learning process is a single-head action process with all actions correlated, we observe that the trainer may be trapped by specific wrong actions and never recover. To resolve this issue, we build an ensemble trainer, which includes multiple trainers that take different actions in their training process. We adopt a particular memory sharing mechanism to make sure all trainers have enough real data to train their target controllers without incurring additional sampling cost. We also rank the rewards of different trainers to determine the best trainer, which is set to be the reference trainer in sampling real data. We will use this reference trainer to sample data such that the data samples come from well-trained agents with a high probability. To recover underperformed trainers that are trapped by unfavorable actions, we design a weight copy process to reboot those trainers. All these ensemble operations are carried out only with the training data the trainers received, and no additional test data are used to guide the training or ensemble process.

We extensively test the proposed framework using five representative tasks of OpenAI gym \([10]\). Compared with a fixed parameter model-based RL baseline algorithm, study different approaches to implement the trainer. We primarily investigate a DQN based intelligent trainer and the ensemble trainer. The results show that the DQN trainer can achieve good performance. However, DQN trainer may encounter performance degradation when the target controller is sensitive to the cyber data. With the designed ensemble trainer, we can solve these problems, and it delivers performance close to manually optimized training algorithms for all cases. The contributions of the paper are as follows:

Firstly, we propose a "reinforce on reinforce" trainer framework for model-based RL. With this framework, the cyber model related setting in the training and sampling process of general model-based RL algorithm is decoupled from the training algorithm of the target controller, which provides high flexibility for the design of the higher level optimization of these settings.

Secondly, we provide an example design of the intelligent trainer based on DQN. With the DQN trainer, we can learn a proper training and sampling setting of model-based RL algorithm in an online fashion, without incurring sampling cost.

Thirdly, we design an ensemble trainer that can boost the performance of the single-head intelligent trainer. With the proposed memory sharing, reference sampling, and weight copy mechanism, the ensemble trainer can successfully learn the best control setting of different scenarios and can perform quite well even the training of the target controller is extremely sensitive to the cyber data.

In summary, the proposed framework aims to reduce the algorithm tweak cost with model-based RL algorithm and to make model-based RL algorithm more applicable in practice. We have open-sourced our training framework at https://bitbucket.org/RLinRL/intelligenttrainerpublic, to facilitate the research in model-based DRL algorithms.

The remainder of this paper is structured as follows. In Section II, we briefly review the existing literature. In Section III, we provide a detailed description of the proposed trainer framework, including its key components, single-head trainer design, and the ensemble trainer design. In Section IV, we present the numerical evaluation results of the proposed framework and in Section V provide a summary of the paper.

II. RELATED WORKS

Motivated to build intelligent agents that can learn to accomplish various control tasks, researchers have been actively studying reinforcement learning for decades, such as in \([11]\), \([12]\), \([13]\), \([14]\), \([15]\). With the recent advancement of deep learning, deep reinforcement learning (DRL) \([1] \) has demonstrated its strength in various applications. For example, in \([16]\), a DRL agent is proposed to solve financial trading tasks; in \([17]\), a neural RL agent is trained to mimic the human motor learning; in \([18]\), an off-policy RL method is proposed to solve nonlinear nonzero-sum games. Despite the significant performance improvement brought by DRL in various research scenarios, the high sampling cost necessitated by DRL becomes a significant issue in practice.

To tackle this issue, model-based RL aims to learn the system dynamics model to reduce the sampling cost from real environment. In \([6]\), the authors provided an example of the general approach of model-based RL by developing a robot controller training algorithm that samples from both the real physical environment and from a learned cyber emulator. In \([19]\) the authors took a slightly different approach by adapting a model trained for other tasks to train the controller for a new and similar task. This approach can combine prior knowledge learned from previous tasks with the online adaptation of the dynamic model to achieve better performance. In these approaches, the number of samples taken from the cyber environment to train the target controller is either preset or can only be tuned manually, resulting in both sample inefficiency and additional training cost. To address this issue, we propose to design an intelligent trainer to simplify as well as automate this process to make the algorithm more adaptive and practically applicable. In \([20]\), the authors proposed a Model-assisted Bootstrapped DDPG algorithm, which utilizes a variance ratio computed from the multiple heads of the critic network to control whether the cyber data should be used or not. Different to their approach relying on bootstrapped DQN \([21]\), we aim to develop an automated framework which can work for general model-based RL with any target controller.
Instead of treating the cyber model as a data source for training, some approaches focused on utilizing the cyber model to conduct pre-trial tree searches in scenarios where selecting the right action is highly critical in real environment \[7\] \[8\]. In particular, the cyber model is used to prevent selecting unfavorable actions in the real environment to accelerate the learning of the optimal policy. In \[9\], the authors proposed a planning agent with a cyber engine and a manager who decides whether to sample from the cyber engine or to take actions to minimize the training cost. Though this approach is similar to ours in spirit, it only focuses on the tree search part by sampling in the cyber environment from a current real state during training. In comparison, we treat the cyber environment as a general data source as in \[6\] instead of a tool only for tree search. In addition, we propose to decouple the training of the “manager” and target controller by treating them as independently trained entities. Specifically, we train the target controller in a normal RL environment and introduce another RL agent that functions as the intelligent trainer and is placed logically on top of the RL environment for the target controller.

Some recent works focus on integrating model-based and model-free approaches in RL. In \[22\], the authors proposed combining model-based and model-free approach for Building Optimization and Control (BOC), in which a simulator is used to train the agent while the real world test-bed is used to evaluate the performance of the agent. In \[23\], the model-based approach is used to train a controller agent, which is then used to provide weights initialization for a reduced network size model-free DRL approach to reduce training cost. Different from these designs that provide specific methods for integrated model-based and model-free approaches, we focus on developing a sample-efficient model-based RL training framework with an intelligent trainer.

III. Approach

With the goal to optimize the sampling cost of a model-based RL algorithm, we aim to maximize the reward received from the trained target agent in a test environment, with a fixed amount of real data sample. To optimize the performance of the model-based RL, we need to adjust the setting like how much cyber data should be used and how to do the sampling in both real and cyber environment. To optimize these settings, we propose an approach that packaging the training process of a general model-based RL algorithm as a target system to optimize and design an intelligent trainer to operate on this target system.

The proposed framework is illustrated in Fig. 1 in which the physical environment represents the actual system for which we develop a target controller and the cyber environment is an emulator of the physical environment. The emulator can be either learning-based (such as a neural system dynamic model) or knowledge-based. The intelligent trainer is an RL agent that can control the sampling and training process of the target controller. As the trainer itself is an RL agent that is self-learned when controlling the learning process of the target controller inside TPE, the proposed intelligent training framework can be considered as an “reinforcement on reinforcement architecture”.

We note that in our proposed architecture, the trainer is set up as an independent entity from the target controller, as opposed to being treated as part of the target controller in existing approaches \[23\]. This setup allows the trainer and target controller to be decoupled, which makes it easy to implement different trainer designs. In the following, we provide the details of the proposed TPE and intelligent trainer, and introduce an additional ensemble trainer design to boost the performance of the proposed intelligent framework without incurring additional sampling cost.

A. Training Process Environment (TPE)

The key components of TPE include two environments (physical and cyber environment), a state to be observed by an agent outside TPE, an action interface which can be used to control the training process of the target controller inside TPE. The reward data output from TPE can help intelligent trainer to learn better control actions.

TPE implements two types of training data generation environments: the real and the cyber environment. The real environment is the actual physical system for which we train a controller. The cyber environment serves as the digital mirror of the real environment, which can be either entirely data-driven (e.g., a neural network prediction model) or knowledge-based. For knowledge-based model, we can add a learning-based correction layer on top of it. As such, regardless of the underlying approaches, the cyber model can be learned during the training process of the model-based DRL.

TPE has two important functions that combined compose the whole training process of a general model-based RL algorithm:

- **Initialization**: execute the initialization tasks for the model-based RL training process. These tasks include the generalization of the real training environment, the cyber emulator, and the target controller.
- **Step(state, action)**: take one step of training of the model-based RL algorithm. This process is shown in Fig. 2 which includes the sampling process from the real and cyber environment, training of the target controller, and the dynamic model of the cyber emulator. In the TPE implementation, sampling from the real/cyber environment is carried out in a streamlined manner, which means that we keep sampling in the real/cyber environment; when a terminal signal from the environment (e.g., “done” in OpenAI gym environment) is received, we reset the environment and continue the sampling. The sampling process stops when a predefined number of \(K_r\) or \(K_c\).
samples are taken. In our design, we use a pre-determined setting of $K_c$, and use the trainer to optimize the setting of $K_c$ and other sampling and training settings to supervise the learning of the target controller. To make the TPE system able to commutate with an RL agent, we define the three RL components State, Action and Reward of TPE as follows:

- **State**: a TPE state indicates the progress of training. In our experiments we find that simplifying TPE state representation is essential because there are usually very limited training data for the trainer who needs to quickly learn a good action, as will be shown in Section [V]. An example of such simple design is to use a constant to represent the state. We find that this simple design can facilitate intelligent trainer in selecting the appropriate control actions and can work to certain degree, as in this case the trainer can simply evaluate if an action is good for the training or not. There are other more informative state representation designs which can be better in certain cases, such as using the last average sampling reward or normalized sampling count. A comparative study of these different designs are provided in Section [V].

- **Action**: the action space includes three controllable parameters used in the training process. We represent these parameters as probability values, so that different test tasks share the same action range $[0, 1]$. Such representation greatly simplifies and accelerates the training process. Details of the three control parameters will be introduced below.

- **Reward**: the reward function reflects whether a good progress has been made from sampling the real physical system. With the goal of maximizing the sampling reward from the real environment, we define the reward as (1).

$$R(s_t, a_t, s_{t+1}) = \text{sign} (\hat{r}_{t+1} - \hat{r}_t),$$

where $\hat{r}_{t+1}$ and $\hat{r}_t$ are the average sampling reward of the target controller in step $t + 1$ and $t$ of the TPE on real environment, respectively. The rationale behind such simple reward design is that as long as the reward is increasing, we can accept the current training action. In the training process, the reward data will be noisy, to tackle this problem the trainer is equipped with a memory buffer for training, details will be shown in Section [III-B].

We note that using this reward function we can gauge the quality of the samples from real environment with no additional cost. Though such a simple design cannot cover all cases that may happen in practice, i.e., the case that the cyber data does not cause performance degeneration but only slows down the convergence, it allows the trainer to quickly learn the settings. A more reasonable rank-based reward design will be used in the ensemble trainer in Section [III-C].

The action space of the TPE has the following three components:

- **Action** $a_0$ is the ratio of sampled real data to the total data sampled (real and cyber). The number of data points sampled from the cyber environment $K_c$ can be calculated as:

$$K_c = K_r/a_0 \cdot (1 - a_0).$$

Action $a_0$ is also used to control the probability to train the target controller with the real data. It represents the probability to get a mini-batch from the real data memory in training the target controller. Equivalently, with a probability $1 - a_0$ we use samples stored in the cyber data buffer for training. If we aim to train with real data for $T_r$ batches, the total number of training batches $T$ in this step is determined by:

$$T = T_r + T_r/a_0 \cdot (1 - a_0).$$

The reason we use only one action to control both the sampling and training process is that for some DRL algorithms, such as TRPO, the sampling and training process cannot be decoupled and in each training epoch all sampled data will be used in training. To generalize, we choose to use one action to control both procedures.

- **Action** $a_1$ is related to the selection of a starting state of a new episode when we sample from the cyber environment and controls how much we favor exploitation over exploration during the sampling process. For sampling from the cyber environment, on the one hand, we can select a starting state from a buffer that stores data previously collected from the physical system. In this case, the subsequent sampling process will be a local search process similar to the imagination process used in [8]. On the other hand, we can use a data point randomly selected from the state space. In our design, $a_1$ represents the probability of choosing starting state from the real data buffer.

- **Action** $a_2$ is related to the selection of a starting state of a new episode when we sample from the real environment. Similar to sampling from the cyber environment, we use $a_2$ to control how much we favor exploitation over exploration. Favoring exploitation in the real environment can be done by picking an initial state that has higher Q value, where Q is the critic network of the target controller. We propose to utilize parameter $a_2$ to control this process with the following sampling quality function for the real environment:

$$\Phi_{real}(s) = a_2 \cdot Q(s, \pi(s)) + (1 - a_2) \cdot \text{rand}(),$$

where $\text{rand}()$ is a uniformly distributed random number drawn from $[0, 1]$. A higher value of $a_2$ indicates that we favor selecting initial states with a higher Q value which can accelerate the convergence of the critic network. Combined with the cyber environment, when $a_1$ or $a_2$ approaches 1, the optimization process will favor exploitation; when $a_1$ or $a_2$ approaches 0, the optimization process will favor exploration.

### B. Design of the Intelligent Trainer

The intelligent trainer is designed to learn the best control action during the training of the target DRL agent. It is trained
in an online and on-policy manner. Each time the trainer samples from TPE, only one sample is taken and the TPE advances for one time step (as shown in Algorithm 1). We refer to this as single-head design as we can neither sample multiple actions nor try for multiple episodes. For such reason, the trainer needs to learn quickly with limited training time steps and samples. Different trainer learning algorithms like DQN and REINFORCE can be used to tackle this optimization problem. In the following, we use an DQN controller to illustrate a trainer design. A comparison of different trainer designs is shown in Section 5.1.3.

We implement a specialized DQN controller that carries out discretized control actions with a relatively small-scale Q network. Each time when trainer making a decision, it evaluates all the actions with the Q network and outputs the action with the highest Q value. To add more diversity to the action output, we add noise to the output of the DQN, as shown in (5):

\[
DQN(s, a) = 0.9 \cdot Q_{\text{Trainer}}(s, a) + 0.1 \cdot \text{rand}(0, 1) \cdot (\max_{\hat{a} \in A}(Q_{\text{Trainer}}(s, \hat{a})) - \min(Q_{\text{Trainer}}(s, \hat{a})))
\]

where \(A\) is the action set of the trainer, \(\text{rand}(0, 1)\) a random number uniformly draw from \([0, 1]\).

The training of the DQN controller follows standard epsilon-greedy exploration strategy. To enhance the training stability, the DQN controller is equipped with a memory like the replay buffer in DDPG, such that the trainer can extract the real good actions from the noisy data it received from TPE. During our experiments, we notice that the buffer can be saturated with samples all resulted from one single action. The lack of diversity in actions will slow or even halt the training of DQN. To resolve this issue, we limit the total number of the samples diversity in actions will slow or even halt the training of DQN.

Algorithm 1 Intelligent Trainer Enhanced Model-Based DRL Training Algorithm

1: Initialization: initialize the trainer agent (with a DQN network), the training process environment, and the target controller. Initialize real data memory and cyber data memory as an empty set. Sample a small data set of size \(o\) to initialize the cyber emulator and initialize the real environment.
2: Set number of total samples generated from real environment \(n = 0\). Set the maximum number of samples allowed to use as \(N\).
3: //Training Process:
4: while \(n < N\) do
5: Generate action \(a\) from the trainer agent.
6: //One step in TPE:
7: Train the target controller if there is enough data in its memory buffer.
8: Sample \(K_r\) data points from real environment according to the sampling reset Algorithm 2 and append the data to the real data memory.
9: Sample \(K_c\) data points from the cyber environment, and append the data to the cyber data memory.
10: Train the dynamic model.
11: Update \(n\).
12: Collect the state, action and reward data of TPE.
13: Update the trainer agent.
14: end while

Algorithm 2 Sampling Reset Procedure

1: if the current sampling environment is the real environment then
2: Initialize data set \(D = \emptyset\), quality set \(G = \emptyset\).
3: for \(i = 1 : 50\) do
4: Generate one initial state \(s_0\) and compute its quality \(\Phi_{\text{real}}(s_0)\).
5: Append \(s_0\) to \(D\) and append \(\Phi_{\text{real}}(s_0)\) to \(G\).
6: if \(i > 5\) and \(\Phi_{\text{real}}(s_0) > 0.999 \cdot (\max(G) - \min(G)) + \min(G)\) then
7: Break.
8: end if
9: end for
10: Return the last state of \(D\).
11: else
12: if \(\text{rand}() < a_2\) then
13: Randomly select a state \(s\) from the real data memory.
14: Set the cyber environment to state \(s\).
15: Return \(s\).
16: else
17: Randomly initialize the cyber environment.
18: Return the current state of the cyber environment.
19: end if
20: end if
C. Ensemble Trainer

In this subsection, we present a further enhancement of the above presented intelligent trainer framework. The enhancement is motivated from the observation that the single-head trainer cannot adequately assess the quality of an action in certain cases. Here single-head means that only one target controller is involved in the training and thus all the actions are tested on this target controller. It is probable that one action may degrade the subsequent training process. In another word, the actions can be correlated and then it is hard to differentiate the quality of the actions. Also, the reward design in Section 1 cannot accurately assess the impact of an action if it produces non-negative reward but only reduces learning speed or leads to locally optimized policy. To solve these issues, we propose an ensemble trainer to utilize a multi-heads training process, which is similar to the boosted DQN [21] yet ensemble at the trainer level. The design rationale is to compare different actions on different trainer agents to find out the good action without additional sampling cost.

The detailed flow of the proposed ensemble trainer is shown in Fig. 3. Three different trainers are constructed at the initialization stage. For better ensemble effects, distinct trainer settings are used for the three trainers by defining the trainer action: for Trainers 0, its actions are predicted by the intelligent trainer; for Trainers 1, its actions are predicted by a random trainer; for Trainer 2, it always uses the pure real data configuration, which means setting the three actions to 0, 0, and 1. Trainer 2 is corresponding to a normal DRL training process with no dynamic model used in training, while Trainers 0 and 1 take different settings to help exploit and explore the trainer action space.

In the ensemble learning process, each trainer takes steps in its independent TPE, in which the sampling and training process of its corresponding target controller is carried out. That means three target controllers are trained. To make sure no additional sampling cost is introduced, we split the number of real samples among the three trainers evenly. An issue to solve here is how to make sure all three target controllers are well trained with only one third of new real samples received. We propose the memory sharing mechanism. As shown in Fig. 4, memory sharing is done before the training of the target controllers. The sharing process is done by a pseudo sampling process for each trainer, to collect the recent real data samples from the other two trainers (from their corresponding real data memory). After the memory sharing process, for each trainer, they can receive \( K_r \) new real data samples in each step, which is the same to a non-ensemble training approach.

After the training process of each trainer, the reward information from each trainer will be calculated in the following manner. Different to the reward in the single-head version, for the ensemble algorithm we use the comparison of the average sampling reward of the target controllers to set the reward. Denote the average sampling reward for the three target controllers as \( r_0, r_1 \) and \( r_2 \), then the reward \( r_i^{\text{rank}} \) for trainer \( i \) is set to be its rank in \( r_i \), i.e., the index of \( r_i \) in sorted sequences of \( r \) with ascending order. The rationale to set it in this manner is that if the action is good, it should help the trainer to achieve best training performance of its target controller. With the above reward design, the three trainers will generate three data samples at the trainer level, and all these data will be used to update the intelligent trainer. As in this case, the trainer memory will never be flooded by one kind of action due to the actions from Trainer 1 and 2, we do not use the special memory insertion mechanism used in Section III-B when using DQN controller in trainer.

After collection of the trainer reward data, we add a particular weight copy mechanism to solve the issue that some target agent may fail due to unfavorable trainer actions. The rationale is that when we have collected the reward information for a certain large number of steps, we can judge which trainer is currently the best trainer with high confidence. In this case, we can transfer the best target agent to the other trainers, such that those trainers who fall behind can restart from a good position. We propose to do weight copy at the target agent level to achieve this. More specifically, after the trainer data are collected, we examine the current trainer steps \( n_c \) that have been taken from the last weight copy; if \( n_c \) is large than a threshold \( T_c \), then we compute the reward data for each trainer in the last \( T_c \) steps, and compute a cumulated reward for each trainer as:

\[
R_c(i) = \sum_{j \in \{T_c-1, \ldots, 0\}} R_{i-j}(i),
\]

where \( i \) is the index of current trainer step. The trainer with maximum \( R_c \) will be set as the best trainer. All trainers that are not the best will update its target agent to be the same to the best trainer in the weight copy procedure. In this way, we can recover those early failed trainers.

Another issue we observe in experiments is that even with the weight copy mechanism, the ensemble trainer may still fail due to the quality of the sampled real data may be affected by the target agents. For example, the training may be corrupted by the data sampled from a underperformed target agent. To resolve this issue, we propose a reference sampling strategy. For every three trainer steps, the second and third step we will use the best trainer in sampling with a probability \( p_{ref} \). This probability will be controlled by a skewness measure of the distribution of the trainer reward data as shown in Algorithm 4. The skewness measures the confidence that there is a strong best player which is significantly better than the other two. With more confidence, we are more likely to use the reference trainer in the sampling process. As the reference sampling mechanism will affect the assessing accuracy for the actions taken by different trainers, we add a scaling factor to the rank reward for each trainer agents in the following manner:

\[
r_i^{\text{rank}} = c \cdot r_i^{\text{rank}},
\]

where \( c \) is determined by

\[
c = \begin{cases} (1 - p_{ref})^2, & \text{if } (1 - p_{ref})^2 \geq 0.95 \\ 0, & \text{otherwise} \end{cases}
\]

In this case, the reward will be set to zero even we use reference sampling with a low probability such that we can make sure all the different rewards we received are really
useful to assess the differences of actions. Algorithm 3 shows the overall flow of the ensemble trainer. To summary, the ensemble trainer is designed to detect the quality of the actions by comparison; can maintain the quality of training by memory sharing without incurring additional sampling cost; can recover underperformed trainer from harmful actions, and can maintain the quality of the samples with reference sampling. In a trade-off of sampling cost and training cost, the ensemble trainer will triple the training time, although this can be relieved by the early stop of some underperformed trainers.

Algorithm 3 Ensemble Trainer Algorithm

1: Initialization: initialize the three trainer agents and the corresponding training process environments, along with the target controllers. Run the initialization process for each trainer. Initialize the best player to be NoDyna trainer and the probability to use best player to sample is $p_{sam}$.  
2: Set number of total samples generated from real environment $n = 0$. Set maximum number of samples allowed to use as $N$. 
3: //Training Process:  
4: while $n < N$ do  
5:    for trainer $i \in 0, 1, 2$ do  
6:        Generate action $a$ from the trainer agent.  
7:        //One step in TPE:  
8:        Execute memory sharing procedure.  
9:        Train the target controller if there is enough data in its memory buffer.  
10:       Sample $K_i/3$ data points from real environment according to the sampling reset Algorithm 2 and append the data to the real data memory. When sampling the data, if the step count cannot be divided by 3, with a probability $p_{ref}$, the actions used will be generated by the target agent trained by the best trainer.  
11:       Sample data from cyber environment according to the trainer action, and append the data to the cyber data memory.  
12:       Train the dynamic model of the current trainer.  
13:       Update $n$.  
14:       Collect the state, action and raw reward data of TPE.  
15:    end for  
16:    Compute reward for each trainer from the raw reward data and append the reward data into the reward historical list $R_{his}$ for trainers $i = 0, 1, 2$.  
17:    Store TPE data of all three trainers into the DQN memory to train the intelligent trainer.  
18:    Update the trainer agents.  
19:    Execute performance skewness analysis procedure Algorithm 4 to update $p_{ref}$ and do weight copy.  
20: end while

$$p_{ref} = \min((((\phi - 0.5)/0.2)^2), 1). \quad (9)$$

IV. EXPERIMENTS

In this section, we evaluate the proposed intelligent trainer and ensemble trainer on five different tasks with continuous control variables from OpenAI gym: Pendulum (V0), Mountain Car (Continuous V0), Reacher (V1), Half Cheetah (Special Version used in [24]), and Swimmer (V1).

A. Experiment Configuration

For the five test cases Pendulum, Mountain Car, Reacher, Half Cheetah, and Swimmer, different target controllers with promising published results are used here. The target controllers for the first two tasks are implemented via DDPG while the target controllers of the last three tasks are all implemented by TRPO. For the hyper-parameter settings of the target controller (including $K_i$ and $T_i$ which controls how much to sample and train with the real environment), they are set to well-tuned parameters used in published codes [25] [26].

As for the cyber models used in these tasks, they are implemented by simple neural networks with the guideline provided in [24]. From our experiments, we note that normalization to both the input and output for the dynamic model can be very useful. Specifically, we use the normalization method provided by [25], in which the mean and standard deviation information of the data is updated during the training process.

B. Comparing Single-Head Intelligent Trainer with Baseline Algorithms

We compare multiple variants of the single-head intelligent trainer with baseline algorithms including a “NoCyber” trainer that uses no cyber data in training, a fixed action trainer “Fixed” that executes only fixed control actions, and a “Random” trainer which outputs random actions. Detailed setting of these algorithms is shown below.

The first intelligent trainer we study here is termed as “DQN” with a DQN controller in which we use the following

Algorithm 4 Performance Skewness Analysis Procedure

1: if $\text{length}(R_{his}^i) > 5$ then  
2:    Compute accumulated reward of trainer $i$ as $R_{acc}^i$ for $i = 0, 1, 2$.  
3:    Update best trainer index as $\arg\max_i (R_{acc}^i)$.  
4:    Compute the skewness ratio $\phi$ for the best player to the other players as $(R_{best}^i - R_{median}^i)/(R_{best}^i - R_{worst}^i)$, where $worst$ and $median$ is the index of the trainer with the worst and median accumulated reward.  
5:    Update best player reference probability $p_{ref}$ according to (9).  
6: if $\phi > 0.9$ then  
7:    Do weight copy.  
8:    Clear $R_{his}^i$ for $i = 0, 1, 2$.  
9: end if  
10: end if

| TABLE I | NUMBER OF TOTAL TPE STEPS FOR DIFFERENT TASKS. |
|---------|-----------------------------------------------|
|         | Pendulum | Mountain Car Continuous | Reacher | Half Cheetah | Swimmer |
| TPE Steps | 1000 | 30000 | 1000 | 400 | 200 |
setting for all five tasks. We discretize each dimension of the action of the trainer into two values 0.2 and 1.0 to simplify the learning process, as for some tasks, the total number of steps of the TPE is only 200, as shown in Table 1. The DQN controller is trained with a memory buffer of size 32, and each time four randomly selected batches of batch size eight are used to update the controller. For exploration purpose, the epsilon-greedy method is utilized here with the first 10% of the trainer steps used for epsilon-greedy exploration by setting final epsilon to 0.1.

To test what if we use more action values in the action-discretization, we study the second trainer termed “DQN-5 actions” in which five values between 0.2 and 1 are used. Also to test if we can use a larger memory size (2000) for the DQN training, which should be more stable yet slow to react to the changes of the training states, are studied here as the third intelligent trainer termed as “DQN-larger memory”.

The fourth intelligent trainer we study here is the REINFORCE trainer with a REINFORCE controller with the same action splitting setting like DQN. One thing notable is that as REINFORCE requires data of multiple episodes to train, we manually set the training steps in one epoch to be 5 which means that we take 5 steps of TPE as an episode.

For the baseline algorithms, the NoCyber trainer is a standard DRL training process without using cyber data; the Fixed trainer follows the same setting of DQN trainer except that we replace the trainer DQN with a fixed probability 0.6 to sample and train from the real environment. In this case, the baseline algorithm is a standard model-based RL approach with fixed sampling and training settings. For the Random trainer, it outputs random action 0.2 or 1.0 uniformly for all actions.

The comparison results of the intelligent trainers and the baseline trainers are shown in Fig. 3. The test results are obtained by periodically evaluating the target controller, on an isolated test environment to make sure that it will not affect the training process and none of the data collected from the test environment are used in the training. We can observe that:

- For the five cases we study here, the first three cases Pendulum, Mountain Car and Reacher can benefit from cyber data used in training; while for the left cases Half Cheetah and Swimmer, the NoCyber trainer is significantly better than the trainer with cyber data used. This proves the statement that utilizing the cyber data may not always useful. It is thus essential to properly utilize the cyber model.
- Comparing with the performance of the Fixed trainer, we see that the intelligent trainer is better in most cases. For example, the DQN trainer with five actions is better than the Fixed trainer on cases Mountain Car, Reacher, and Half Cheetah, with similar performance on the left cases. We show the saving of real data samples in Table 4. This proves that the intelligent trainer can indeed work to a certain degree.
- For cases Pendulum and Mountain Car, the Random trainer performs the best. The reason is that for these cases, the training algorithm can perform better if we add more noises to encourage exploration. For example, the Mountain Car case requires more exploration than other cases because there is a local optimum which can lead the target agent to a negative action searching direction. In this case, the Random trainer can provide more exploration by adding more stochasticity into the training process. We can also observe that the DQN-5 actions is more stable than DQN, due to the larger number of actions which can increase the training diversity.
- For cases Half Cheetah and Swimmer, the cyber data enhanced algorithms degenerate to certain degree. For Half Cheetah case, the performance degeneration is mainly because with the cyber data, the performance becomes unstable, resulting in higher variance and low mean reward in the ten independent tests. For the Swimmer case, most trainers with cyber data perform quite bad. This is because the Swimmer environment is a partially observed Markov Decision Process (POMDP). The model learned in this case is unable to predict the correct system state transition. Our results show that even incorporating cyber data in training with a chance of 10%, significant performance degeneration happens. For this case, once an action is taken to utilize cyber data, the target controller
will be trapped by a local optimum and hard to recover. We leave this issue to the ensemble trainer to tackle.

To analyze the behavior of the trainer, we show the actions taken by the DQN trainer on cases Reacher, Swimmer and Mountain Car in Fig. [4](f). The curves show how the mean value of the actions changes during the training process. We can observe that for the Mountain Car case, the actions taken by the DQN fluctuates around 0.5, which is reasonable as in this case random actions are best. For Reacher and Swimmer cases, the trainer quickly learns to use more of the cyber data with the mean value of action $a_0$ larger than 0.6, which proves that the trainer can learn to a certain degree. Note that for the Swimmer case, we observe that even the mean value of action $a_0$ is larger than 0.6, the performance of the target controller is still very bad due to the reason that the training process is very sensitive to the cyber data. It proves the necessity to design a mechanism to help the trainer recover from such a situation, which we rely on the ensemble trainer.

C. Sensitivity Analysis on Various Trainer and TPE Designs

In this subsection, we test different variants of the trainer and TPE design to show that with a general reasonable design, the trainer framework can output stable acceptable performance. For practical application, such generalization power is essential.

To examine the performance sensitivity to different designs, we further test two DQN trainers with different TPE state designs. The first TPE design V1 is to use the last average sampling reward of the target agent as the state of TPE, and the second variant V2 sets the state of TPE to the number of real samples used (normalized into the range $[0, 1]$). Along with the constant state design shown above, we compare three different designs of TPE states. For the trainer design, we further test two DQN trainers with different TPE state designs. The accumulated rewards of in total five variants are shown in Table [III] We can observe that:

- For cases Mountain Car, Reacher and Half Cheetah in which the intelligent trainer is better than the fixed baseline algorithm, all variants achieve certain improvement
than the simplest DQN and TPE design. This proves that for certain cases, it is possible to achieve better performance by setting the TPE state more informative or increasing the DQN trainer memory and control action variations. Note that complicating the design of trainer or TPE may aggravate the situation of lack of training data. • For case Swimmer, we observe that none of the tested variants of DQN or TPE can achieve satisfying performance. This is because for the Swimmer case, the usage of a very small amount of cyber data can lead the target controller to a local minimum that cannot be recovered. The single-head trainer cannot tackle this issue as in the beginning stage the trainer has to try different actions to find the optimal action, which will trap the trainer in the beginning.

D. Tackling Action Correlation with multi-heads Ensemble Trainer

In this subsection, we verify the ensemble trainer design by comparing its performance with the single-head trainers. The task of ensemble trainer is to overcome the action correlation problem in the single-head trainer. The ensemble trainer is configured to be with one NoCyber trainer, one DQN trainer (with larger memory setting and TPE state design V2) and one Random trainer. Following the design in Section III-C, these three trainers work together to sample and train three trainers: NoCyber, DQN, and Random, to find the optimal action.
independent target controllers. The target controller of the best trainer will be used in the test. Note that the best trainer is learned only with the training data.

We show the result in Fig. 5. The results show that with the ensemble trainer, we can achieve overall good performance even in the case that the single-head trainer fails. For example, the ensemble trainer can perform almost as good as the DQN or Random trainer on cases Pendulum, Mountain Car, while can perform as good as the NoCyber trainer on the Swimmer case. Results on Reacher show that the ensemble trainer converges faster in the early stage yet achieve lower learning precision in the end, this is because the DQN trainer applies large memory which is not that good compared with DQN as shown in Table III. In Fig. 5 (f) we can see that the action $a_0$ taken by the DQN trainer in the ensemble trainer is changed significantly. We can observe that for case Swimmer, the action $a_0$ quickly evolve towards 1, which proves that the proposed ensemble trainer can assess the control actions better than the single-head trainer.

To show how different trainers are interplayed with each other in the ensemble trainer, we show the separate test results of different trainers in ensemble trainer on Reacher, Half Cheetah and Swimmer in Fig. 6. For cases Mountain Car and Reacher, we observe significant fusion of different trainers, while before the NoCyber is significantly worse than the DQN or Random trainer. In this case, the NoCyber trainer is improved, but it also affects the performance of the other trainers. For case Swimmer, we observe that the NoCyber trainer performs the best, while the DQN trainer can now perform as good as the NoCyber trainer, which shows that the intelligent trainer can learn better with the ensemble memory. Also, we found that the Random trainer shows significant performance improvement, which is because it receives a lot of training data sampled by the other trainers through memory sharing and reference sampling; yet its performance is still far behind NoCyber, which again proves its sensitivity to cyber data, even with weight copy it’s performance is not satisfying. Overall the ensemble trainer can achieve a satisfying trade-off among different parties in the ensemble process and achieve close to optimal performance without incurring additional sampling cost.

To examine how the memory sharing and reference sampling mechanism can help, we compare the performance of an ensemble trainer without memory sharing and an ensemble trainer without reference sampling process on case Swimmer, as shown in Fig. 7. The results show that without memory sharing, all three agents will be trained with 1/3 of the original data samples, which naturally causes serious performance degeneration. For the variant without reference sampling, we observe that its performance is very similar to the DQN trainer. This is because without reference sampling, most of the real data samples will be sampled by underperformed target controllers from the DQN trainer and Random trainer, which will contaminate the learning process of the NoCyber trainer. The results prove that the ensemble trainer design is indeed helpful to make it to adapt to different cases.

V. CONCLUSION

In this paper we propose an intelligent trainer for general model-based reinforcement learning algorithm. The proposed approach treats the training process of model-based RL as the target system to optimize, and utilizes a trainer that monitors the training process takes the task to control the sampling and training process. Further more, an ensemble trainer that can boost the performance of the trainer without incurring additional sampling cost is used to solve the problem of limited and correlated training data for the trainer. With the proposed trainer framework, the model-based RL is more likely to work in practice to reduce the sampling cost while achieve close-to-optimal performance.

For the future work, the proposed trainer framework can be further improved by adding more control actions to ease algorithm tweak cost. An even bolder idea is to utilize one trainer to train different DRL controllers for multiple tasks, which can learn the common knowledge shared by different DRL algorithms for these tasks. Note that this is different from using the same controller for multiple tasks; this will be more like using the same brain for multiple tasks.

REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” arXiv preprint arXiv:1312.5602, 2013.
[2] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” arXiv preprint arXiv:1509.02971, 2015.
[3] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in International Conference on Machine Learning, 2015, pp. 1889–1897.
[4] M. Hausknecht and P. Stone, “Deep reinforcement learning in parameterized action space,” in Proceedings of the International Conference on Learning Representations (ICLR), May 2016.
[5] M. Wiering and M. Van Otterlo, “Reinforcement learning,” Adaptation, learning, and optimization, vol. 12, 2012.
[6] M. P. Deisenroth, C. E. Rasmussen, and D. Fox, “Learning to control a low-cost manipulator using data-efficient reinforcement learning,” 2011.
[7] X. Guo, S. Singh, H. Lee, R. L. Lewis, and X. Wang, “Deep learning for real-time atari game play using offline monte-carlo tree search planning,” in Advances in neural information processing systems, 2014, pp. 3338–3346.
[8] T. Weber, S. Racanière, D. P. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li et al., “Imagination-augmented agents for deep reinforcement learning,” arXiv preprint arXiv:1707.06203, 2017.
[9] R. Pascanu, Y. Li, O. Vinyals, N. Heess, L. Buesing, S. Racanière, D. Reichert, T. Weber, D. Wierstra, and P. Battaglia, “Learning model-based planning from scratch,” arXiv preprint arXiv:1707.06170, 2017.
[10] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, “Openai gym,” 2016.
[11] A. G. Barto, R. S. Sutton, and C. W. Anderson, “Neuronlike adaptive elements that can solve difficult learning control problems,” IEEE transactions on systems, man, and cybernetics, no. 5, pp. 834–846, 1983.
[12] F. L. Lewis and D. Liu, Reinforcement learning and approximate dynamic programming for feedback control. John Wiley & Sons, 2013, vol. 17.
[13] D. Liu and Q. Wei, “Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems,” IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 3, pp. 621–634, 2014.
[14] B. Luo, H.-N. Wu, and T. Huang, “Off-policy reinforcement learning for $h_{\infty}$ control design,” IEEE transactions on cybernetics, vol. 45, no. 1, pp. 65–76, 2015.
[15] D. Liu, X. Yang, D. Wang, and Q. Wei, “Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints,” IEEE transactions on cybernetics, vol. 45, no. 7, pp. 1372–1385, 2015.
[16] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, “Deep direct reinforcement learning for financial signal representation and trading,” IEEE transactions on neural networks and learning systems, vol. 28, no. 3, pp. 653–664, 2017.
[17] Y. Pan and H. Yu, “Biomimetic hybrid feedback feedforward neural-network learning control,” IEEE transactions on neural networks and learning systems, vol. 28, no. 6, pp. 1481–1487, 2017.
[18] R. Song, F. L. Lewis, and Q. Wei, “Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games,” IEEE transactions on neural networks and learning systems, vol. 28, no. 3, pp. 704–713, 2017.
[19] S. Levine, N. Wagener, and P. Abbeel, “Learning contact-rich manipulation skills with guided policy search,” in Robotics and Automation (ICRA), 2015 IEEE International Conference on. IEEE, 2015, pp. 156–163.
[20] G. Kalweit and J. Boedecker, “Uncertainty-driven imagination for continuous deep reinforcement learning,” in Conference on Robot Learning, 2017, pp. 195–206.
[21] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, “Deep exploration via bootstrapped DQN,” in Advances in neural information processing systems, 2016, pp. 4026–4034.
[22] S. Baldi, I. Michailidis, C. Ravanis, and E. B. Kosmatopoulos, “Model-based and model-free plug-and-play building energy efficient control,” Applied Energy, vol. 154, pp. 829–841, 2015.
[23] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, “Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning,” arXiv preprint arXiv:1708.02596, 2017.
[24] https://github.com/berkeleydeeprlcourse/homework/tree/master/hw4, accessed: 2018-05-06.
[25] https://github.com/pat-coady/trpo, accessed: 2018-05-06.
[26] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Pfaff, A. Radford, J. Schulman, S. Sidor, and Y. Wu, “Openai baselines,” https: //github.com/openai/baselines, 2017.