

Environment Reconstruction with Hidden Confounders for Reinforcement Learning based Recommendation*

Wenjie Shang
National Key Laboratory for Novel
Software Technology
Nanjing University
shangwj@lamda.nju.edu.cn

Yang Yu
National Key Laboratory for Novel
Software Technology
Nanjing University
yuy@nju.edu.cn

Qingyang Li
AI Labs, Didi Chuxing
qingyangli@didiglobal.com

Zhiwei Qin
AI Labs, Didi Chuxing
qinzhiwei@didiglobal.com

Yiping Meng
AI Labs, Didi Chuxing
mengyipingkitty@didiglobal.com

Jieping Ye
AI Labs, Didi Chuxing
yejieping@didiglobal.com

ABSTRACT

Reinforcement learning aims at searching the best policy model for decision making, and has been shown powerful for sequential recommendations. The training of the policy by reinforcement learning, however, is placed in an environment. In many real-world applications, however, the policy training in the real environment can cause an unbearable cost, due to the exploration in the environment. Environment reconstruction from the past data is thus an appealing way to release the power of reinforcement learning in these applications. The reconstruction of the environment is, basically, to extract the casual effect model from the data. However, real-world applications are often too complex to offer fully observable environment information. Therefore, quite possibly there are unobserved confounding variables lying behind the data. The hidden confounder can obstruct an effective reconstruction of the environment. In this paper, by treating the hidden confounder as a hidden policy, we propose a *deconfounded multi-agent environment reconstruction* (DEMER) approach in order to learn the environment together with the hidden confounder. DEMER adopts a multi-agent generative adversarial imitation learning framework. It proposes to introduce the confounder embedded policy, and use the compatible discriminator for training the policies. We then apply DEMER in an application of driver program recommendation. We firstly use an artificial driver program recommendation environment, abstracted from the real application, to verify and analyze the effectiveness of DEMER. We then test DEMER in the real application of Didi Chuxing. Experiment results show that DEMER can effectively reconstruct the hidden confounder, and thus can build the environment better. DEMER also derives a recommendation policy with

*This work is supported by the National Key R&D Program of China (2017YFB1002201), NSFC (61876077), Jiangsu SF (BK20170013), and Collaborative Innovation Center of Novel Software Technology and Industrialization. This work is done during the first author's internship in Didi Chuxing. Yang Yu is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6201-6/19/08...\$15.00

<https://doi.org/10.1145/3292500.3330933>

a significantly improved performance in the test phase of the real application.

CCS CONCEPTS

• **Applied computing** → **Transportation**; • **Computing methodologies** → **Reinforcement learning**; **Simulation environments**.

KEYWORDS

reinforcement learning, environment reconstruction, hidden confounder, imitation learning, recommendation

ACM Reference Format:

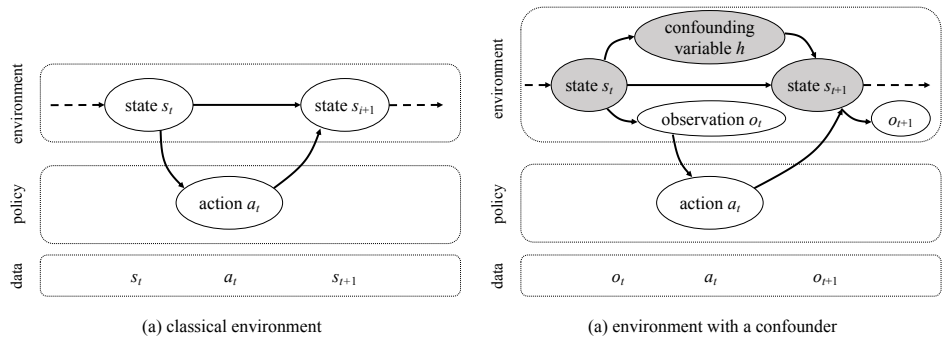
Wenjie Shang, Yang Yu, Qingyang Li, Zhiwei Qin, Yiping Meng, and Jieping Ye. 2019. Environment Reconstruction with Hidden Confounders for Reinforcement Learning based Recommendation. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19)*, August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3292500.3330933>

1 INTRODUCTION

In sequential recommendation problems [20, 21], where the system needs to recommend multiple items to the user while responding to the user's feedback, there are multiple decisions to be made in sequence. For example, in our application of program recommendation to taxi drivers, the system recommends a personalized driving program to each driver, and a program consists of multiple steps, where each step is recommended according to how the previous steps was followed. Therefore, recommending the program steps is a sequential decision problem, and it can be naturally solved by reinforcement learning [19].

As a powerful tool for learning decision-making policies, reinforcement learning learns from interactions with the environment via trial-and-errors [19]. In digital worlds where interactions with the environment are feasible and cheap, it has made remarkable achievements, e.g., [10, 18]. When it comes to real-world applications, the convenience of available digital environments does not exist. It is not practical to interact with the real-world environment directly for training the policy, because of the high interaction cost and the huge amount of interactions required by the current reinforcement learning techniques. A recent study [17] disclosed a viable option to conduct the reinforcement learning on real-world tasks, which is by reconstructing a virtual environment from the

Figure 1: Illustration of the graph structure and the collected data (a) in the classical environment that assumes fully observable, and (b) in the more realistic environment with an unobserved confounder.



past data. As a result, the reinforcement learning process could be more efficient by interacting with the virtual environment, and the interaction cost could be avoided as well.

The environment reconstruction can be done by treating the environment as a policy that makes responses to the interactions, and employing the imitation learning [1, 15] to learn the environment policy from the past data, which has drawn a lot attentions recently. Comparing with using supervised learning, i.e., behavior clone, to learn the environment policy, a more promising solution in [17] is to formulate the environment policy learning as an interactive process between the environment and the system in it. Take the example of the commodity recommendation system, the user and the platform could be viewed as two agents interacting with each other, where user agent views the platform as the environment and the platform agent views the user as the environment. By this multi-agent view, [17] proposed a multi-agent imitation method MAIL, extending the GAIL framework [6], which learns the two policies simultaneously by beating the discriminator that finds the difference between the generated and the real interaction data.

However, the MAIL method [17] is under the assumption that the whole world consists of the two agents only. From the perspective of the real users, they can receive much more information from the real-world that is not recorded in the data. Therefore, it is still quite challenging to reconstruct the environment in real-world applications, since the real-world scenario is too complex to offer a fully observable environment, which means that it might exist the hidden confounders. As shown in Figure 1, in the classical setting, the next state depends on the previous state and the executed action. While in most of real-world scenarios, the next state could be extra influenced by some hidden confounders. If we follow the assumption of a fully observable world, the reconstruction may be misled by the appeared fake associations in the data, due to the unawareness of the possible hidden causes. Thus, it is essential to take hidden confounders into consideration.

Originally, confounder is a casual concept [11]. It can affect both the treatment and the outcome in an experiment and cause a spurious association in observational data [7]. Similarly, in reinforcement learning, hidden confounders can affect both actions and rewards as an agent interacts with the environment. When it comes to such real-world applications, it is necessary to involve the confounder into the learning task because of the confounding effect. Yet, little work has been done in this promising area [2, 4]. To the best of our knowledge, this is the first study in reinforcement learning to reconstruct an environment together with hidden confounders.

To involve hidden confounders into the environment reconstruction, we propose a *deconfounded multi-agent environment reconstruction* method, named DEMER. Firstly, we formulate two representative agents, π_a and π_b , interacting with each other. Then, in order to simulate the confounding effect of hidden confounders, we add a confounding agent π_h into the formulation. According to the casual relationship, the confounding agent π_h interacts with the other two agents. Based on the formulation, we learn each policy of three agents from the historical data by imitation learning. Since the hidden confounder is unobservable, to learn the policy of it, we propose two techniques: the confounder embedded policy and the compatible discriminator under the framework of GAIL [6]. The confounder embedded policy involves the confounding policy into the observable policy. The compatible discriminator is designed to discriminate the state-action pairs of the two observable policies so as to provide the respective imitation reward. As the training converges, the deconfounded environment is reconstructed.

To verify the effectiveness of DEMER, we firstly use an artificial environment abstracted from the real application. Then, we apply DEMER to a large-scale recommender system for ride-hailing driver programs in Didi Chuxing. Through comparative evaluations, DEMER shows significant improvements in this real application.

The contribution of this work is summarized as follows:

- We propose a novel environment reconstruction method to tackle the practical situation where hidden confounders exist in the environment. To the best of our knowledge, this is the first study to reconstruct environment with taking hidden confounders into consideration.
- By treating the hidden confounder as a hidden policy, we formulate the confounding effect into a multi-agent interactive environment. We propose an imitation learning framework by considering the interaction among two agents and the confounder. We define the confounder embedded policy and the compatible discriminator to learn policies effectively.
- We deploy the proposed framework to the driver program recommendation system on a large-scale riding-hailing platform of Didi Chuxing, and achieve significant improvements in the test phase.

The rest of this paper is organized as follows: we introduce the background in Section 2 and the proposed method DEMER in Section 3. We describe the application for the scenario of driver program recommendation in Section 4. Experiment results are reported in Section 5. Finally, we conclude the paper in Section 6.

2 REINFORCEMENT LEARNING AND ENVIRONMENT RECONSTRUCTION

2.1 Reinforcement Learning

The problem to be solved by Reinforcement Learning (RL) can usually be represented by a Markov Decision Processes (MDP) quintuple (S, A, T, R, γ) , where S is the state space and A is the action space and $T: S \times A \mapsto S$ is the state transition model and $R: S \times A \mapsto \mathbb{R}$ is the reward function and γ is the discount factor of cumulative reward. Reinforcement learning aims to optimize policy $\pi: S \mapsto A$ to maximize the expected γ -discounted cumulative reward $\mathbb{E}_\pi[\sum_{t=0}^T \gamma^t r_t]$ by enabling agents to learn from interactions with the environment. The agent observes state s from the environment, selects action a given by π to execute in the environment and then observes the next state, obtains the reward r at the same time until the terminal state is reached. Consequently, the goal of RL is to find the optimal policy

$$\pi^\star = \arg \max_{\pi} \mathbb{E}_\pi[\sum_{t=0}^T \gamma^t r_t], \quad (1)$$

of which the expected cumulative reward is the largest.

Imitation Learning. Learning a policy directly from expert demonstrations has been proven very useful in practice, and has made a significant improvement of performance in a wide range of applications [13]. There are two traditional imitation learning approaches: behavioral cloning, which trains a policy by supervised learning over state-action pairs of expert trajectories [12], and inverse reinforcement learning [14], which learns a cost function that prioritizes the expert trajectories over others. Generally, common imitation learning approaches can be unified as the follow formulation: training a policy π to minimize the loss function $l(s, \pi(s))$, under the discounted state distribution of the expert policy: $P_{\pi_e}(s) = (1 - \gamma) \sum_{t=0}^T \gamma^t p(s_t)$. The object of imitation learning is represented as

$$\pi = \arg \min_{\pi} \mathbb{E}_{s \sim P_{\pi_e}} [l(s, \pi(s))], \quad (2)$$

Confounding Reinforcement Learning. Originally, confounding is a concept in casual inference [11]. Confounder is a variable that influences both the treatment and the outcome, naturally corresponding to the action and the reward in reinforcement learning. From the perspective of traditional reinforcement learning, the state is a confounder between the action and the reward. Although there are inherent similarities between causal inference and reinforcement learning, little work has been done in reinforcement learning that confounders exist in the environment [2, 4]. Only recently, Lu et al. [8] proposed the deconfounding reinforcement learning to adapt to the confounding setting, while the model of confounder is stationary at each time step which actually can be dynamic.

2.2 Environment Reconstruction

Reinforcement learning relies on an environment. However, when it comes to real-world applications, it is not practical to interact with the real-world environment directly to optimize the policy because of the low sampling efficiency and the high-risk uncertainty, such as online recommendation in E-commerce and medical diagnosis. A viable option is to reconstruct a virtual environment [17]. As a result, the learning process could be more efficient by interacting

with the virtual environment and the interaction cost could be avoided as well.

Generative Adversarial Nets. Generative adversarial networks (GANs) [5] and its variants are rapidly emerging unsupervised machine learning techniques. GANs involve training a generator G and discriminator D in a two-player zero-sum game:

$$\arg \min_G \arg \max_{D \in (0,1)} \mathbb{E}_{x \sim p_E} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))], \quad (3)$$

where p_z is some noise distribution. In this game, the generator learns to produce samples (denoted as x) from a desired data distribution (denoted as p_E). The discriminator is trained to classify the real samples and the generated samples by supervised learning, while the generator G aims to minimize the classification accuracy of D by generating samples like real ones. In practice, the discriminator and the generator are both implemented by neural networks, and updated alternately in a competitive way. The training process of GANs can be seen as searching for a Nash equilibrium in a high-dimensional parameter space. Recent studies have shown that GANs are capable of generating faithful real-world images [9], demonstrating their applicability in modeling complex distributions.

Generative Adversarial Imitation Learning. GAIL [6] has become a popular imitation learning method recently. It was proposed to avoid the shortcoming of traditional imitation learning, such as the instability of behavioral cloning and the complexity of inverse reinforcement learning. It adopts the GAN framework to learn a policy (i.e., the generator G) with the guidance of a reward function (i.e., the discriminator D) given expert demonstrations as real samples. GAIL formulates a similar objective function like GANs, except that here p_E stands for the expert's joint distribution over state-action pairs:

$$\arg \min_{\pi} \arg \max_{D \in (0,1)} \mathbb{E}_\pi [\log D(s, a)] + \mathbb{E}_{\pi_e} [\log(1 - D(s, a))] - \lambda H(\pi), \quad (4)$$

where $H(\pi) \triangleq \mathbb{E}_\pi [-\log \pi(a|s)]$ is the entropy of π .

GAIL allows the agent to execute the policy in the environment and update it with policy gradient methods [16]. The policy is optimized to maximize the similarity between the policy-generated trajectories and the expert ones measured by D . Similar to the equation (2), the policy π is updated to minimize the loss function

$$l(s, \pi(s)) = \mathbb{E}_\pi [\log D(s, a)] - \lambda H(\pi) \cong \mathbb{E}_{\tau_i} [\log \pi(a|s) Q(s, a)] - \lambda H(\pi). \quad (5)$$

where $Q(s, a) = \mathbb{E}_{\tau_i} [\log(D(s, a)) | s_0 = s, a_0 = a]$ is the state-action value function. The discriminator is trained to predict the conditional distribution: $D(s, a) = p(y|s, a)$ where $y \in \{\pi_e, \pi\}$. In other words, $D(s, a)$ is the likelihood ratio that the pair (s, a) comes from π rather than from π_e . GAIL is proven to achieve similar theoretical and empirical results as IRL [3] while it is more efficient. Recently, the multi-agent extension of GAIL [17] has been proven effective to reconstruct an environment.

Shi et al. [17] proposed to virtualize an online retail environment by extending the GAIL framework to a multi-agent approach, MAIL, that learns the interacting factors simultaneously. They showed that the multi-agent method leads to a better generalizable environment.

3 DECONFOUNDED MULTI-AGENT ENVIRONMENT RECONSTRUCTION

To reconstruct environments where hidden confounders exist, we propose a novel *deconfounded multi-agent environment reconstruction* (DEMER) method.

In this study, by treating the hidden confounder as a hidden policy, we formulate the deconfounding environment reconstruction as follows: there are two agents A (known as the policy agent) and B (known as the environment), interacting with each other and both of them are confounded by a hidden confounder H . Specifically, the dynamic effect of the hidden confounder H is modeled as a hidden policy π_h . The observation and action of each agent are defined as follows: Given o_A as the observation of agent A , it takes an action $a_A = \pi_a(o_A)$. The observation o_H of the hidden policy is formatted as the concatenation $o_H = \langle o_A, a_A \rangle$, and action $a_H = \pi_h(o_H)$ has the same format as a_A . The observation o_B of agent B is formatted as the concatenation $o_B = \langle o_A, a_A, a_H \rangle$, and its action is $a_B = \pi_b(o_B)$ which can be used to move forward to the next state. The objective is to use only observed interactions, that is, trajectories $\{(o_A, a_A, a_B)\}$, to imitate the policies of A, B and recover the potential effect of H by inferring the hidden policy π_h . The objective function of multi-agent imitation learning is then defined analogy to equation (2):

$$(\pi_a, \pi_b, \pi_h) = \arg \min_{(\pi_a, \pi_b, \pi_h)} \mathbb{E}_{s \sim P_{\tau_{real}}} [L(s, a_A, a_B)], \quad (6)$$

where a_A, a_B depend on three policies. By adopting the GAIL framework, according to equation (5), we can get

$$L(s, a_A, a_B) = \mathbb{E}_{\pi_a, \pi_h, \pi_b} [\log D(s, a_A, a_B)] - \lambda \sum_{\pi \in \{\pi_a, \pi_h, \pi_b\}} H(\pi) \quad (7)$$

and observe that π_a is independent with π_h and π_b given s and a_A , then

$$\begin{aligned} D(s, a_A, a_B) &= p(\pi_a, \pi_h, \pi_b | s, a_A, a_B) \\ &= p(\pi_a | s, a_A, a_B) p(\pi_h, \pi_b | s, a_A, a_B) \\ &= p(\pi_a | s, a_A) p(\pi_h, \pi_b | s, a_A, a_B) \\ &= D_a(s, a_A) D_{hb}(s, a_A, a_B). \end{aligned} \quad (8)$$

Combining equations (7) and (8), we can get the formulation as

$$\begin{aligned} L(s, a_A, a_B) &= \mathbb{E}_{\pi_a, \pi_h, \pi_b} [\log D_a(s, a_A) D_{hb}(s, a_A, a_B)] - \\ &\quad \lambda \sum_{\pi \in \{\pi_a, \pi_h, \pi_b\}} H(\pi) \\ &= \mathbb{E}_{\pi_a} [\log D_a(s, a_A)] - \lambda H(\pi_a) + \\ &\quad \mathbb{E}_{\pi_h, \pi_b} [\log D_{hb}(s, a_A, a_B)] - \lambda \sum_{\pi \in \{\pi_h, \pi_b\}} H(\pi) \\ &= l(s, \pi_a(s)) + l((s, a_A), \pi_b \circ \pi_h((s, a_A))) \end{aligned} \quad (9)$$

which indicates that the optimization can be decomposed as optimizing policy π_a and joint policy $\pi_{hb} = \pi_b \circ \pi_h$ individually by minimizing

$$\begin{aligned} l(s, \pi_a(s)) &= \mathbb{E}_{\pi_a} [\log D_a(s, a_A)] - \lambda H(\pi_a) \\ &\cong \mathbb{E}_{\tau_i} [\log \pi_a(a_A | s) Q(s, a_A)] - \lambda H(\pi_a), \end{aligned} \quad (10)$$

where $Q(s, a_A) = \mathbb{E}_{\tau_i} [\log(D(s, a_A)) | s_0 = s, a_0 = a_A]$ is the state-action value function of π_a , and

$$\begin{aligned} l((s, a_A), \pi_{hb}((s, a_A))) &= \mathbb{E}_{\pi_h, \pi_b} [\log D_{hb}((s, a_A), a_B)] - \\ &\quad \lambda \sum_{\pi \in \{\pi_h, \pi_b\}} H(\pi) \\ &\cong \mathbb{E}_{\tau_i} [\log \pi_{hb}(a_B | s, a_A) Q(s, a_A, a_B)] - \\ &\quad \lambda \sum_{\pi \in \{\pi_h, \pi_b\}} H(\pi), \end{aligned} \quad (11)$$

where $Q(s, a_A, a_B) = \mathbb{E}_{\tau_i} [\log(D((s, a_A), a_B)) | s_0 = s, a_{A0} = a_A, a_{B0}]$ is the state-action value function of π_{hb} .

Based on this result, we propose the confounder embedded policy and the compatible discriminator to achieve the goal of imitating policies of each agent, thus obtaining the DEMER approach.

3.1 Confounder Embedded Policy

In this study, the interaction between the agent A (known as the policy agent) and the agent B (known as the environment) could be observed, while the policy and data of the agent H (known as hidden confounders) are unobservable. Thus, we combine the confounder policy π_h with policy π_b as a joint policy named $\pi_{hb} = \pi_b \circ \pi_h$. Together with the policy π_a , the generator is formalized as an interactive environment of two policies as shown in the top of Figure 2. The joint policy can actually be expressed as $\pi_{hb}(o_A, a_A) = \pi_b(o_A, a_A, \pi_h(o_A, a_A))$ in which the input o_A, a_A and the output a_B are both observable from the historical data. Therefore, we can use imitation learning methods to train these two policies by imitating the observed interactions.

The policies in generator are updated respectively for each training step: firstly the joint policy π_{hb} is updated with the reward r^{HB} given by the discriminator, secondly the policy π_a is updated with the reward r^A . Though there is no explicit updating step for the hidden confounder policy π_h , it has been optimized iteratively by these two steps. Intuitively, the generated hidden policy π_h is just like a by-product along with the process of optimizing policies π_a and π_{hb} towards the truth and in consequence it can recover the real confounding effect to some extent. To make the training process more stable, we employ TRPO to update policies mentioned above.

3.2 Compatible Discriminator

In most of generative adversarial learning frameworks, there is only one task to model and learn in the generator. In this study, it is essential to simulate and learn different reward functions for the two policies π_a, π_{hb} in the generator respectively. Thus, we design the discriminator compatible with two classification tasks. As Figure 2 illustrates, one task is designed to classify the real and generated state-action pairs of π_{hb} while the other one is to classify the state-action pair of π_a . Correspondingly, the discriminator has two kinds of input: the state-action pair (o_A, a_A, a_B) of policy π_{hb} and the zero-padded state-action pair $(o_A, a_A, \mathbf{0})$ of policy π_a . This setting indicates that the discriminator splits not only the policy π_{hb} 's state-action space, but also the policy π_a 's. The loss function of each task is defined as

$$E_{\tau_{sim}} [\log(D_{\sigma}(o_A, a_A, a_B))] + E_{\tau_{real}} [\log(1 - D_{\sigma}(o_A, a_A, a_B))] \quad (12)$$

for π_{hb} , and

$$E_{\tau_{sim}} [\log(D_{\sigma}(o_A, a_A, \mathbf{0}))] + E_{\tau_{real}} [\log(1 - D_{\sigma}(o_A, a_A, \mathbf{0}))] \quad (13)$$

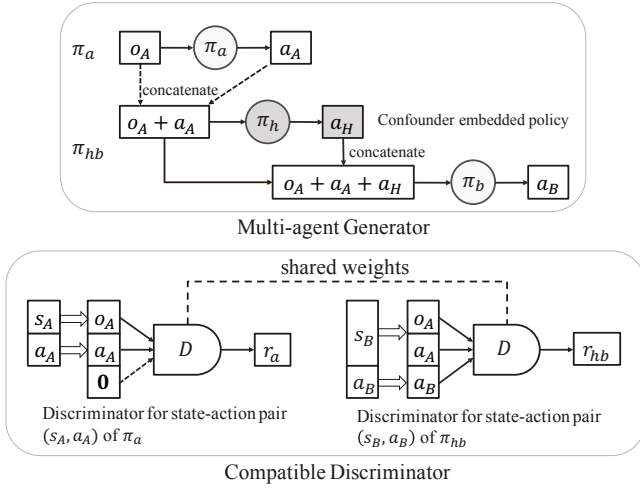


Figure 2: The generator and discriminator in DEMER.

for policy π_a .

The output of the discriminator is the probability that the pair data comes from the real data. The discriminator is trained with supervised learning by labeling the real state-action pair as 1 and the generated fake state-action pair as 0. Then it is used as a reward giver for the policies while simulating interactions. The reward function for policy π_{hb} can be formulated as:

$$r^{HB} = -\log(1 - D(o_A, a_A, a_B)), \quad (14)$$

and the reward function for policy π_a is

$$r^A = -\log(1 - D(o_A, a_A, 0)). \quad (15)$$

3.3 Simulation

We simulate interactions in the generator module. The policy-generated trajectory is generated as follows: Firstly, we randomly sample one trajectory from the observed data and set its first state as the initial observation o_0^A . Then we can use the two policies π_a, π_{hb} to generate a whole trajectory triggered from o_0^A . Given the observation o_t^A as the input of π_a , the action a_t^A can be obtained. In consequence, the action a_t^B can be obtained from the joint policy π_{hb} with the concatenation $\langle o_t^A, a_t^A \rangle$ as input. Then we can get the simulation reward r_t^{HB} by equation (14) and r_t^A by equation (15) which would be used for updating policies in the adversarial training step. Next, we can get the next observation o_{t+1}^A given o_t^A and a_t^B by the predefined transition dynamics. This step is repeated until the terminal state and a fake trajectory is generated.

3.4 DEMER Algorithm

Based on the confounder embedded policy and the compatible discriminator, we propose the DEMER method to achieve the goal of reconstructing environment with hidden confounders from the observed data.

Algorithm 1 shows the details of DEMER. The whole algorithm adopts the generative adversarial training framework. In each iteration, firstly the generator simulates interactions using policies

Algorithm 1 DEMER

Input: $D_{real} = \{\tau_1, \tau_2, \dots, \tau_n\}$: The observed real trajectories over T steps;
 N : Number of trajectories generated in each iteration;
 K : Steps of generator per discriminator step;
1: Initialize parameters θ^{hb} and θ^a of policy π_{hb} and π_a , parameters σ of discriminator D ;
2: **for** $i = 1, 2, \dots$ **do**
3: **for** $k = 1, 2, \dots, K$ **do**
4: $\tau_{sim} = \emptyset$;
5: **for** $j = 1, 2, \dots, N$ **do**
6: $\tau_j = \emptyset$;
7: Randomly sample one trajectory τ_r from D_{real} and set its first state as the initial observation o_0^A ;
8: **for** $t = 0, 2, \dots, T - 1$ **do**
9: Simulate the action $a_t^A = \pi_a(o_t^A)$;
10: Simulate the action $a_t^B = \pi_{hb}(o_t^A, a_t^A)$;
11: Get reward r_t^A according to Equation (15);
12: Get reward r_t^{HB} according to Equation (14);
13: Get next observation o_{t+1}^A given o_t^A, a_t^B by predefined transition;
14: Add $\{o_t^A, a_t^A, a_t^B, r_t^A, r_t^{HB}\}$ to τ_j ;
15: **end for**
16: Add τ_j to τ_{sim} ;
17: **end for**
18: TRPO update θ^a and θ^{hb} with simulation trajectories τ_{sim} according to the equation (10) and (11) respectively;
19: **end for**
20: Update the parameters σ of the discriminator D by minimizing the losses in equation (12) and (13) in turn;
21: **end for**
22: **return** the trained policies π_a, π_b, π_h .

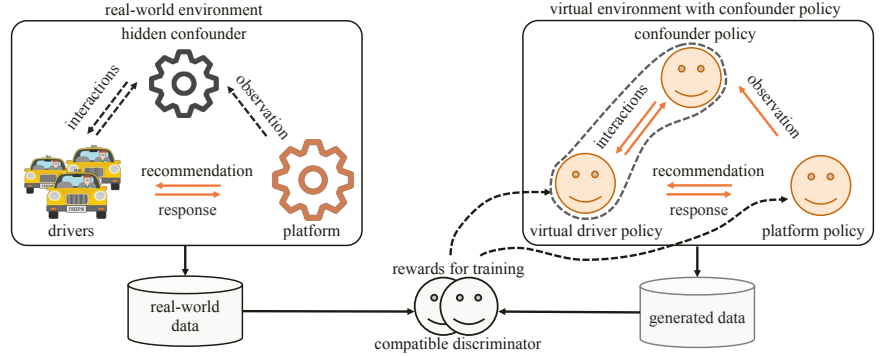
π_a, π_{hb} to collect the trajectory set τ_{sim} corresponding to the line 5 to line 17. Then the policy π_a and π_{hb} are updated in turn using TRPO with generated trajectories τ_{sim} in line 18. After K generator steps, the compatible discriminator is trained by two steps as shown in line 20. Specifically, the predefined transition dynamics in line 13 depends on specific tasks. The DEMER method can effectively imitate the policies of observed interactions and recover the hidden confounder beyond observations.

4 APPLICATION IN DRIVER PROGRAM RECOMMENDATION

4.1 Driver Program Recommendation

We have witnessed a rapid development of on-demand ride-hailing services in recent years. In this economic pattern, the platform often recommends programs to drivers, aimed to help them finish more orders. Specifically, the platform would select the appropriate program to recommend the drivers to participate every day, and then adjust the program content according to the drivers' feedback behavior. This is a typical reinforcement learning task. However, since the behavior of drivers is not only affected by the recommended programs, but also affected by some other unobservable

Figure 3: DEMER framework applied in the driver program recommendation. While real-world data only collects the interactions between the drivers and the Didi Chuxing platform, the virtual environment contains three policies simulating the drivers, the platform, and the confounding variable.



factors, such as the response to special events and so on, that is, hidden confounders. In order to achieve the goal of policy optimization, it is essential to take into account the potential influence of hidden factors when recommending programs.

However, traditional reinforcement learning approaches are applied in these problems without exploring the impact of hidden confounders, which would consequently degrade the learning performance. Thus, a more adaptive approach such as DEMER proposed in this paper is desirable to tackle these problems.

4.2 DEMER based Driver Program Recommendation

As for the driver program recommendation, we apply DEMER to build a virtual environment with hidden confounders by using historical data. As shown in Figure 3, there are three agents in the environment, representing driver policy π_d , platform policy π_p and confounder policy π_h . We can see that the driver policy and the platform policy have the nature of "mutual environment" from the perspective of MDP. From the platform's point of view, its observation is the driver's response, and its action is the recommendation program to the driver. Correspondingly, from the driver's point of view, its observation is the platform's recommendation program, and its action is the driver's response to the platform. The hidden confounder is modeled as a hidden policy same as DEMER, so as to make a dynamic effect at each time step.

Data preparation. Based on the above scenario, we integrate the historical data and then construct historical trajectories $D_{hist} = \{\tau_1, \dots, \tau_i, \dots, \tau_n\}$ representing trajectories of n drivers. Each trajectory $\tau_i = \{o_0^P, a_0^P, o_0^D, o_1^P, \dots, o_t^P, a_t^P, a_t^D, o_{t+1}^P, \dots, o_T^P\}$ represents the T steps of interactions of driver d_i .

Definition of policies. According to the interaction among agents in this scenario, the observation and action of each agent policy are defined as follows:

- platform policy π_p : The observation o_t^P consists of the driver's static characteristics (using real data) and the simulated response behavior a_{t-1}^D ; the action a_t^P is the program information recommended for the driver.
- hidden policy π_h : The observation o_t^H consists of o_t^P and a_t^P ; the action a_t^H is the same format as a_t^P .
- driver policy π_d : The observation o_t^D consists of o_t^P , a_t^P and a_t^H ; the action a_t^D is the simulated driver's behavior at the current step.

Similar to the DEMER setting, we further combine the policies π_h , π_d into a joint policy named π_{joint} . We then apply DEMER to train π_{joint} and π_p . Afterwards, the deconfounding environment of driver program recommendation is reconstructed.

RL in virtual environment. Once the deconfounding virtual environment is built, we perform reinforcement learning efficiently to optimize the policy π_p by interacting with the environment. Due to the simulated confounders in the environment, the reinforcement learning approach could learn a deconfounding policy with improved performance in the real world.

5 EXPERIMENTS

In this section, we conduct two groups of experiments to validate the effect of DEMER method. One is a toy experiment in which we design an environment with predefined rules, the other is a real-world application of driver program recommendation on a large-scale ride-hailing platform Didi Chuxing.

5.1 Artificial Environment

We firstly hand-craft an artificial environment, consisting of the artificial platform policy π_p , the artificial driver policy π_d , and the artificial confounder π_h , with deterministic rules to mimic the real environment. Then we use DEMER to learn the policies and compare with the real rules. Besides, we conduct the MAIL method, without modeling hidden confounders, as a comparison.

Description of the artificial environment. Similar to the interaction in the scenario of driver program recommendation, we define a triple-agents environment to simulate a Markov decision process. The semantic drawing of this toy scenario is shown in Figure 4. In a Markov decision process, the key variant v (denotes the driver's response) is affected by three policies at each time step. The policy π_d has an intrinsic evolution trend on the variant v in the period of 7 time steps, as defined in equation (19). The policy π_p has a positive effect on the variant v if the value of v is under the green line else no effect. Oppositely, the policy π_h has a negative effect on the variant v if the value of v is above the blue line else no effect. The green and blue lines can be seen as the thresholds of π_p and π_h to make effect on the evolution trend of v . Here we set the policy π_h as a role of hidden confounders in this environment, of which the effect on the interaction would not be observed.

MDP definition. The observation o is a tuple (tw, r, v) , in which $tw \in \{1, 2, \dots, 7\}$ is the time step in one period, r is a static factor

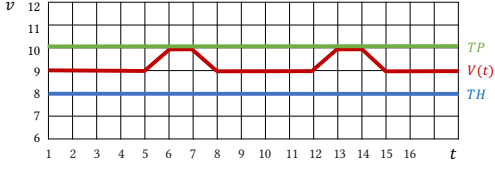


Figure 4: Schematic drawing of interaction in the toy environment: t represents the time step and v is a variant affected by all three policies. TP and TH are the thresholds for policies taking effect, and $V(t)$ describes the intrinsic evolution trend of the artificial driver policy π_d .

used to make a difference on the effect of each agent and v is the key variant in the interaction process. The initial value v_0 is sampled from a uniform distribution $U(9 + wave, 9 - wave)$, $wave = 1.2$. We add the static factor $r = 1 - 0.5 \times \frac{v_0 - 9}{wave}$ into the state to make the episodes generated by this setting more diverse.

The *action* is defined as the output of the deterministic policy. The thresholds of green line TP and blue line TH are 10 and 8 correspondingly. Then we define the deterministic policy rule of each agent as follows:

$$a_p = \pi_p(tw, r, v) = \max(0, \min(1, r \times (TP - v) \times \frac{tw}{7})), \quad (16)$$

$$a_h = \pi_h(tw, r, v, a_p) = \max(-1, \min(0, r \times (TH - v - \frac{a_p}{2}) \times \frac{tw}{7})), \quad (17)$$

$$a_d = \pi_d(tw, r, v, a_p, a_h) = \Delta V(tw) + a_p + a_h. \quad (18)$$

where

$$\Delta V(tw) = \begin{cases} 1 & \text{if } tw = 5; \\ -1 & \text{else if } tw = 7; \\ 0 & \text{otherwise.} \end{cases} \quad (19)$$

The *transition dynamics* is simply defined as: $v_{t+1} = v_t + a_d^t$ and r is a constant once initialized. tw is a timestamp indicator cycling in the sequence $[1, 2, \dots, 7]$. In this experiment, we set the length of trajectory T to 8.

By running in the toy environment, we collect many episodes as training data. Each episode is formatted as $\{o_p^0, a_p^0, a_d^0, o_p^1, \dots, o_p^T\}$. Note that there is no action of policy π_h in the episode.

Implementation details. We conduct two training settings on this artificial environment: DEMER and MAIL. The major difference is that there is no confounding policy in the MAIL setting. We aim to compare the similarity between the generated policies and the defined rules. In detail, each policy is embodied by a neural network with 2 hidden layers and combined sequentially into a joint policy network illustrated in Figure 2. There are 64 neurons in each hidden layer activated by tanh functions. To control the same complexity of the policy model, the joint policy network in MAIL has the same number of hidden layers as DEMER. The discriminator network adopts the same structure as each policy network. Different from GANs training, we perform $K = 3$ generator steps per discriminator step, and sample $N = 200$ trajectories per generator step. The detail of the training process is described in Section 3.

Results. The generated policy functions trained by DEMER and MAIL are shown in Figure 5.

First of all, from the perspective of the two observable policies, the policy function maps of π_p and π_d produced by DEMER are both more similar to the real function space than those by MAIL, as shown in Figures 5 [a] and [c]. MAIL produces sharp distortion shape locally when r is large. We believe that this is because the hidden confounder has a greater impact on the interaction as r increases, and a large confounding bias has reached a point where it cannot be neglected.

Then we further compare the similarity between the confounder policy generated by DEMER and the true policy π_h . In Figure 5 [b], it can be seen the generated confounder policy can describe threshold effects well and match the real function map roughly, although it is difficult under the setting of fully unobservable confounders. Our results show the great potential of using observational data to infer the hidden confounder model.

5.2 Real-world Experiment

In this part, we apply DEMER to a real-world scenario of driver program recommendation as introduced in Section 4.1. Firstly, we use historical data to reconstruct four virtual environments by four comparative methods. Next, we evaluate these environments from various statistical measures. Finally, we train four recommendation policies in these environments by the same training method and evaluate these policies in offline and online environments.

Specifically, we include four methods in our comparison:

- **SUP:** Supervised learning of the driver policy with historical state-action pairs, i.e., behavioral cloning;
- **GAIL:** GAIL to learn the driver policy, given the historical record of program recommendation as a static environment;
- **MAIL:** Multi-agent adversarial imitation learning, without modeling the hidden confounder.
- **DEMER:** The proposed method in this study;

We evaluate the models by different statistical metrics.

Log-likelihood of real data on models. We evaluate the policy distribution of four different models by the mean log-likelihood of real state-action pairs on both training set and testing set. As shown in Table 1, the model trained by DEMER achieves the highest mean log-likelihood on both data sets. Since the evaluation is on the view of each state-action pair, the behavioral cloning method SUP achieves a better performance than MAIL. While our method DEMER makes a significant improvement on MAIL, which indicates the positive influence of our confounder setting.

Correlation of key factors trend. Another important measurement of generalization performance is the trend of drivers' response. We use two indicators' trend lines to compare different simulators: Number of finished orders (FOs) and Total Driver Incomes (TDIs). The same as above, we apply the simulator to a subsequent testing data and simulate the trend of FOs and TDIs. Then we calculate the Pearson correlation coefficient between the simulation trend line and the real. As shown in Table 2, the simulation trend lines of two indicators by DEMER and MAIL achieve high correlations to the real with Pearson correlation coefficient of 0.8 approximately. While the methods SUP and GAIL, trained directly with real data, get bad performance in this evaluation.

Distribution of program response. To compare the generalization performance of models, we apply the built simulators to

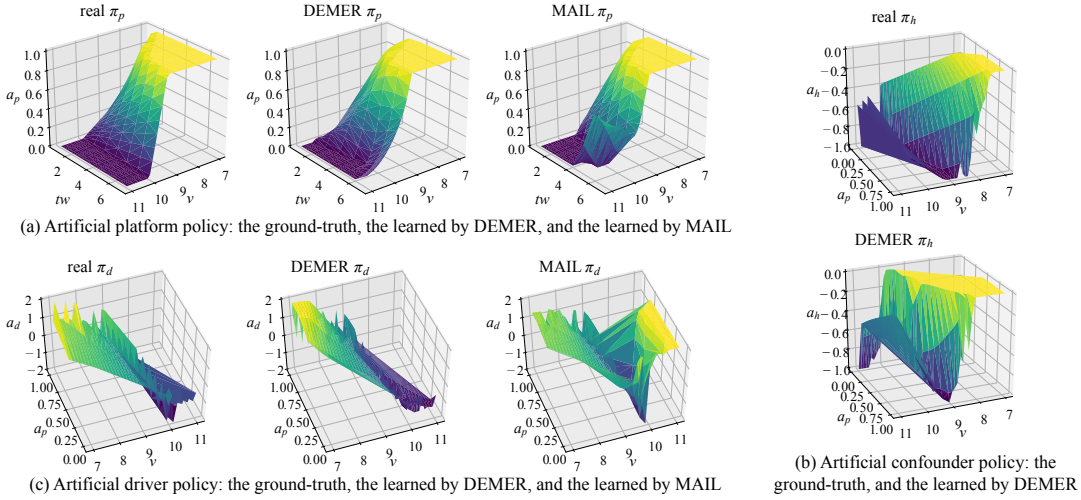


Figure 5: Visualization and comparison of policy functions, with $r = 1.3$. More visualizations with various of r values are presented in the supplement material.

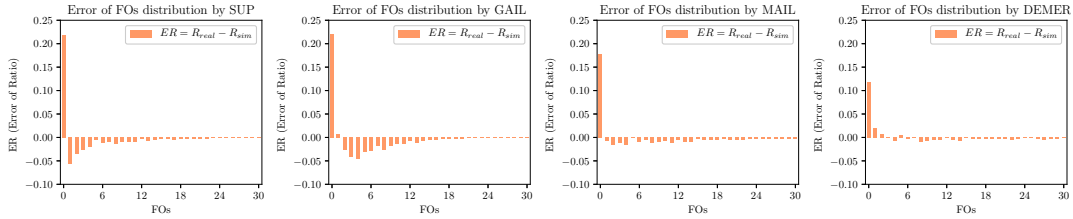


Figure 6: Error of FOs distribution generated by four different methods on testing data. Y-axis is the error of FOs distribution between the simulation and the real data. The original distribution is presented in the supplement material.

Table 1: Comparison of test log-likelihood on real data.

Methods	Training set	Testing set
SUP	17.09	18.00
GAIL	18.43	17.85
MAIL	15.27	14.52
DEMER	21.74	21.21

Table 2: Comparison of Pearson correlation coefficients on FOs and TDIs trend lines.

Methods	FOs	TDIs
SUP	-0.0213	0.0010
GAIL	0.4987	0.4252
MAIL	0.8129	0.7861
DEMER	0.7945	0.8596

subsequent program recommendation records. We simulate the drivers' responses by using real program records on testing data, then compare the simulation distribution of drivers' responses with the real distribution. Here we use FOs as an indicator. Figure 6 shows the error of FOs distributions simulated in four simulators. The

simulation distributions by SUP and GAIL are biased apparently when FOs is low. The reason is that these two methods use whole or partial real data directly for building simulators, which limits the generalization performance of simulators, and the lower FOs means the higher uncertainty, especially zero. Furthermore, the FOs distribution by DEMER is exactly closer to the real than by MAIL, where the confounder setting makes difference explicitly.

Policy evaluation results in semi-online tests. In this part, we evaluate the effect of different simulators for reinforcement learning. Firstly, we use policy gradient method to train a recommendation policy in each simulator. Then we apply MAIL and DEMER respectively to build a virtual environment using testing data for policy evaluation, namely *EvalEnv-MAIL* and *EvalEnv-DEMER*. Given these two environments, we execute the optimized policies and compare the improvement of FOs. As shown in Figure 7, the policy π_{DEMER} optimized in the simulator built by DEMER achieves best performance on both *EvalEnv-MAIL* and *EvalEnv-DEMER*, while the control policies π_{SUP} and π_{GAIL} perform bad on both environments. The promotion by π_{DEMER} compared to π_{MAIL} can further verify that a virtual environment with hidden confounders can bring better performance to traditional reinforcement learning. Besides, the performance of policies π_{SUP} , π_{GAIL} shows a significant degradation in *EvalEnv-DEMER*, while not shown up

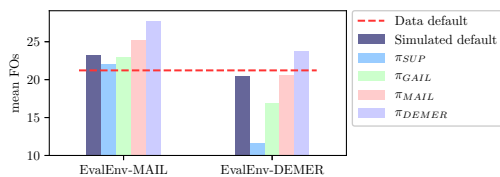


Figure 7: Comparison of different policies trained from different simulators in *EvalEnv-MAIL* and *EvalEnv-DEMER*. Y-axis is the mean FOs by executing different policies. The Data default is the mean FOs in the real testing data. The Simulated default is the mean FOs of the original simulation in each evaluation environments.

Table 3: Results of online A/B tests on the platform of Didi Chuxing. Improvements of FOs and TDIs by policy π_{DEMER} .

Cities	Δ FOs(%)	Δ TDIs(%)
City A	+10.73	+6.16
City B	+10.16	+9.38
City C	+18.47	+17.84
Total	+11.74	+8.71

in *EvalEnv-MAIL*, which indicates that the environment built by DEMER can recover the real environment more precisely.

Policy evaluation results in online A/B tests. We further conduct online A/B tests to evaluate the effect of the policy π_{DEMER} . The online tests are conducted in three cities of different scale. The drivers in each city are divided randomly into two groups of equal size, namely control group and treatment group. The programs for the drivers in the control group are recommended by an existing recommendation policy, which can be viewed as a baseline policy. The drivers in the treatment group are recommended by π_{DEMER} . The results of online A/B tests are shown in Table 3. The proposed policy π_{DEMER} achieves significant improvements on FOs and TDIs in all three cities, and the overall improvements are 11.74% and 8.71% respectively.

6 CONCLUSION

This paper explores how to construct a virtual environment with hidden confounders from observed interactions. We propose the DEMER method following the generative adversarial training framework. We design the confounder embedded policy as an important part of generator and make the discriminator compatible with two different classification tasks so as to guide the optimization of each policy precisely. Further, we apply DEMER to build a virtual environment of driver program recommendation task on a large-scale ride-hailing platform, which is a highly dynamic and confounding environment. Experiment results verify that the policies generated by DEMER can be very similar to the real ones and have better generalization performance in various aspects. Furthermore, the simulator built by DEMER can produce better policy. It is worth noting that the proposed method DEMER can be used not only in this task, but also in many real-world dynamic environments with hidden confounders and can lead to better learning performance.

ACKNOWLEDGMENTS

We would like to thank Prof. Yuan Jiang for her constructive suggestions to this work.

REFERENCES

- [1] Brenna Argall, Sonia Chernova, Manuela M. Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57, 5 (2009), 469–483.
- [2] Elias Bareinboim, Andrew Forney, and Judea Pearl. 2015. Bandits with Unobserved Confounders: A Causal Approach. In *Advances in Neural Information Processing Systems* 28. 1342–1350.
- [3] Chelsea Finn, Paul F. Christiano, Pieter Abbeel, and Sergey Levine. 2016. A Connection between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models. *arXiv abs/1611.03852* (2016).
- [4] Andrew Forney, Judea Pearl, and Elias Bareinboim. 2017. Counterfactual Data-Fusion for Online Reinforcement Learners. In *Proceedings of the 34th International Conference on Machine Learning*. 1156–1164.
- [5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems* 27. 2672–2680.
- [6] Jonathan Ho and Stefano Ermon. 2016. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems* 29. 4565–4573.
- [7] Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard S. Zemel, and Max Welling. 2017. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems* 30. 6449–6459.
- [8] Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. 2018. Deconfounding Reinforcement Learning in Observational Settings. *arXiv abs/1812.10576* (2018).
- [9] Jacob Menick and Nal Kalchbrenner. 2018. Generating High Fidelity Images with Subscale Pixel Networks and Multidimensional Upscaling. *arXiv abs/1812.01608* (2018).
- [10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmash Kumar, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533.
- [11] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
- [12] Dean Pomerleau. 1991. Efficient Training of Artificial Neural Networks for Autonomous Navigation. *Neural Computation* 3, 1 (1991), 88–97.
- [13] Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 627–635.
- [14] Stuart J. Russell. 1998. Learning Agents for Uncertain Environments (Extended Abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. 101–103.
- [15] Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences* 3, 6 (1999), 233–242.
- [16] John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. 2015. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 1889–1897.
- [17] Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and Anxiang Zeng. 2018. Virtual-Taobao: Virtualizing Real-world Online Retail Environment for Reinforcement Learning. *arXiv abs/1805.10000* (2018).
- [18] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.
- [19] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction (2nd Edition)*. MIT Press.
- [20] Zeyang Ye, Keli Xiao, Yong Ge, and Yuefan Deng. 2019. Applying Simulated Annealing and Parallel Computing to the Mobile Sequential Recommendation. *IEEE Transactions on Knowledge and Data Engineering* 31, 2 (2019), 243–256.
- [21] Zeyang Ye, Lihao Zhang, Keli Xiao, Wenjun Zhou, Yong Ge, and Yuefan Deng. 2018. Multi-User Mobile Sequential Recommendation: An Efficient Parallel Computing Paradigm. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2624–2633.

A SUPPLEMENT MATERIAL

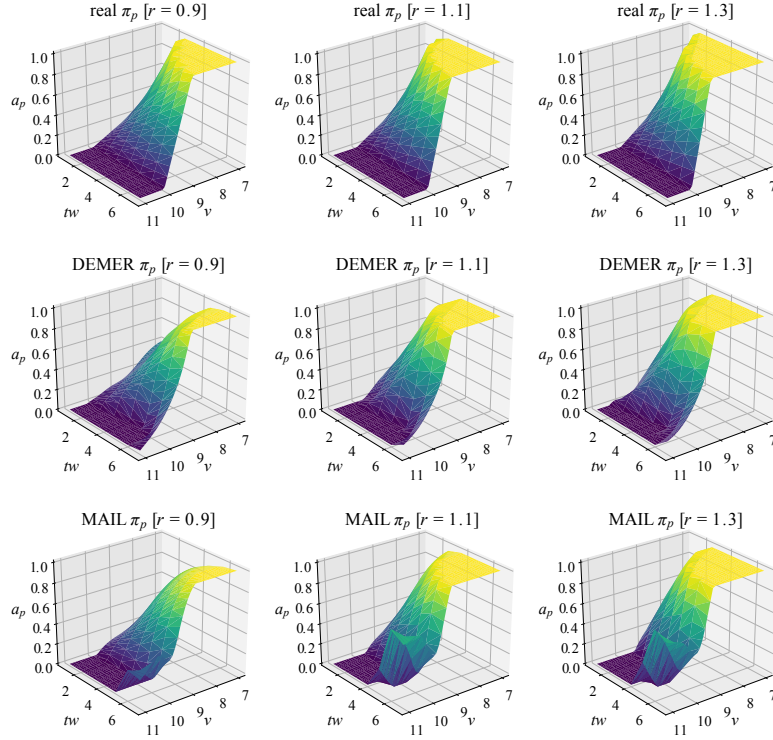


Figure 8: Visualization of the artificial platform policy function π_p with respect to v and tw on different values of r . The first line is the ground-truth rule function. The second line is the policy function generated by DEMER and the third line corresponds to MAIL.

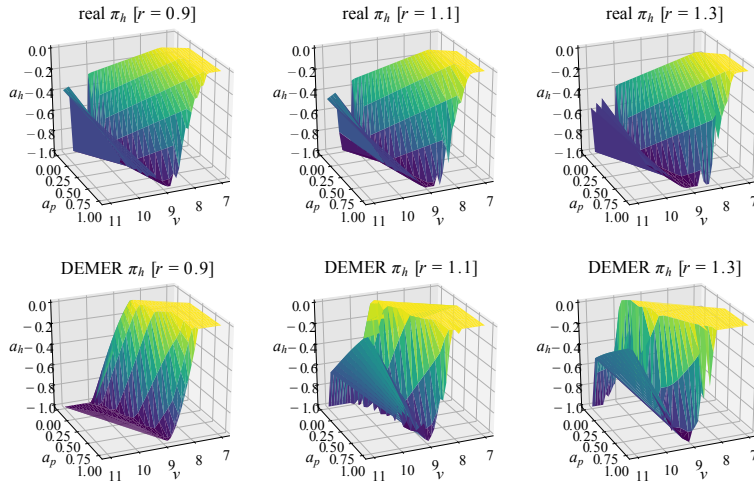


Figure 9: Visualization of the artificial confounder policy function π_h with respect to v and a_p on different values of r . The first line is the ground-truth rule function. The second line is the policy function generated by DEMER.

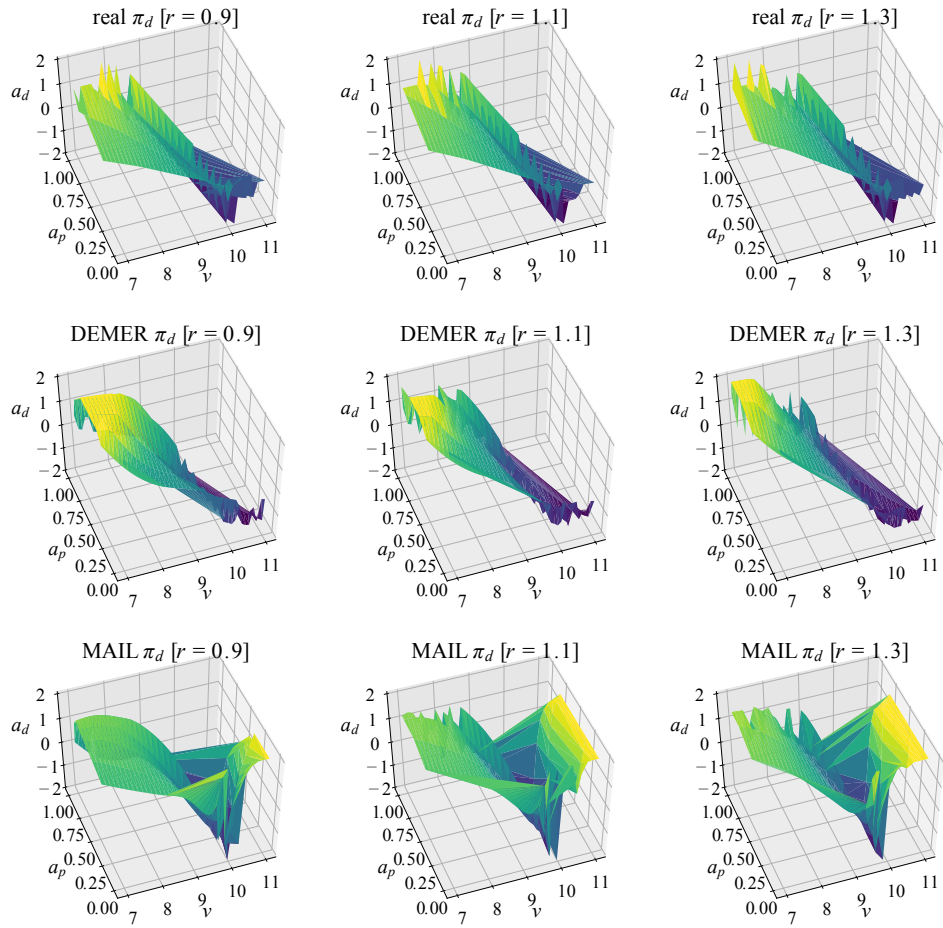


Figure 10: Visualization of the artificial driver policy function π_d with respect to v and a_p on different values of r . The first line is the ground-truth rule function. The second line is the policy function generated by DEMER and the third line corresponds to MAIL.

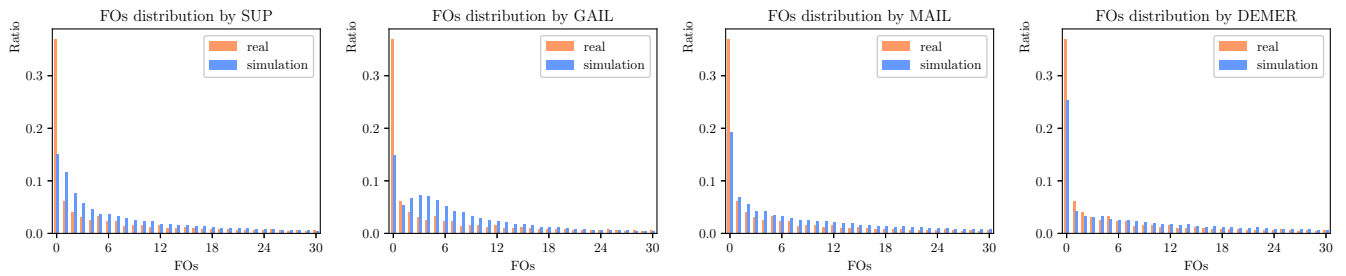


Figure 11: The original FOs distribution generated by four different methods on testing data. Y-axis is the ratio of FOs.