

RESEARCH ARTICLE

Membership Inference Attacks Against Temporally Correlated Data in Deep Reinforcement Learning

MAZIAR GOMROKCHI^{1,2,*}, SUSAN AMIN^{1,2,*}, HOSSEIN ABOUTALEBI^{3,*},
ALEXANDER WONG^{3,4}, (Senior Member, IEEE), AND DOINA PRECUP^{1,2}

¹Department of Computer Science, McGill University, Montréal, QC H3A 2A7, Canada

²Mila-Québec AI Institute, Montréal, QC H2S 3H1, Canada

³VIP Laboratory, Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

⁴DarwinAI Corporation, Waterloo, ON N2V 1K4, Canada

Corresponding author: Maziar Gomrokchi (gomrokma@mila.quebec)

*Maziar Gomrokchi, Susan Amin, and Hossein Aboutalebi contributed equally to this work.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

ABSTRACT While significant research advances have been made in the field of deep reinforcement learning, there have been no concrete adversarial attack strategies in literature tailored for studying the vulnerability of deep reinforcement learning algorithms to membership inference attacks. In such attacking systems, the adversary targets the set of collected input data on which the deep reinforcement learning algorithm has been trained. To address this gap, we propose an adversarial attack framework designed for testing the vulnerability of a state-of-the-art deep reinforcement learning algorithm to a membership inference attack. In particular, we design a series of experiments to investigate the impact of temporal correlation, which naturally exists in reinforcement learning training data, on the probability of information leakage. Moreover, we compare the performance of *collective* and *individual* membership attacks against the deep reinforcement learning algorithm. Experimental results show that the proposed adversarial attack framework is surprisingly effective at inferring data with an accuracy exceeding 84% in individual and 97% in collective modes in three different continuous control Mujoco tasks, which raises serious privacy concerns in this regard. Finally, we show that the learning state of the reinforcement learning algorithm influences the level of privacy breaches significantly.

INDEX TERMS Adversarial attack, deep reinforcement learning, membership inference attack, privacy.

I. INTRODUCTION

Despite the recent advancements in the design and performance of deep reinforcement learning (deep RL) algorithms in complex domains ([1], [2], [3]), the vulnerability of these models to privacy breaches has only begun to be explored in the literature. In particular, while there have been a few studies on the vulnerability of deep RL models to adversarial attacks [4], [5], [6], there has been no study on the potential membership leakage of the data directly employed in training deep RL models, which is known as membership inference attacks (MIAs). The potential success of such MIAs can have

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Yan¹.

serious security ramifications in the deployment of models resulting from deep RL.

One of the major challenges in the implementation of MIAs in deep RL settings is the sequential and correlated nature of deep RL data points. Unlike in deep supervised settings, a data point in deep RL algorithms may consist of hundreds of correlated components in the form of tuples, all together forming a single trajectory. A successful MIA algorithm against a deep RL model should be able to learn not only the relation between the training and output trajectories but also the correlation between the tuples within each trajectory (data point). Another complication in this regard concerns the relationship between the training and prediction data points. In deep RL settings, batches of collected input

data are used for training the deep RL policy. Thus, each output data point corresponds to every single data point in the training batches. This feature is in contrast with, for instance, that of data points in text generation problems (e.g. machine translation or dialogue generation systems), where there is a direct (usually one-to-one) correspondence between the input and output sequential data points. Finally, RL algorithms are learning systems where the concept of labels is not defined as it is in supervised learning methods. Instead, during the learning phase, the deep RL agent receives reinforcement (aka rewards) from the environment as the outcome of the selected action. The deep RL agent uses the obtained rewards to learn the task and optimize its learning policy. These factors lead to complications in defining input-output pairs in training attack classifiers and subsequently establishing a meaningful relationship between the pair constituents.

Deep RL methods have a unique structural difference compared to deep supervised or unsupervised methods, *i.e.* learning based on temporal correlation between the tuples in each trajectory and partial reinforcements the model receives upon interaction with the underlying environment. Even though deep RL models decorrelate input trajectories through an intermediate mechanism called *replay buffer* (for more information, refer to the Background section), the inherent correlation between transition tuples still plays a significant role in the feature representation learned by the deep RL model [7], hence the behaviour of the output policy. In this regard, two natural questions arise:

- 1) How much information (concerning the training data points) can an adversary extract from the output of a trained deep RL model?
- 2) To what extent can an adversary benefit from feature correlation in the learned policy?

This study presents the first black-box MIA against a deep RL agent to address these two questions. In our proposed adversarial attack framework, the target model is considered a black box; thus, the attacker does not have access to the internal structure of the deep RL agent. In particular, the attacker can only access the model output in the form of trajectories τ_T^{out} resulting from the trained policy π_f . We use *batch off-policy reinforcement learning* setting, where the common practice is that an (unknown) *exploration policy* (behaviour policy) π_b collects private data points in the form of a batch of trajectories. The batch data is thereafter delivered to the deep RL algorithm in the form of independent trajectories (Markov chains) to train the target policy. In this setting, the RL agent decouples the data collection phase from the policy training phase (*i.e.* off-policy). In the off-policy setting, the learning system is not tied to a particular exploration algorithm and ensures disjointedness between the training data sets provided for the RL algorithm in different settings. Off-policy setup is particularly preferred in designing MIA frameworks in black-box settings, where neither the internal structure of the target model nor the exploration policy used to collect the training trajectories is known to the adversary.

Our proposed attack framework tests the vulnerability of a state-of-the-art off-policy deep RL model to MIA in two modes: *individual* and *collective*. In the individual mode, the attacker's goal is to train a probabilistic model that infers the membership probability of a single trajectory τ_T^{in} given the trained policy π_f and the initial state s_0 . In this case, the goal is to measure the extent to which the adversary can exploit trajectory-level temporal correlation to reveal the presence of a trajectory in the training set. In the collective mode, the attacker's target is to predict the membership probability of a collection of data points. In this mode, the goal is to measure the extent to which the adversary is capable of exploiting not only the trajectory-level temporal correlation but also the batch-level correlation to reveal the presence of a trajectory in the training set. We show that the deep RL model is more vulnerable to collective MIA as in this mode, the attack classifier has access to more information.

Moreover, we assess the vulnerability of the RL algorithm to MIA in terms of the learning state of the algorithm. Our results show that the cumulative amount of reinforcement the RL agent obtains in the course of training the policy is proportional to the level of its vulnerability to MIAs. Finally, to determine the role of data correlation in the vulnerability of the deep RL model to MIA, we disturb the correlation within the data points used to train the attack classifier and subsequently compare the impact of training the attacker with the resulting decorrelated trajectories on the performance of the attack classifier. We observe that the presence of correlation within the trajectories helps the adversary discern between the member and non-member data points with higher probability compared with the results obtained from the decorrelated case.

The contributions of this study summarize as follows:

- We present the first black-box membership inference attack against a deep RL agent.
- The proposed model utilizes batch off-policy reinforcement learning setting for data collection, which ensures disjointedness between training data sets and is preferred in black-box MIA frameworks.
- We introduce two modes of membership inference attack: individual and collective, to measure the extent of exploiting temporal and batch-level correlation in revealing training set presence.
- This study demonstrates that the deep RL model is more vulnerable to collective MIA, as the attack classifier has access to more information in this mode.
- In this study, we assess the vulnerability of the deep RL algorithm to MIAs in terms of the learning state, demonstrating that the cumulative amount of reinforcement obtained during policy training is proportional to the vulnerability of the deep RL model.
- Finally, this study investigates the impact of data correlation on vulnerability to MIAs and shows that the presence of correlation within trajectories helps the adversary discern between member and non-member data points with higher probability.

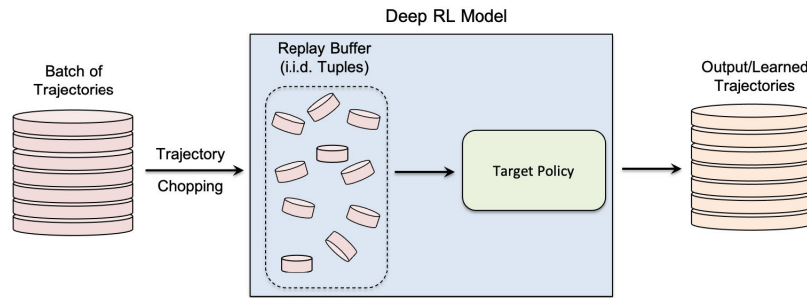


FIGURE 1. Batch off-policy RL learning architecture. An external behaviour policy generates a batch of trajectories composed of transition tuples (state, action, reward, new state). The trajectories are subsequently passed to the deep RL model. The replay buffer mechanism, as an internal part of the model, decorrelates each trajectory into a collection of i.i.d. transition tuples and then uses them in the form of mini-batches to train the target policy.

II. BACKGROUND

In this section, we provide the background information in two parts: *i*) a general introduction to reinforcement learning systems, and *ii*) membership inference attacks.

A. REINFORCEMENT LEARNING

In reinforcement learning (RL) systems, an agent learns a task through a sequence of trial and error and receives rewards through environmental interactions. The agent's task is formalized as a stochastic process that is described by a Markov Decision Process (MDP). An MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, p_0 \rangle$ consisting of a set of states \mathcal{S} , a set of actions \mathcal{A} , a transition probability kernel $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \text{Pr}(\mathcal{S})$, a reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$, and an initial state distribution p_0 that characterizes the initial state of each episode. At each time step $t = 0, 1, 2, \dots, T - 1$, the agent is at the environment state $s_t \in \mathcal{S}$ and selects action $a_t \in \mathcal{A}$ according to the policy $\pi(a_t|s_t)$. The policy $\pi : \mathcal{S} \rightarrow \text{Pr}(\mathcal{A})$ is the agent's action-selection strategy, which maps the current state to a distribution over actions and is updated throughout the learning process. Upon taking action a_t , the environment determines the agent's next state s_{t+1} via the transition probability kernel $\mathcal{P}(s_{t+1}|s_t, a_t)$ and returns the reward r_t computed by the reward function $\mathcal{R}(s_t, a_t)$.

The RL agent's goal is to maximize the rewards received in the long run. The cumulative reward that the RL agent receives after time step t is called *return*, defined as $G_t^\pi := \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, where the discount factor $\gamma \in [0, 1]$ determines the weight of the future rewards. The value of each state at time t under policy π is called the *state-value function* $V^\pi(s_t)$, and is defined as the expected return when the agent starts at s_t and follows the policy π :

$$V^\pi(s_t) = \mathbb{E}_\pi \{G_t | s_t\}. \quad (1)$$

Similarly, we can determine the value of a state s_t and action a_t taken at time t (on the condition that we follow the policy π afterwards) using the notion of *action-value function* $Q^\pi(s_t, a_t)$, defined as

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi \{G_t | s_t, a_t\}. \quad (2)$$

The ultimate goal of the RL agent is to learn an effective policy that maximizes the value functions.

In the context of RL, a data point in a batch of data is a sequence of temporally correlated tuples (s_t, a_t, r_t, s_{t+1}) that denote the history of the RL agent's interaction with the environment. This sequence of tuples is often referred to as *trajectory*, and is denoted as

$$\tau_T = (s_0, a_0, r_1, s_1), \dots, (s_{T-2}, a_{T-2}, r_{T-1}, s_{T-1}). \quad (3)$$

RL agents do not have knowledge of the environment at the very initial stage of learning and acquire the necessary experience through continued interactions with the environment. An RL agent can acquire the necessary information in two ways: *on-policy* and *off-policy*. In the on-policy setting, the agent uses the target policy that is trained so far to obtain data through interaction with the environment. In the off-policy setting, on the other hand, the agent receives an input batch of data in the form of trajectories provided by an exploratory agent (*i.e.* *behaviour policy* π_b), and subsequently uses the acquired data to train the *target policy* π_f (exploitation). The output of the trained target policy consists of data points (trajectories) produced as a result of the interaction between the target policy and the environment (Figure 1). From the privacy point of view, since private data is assumed to exist a priori, off-policy methods are natural choices to be analyzed in this regard. Figure 1 presents a schematic of off-policy deep RL architecture.

The fact that the input trajectories in off-policy deep RL models are temporally correlated necessitates the use of a mechanism that converts the input data to *i.i.d.* samples before passing it to the deep network. A widespread and fundamental data management mechanism that has become an inevitable part of the existing off-policy deep RL models is *experience replay buffer* or *replay buffer*. The main intuition behind the application of replay buffer in deep RL lies at the heart of RL theory. It is well-studied that fundamental RL algorithms (*e.g.* Q-learning) easily diverge in the case of linear function approximation [8], [9]. The solution that replay buffer offers to the divergence problem of these algorithms is to decorrelate the input trajectories and subsequently treat

each transition tuple as an i.i.d. sample point (Figure 1). This intermediate decorrelation step significantly improves data efficiency and helps the deep RL algorithm converge to the optimal policy according to the law of large numbers. Moreover, it allows the deep RL algorithm to benefit from mini-batch training and shuffling techniques, which are proven to improve the performance of deep RL algorithms significantly [10], [11], [3], [12], [13], [14].

B. MEMBERSHIP INFERENCE ATTACK

In machine learning, a membership inference attack (MIA) or tracing attack [15], [16], [17] is a form of adversarial attack that is designed to infer the presence of a particular data point in the training set of a target model. For example, an attacker can collect a set of emails, some of which are part of the original training dataset used to train a spam classification model and use the trained model's predictions to train a binary classifier that distinguishes between in-dataset and out-of-dataset emails. By inputting new emails into the binary classifier, the attacker can infer if the new email was likely part of the training dataset or not, raising privacy concerns about the potential leakage of sensitive data used to train the model.

The central intuition in the design of MIAs is that publicly available trained models tend to exhibit higher confidence in their predictions of the individuals who participated in the training data. Consequently, the members of training sets are vulnerable to privacy threats [16], [18]. The main challenge for the adversary in MIAs is to design a classifier compatible with the target model domain setting and decide whether a particular data point was part of the training set given the training target model's output. Attackers employ different MIA design strategies based on: i) the adversary's knowledge level of the parameters in the target model (Label-only strategy) and ii) the adversary's knowledge level of the training data (Shadow model technique).

In the *Label-only* strategy [19], [20], the attacker only relies on model predictions and discards the model's confidence scores. In this technique, the attacker uses the generalization gap (the difference between the train and test accuracy) in the attack model as the main driver in inferring the membership of individuals used in training the target model. The label-only technique was first introduced by Yeom et al. [19] and was subsequently extended by Choquette et al. [20] to show how the label-only technique can improve the existing attack baselines. As the notion of the label is not defined in the general RL setting, the label-only technique cannot be applied here in devising MIAs against RL models.

Shadow model technique [16] is known as an effective and practical approach for designing MIA models. Shadow models are parallel local models trained on data sets often sampled from the same distribution as the underlying distribution of the private data. In this method, the adversary trains the shadow models with complete knowledge of the training set. Thus, using the auxiliary membership information and

the trained shadow models, the adversary can build a membership classifier that identifies whether an individual has participated in the training of similarly trained models.

In the training phase in both *label-only* and *shadow model* techniques, the adversary should have access to the model output labels and the training data true labels. However, the sequential nature of the training and output data points and the temporal nature of model training make the design of MIAs for RL models fundamentally different. Moreover, the presence of a replay buffer as an inevitable part of off-policy deep RL models adds another level of complexity to the design of MIAs, as this intermediate transformation phase adds a new source of noise to the data from the attacker's perspective.

III. RELATED WORK

MIAs were used for the first time against machine learning systems by Shokri et al. [16]. In the following years, extensive studies were performed on the application of MIAs against supervised ([16], [18], [19], [21], [22]) and unsupervised ([23], [24], [25]) machine learning models, surveyed comprehensively by Hu et al. [26], and Rigaki and Garcia [27]. This section reviews the existing attack models against supervised and unsupervised models trained on sequential data.

MIAs have been executed against aggregate location time-series [28], [29], [30]. For the first time, Pyrgelis et al. [30] studied the impact of different spatial-temporal factors that contribute to the vulnerability of time-series-based algorithms to MIAs. MIAs have also been studied in the context of text generation problems [31], [32], where the attacker's goal is to identify whether or not a specific sequence-to-sequence or sequence-to-word pair is part of the input training data of a machine translation engine, a dialogue system or a sentimental recommendation system. The structure of machine learning algorithms with sequential data differs from that of classic classification tasks in the input and prediction types. While inputs and outputs in standard classification problems have fixed sizes, they are chains of correlated elements with variable lengths in sequence generation tasks. This difference poses a fundamentally different approach to designing MIAs against sequence generation tasks. The knowledge of output space distribution is no longer valid for the attack classifier since the output length may vary from one model to another. To tackle this challenge, Song and Shmatikov [31] assume access to a probability distribution over output-space vocabularies. They [31] split their proposed attack model into two phases, shadow model training and audit model training. In the shadow model training phase, the attacker trains multiple shadow models assuming that the attacker has access to a generative model that generates a sequence of vocabularies. In the audit training phase, the attacker uses the rank of the words produced by the target model instead of the output probability distribution. The central assumption is that the gap observed between the trained model rank predictions depends on word frequencies in the training and test sequences. In a similar study, Hisamoto et al. [32] address

MIA against sequence-to-sequence models in a setting where the adversary is agnostic to the word sequence distribution. In their work, the attacker is equipped with a generative model for different translation subcorpora, an alternative for output word sequence distribution.

While in models trained on sequential data, the input-output relation is well defined and deterministic, in deep RL models, the output data are generated through the trained policy; thus, each output sequence can be considered as evidence for the entire input dataset. Therefore, one requires a fundamentally different approach in designing MIAs against RL algorithms. To the best of our knowledge, there is no prior work in the context of deep RL that addresses the problem of membership inference at a microscopic level, where the attacker infers the membership of a particular data point in the training set of deep RL models [26], [27].

IV. METHODS AND EXPERIMENTS

In our proposed adversarial attack framework, we successfully conduct MIA against deep RL in a black-box setting, where only the model output is accessible to external users. The deep RL model interacts with an environment whose distribution of initial states, state space \mathcal{S} and action space \mathcal{A} are common knowledge, an assumption widely accepted in the RL community [33], [34], [35]. In this section, we first explain the general setting of the problem and subsequently introduce our attack platform and our proposed method of data formatting for training the attack models. We further mention the different settings we have considered in our experimental design. Finally, we provide our choices of performance measures to assess the behaviour of the attack model.

A. GENERAL SETUP

We propose an adversarial attack method for studying the vulnerability of the deep RL algorithm to MIA in a black-box setting, where the attacker's access to the model is limited to the output trajectories of the model trained on a private batch of input trajectories. Figure 2 depicts the general framework of our proposed black-box attack against deep RL algorithms.

The two important oracles that should accompany the end-to-end design of a black-box attack model in off-policy deep RL are i) the data oracle $\mathcal{O}_{\text{data}}$, and ii) the model trainer oracle $\mathcal{O}_{\text{train}}$. The data oracle interacts with the environment and returns a set of independent and identically distributed (i.i.d.) training trajectories (Markov chains) for the model trainer oracle $\mathcal{O}_{\text{train}}$ (see Figures 2 (a, b)). The data oracle is a black box which is equipped with a set of unknown exploration policies. To train the target model, whose training input is of the adversary's interest, the data oracle is initialized privately (see Figure 2 (a)), leading to the generation of a batch of private training data points in the form of trajectories. The model trainer oracle is agnostic to the exploration policy used for the data collection. The training data batch is passed to the deep RL trainer oracle, and the resulting trained model is made publicly available

for data query. Our experimental framework can adopt any of the existing off-policy batched deep RL models as the deep RL trainer oracle. In this study, we choose to work with the state-of-the-art Batch-Constrained deep Q-learning (BCQ) [36] model, which is widely used as the basis of other deep RL algorithms and exhibits remarkable performance in complex control tasks. Structurally, BCQ trains a generative model on the input trajectories such that the model learns the relationship between the visited states in the input trajectories and the corresponding actions taken. The BCQ algorithm subsequently uses the developed generative model to train a deep Q-network, which ultimately learns to sample the highest-valued actions similar to the ones in the input trajectories.

We use the *shadow model* [16] training technique to acquire the data needed for training the attack classifier. In this method, through the data oracle, the attacker provides the deep RL trainer oracle with a set of *non-private* training trajectories (Figure 2 (b)), on which the deep RL model is trained. The attacker subsequently queries output trajectories from the trained deep RL model and passes the training and output trajectories to the data formatter (Figure 2 (c)). In this step, the training-output trajectories are augmented into pairs and are subsequently labelled as 1 in positive pairs and 0 in negative ones depending on whether or not the trajectories belong to the same trained model. Finally, the *attack trainer* trains a probabilistic classifier that takes as input the pairs of trajectories prepared by the data formatter and returns a trained probabilistic attack classifier that is subsequently used to infer the membership of target input trajectories (Figure 2 (d)).

Since the attack training data collected by the data oracle $\mathcal{O}_{\text{data}}$ and prepared by the data formatter is of a sequential nature, we need to adopt an attack model that is compatible with time-series data. The classifier should minimize the expected loss, defined as

$$\mathbb{E}_{\mathcal{D}} [l(A_{D,\theta}(\cdot, \pi_f), g(\cdot))] \approx \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} l(A_{\theta}(\tau, \pi_f), g(\tau, \pi_f)), \quad (4)$$

where $g(\cdot)$ is the function that assigns labels to the formatted pairs, $A(\cdot)$ is the parameterized classifier, and $l(\cdot)$ is the loss function adopted by A . The dataset \mathcal{D} contains a set of i.i.d. trajectories drawn from \mathcal{D} , and π_f denotes the policy trained on \mathcal{D} . The goal of the attacker is to train a classifier that learns a parameter vector (or network) θ^* that minimizes the loss function. The following sections provide more details regarding the data formatter and attack classifier.

B. EXPERIMENTAL SETUP

In our experimental design, we study the vulnerability of the deep RL model to MIAs in terms of the following factors:

1) *the membership inference mode (collective vs. individual MIA)* - In the individual mode, the adversary's goal is to infer the membership of *single* training data points (trajectories), while in the collective mode, the adversary's target is

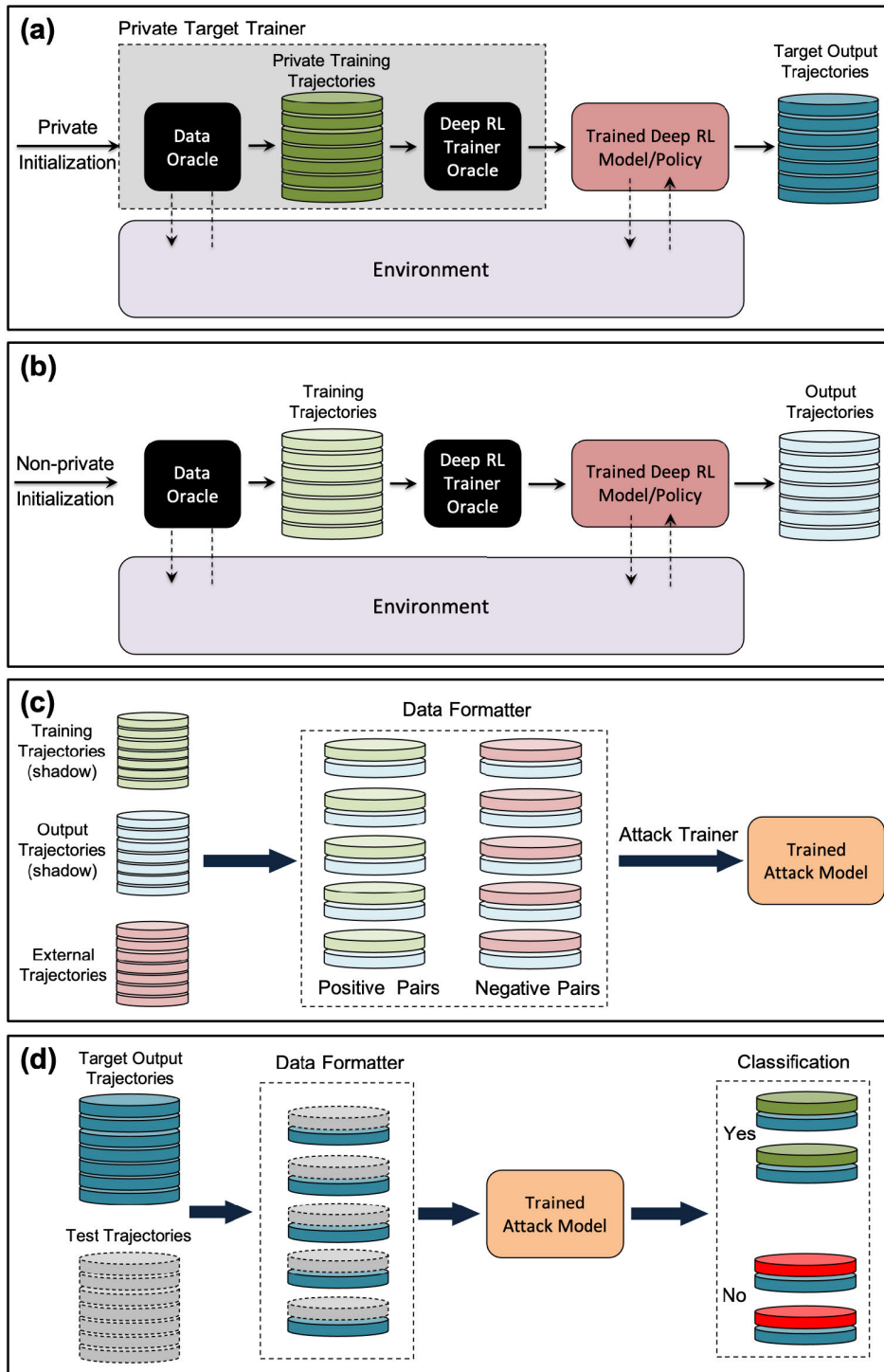


FIGURE 2. The proposed black-box MIA architecture. (a) Private deep RL model training: the black-box exploration engine (data oracle) interacts with the environment and provides private training trajectories for the black-box deep RL model trainer. The trained deep RL model is subsequently used to output target trajectories via interaction with the environment. (b) Shadow training: the data oracle is used to produce non-private training trajectories to be used as input to train the deep RL trainer oracle. The trained deep RL model subsequently generates the output trajectories. (c) Training the attack classifier: the input and output trajectories obtained in part (b) are paired together in *data formatter* to provide *positive* training pairs. Another set of trajectories, which has not been used in training the shadow model, is used with the output trajectories from part (b) to create *negative* training pairs. The attack model is subsequently trained using the paired trajectories with the corresponding *positive* and *negative* labels. (d) Membership inference attack: the target output trajectories from part (a) are paired with sample test trajectories in the *data formatter*. The trained attack model subsequently uses the pairs to infer the test set trajectories that were used to train the private deep RL model.

a *batch* of trajectories used in the training of the deep RL model. In this experimental setup, we aim to address two scenarios in which a participant's identity is revealed. In the individual mode, a trajectory reveals a user's identity, while in the collective mode, a collection of trajectories represents the user's identity.

2) *the maximum trajectory length T_{max} within each episode* - The value of T_{max} is determined and fixed by the environment during data collection and model training. In particular, the RL agent's trajectory in each episode ends when either the agent arrives at an absorbing state at $T < T_{max}$ or the number of time steps $T = T_{max}$. Larger T_{max} corresponds to larger values of return (cumulative reward), thus an improved deep RL policy.

3) *the level of correlation within the input trajectories used to train the attack classifier* - In the case of individual MIA, we study the performance of our proposed attack classifier in two modes: 1) *correlated* mode, where the adversary is trained on pairs with undisturbed input trajectories, 2) *decorrelated* mode, where in the attack training phase, the input trajectory is formed by sampling tuples at random from the whole batch. This set of experiments provides useful information regarding the effect of the correlation level within the input trajectories on the performance of the attack model.

Below is a detailed description of the environments used in our experimental design, the data formatting technique, and the attack architecture.

1) ENVIRONMENTS AND RL SETTING

We assess the algorithm on OpenAI Gym environments [37] powered by MuJoCo physics engine [38], which are standard tasks adopted by many recent RL studies [39], [40], [41], [42], [43]. Gym provides a variety of simulated locomotion tasks with different action and state space dimensionalities. Here, we train the deep RL agent on three high-dimensional continuous control tasks: *Hopper-v2* ($\mathcal{A} \subset \mathbb{R}^3$ and $\mathcal{S} \subset \mathbb{R}^{11}$), *Half Cheetah-v2* ($\mathcal{A} \subset \mathbb{R}^6$ and $\mathcal{S} \subset \mathbb{R}^{17}$), and *Ant-v2* ($\mathcal{A} \subset \mathbb{R}^8$ and $\mathcal{S} \subset \mathbb{R}^{11}$). Starting from virtually zero knowledge of how each task works, the deep RL model's goal is to teach the Hopper how to hop, the HalfCheetah how to run, and the Ant how to walk as fast as possible. We use the Deep Deterministic Policy Gradient (DDPG) algorithm [39] as the data oracle \mathcal{O}_{data} and Batch-Constrained Deep Q-Learning (BCQ) [36] as the batch off-policy deep RL method used in the trainer oracle \mathcal{O}_{train} .

2) DATA AUGMENTATION

Each trajectory starts with an initial state s_0 drawn from the available distribution of initial states in the environment, followed by action a_0 selected and taken by the RL agent. The environment subsequently takes the agent to the next state s_1 and returns the reward r_1 . The agent's next choice of action is based on s_1 , and this cycle continues until the trajectory ends at s_T . In other words, the initial state s_0 plays a significant role in determining the sequence of actions taken

by the RL policy and the consequent states and rewards. Thus, to prepare training pairs to train the attack classifier, we pair the training and output trajectories that have the same initial states, fixing the starting point of the two trajectories in a pair. Moreover, as the RL agent interacts with MDP, the resulting trajectory is a Markov chain, *i.e.* every state and reward in the trajectory is the direct consequence of the previous state and action. Therefore, we choose to remove states and rewards from the trajectories, keep the actions in the trajectory, and use them in the pairing process.

Each task is equipped with a set of absorbing states $\mathcal{B} \in \mathcal{S}$. An absorbing state is a state that leads to the termination of an agent's chain of interactions with an environment. Due to the presence of absorbing states in the environment, the generated trajectories have different lengths. To pair the training and output action trajectories obtained from the deep RL model, we need to either increase the length of shorter action trajectories to match that of the longest one or clip longer action trajectories to a pre-determined length. Based on the desired length, we choose to repeat the last action in shorter action trajectories for the required number of times and trim longer trajectories. Each action trajectory is a $d^A \times T$ dimensional array, where d^A is the dimension of action space, and T is the total number of actions in the trajectory. Every output action trajectory is concatenated with a training trajectory such that the resulting pair is a $2d^A \times T$ dimensional array. The pairs are subsequently passed to the attack classifiers in multi-dimensional arrays $\mathbb{R}^{2d^A \times T}$ and $\mathbb{R}^{2d^A \times T \times m}$ in individual and collective modes, respectively. The value m refers to the number of pairs in each batch in the collective mode.

3) ATTACK CLASSIFIER ARCHITECTURE

We use Temporal Convolutional Networks (TCNs) [44] as the classifier for individual MIA, and Residual Network (ResNet) [45] deep architecture for collective MIA. Figure 3 shows a schematic of TCN (Figure 3(a)) and ResNet (Figure 3(b)) architectures.

Individual-Mode Attack Classifier Architecture- As both training and output trajectories of RL models are composed of temporally correlated transition tuples, the choice of attack classifier must utilize the input-level temporal correlation in its feature representation. TCNs are structurally designed to utilize the inherent temporal correlation in the training data through a hierarchy of temporal convolutions architecture. In this regard, TCN employs a 1D fully-convolutional network (FCN) architecture [46], where each of its hidden layers has the same length as the input layer (Figure 3(a)). The main advantage of TCN is its ability to use dilation in convolution layers to keep the long-range temporal dependency and increase the receptive field of the convolutional layers. In the individual MIA mode, since the input data to the classifier is a pair of temporally correlated tensors (*i.e.* $\mathbb{R}^{2d^A \times T}$), the long-range correlation between input tuples within each trajectory is well-aligned with the input structure

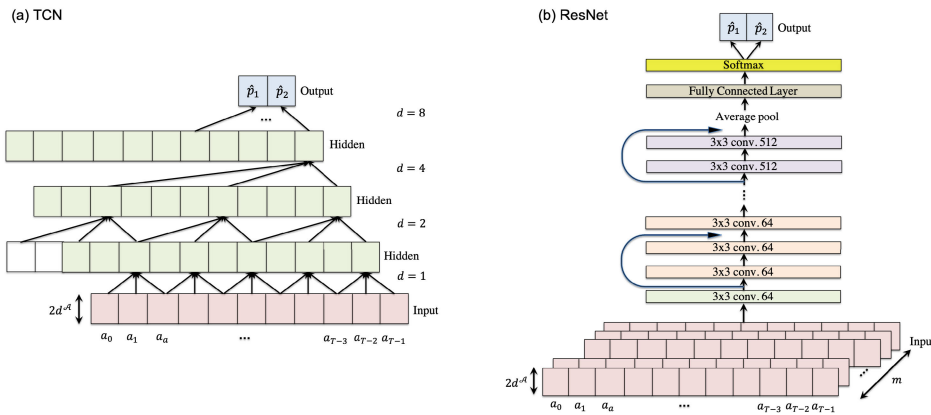


FIGURE 3. The network architecture of TCN (a) and ResNet (b) used in the individual and collective MIAs, respectively.

of TCNs. For more information on the internal structure of TCN architecture, refer to the Supplementary Information.

Collective-Mode Attack Classifier Architecture- In this case, while more information is accessible to the attacker, it requires a more complex learning architecture and more sophisticated hyper-parameter tuning to exploit the cross-correlation among the training trajectories and the temporal correlation within a trajectory. In the collective mode, our input is in the form of a three-dimensional tensor (e.g. $\mathbb{R}^{2d^A \times T \times m}$). Unlike the individual MIA mode, which involves 2-dimensional inputs, in the collective MIA mode, we have another dimension m for the number of trajectories in each batch of trajectories, similar to the data structure used in image classification problems [45], [47], [48]. Thus, we use the deep residual network (ResNet) architecture [45] because of its inherent compatibility with data sets with temporally deep structures. ResNets are popular for solving standard computer vision problems [45], [49], [50].

C. PERFORMANCE METRICS

We adopt the standard performance metrics used in the classification literature [51] to evaluate the performance of our proposed attack models. We measure the performance of the attack classifier with the following metrics:

Overall accuracy (ACC), which captures the overall performance of the attack classifier and is calculated as follows,

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

where TP (true positives) denotes the number of correctly recognized positives, and TN (true negatives) shows the number of correctly recognized negative ones. The two other quantities, false positives FP and false negatives FN indicate the number of incorrectly recognized positives and negatives, respectively.

Precision (PR), which shows the fraction of pairs classified as matching pairs that are indeed coming from the same model, and is written as, $PR = TP/TP + FP$.

Recall (RE), which measures the fraction of matching pairs that the attack classifier can infer correctly, and is computed as, $RE = TP/TP + FN$.

F1 score (F1), which is the harmonic mean of the precision (PR) and recall (RE), and is calculated as $F1 = (2 \cdot PR \cdot RE)/(PR + RE)$.

Matthews Correlation Coefficient (MCC) [52], which calculates the correlation between the predicted and the true classification labels, and is defined as,

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{6}$$

MCC is an effective and meaningful combination of all four quantities TP, TN, FP, and FN, and ranges from -1 to 1 . The closer MCC is to 1 , the better the model performs [53]. $MCC = 0$ shows that the model is a random guesser. The other evaluation metrics ACC, PR, RE, and F1 vary in the $[0, 1]$ range. In a well-performing model, all of these evaluation metrics have values close to 1 . Finally, to show the performance of our proposed MIA classifiers in individual and collective modes at different classification thresholds θ , we plot the receiver operating characteristic (ROC) curve, which shows the changes of recall RE as a function of False Positive Rate $FPR = FP/(FP + TN)$ for different values of θ .

V. RESULTS AND DISCUSSION

This section presents and discusses the results of different experimental scenarios to capture the interdependence between different parameters that affect the accuracy of membership inference in deep RL settings.

A. COLLECTIVE VS. INDIVIDUAL MIAs

Using different classification metrics, we assess the behaviour of TCN and ResNet attack classifiers in predicting the membership probability of individual and collective data points, respectively. Figure 4 presents the performance of the classifiers TCN and ResNet in Hopper-v2 (Figure 4(a)),

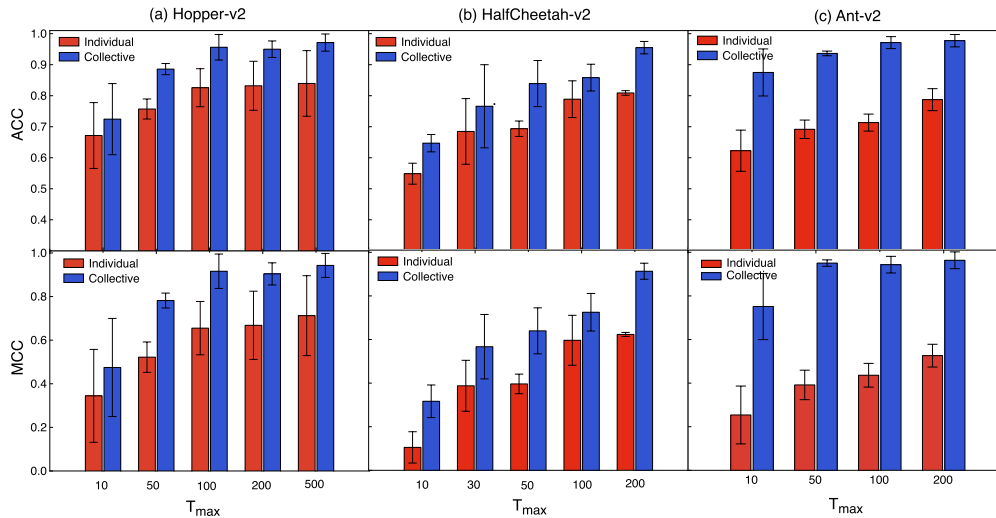


FIGURE 4. The performance of the attack classifiers in tasks Hopper-v2 (a), HalfCheetah-v2 (b), and Ant-v2 (c) in individual and collective attack modes. Each data point is determined from the average result of 5 separate runs. The error bars depict the error on the mean for ACC (top) and MCC (bottom) in the corresponding runs. The batch size $m = 50$ in the collective mode.

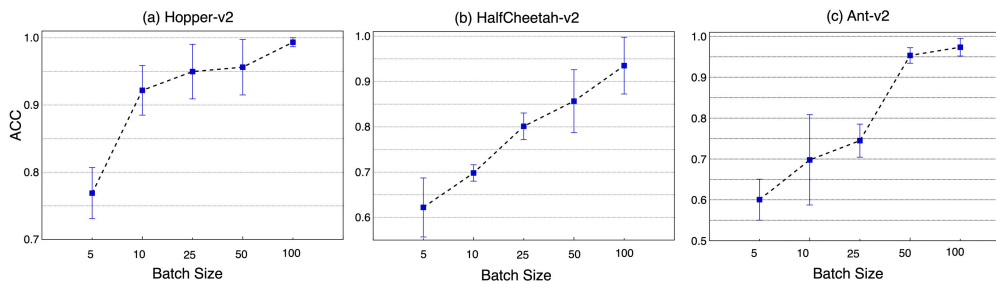


FIGURE 5. The MIA accuracy in Hopper-v2 (a), HalfCheetah-v2 (b), and Ant-v2 (c) in the collective attack mode for different batch sizes. Each data point is determined from the average result of 5 separate runs. The error bars depict the error on the mean. The maximum trajectory length $T_{max} = 100$.

HalfCheetah-v2 (Figure 4(b)), and Ant-v2 (Figure 4(c)) in terms of ACC and MCC for different maximum trajectory lengths T_{max} . The full report of their performance in these three tasks is provided in Section II of the Supplementary Information file. The results show that our proposed attack framework can infer the RL model training data points with high accuracy (e.g. > 0.8 in the individual and > 0.9 in the collective mode for $T_{max} \geq 100$ in Hopper-v2), indicating a high risk of privacy invasion. Moreover, the results reveal that for a fixed T_{max} , the adversary infers collective data points with significantly higher accuracy than the accuracy value in the individual mode. For example, in the Hopper-v2 task, the membership inference accuracies in the collective mode for T_{max} are more than 12% higher than those in the individual mode. This observation shows that the deep RL algorithm is more vulnerable to MIA in the collective mode, which is expected since more information is provided to the attack classifier through a batch of data points instead of one. In particular, in the collective mode, the adversary can capture the collective properties of the training data points and their relationship with the output trajectories, which could be veiled in one individual trajectory.

To further study the effect of *batch size* on the performance of MIA in the collective mode, we conduct MIAs against the deep RL agent for different batch sizes (Figure 5). A closer analysis of the two figures (Figure 4 and Figure 5) reveals that while larger batch sizes correspond to a higher level of deep RL training members' vulnerability to MIA, batch sizes $m \leq 5$ in Hopper-v2, $m < 25$ in HalfCheetah-v2, and $m \leq 10$ in Ant-v2 lead to smaller values of inference accuracy compared with those in the individual mode. We believe that this difference in the performance of the adversary between the two cases corresponds to the different structures used in the individual and collective modes (i.e. TCN and ResNet). In particular, our results show that the effectiveness of the ResNet classifier in inferring the membership of the data points surpasses that of TCN in larger batch sizes.

B. THE IMPACT OF T_{max}

We test the performance of attack classifiers against the target model for different values of T_{max} in a set of experiments. As the environment is unvarying, the value of T_{max} remains unchanged throughout each experiment. Our observations presented in Figure 4 show that as T_{max} increases, the

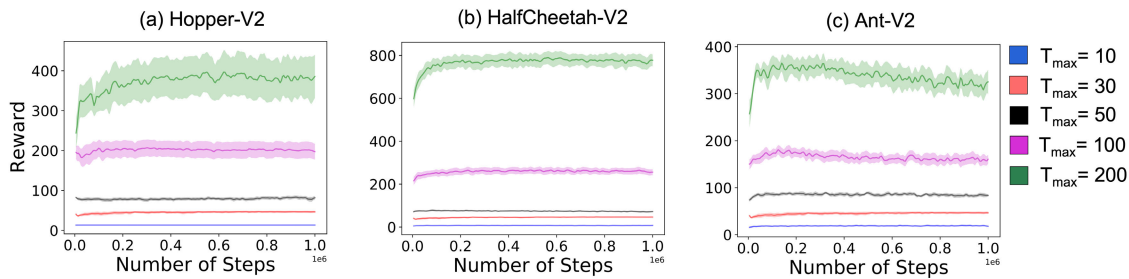


FIGURE 6. Deep RL curves in three high-dimensional locomotion tasks Hopper-v2 (a), HalfCheetah-v2 (b), and Ant-v2 (c). The graphs depict the performance of the deep RL model as a function of time for different maximum trajectory lengths T_{max} . The plots are averaged over 5 random seeds. The performance of the deep RL policy is assessed every 5000 step over 100000 time steps.

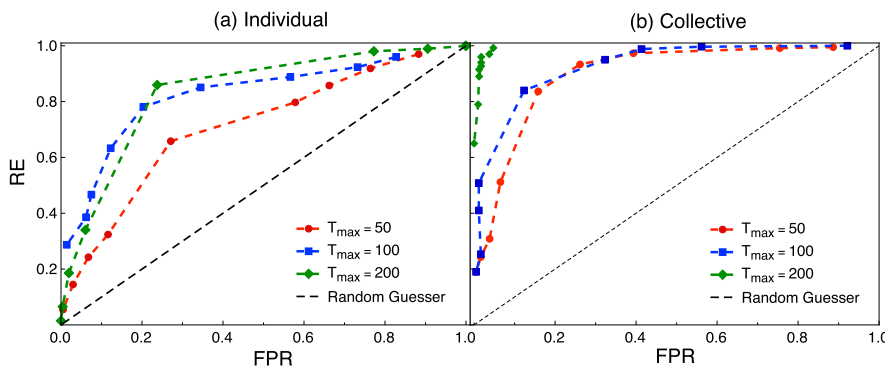


FIGURE 7. The receiver operator characteristic (ROC) curves of the MIA in HalfCheetah-v2 in the individual (a) and collective (b) modes for different values of T_{max} .

accuracy ACC of the attack classifiers in inferring the training data points in both individual and collective modes improves. Moreover, our results show consistent improvement of MCC as a function of T_{max} in all three environments Hopper-v2, Half Cheetah-v2, and Ant-v2, which is consistent with the changes in ACC. Note that as MCC utilizes all four values in the confusion matrix, it provides a more reliable and robust measure compared to the other metrics (for additional results and comparison, refer to the tables provided in the Supplementary Information file).

Maximum trajectory length T_{max} plays a significant role in the performance of deep RL models. Figure 6 illustrates the learning curves for the deep RL model in Hopper-v2 (Figure 6(a)), HalfCheetah-v2 (Figure 6(b)), and Ant-v2 (Figure 6(c)) for different values of T_{max} . The plots show that as T_{max} increases, the deep RL policy presents a consistently improved behaviour. As RL policy is a function that maps the visited states to the selected actions, a closer deep RL policy to the optimal policy corresponds to a more predictable relationship between the training and the output trajectories. We argue that this feature of deep RL policies contributes to the higher level of vulnerability of the deep RL models that are trained with larger values of T_{max} .

As the attack classifiers output membership probabilities, we determine the predicted binary label for a range of acceptance thresholds $\theta = 0.1, 0.2, \dots, 0.9$, and subsequently

choose the threshold θ , at which the classifier shows the highest performance. Figure 7 depicts the sample ROC curves for HalfCheetah-v2 in individual (Figure 7(a)) and collective (Figure 7(b)) modes. The plots show that larger values of T_{max} lead to better performance of the attack classifiers. The best result is obtained at $T_{max} = 200$ in both individual and collective modes. We find that the acceptance threshold $\theta = 0.5$ yields the highest performance throughout all of our experiments.

C. TEMPORAL CORRELATION

The results presented so far exhibit the performance of the MIAs against deep RL as a result of training the attack classifiers on the temporally correlated data collected from the training set and the output of the deep RL model. Considering that the training trajectories are decorrelated in the replay buffer as the first step after entering the deep RL trainer oracle, a question arises as to what role the temporal correlation in the data set plays in the vulnerability of deep RL models to MIAs. To answer this question, we have performed a set of experiments where prior to the data augmentation phase, the temporal correlation between the deep RL training trajectories is broken. In particular, we decorrelate the training trajectories by shuffling their constituent tuples. We subsequently store the decorrelated transition tuples in an auxiliary buffer. In the next step, we generate trajectories of

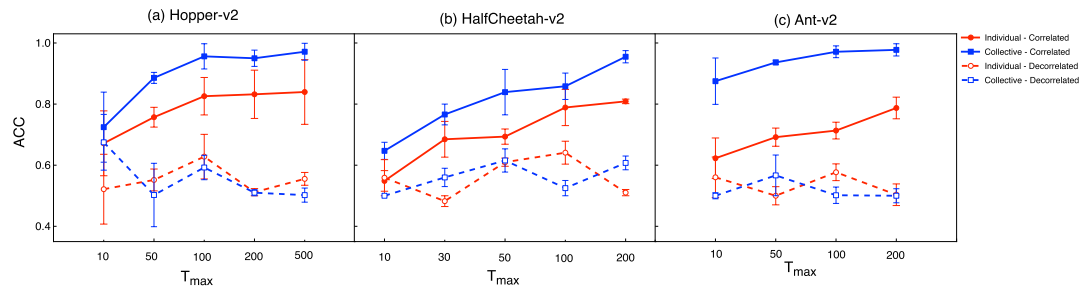


FIGURE 8. Comparison of the MIA accuracy between correlated and decorrelated settings for Hopper-v2 (a), HalfCheetah-v2 (b), and Ant-v2 (c).

the desired length by sampling actions uniformly at random from the buffer. Finally, we pass the collection of decorrelated training trajectories together with output trajectories to the data augmentation mechanism and train the attack classifiers with the paired trajectories in the individual and collective modes. Figure 8 compares the accuracy of the MIA in the correlated mode with that in the decorrelated mode for the three tasks. The plots depict that the adversary's accuracy in inferring RL training members decreases significantly upon decorrelating the training trajectories. The results show that despite the inevitable input decorrelation imposed by the replay buffer mechanism in the training phase of off-policy deep RL models, the temporal correlation in the training trajectories is channelled to the model output data points. Thus, the attack classifiers trained on temporally correlated training data points exhibit higher accuracy than those trained on decorrelated trajectories.

VI. CONCLUSION

In this study, we design and evaluate the first membership inference attack (MIA) framework against off-policy deep RL in collective and individual membership inference modes by exploiting the inherent and structural temporal correlation present in deep RL data points. We demonstrate the performance of the proposed adversarial attack framework in complex high-dimensional locomotion tasks for different maximum trajectory lengths. Our proposed attack framework reveals the substantial vulnerability of a state-of-the-art off-policy deep RL model to the black-box MIAs. We show that it is significantly more vulnerable to MIA in the collective setting when compared to its vulnerability in individual MIAs. Moreover, our results demonstrate a consistent increase in the accuracy of the membership inference as a function of batch size in the collective mode. Furthermore, our experimental results reveal that the maximum trajectory length (in the episodic RL setting), which is set by the environment, plays a significant role in the vulnerability of the training data used in the deep RL model to the MIA. We show that a longer maximum trajectory length leads to an improved deep RL policy, thus a more defined relationship between the training and output trajectories, and consequently less private training data. Moreover, our results reveal the determinative role of temporal correlation in obtaining high MIA performance,

which the attacker can utilize to design high-accuracy MIAs against deep RL. Despite the existence of replay memory as an intermediate data decorrelation mechanism at the heart of deep RL models, the trained policy still fully exploits the inherent correlation in learning feature representation, which poses a significant privacy concern in the deployment of trained RL policies at the industrial scale. Finally, the results from this study highlight serious privacy concerns in the widespread deployment of similar deep RL models, which demand more investigation of this matter to offer solutions in future studies. The tasks employed in the current study are under the umbrella of robotics simulation tasks that motivate the extension of experiments to real-world robot learning tasks. Moreover, dialogue systems such as Amazon Alexa, Apple Siri, and Google Assistant are other interesting future platforms to apply RL-based MIAs on. In virtual dialogue systems, a data point is presented by a collection of interaction trajectories between the chatbot and the end user. A chatbot in this setting is the trained RL policy, and the user interactions with the bot form the training trajectories. In such settings, the collective mode is the natural inference setting since a collection of user interactions with the bot represents individual identity in the training set. In other words, the user's presence in the training set can be inferred by the adversary if and only if the attacker correctly infers a batch of trajectories representing the individual in the training set. Another extension to this line of research is to investigate MIAs against Deep RL models in a white-box setting, where the internal structure of the target policy is also known to the adversary.

ACKNOWLEDGMENT

The authors would like to thank Hamidreza Ghafghazi and Spencer Main for their valuable contribution to the design and development of the preliminary version of the codebase. Computing resources were provided by Compute Canada, Calcul Québec, and the VIP Laboratory at the University of Waterloo throughout the project, which the authors appreciate.

REFERENCES

- [1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, and K. Tunyasuvunakool, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

- [2] P. R. Wurman, "Outracing champion Gran Turismo drivers with deep reinforcement learning," *Nature*, vol. 602, no. 7896, pp. 223–228, 2022.
- [3] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, and T. Hubert, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.
- [4] X. Pan, W. Wang, X. Zhang, B. Li, J. Yi, and D. Song, "How you act tells a lot: Privacy-leakage attack on deep reinforcement learning," 2019, *arXiv:1904.11082*.
- [5] A. Gleave, M. Dennis, N. Kant, C. Wild, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," in *Proc. ICLR*, 2020, pp. 1–16.
- [6] X. Wu, W. Guo, H. Wei, and X. Xing, "Adversarial policy training against deep reinforcement learning," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 1883–1900.
- [7] B. Mavrin, H. Yao, and L. Kong, "Deep reinforcement learning with decorrelation," 2019, *arXiv:1903.07765*.
- [8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
- [9] M. Fairbank and E. Alonso, "The divergence of reinforcement learning algorithms with value-iteration and function approximation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–8.
- [10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, and A. Graves, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [11] S. Zhang and R. S. Sutton, "A deeper look at experience replay," 2017, *arXiv:1712.01275*.
- [12] R. Liu and J. Zou, "The effects of memory replay in reinforcement learning," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 478–485.
- [13] W. Fedus, P. Ramachandran, R. Agarwal, Y. Bengio, H. Larochelle, M. Rowland, and W. Dabney, "Revisiting fundamentals of experience replay," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3061–3071.
- [14] A. Seyyedabbasi, R. Aliyev, F. Kiani, M. U. Gulle, H. Basyildiz, and M. A. Shah, "Hybrid algorithms based on combining reinforcement learning and metaheuristic methods to solve global optimization problems," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 107044.
- [15] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! A survey of attacks on private data," *Annu. Rev. Statist. Appl.*, vol. 4, no. 1, pp. 61–84, 2017.
- [16] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.
- [17] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2022, pp. 3093–3106.
- [18] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2019, pp. 11–14.
- [19] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE 31st Comput. Secur. Found. Symp. (CSF)*, Jul. 2018, pp. 268–282.
- [20] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1964–1974.
- [21] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "A pragmatic approach to membership inferences on machine learning models," in *Proc. IEEE Eur. Symp. Secur. Privacy*, Sep. 2020, pp. 521–534.
- [22] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 1–6.
- [23] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Membership inference attacks against generative models," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 1, pp. 133–152, 2019.
- [24] B. Hilprecht, M. Härterich, and D. Bernau, "Monte Carlo and reconstruction membership inference attacks against generative models," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 4, pp. 232–249, Oct. 2019.
- [25] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 343–362.
- [26] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," 2021, *arXiv:2103.07853*.
- [27] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," 2020, *arXiv:2007.07646*.
- [28] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "What does the crowd say about you? Evaluating aggregation-based location privacy," *Proc. Privacy Enhancing Technol.*, vol. 4, pp. 76–96, Jan. 2017.
- [29] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "Knock knock, who's there? Membership inference on aggregate location data," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, 2018, pp. 1–16.
- [30] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "Measuring membership privacy on aggregate location time-series," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 4, no. 2, pp. 1–28, Jun. 2020.
- [31] C. Song and V. Shmatikov, "Auditing data provenance in text-generation models," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 196–206.
- [32] S. Hisamoto, M. Post, and K. Duh, "Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system?" *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 49–63, Dec. 2020.
- [33] R. S. Sutton, "Temporal credit assignment in reinforcement learning," Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, 1984.
- [34] G. Vietri, B. Balle, A. Krishnamurthy, and S. Wu, "Private reinforcement learning with PAC and regret guarantees," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9754–9764.
- [35] C. Szepesvári, *Algorithms for Reinforcement Learning*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2010.
- [36] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2052–2062.
- [37] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," 2016, *arXiv:1606.01540*.
- [38] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 5026–5033.
- [39] 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," 2015, *arXiv:1509.02971*.
- [40] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.
- [41] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
- [42] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–4.
- [43] V. Francois-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," 2018, *arXiv:1811.12560*.
- [44] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
- [45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [46] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [48] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [49] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2021.
- [50] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

- [51] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.
- [52] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica Biophysica Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, 2016.
- [53] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, no. 1, pp. 1–22, Feb. 2021.



MAZIAR GOMROKCHI is currently a Senior Machine Learning Scientist with System1 and a Research Affiliate with Mila–Québec AI Institute. Before joining System1, he was honored to serve as a Researcher and the Ph.D. Student with Mila–Québec AI Institute and McGill University. His current research interests include the intersection of reinforcement learning and data privacy.



SUSAN AMIN received the Ph.D. degree in physics from McGill University. She is a Postdoctoral Fellow with McGill University and Mila–Québec AI Institute. She has a strong interest in using concepts from statistical physics in machine learning. Her current research interests include the connection between statistical physics and reinforcement learning.



HOSSEIN ABOUTALEBI is currently pursuing the Ph.D. degree with the University of Waterloo. His recent works focus on adversarial attacks and defenses. Previously, he was with Cerebras, as a Machine Learning Engineer. His main research interests include computer vision and natural language processing applications.



ALEXANDER WONG (Senior Member, IEEE) is the Co-Director of the Vision and Image Processing Research Group and a Professor with the Department of Systems Design Engineering, University of Waterloo. He has published over 600 refereed journals, conference papers, and patents, with a particular interest in AI explainability, responsible AI, adversarial machine learning, and generative AI. He is the Canada Research Chair of artificial intelligence and medical imaging, a fellow of the Institution of Engineering and Technology, and a member of the College of the Royal Society of Canada.



DOINA PRECUP holds several positions in the field of artificial intelligence. She is a Canada CIFAR AI Chair and a Research Fellow in the Machines and Brains Program, CIFAR. She is also an Associate Professor of computer science with McGill University and a Core Member of Mila–Québec AI Institute. She also leads the research team with DeepMind.

...