THEORETICAL UNDERSTANDING OF ADVERSARIAL REINFORCEMENT LEARNING VIA MEAN-FIELD OPTIMAL CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial reinforcement learning has been shown promising in solving games in adversarial environments, while the theoretical understanding is still premature. This paper theoretically analyses the convergence and generalization of adversarial reinforcement learning under the mean-field optimal control framework. A new mean-field Pontryagin's maximum principle is proposed for reinforcement learning with implicit terminal constraints. Applying Hamilton-Jacobi-Issacs equation and mean-field two-sided extremism principle (TSEP), adversarial reinforcement learning is modeled as a mean-field quantitative differential game between two constrained dynamical systems. These results provide the necessary conditions for the convergence of the global solution to the mean-field TSEP. The global solution is also unique when the terminal time is sufficiently small. Moreover, two generalization bounds are delivered via Hoeffding's inequality and algorithmic stability. Both bounds do not explicitly depend on the dimensions, norms, or other capacity measures of the parameter, which are usually prohibitively large in deep learning. The bounds help characterize how the algorithm randomness facilitates the generalization of adversarial reinforcement learning. Moreover, the techniques may be helpful in modeling other adversarial learning algorithms.

1 INTRODUCTION

Adversarial reinforcement learning (Uther and Veloso, 1997) has been successfully deployed in many application areas, including autonomous driving (Behzadan and Munir, 2019; Pan et al., 2019) and AI gaming (Mandlekar et al., 2017; Pinto et al., 2017; Zhang et al., 2020). Adversarial neural networks are employed for solving the games in the adversarial environments (Mandlekar et al., 2017). Moreover, the adversarial neural networks can also improve feature robustness and sample efficiency (Ma et al., 2018). However, the theoretical understanding of the convergence and generalization in adversarial reinforcement learning is still premature.

In this paper, we establish the theoretical foundations of adversarial reinforcement learning under the mean-field optimal control framework (Bensoussan et al., 2013). Our contributions are summarized as follows.

(i) Reinforcement learning is modeled from the view of the dynamical system. A new mean-field Pontryagin's maximum principle (PMP) (Pontryagin, 1987) is proposed to give the necessary condition for optimality of this dynamical system. This PMP is extended from E et al. (2019) to cover the dynamical systems with constraints on the terminal time. This extension helps generalize the applicable domains to cover control problems with constraints, such as controlling target (e.g., a vehicle) to reach a certain area.

(ii) Adversarial reinforcement learning is modeled as a mean-field quantitative differential game (Pontryagin, 1985); and thus, its corresponding training process is regarded as how to achieve the optimal control of this game. The mean-field two-sided extremism principle (TSEP) (Guo et al., 2005) is then presented, which relies on the loss function and terminal constraints. This mean-field TSEP serves as the necessary conditions of the convergence (or equivalently, the optimality) of the mean-field quantitative differential game; when the terminal time is small

enough, this mean-field TSEP is also a unique solution, and thus serves as the sufficient conditions of the convergence.

(iii) The learned model of adversarial reinforcement learning is characterized by the viscosity solution (E et al., 2019) of a mean-field Hamilton-Jacobi-Issacs (HJI) equation (Guo et al., 2005). We then prove that this viscosity solution is unique. The HJI equation gives a global characterization of adversarial reinforcement learning, while the previously given mean-field TSEP is a local special case.

(iv) Two generalization error bounds are proved, which characterize the gap between the expected mean-field solution and the learned model. The two bounds are obtained from Hoeffding's inequality (Pinelis and Sakhanenko, 1986) and algorithmic stability (Mou et al., 2018). The bounds are of order $\mathcal{O}\left(e^{-N}\right)$ and $\mathcal{O}(1/N)$, respectively, where N is the number of samples. They do not explicitly rely on the dimensions, norms, or other capacity measures of the network parameter, which are usually prohibitively large in deep learning. Moreover, the latter bound helps characterize how the randomness in the training algorithms and the aggregated step sizes influence the generalization.

Some previous works have been devoted to establish the theoretical foundations of deep learning by dynamical system viewpoint, since E (2017). Based on the PMP and the method of successive approximation (Kantorovitch, 1939), new optimization methods are developed by Li et al. (2018); Li and Hao (2018). Sonoda and Murata (2017) study the continuum limit of training neural networks, and Chang et al. (2018b;a); Haber and Ruthotto (2017) contribute to the design of network architecture based on dynamical systems and differential equations. E et al. (2019) propose to employ mean-field optimal control formulation for explaining deep learning. They prove the mean-field optimality conditions of both the Hamilton-Jacobi-Bellman type and the Pontryagin type (Pontryagin, 1987). Similar results are given by Persio and Garbelli (2021) through associating deep learning with stochastic optimal control (Guo et al., 2005) from the perspective of mean-field games (Lasry and Lions, 2007). These mean-field results reflect the probabilistic nature of the reinforcement learning.

To the best of our knowledge, this is the first work on developing theoretical foundations for adversarial reinforcement learning. Compared with the existing studies on deep learning theory from the mean-field optimal control view, this work models a machine learning algorithm as a mean-field quantitative differential game between two dynamical systems, rather than a single dynamical system. Our work may inspire novel designs of optimization methods for adversarial reinforcement learning. Moreover, the techniques may be of independent interest in modeling other adversarial learning algorithms, including generative adversarial networks (Goodfellow et al., 2020; Liu and Tuzel, 2016; Mao et al., 2017), and solving partial differential equations (Zang et al., 2020).

2 PRELIMINARIES

In this section, we bridge reinforcement learning with mean-field optimal control problems. We first present the optimal control formulation of deep learning as introduced in (Li et al., 2018; Li and Hao, 2018; E, 2017). A deep residual network with K layers can be represented by

$$x(k+1) = x(k) + f(x(k), \theta(k)), \quad k = 0, \cdots, K-1, \quad x(0) = x_0, \quad \theta(k) \in \Theta, \tag{1}$$

where we get rid of explicit k dependence in f via the usual trick, $x_0 \in \mathbb{R}^n$ is the input data, and $\theta(k)$ is the collection of the parameters in the k-th layer of the deep residual network. In deep reinforcement learning, the action $a(k)$ is actually a function of the state $x(k)$ and parameters $\theta(k)$, that is $a(x(k), \theta(k))$. The state and action of each time step will affect the state of the next time step simultaneously, and will bring a reward function $L(x(k), a(x(k), \theta(k)))$. By absorbing $a(\cdot)$ in L, the reward can be expressed as $L(x(k), \theta(k))$. The final output $x(K)$ of the network may be constrained by condition $g(x(K)) = 0$ in some practical problems, and there might be a terminal cost function $\Phi(x(K), y_0)$, where y_0 represents some known variables corresponding to x_0. In deep learning, the pair (x_0, y_0) is usually a data point sampled from a distribution μ. Hence, the reinforcement learning problem seeks the solution of the following problem

$$\begin{aligned} &\inf_{(\theta(0), \cdots, \theta(K-1)) \in \Theta^K} \mathbb{E}_{(x_0, y_0) \sim \mu} \left[\Phi(x(K), y_0) + \sum_{k=0}^{K-1} L(x(k), \theta(k)) \right] \\ &s.t.\ x(k+1) = x(k) + f(x(k), \theta(k)), \quad k = 0, \cdots, K-1, \quad x(0) = x_0, \quad g(x(K)) = 0. \end{aligned} \tag{2}$$

To avoid the repeated compositional structure in the difference equation in equation 2, we introduce the dynamical systems viewpoint and replace the discrete dynamics equation 1 by a continuous dynamical system in the following. Consider the dynamical system with terminal constraint, the state equation and target set are

$$\dot{x} = f(x, \boldsymbol{\theta}), \quad x(0) = x_0 \in \mathbb{R}^n, \quad \mathcal{S} := \{x \big| g(x(t_f)) = 0\}, \tag{3}$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r$ is a vector. Then we can define the continuous case of the objective function in equation 2 as follows

$$J(\theta) = \mathbb{E}_{(x_0, y_0)\sim\mu}\left[\Phi(x(t_f), y_0) + \int_0^{t_f} L(x(t), \theta(t))dt\right], \tag{4}$$

where $f : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n$, $\theta : [0, t_f] \to \mathbb{R}^r$, $g : \mathbb{R}^n \to \mathbb{R}^p$, $p \leq n$, $y_0 \in \mathbb{R}^m$, $\Phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $L : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ and t_f is the terminal time. We say θ is admissable if $\theta(t) \in \Theta$ for all $t \in [0, t_f]$. The optimal control problem is

$$\inf_{\theta} J(\theta) \quad s.t.\ \dot{x} = f(x, \theta), \quad x(0) = x_0, \quad g(x(t_f)) = 0, \tag{5}$$

and the optimal strategy is $\theta^* = \arg\inf_{\theta\in\mathcal{U}} J(\theta)$, where

$$\mathcal{U} := \{\theta : [0, t_f] \to \Theta \big| \theta \text{ is bounded and piecewise continuous, } x(t_f; \theta) \in \mathcal{S}\}.$$

3 MEAN-FIELD OPTIMAL CONTROL VIEW OF REINFORCEMENT LEARNING

In this section, we prove the mean-field Pontryagin's maximum principle with terminal constraints, which is the necessary condition for the optimality of dynamical systems. Define the Hamilton function as $H(x, \theta, \psi) := -L(x, \theta) + \psi^T f(x, \theta)$, where $H : \mathbb{R}^n \times \Theta \times \mathbb{R}^n$ and $\psi \in \mathbb{R}^n$, then we have the following theorem.

Theorem 3.1 *Under the assumption*

- *f is bounded, and f and L are continuous w.r.t. θ;*
- *f, L and Φ are continuously differentiable w.r.t x, and μ has bounded support.*

Let θ^* be the optimal strategy, and $x^*(t)$ is the corresponding optimal trajectory, then there exists $\psi^* : [0, t_f] \to \mathbb{R}^n$ and $\xi \in \mathbb{R}^p$ such that

$$\begin{aligned}
&\dot{x}^*(t) = f(x, \theta^*(t)), \qquad x^*(0) = x_0, \\
&\dot{\psi}^*(t) = -\nabla_x H(x^*(t), \theta^*(t), \psi^*(t)),\ \psi^*(t_f) = -\nabla_x \Phi(x^*(t_f), y_0) - \xi^T \nabla_x g(x^*(t_f)), \\
&\mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta^*(t), \psi^*(t)) = \sup_{\theta\in\Theta} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta, \psi^*(t)) \quad a.e.\ t \in [0, t_f].
\end{aligned} \tag{6}$$

Theorem 4.1 provide the necessary condition for optimality, which relies on the loss function and terminal constraints. There is a feedforward ODE, describing the state dynamics under optimal controls (θ_z^*, θ_d^*). The evolution of the co-state variable Ψ is defined, characterizing the evolution of an adjoint variational condition backward in time.

Comparison with existing works. This theorem is a more practical form of Theorem 3 in (E et al., 2019), which contains implicit terminal constraints. The optimal strategy must globally maximize the Hamiltonian function for a.e. $t \in [0, t_f]$. In fact, here a.e. $t \in [0, t_f]$ represents the set of t that makes the optimal strategy continuous with respect to t.

Proof sketch. The proof has two parts: (1) transform the problem into an unconstrained problem by the Lagrange multiplier method; (2) apply Theorem 3 in (E et al., 2019) to the transformed problem. A detailed proof is omitted here and is given in Appendix A.1.

4 MEAN-FIELD QUANTITATIVE DIFFERENTIAL GAME OF ADVERSARIAL REINFORCEMENT LEARNING

In this section, we model adversarial reinforcement learning as a mean-field quantitative differential game. In addition to the original deep neural network $NN_z(\cdot;\hat{\theta}_z) : \mathbb{R}^{n_1} \to \mathbb{R}^{n_1}$, an adversarial neural network $NN_d(\cdot;\hat{\theta}_d) : \mathbb{R}^{n_2} \to \mathbb{R}^{n_2}$ is usually added in adversarial learning, where $\hat{\theta}_z \in \hat{\Theta}_z$ and $\hat{\theta}_d \in \hat{\Theta}_d$ are parameters of NN_z and NN_d respectively. Assume that both NN_z and NN_d are deep residual networks with K layers, thus $\hat{\theta}_z = (\theta_z(0), \cdots, \theta_z(K-1)) \in \hat{\Theta}_z = \Theta_z^K$ and $\hat{\theta}_d = (\theta_d(0), \cdots, \theta_d(K-1)) \in \hat{\Theta}_d = \Theta_d^K$. Similar to Section 3, define the penalty function $L(x_z(k), x_d(k), \theta_z(k), \theta_d(k))$ and the terminal cost function $\Phi(NN_z(x_{z_0};\hat{\theta}_z), NN_d(x_{d_0};\hat{\theta}_d), y_0)$, where $y_0 \in \mathbb{R}^m$ represents some known variable corresponding to x_{z_0} and x_{d_0}.

4.1 MEAN-FIELD TWO-SIDED EXTREMISM PRINCIPLE

In adversarial reinforcement learning, NN_z is trained to maximize the loss, while NN_d is trained to minimize the loss. Assume that both NN_z and NN_d are deep residual networks with the same number of layers, thus we can formulate the adversarial learning problem as

$$\begin{aligned}
&\inf_{\hat{\theta}_z \in \hat{\Theta}_z} \sup_{\hat{\theta}_d \in \hat{\Theta}_d} \mathbb{E}_{(x_{z_0}, x_{d_0}, y_0)\sim\mu}\left[\Phi(x_z(K), x_d(K), y_0) + \sum_{k=0}^{K-1} L(x_z(k), x_d(k), \theta_z(k), \theta_d(k))\right] \\
&subject\ to \\
&x_z(k+1) = x_z(k) + f_z(x_z(k), \theta_z(k)), \quad k = 0, \cdots, K-1, \quad x_z(0) = x_{z_0}, \quad g_z(x_z(K)) = 0, \\
&x_d(k+1) = x_d(k) + f_d(x_d(k), \theta_d(k)), \quad k = 0, \cdots, K-1, \quad x_d(0) = x_{d_0}, \quad g_d(x_d(K)) = 0.
\end{aligned} \tag{7}$$

Now we consider the dynamical systems viewpoint and translate problem equation 7 into a continuous form. Let the objective function is

$$J(\theta_z, \theta_d) = \mathbb{E}_{(x_{z_0}, x_{d_0}, y_0)\sim\mu}\left[\Phi(x_z(t_f), x_d(t_f), y_0) + \int_0^{t_f} L(x_z(t), x_d(t), \theta_z(t), \theta_d(t))dt\right], \tag{8}$$

where $\boldsymbol{\theta_z} \in \Theta_z \subset \mathbb{R}^{r_1}$, $\boldsymbol{\theta_d} \in \Theta_d \subset \mathbb{R}^{r_2}$, $x_z : [0, t_f] \to \mathbb{R}^{n_1}$, $x_d : [0, t_f] \to \mathbb{R}^{n_2}$, $\theta_z : [0, t_f] \to \mathbb{R}^{r_1}$, $\theta_d : [0, t_f] \to \mathbb{R}^{r_2}$, $g_z : \mathbb{R}^{n_1} \to \mathbb{R}^{p_1}$, $g_d : \mathbb{R}^{n_2} \to \mathbb{R}^{p_2}$, Φ, L and f are all functions of appropriate input and output dimensions. Let $x = (x_z^T, x_d^T)^T$, $x(0) = (x_z^T(0), x_d^T(0))^T$, $f(x, \theta_z, \theta_d) = (f_z^T(x, \theta_z), f_d^T(x, \theta_d)^T)$, $g(x(t_f)) = (g_z^T(x_z(t_f)), g_d^T(x_d(t_f)))^T$. Then the state equation, target set and objective functionals of quantitative differential game are

$$\begin{aligned}
\dot{x} = f(x, \theta_z, \theta_d), \quad x(0) = x_0, \quad \mathcal{S} := \{x | g(x(t_f)) = 0\}, \\
J(\theta_z, \theta_d) = \mathbb{E}_{(x_0, y_0)\sim\mu}\left[\Phi(x(t_f), y_0) + \int_0^{t_f} L(x(t), \theta_z(t), \theta_d(t))dt\right],
\end{aligned} \tag{9}$$

where $x : [0, t_f] \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^p$, $n = n_1 + n_2$ and $p = p_1 + p_2$. Define $\mathcal{U}_z$ ($\mathcal{U}_d$) as the set of all admissable θ_z (θ_d) that satisfy the terminal constraint $g_z(x_z(t_f)) = 0$ ($g_d(x_d(t_f)) = 0$). Our goal is to find the optimal strategy (θ_z^*, θ_d^*) of equation 9 and the corresponding optimal trajectory $x^*(t)$ satisfies

$$J(\theta_z^*, \theta_d) \le J(\theta_z^*, \theta_d^*) \le J(\theta_z, \theta_d^*), \quad \forall (\theta_z, \theta_d) \in \mathcal{U}_z \times \mathcal{U}_d, \tag{10}$$

where equation 10 is called the saddle point condition. Now define the Hamilton function of equation 9 as $H(x(t), \theta_z(t), \theta_d(t), \psi(t)) := -L(x(t), \theta_z(t), \theta_d(t)) + \psi^T(t) f(x(t), \theta_z(t), \theta_d(t))$, then we have the following mean-field TSEP.

Theorem 4.1 *Under the assumption,*

- *f is bounded and f, L are continuous w.r.t. θ_z, θ_d;*
- *f, L and Φ are continuously differentiable w.r.t x, and μ has bounded support.*

Let $(\theta_z^*, \theta_d^*) \in \mathcal{U}_z \times \mathcal{U}_d$ is the optimal strategy of problem equation 9, $x^*(t)$ is the corresponding optimal trajectory, then there exists $\psi^* : [0, t_f] \to \mathbb{R}^n$ and $\xi \in \mathbb{R}^p$ such that

i)

$$\begin{aligned}&\dot{x}^*(t) = f(x, \theta_z^*, \theta_d^*), \qquad x^*(0) = x_0,\\ &\dot{\psi}^*(t) = -\nabla_x H(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi^*(t)),\ \psi^*(t_f) = -\nabla_x \Phi(x^*(t_f), y_0) - \xi^T \nabla_x g(x^*(t_f)),\end{aligned} \tag{11}$$

ii)

$$\begin{aligned}&\mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi^*(t))\\ =& \sup_{\theta_z\in\Theta_z} \inf_{\theta_d\in\Theta_d} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t))\\ =& \inf_{\theta_d\in\Theta_d} \sup_{\theta_z\in\Theta_z} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t)), \quad a.e.\ t\in[0,t_f].\end{aligned} \tag{12}$$

Theorem 4.1 introduces the necessary conditions for the convergence of the unique global solution to the mean-field TSEP, relying on the loss loss function and terminal constraints. Comparing Theorem 4.1 to Theorem 3.1, we can see the main difference is that the maximization condition is turned into the saddle point condition.

Comparison with existing works. This theorem is a mean-field form of Theorem 7.4.1 in (Guo et al., 2005) gives general necessary conditions for optimality for our problem.

Proof sketch. The proof has two parts: (1) Fix the optimal strategy θ_z^* and θ_d^* respectively, and transform the problem into two optimal control problems; (2) apply Theorem 3.1 to these two transformed problems and use max-min inequality. A detailed proof is omitted here and is given in Appendix A.2.

4.2 Small-time uniqueness

Since the necessary condition for optimality has been provided by the TSEP, a natural question is to understand when the sufficient conditions for optimality can be also provided. In this subsection, we will establish assumptions which are required for a unique solution of the mean-field TSEP equations.

Theorem 4.2 *Suppose that*

- *f is bounded, g is continuously differentiable w.r.t x with bounded and Lipschitz partial derivatives, μ has bounded support in $\mathbb{R}^n \times \mathbb{R}^m$;*
- *f, L and Φ are twice continuously differentiable w.r.t x, θ_z and θ_d with bounded and Lipschitz partial derivatives, and $\partial f/\partial\theta_z\partial\theta_d, \partial L/\partial\theta_z\partial\theta_d \equiv 0$;*
- *$H(x, \theta_z, \theta_d, \psi)$ is strongly concave in θ_z, strongly convex in θ_d and uniformly in $x \in \mathbb{R}^n$, $\psi \in \mathbb{R}^n$.*

Then for sufficiently small t_f, if (θ_z^1, θ_d^1) and (θ_z^2, θ_d^2) are solutions of the mean-field TSEP derived in Theorem 4.1 and are continuously w.r.t time t, then $(\theta_z^1, \theta_d^1) = (\theta_z^2, \theta_d^2)$.

Theorem 4.2 shows that the small t_f roughly corresponds to the regime where the reachable set of the forward dynamics is small, hence the solution is unique. We assume the continuity of $\theta_z^1, \theta_d^1, \theta_z^2, \theta_d^2$ with respect to t in Theorem 4.2. In fact, when $\theta_z^1, \theta_d^1, \theta_z^2, \theta_d^2$ are discontinuous on at most a set with zero measure, we can also conclude that $(\theta_z^1(t), \theta_d^1(t)) = (\theta_z^2(t), \theta_d^2(t))$ for a.e. $t \in [0, t_f]$.

Comparison with existing works. To the best of our knowledge, there is no existing method to prove the uniqueness of TSEP's solution.

Proof sketch. The proof has two parts: (1) bound the difference of flow-maps driven by two different controls; (2) apply the first-order optimality condition for $(\theta_z^1(t), \theta_d^1(t))$ and $(\theta_z^2(t), \theta_d^2(t))$. A detailed proof is omitted here and is given in Appendix B.

4.3 DERIVATIVE IN WASSERSTEIN SPACE

Now we propose the mean-field Hamilton-Jacobi-Issacs equation, whose solution is a real value function satisfying the saddle point condition. To begin with, we first introduce the Wasserstein space and its derivation rules. Let D represent the Fréchet derivative on Banach spaces. Namely, if $F : U \to V$ is a mapping between two Banach spaces $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$, then $DF(x) : U \to V$ is a linear operator satisfies

$$\frac{\|F(x+y) - F(x) - DF(x)(y)\|_V}{\|y\|_U} \to 0, \qquad as\ \|y\|_U \to 0. \tag{13}$$

Denote $X \in \mathbb{R}^{n+m}$ as a random variable, we use the shorthand $L^2(\Omega, \mathbb{R}^{n+m})$ for $L^2((\Omega, \mathcal{F}, \mathbb{P}), \mathbb{R}^{n+m})$ to represent the set of $\mathbb{R}^{n+m}$-valued square integrable random variables with respect to a probability measure $\mathbb{P}$. Then we equip this Hilbert space with the norm $\|X\|_{L^2} := (\mathbb{E}\|X\|^2)^{1/2}$. As we assumed in the previous section, $x_0 \in \mathbb{R}^n$, $y_0 \in \mathbb{R}^m$ are random variables and $(x_0, y_0) \sim \mu \in \mathcal{P}_2(\mathbb{R}^{n+m})$, where $\mathcal{P}_2(\mathbb{R}^{n+m})$ denotes the integrable probability measure defined on the Euclidean space $\mathbb{R}^{n+m}$. The space $\mathcal{P}_2(\mathbb{R}^{n+m})$ can be equipped with a metric by 2-Wasserstein distance

$$\mathcal{W}_2(\mu, \nu) := \inf\left\{ \|X - Y\|_{L^2} \middle| X, Y \in L^2(\Omega, \mathbb{R}^{n+m}) \text{ with } \mathbb{P}_X = \mu, \mathbb{P}_Y = \nu \right\}.$$

For $\mu \in \mathcal{P}_2(\mathbb{R}^{n+m})$, define $\|\mu\|_{L^2} := (\int_{\mathbb{R}^{n+m}} \|w\|^2 \mu(dw))^{1/2}$. Now the variable $X \in L^2(\Omega, \mathbb{R}^{n+m})$ if and only if its law $\mathbb{P}_X \in \mathcal{P}_2(\mathbb{R}^{n+m})$. For any function $u : \mathcal{P}_2(\mathbb{R}^{n+m}) \to \mathbb{R}$, we can lift it into its "extension" $U \in L^2(\Omega, \mathbb{R}^{n+m})$ (Cardaliaguet, 2012) by $U(X) = u(\mathbb{P}_X), \forall X \in L^2(\Omega, \mathbb{R}^{n+m})$. In particular, we have that u is $C^1(\mathcal{P}_2(\mathbb{R}^{n+m}))$, if the lifted function U is Fréchet differentiable with continuous derivatives. Since $L^2(\Omega, \mathbb{R}^{n+m})$ can be identified with its dual, if the Fréchet derivative $DU(X)$ exists, by Riesz' theorem, it can be identified with an element of $L^2(\Omega, \mathbb{R}^{n+m})$,

$$DU(X)(Y) = \mathbb{E}[DU(X) \cdot Y],\ \forall Y \in L^2(\Omega, \mathbb{R}^{n+m}).$$

One may check that the law of $DU(X)$ does not depend on X but only on the law of X, thus the derivative of u at $\mu = \mathbb{P}_X$ is defined as $DU(X) = \partial_\mu u(\mathbb{P}_X)(X)$, for some function $\partial_\mu u(\mathbb{P}_X) : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$.

4.4 MEAN-FIELD HJI EQUATION

Without losing generality, we consider the quantitative differential game (equation 9) with $\mathcal{S} = \mathbb{R}^n$. Motivated by the mean field dynamic programming principle (E et al., 2019) that any last part of an optimal control is optimal, we consider the quantitative different game with any $t_0 \in [0, t_f]$ as initial time

$$\begin{aligned} &\dot{x}(t) = f(x(t), \theta_z(t), \theta_d(t)), \quad x(t_0) = x_{t_0} \in \mathbb{R}^n, \quad \mathcal{S} = \mathbb{R}^n, \\ &J_H(\theta_z, \theta_d, t_0, \mu) = \mathbb{E}_{(x_{t_0}, y_0) \sim \mu}\left[\Phi(x(t_f), y_0) + \int_{t_0}^{t_f} L(x(t), \theta_z(t), \theta_d(t)) dt \right]. \end{aligned} \tag{14}$$

Our goal is to find the optimal strategy $(\theta_z^*, \theta_d^*) \in \mathcal{U}_z \times \mathcal{U}_d$ of equation 14 satisfies

$$J_H(\theta_z^*, \theta_d, t, \mu) \le J_H(\theta_z^*, \theta_d^*, t, \mu) \le J_H(\theta_z, \theta_d^*, t, \mu), \quad \forall t \in [0, t_f], (\theta_z, \theta_d) \in \mathcal{U}_z \times \mathcal{U}_d. \tag{15}$$

Define $v^*(t, \mu) := J_H(\theta_z^*, \theta_d^*, t, \mu)$, we have the following theorem.

Theorem 4.3 *Under the assumption*

- *f, L and Φ are bounded, and the distribution $\mu \in \mathcal{P}_2(\mathbb{R}^{n+m})$;*
- *f, L and Φ are Lipschitz continuous w.r.t x and the Lipschitz constant of f and L are independent of θ_z, θ_d.*

Suppose the optimal value function $v^*(t, \mu)$ of equation 14 exists, then it is the unique viscosity solution (see definition in appendix C) to the following mean-field HJI equation

$$\begin{aligned} &\partial_t v(t, \mu) + \inf_{\theta_z \in \Theta_z} \sup_{\theta_d \in \Theta_d} \left\{ \int_{\mathbb{R}^{n+m}} [\partial_\mu v(t, \mu)(x, y)]^T [f(x, \theta_z, \theta_d), 0] + L(x, \theta_z, \theta_d) d\mu(x, y) \right\} = 0, \\ &v(t_f, \mu) = \int_{\mathbb{R}^{n+m}} \Phi(x, y) d\mu(x, y). \end{aligned} \tag{16}$$

Theorem 4.3 establishes the uniqueness, in the viscosity sense, of the HJI equation and identifies the value function for the mean-field optimal control problem as the unique solution of the HJI equation.

Comparison with existing works. This theorem is analogous to Theorem 7.4.2 in (Guo et al., 2005), the state variables we are dealing with are probability measures rather than Euclidean vectors.

Proof sketch. The proof has two parts: (1) Fix the optimal strategy θ_z^* and θ_d^* respectively, and transform the problem into two optimal control problems; (2) apply Theorem 1 and 2 in (E et al., 2019) to these two transformed problems. A detailed proof is given in Appendix A.2.

4.5 Connection between HJI and TSEP

In what follows, we provide the connection between the HJI equation and TSEP. We will show that the TSEP can be understood as a local result compared to the global characterization of the HJI equation. For the value function $v(t,\mu)$ in deduced HJI (equation 16), consider the lifted function $V(t,X)$, where $X=(x,y)\sim\mu$. We define the Hamiltonian for the lifted HJI equation as

$$\mathcal{H}(X, D_X V(t,X)) == \inf_{\theta_z\in\Theta_z}\sup_{\theta_d\in\Theta_d} \mathbb{E}_\mu\big[D_X V(t,X)^T[f(x,\theta_z,\theta_d),0] + L(x,\theta_z,\theta_d)\big]. \tag{17}$$

Suppose $\theta_z^\dagger(X, D_X V(t,X))$ and $\theta_d^\dagger(X, D_X V(t,X))$ are the corresponding optimal strategies and define $P = D_X V(t,X)$, we have

$$\begin{aligned}\mathcal{H}(X,P) &= \mathbb{E}_\mu\big[P^T[f(x,\theta_z^\dagger(X,P),\theta_d^\dagger(X,P)),0] + L(x,\theta_z^\dagger(X,P),\theta_d^\dagger(X,P))\big],\\ \mathbb{E}_\mu\big[\nabla_{\theta_z,\theta_d}[f(x,\theta_z^\dagger(X,P),\theta_d^\dagger(X,P)),0]P &+ \nabla_{\theta_z,\theta_d}L(x,\theta_z^\dagger(X,P),\theta_d^\dagger(X,P))\big] = 0,\end{aligned} \tag{18}$$

where the last equation follows from the first order optimality condition. Define $X_t = (x_t, y)$, $P_t = D_X V(t, X_t)$, we can apply the characteristic evolution equations (Subbotina, 2006)

$$\dot{X}_t = D_P\mathcal{H}(X_t,P_t), \quad \dot{P}_t = -D_X\mathcal{H}(X_t,P_t). \tag{19}$$

Plugging equation 18 into equation 19, and let $\theta_z^*(t) = \theta_z^\dagger(X_t,P_t)$, $\theta_d^*(t) = \theta_d^\dagger(X_t,P_t)$ and p_t is the first n components of P_t, we have

$$\dot{x}_t = f(x_t,\theta_z^*(t),\theta_d^*(t)), \quad \dot{p}_t = -\nabla_x f(x_t,\theta_z^*(t),\theta_d^*(t))p_t - \nabla_x L(x_t,\theta_z^*(t),\theta_d^*(t)). \tag{20}$$

If we let $\psi = -p$, the first two equalities of equation 11 in Theorem 4.1 is converted to equation 20. The Hamilton equation in TSEP can be regarded as the characteristic equations for the HJI equation originating from μ_0, which justifies the claim that the TSEP constitutes a local condition as compared to the HJI equation.

5 Estimation of generalization bounds

So far, we have focused on the mean-field quantitative differential game and mean-field TSEP. However, the solution of the mean-field TSEP requires a saddle point with respect to an underlying sample distribution. In this section, we will establish generalization bounds from two aspects, global minimum of the loss function and algorithmic stability. We define the loss function of each training sample $X_i := (x_{z_i}, x_{d_i}, y_i),\ i = 1,\cdots,N$ as

$$J^0(\theta_z,\theta_d;X_i) = \Phi(x_z(t_f),x_d(t_f),y_i) + \int_0^{t_f} L(x_z(t),x_d(t),\theta_z(t),\theta_d(t))dt,$$

where $x_z(0) = x_{z_i}, x_d(0) = x_{d_i}$. Now $J(\theta_z,\theta_d) = \mathbb{E}_{X_0\sim\mu}J^0(\theta_z,\theta_d;X_0)$, and we define

$$J_N(\theta_z,\theta_d) = \frac{1}{N}\sum_{i=1}^N J^0(\theta_z,\theta_d;X_i).$$

We first estimate the generalization bounds based on the global minimum of the loss function. The necessary condition of Hamiltonian of sampled version is expressed as

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N H(x^{\theta_z^N,\theta_d^N,i}(t),\theta_z^N(t),\theta_d^N(t),\psi^{\theta_z^N,\theta_d^N,i}(t))\\ =&\inf_{\theta_d\in\Theta_d}\sup_{\theta_z\in\Theta_z}\frac{1}{N}\sum_{i=1}^N H(x^{\theta_z^N,\theta_d^N,i}(t),\theta_z,\theta_d,\psi^{\theta_z^N,\theta_d^N,i}(t)), \quad a.e.\ t\in[0,t_f],\end{aligned} \tag{21}$$

where θ_z^N and θ_d^N are the solution of sampled TSEP. Note that if Θ_z and Θ_d are sufficiently large, e.g. $\Theta_z = \mathbb{R}^{r_1}, \Theta_d = \mathbb{R}^{r_2}$, the solution θ_z^*, θ_d^* of TSEP satisfies

$$F(\theta_z^*, \theta_d^*)(t) := \mathbb{E}_{\mu_0} \nabla_{\theta_z, \theta_d} H(x_t^{\theta_z^*, \theta_d^*}, \psi_t^{\theta_z^*, \theta_d^*}, \theta_z^*(t), \theta_d^*(t)) = 0, \quad a.e.\ t \in [0, t_f], \tag{22}$$

while the solution θ_z^N, θ_d^N of sampled TSEP satisfies

$$F_N(\theta_z^N, \theta_d^N)(t) := \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta_z, \theta_d} H(x_t^{\theta_z^N, \theta_d^N, i}, \psi_t^{\theta_z^N, \theta_d^N, i}, \theta_z^N(t), \theta_d^N(t)) = 0, \quad a.e.\ t \in [0, t_f]. \tag{23}$$

Now, F_N is a random approximation of F and $\mathbb{E}F_N(\theta_z, \theta_d)(t) = F(\theta_z, \theta_d)(t)$ for all θ_z, θ_d and a.e. $t \in [0, t_f]$. Let $(U, \|\cdot\|_U)$, $(V, \|\cdot\|_V)$ be Banach spaces and $F : U \to V$. We first provide the definition of stability, which is the primary condition that ensures the approximation of F_N to F.

Definition 5.1 *For $\rho > 0$ and $x \in U$, $S_\rho(x) := \{y \in U : \|x - y\|_U < \rho\}$. The mapping F is stable on $S_\rho(x)$ if there exists a constant $K_\rho > 0$ such that,*

$$\|y - z\|_U \le K_\rho \|F(y) - F(z)\|_V, \quad \forall y, z \in S_\rho(x).$$

Notice that here not consider θ_z^* and θ_d^* correspond to maximum and minimum, we only care that they both follow the first-order optimality condition. We define $\theta = (\theta_z^T, \theta_d^T)^T$ and redefine $F(\theta_z^*, \theta_d^*)(\cdot)$ and $F_N(\theta_z^N, \theta_d^N)(\cdot)$ as $F(\theta)(\cdot)$, $F(\theta^N)(\cdot)$, respectively. We follow the proof idea of Theorem 6 in (E et al., 2019) and get similar results stated in 5.1, which describes the convergence of sampled solution to mean-field solution as the number of samples increases.

Theorem 5.1 *Assuming that f, L, and Φ are bounded and Lipschitz continuous with respect to x and the Lipschitz constant of f and L are independent of θ_z, θ_d. Let θ_z^*, θ_d^* be a solution of $F = 0$ (equation 22), which is stable on $S_\rho(((\theta_z^*)^T, (\theta_d^*)^T)^T)$ for some $\rho > 0$. Then there exists positive constants s_0, C, K_1, K_2, $\rho_1 < \rho$ and a random variable $\theta^N := ((\theta_z^N)^T, (\theta_d^N)^T)^T \in S_\rho(((\theta_z^*)^T, (\theta_d^*)^T)^T)$, such that*

$$\begin{aligned}
&\mathbb{P}\big[\|\theta_z^* - \theta_z^N\|_{L^\infty} \ge Cs\big] \le 4\exp\left\{-\frac{Ns^2}{K_1 + K_2 s}\right\}, \quad s \in (0, s_0], \\
&\mathbb{P}\big[\|\theta_d^* - \theta_d^N\|_{L^\infty} \ge Cs\big] \le 4\exp\left\{-\frac{Ns^2}{K_1 + K_2 s}\right\}, \quad s \in (0, s_0], \\
&\mathbb{P}\big[|J(\theta_z^*, \theta_d^*) - J(\theta_z^N, \theta_d^N)| \ge s\big] \le 4\exp\left\{-\frac{Ns^2}{K_1 + K_2 s}\right\}, \quad s \in (0, s_0], \\
&\mathbb{P}\big[F_N(\theta^N) \ne 0\big] \le 4\exp\left\{-\frac{Ns_0^2}{K_1 + K_2 s_0}\right\}.
\end{aligned}$$

The loss function is uniformly bounded under the given assumptions, then we can apply the Hoeffding's inequality (Corollary 2 in (Pinelis and Sakhanenko, 1986)). Using Theorem 6 in (E et al., 2019) and rewriting θ as $(\theta_z^T, \theta_d^T)^T$, this theorem can be proved. Theorem 5.1 basically shows that the difference between the optimizer over the whole distribution and the optimizer over finite samples is bounded, and is exponential decay with the total number of samples N. This bound is not rely on the training algorithms.

Now we estimate the generalization bounds from the view of algorithmic stability. In the rest of this section, we redefine the integral form in J, J_N, J^0 as the discrete sum form equation 7 and redefine r_1, r_2 as the total dimension of θ_z, θ_d. We consider the error by taking expectation with respect to randomized algorithm and define

$$er(\theta_z, \theta_d) := \mathbb{E}_{\mathcal{A}}\big[J(\theta_z, \theta_d) - J_N(\theta_z, \theta_d)\big].$$

We update θ_z and θ_d alternately, i.e. from the initial value $(\theta_{z,0}, \theta_{d,0})$, update θ_z by $M_{z,1}$ steps to get $(\theta_{z,M_{z,1}}, \theta_{d,0})$, then update θ_z by $M_{d,1}$ steps to get $(\theta_{z,M_{z,1}}, \theta_{d,M_{d,1}})$. Keep going until the algorithm converges, we can get $(\theta_{z,M_{z,2}}, \theta_{d,M_{d,2}})$, $(\theta_{z,M_{z,3}}, \theta_{d,M_{d,3}})$ $\cdots$.

Consider Stochastic Gradient Langevin Dynamics (SGLD), which is a popular variant of stochastic gradient methods which adds isotropic Gaussian noise in each iteration, e.g.

$$\theta_{z,k+1} = \theta_{z,k} - \eta_k \nabla_{\theta_z} J_N(\theta_{z,k}, \theta_{d,0}) + \sqrt{\frac{2\eta_k}{\beta}} \mathcal{N}(0, I_{r_1}).$$

We follow the proof idea of Theorem 8 in (Mou et al., 2018) and get the following generalization bound in expectation of random draw of training data.

Theorem 5.2 *Supposes $J^0(\theta_z, \theta_d; X)$ is uniformly bounded by C, and $\|\nabla_{\theta_z} J^0(\theta_z, \theta_d; X) - \nabla_{\theta_z} J^0(\theta_z, \theta_d; X')\| \leq L_z$, $\|\nabla_{\theta_d} J^0(\theta_z, \theta_d; X) - \nabla_{\theta_d} J^0(\theta_z, \theta_d; X')\| \leq L_d \ \forall X, X'$, then we have the following generalization bound*

$$\begin{aligned}\mathbb{E}[er(\theta_{z,M_{z,n}}, \theta_{d,M_{d,n}})] \leq & \frac{2}{N} \sum_{i=1}^{n} \min(k_1, M_{z,i} - M_{z,i-1}) + \frac{\sqrt{\beta} L_z C}{N} \sum_{i=1}^{n} (\sum_{j=k_1+1}^{M_{z,i}-M_{z,i-1}} \eta_j)^{1/2} \\ & + \frac{2}{N} \sum_{i=1}^{n} \min(k_2, M_{d,i} - M_{d,i-1}) + \frac{\sqrt{\beta} L_d C}{N} \sum_{i=1}^{n} (\sum_{j=k_2+1}^{M_{d,i}-M_{d,i-1}} \eta_j)^{1/2},\end{aligned} \tag{24}$$

where $M_{z,0} = M_{d,0} = 0$, k_1 and k_2 are chosen to satisfy $\eta_{k_1} \leq \ln 2/\beta L_z^2$, $\eta_{k_2} \leq \ln 2/\beta L_d^2$.

Theorem 5.2 obtains a bound of $\mathcal{O}(1/N)$, which matches the generalization bounds of stochastic gradient descent ascent (SGDA) for minimax problems in (Lei et al., 2021). This bound relies on the aggregated step sizes and do not explicitly depend on the dimensions, norms, or other capacity measures of the parameter.

Comparison with existing works. Unlike the existing methods to analyze the algorithms with $M_{z,i} - M_{z,i-1} = 1$, $M_{d,i} - M_{d,i-1} = 1$, e.g. SGDA, we analyzed the case of taking any $M_{z,i}, M_{d,i}$. As $\{\eta_i\}$ is usually monotonically non-increasing, we find that set $M_{z,1} - M_{z,0} = M_z$, $M_{z,i} - M_{z,i-1} = 0$, $i > 1$ and $M_{d,1} - M_{d,0} = M_d$, $M_{d,i} - M_{d,i-1} = 0$, $i > 1$ will lead to the smallest bound for fixed $M_{z,n} = M_z$, $M_{d,n} = M_d$ in equation 24. This finding gives an inspiration to the training method, training one network to make it converge before training the other one.

Proof sketch. The proof is based on the uniform stability of the loss function and standard symmetrization argument (Hardt et al., 2016). Using Theorem 8 in (Mou et al., 2018) alternately in the training process of the two networks, this theorem can be proved. See details in Appendix D.

6 Conclusion

In this paper, adversarial reinforcement learning was considered as the mean-field quantitative differential game and the convergence and generalization were analyzed under this framework. We first proposed a new mean-field PMP for reinforcement learning with terminal constraints in Theorem 3.1. Then we bridged adversarial reinforcement learning to the mean-field quantitative differential game and proved a mean-field TSEP in Theorem 4.1, which provides the necessary condition for the convergence to the global optimality. Moreover, the uniqueness of the solution to mean-field TSEP is shown in Theorem 4.2, where a sufficient small terminal time is required. After that, the mean-field HJI equation for calculating the optimal loss function is derived in Theorem 4.3. Furthermore, we have shown that the TSEP is actually a local special case compared to the global characterization of the HJI equation. Lastly, we proved two generalization error bounds of order $\mathcal{O}(e^{-N})$ and $\mathcal{O}(1/N)$, which characterize the gap between the expected solution and the learned model. The two bounds are obtained from Hoeffding's inequality (Theorem 5.1) and algorithmic stability (Theorem 5.2). Both bounds do not explicitly rely on the dimensions, norms, or other capacity measures of the network parameter. Moreover, the latter bound characterizes the influence of randomness in the training algorithms and aggregated step sizes on generalization. Our results allow the possibility to develop a learning algorithm without referring to the classical methods of deep learning such as the SGD. A further direction of our ongoing research is based on the study of the TSEP and HJI equation from a discrete-time perspective in order to relate the theoretical framework directly to the applications.

REFERENCES

V. Behzadan and A. Munir. Adversarial reinforcement learning framework for benchmarking collision avoidance mechanisms in autonomous vehicles. *IEEE Intelligent Transportation Systems Magazine*, 2019.

A. Bensoussan, J. Frehse, and P. Yam. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

P. Cardaliaguet. Notes on mean field games. Technical report, from P.-L. Lions' lectures at Collège de France, 2012.

B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert, and E. Holtham. Reversible architectures for arbitrarily deep residual neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a.

B. Chang, L. Meng, E. Haber, F. Tung, and D. Begert. Multi-level residual networks from dynamical systems view. In *International Conference on Learning Representations*, 2018b.

W. E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

W. E, J. Han, and Q. Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):1–41, 2019.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

L. Guo, D. Z. Cheng, and D. X. Feng. *Introduction to control theory: from basic concepts to research frontiers*. Science Press, 2005.

E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1): 014004, 2017.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

L. Kantorovitch. The method of successive approximation for functional equations. *Acta Mathematica*, 71(1):63–97, 1939.

J. Lasry and P. Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

Y. Lei, Z. Yang, T. Yang, and Y. Ying. Stability and generalization of stochastic gradient methods for minimax problems. *arXiv preprint arXiv:2105.03793*, 2021.

Q. Li and S. Hao. An optimal control approach to deep learning and applications to discrete-weight neural networks. In *International Conference on Machine Learning*, pages 2985–2994. PMLR, 2018.

Q. Li, L. Chen, and C. Tai. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18:1–29, 2018.

M. Liu and O. Tuzel. Coupled generative adversarial networks. *Advances in neural information processing systems*, 29:469–477, 2016.

X. Ma, K. Driggs-Campbell, and M. J. Kochenderfer. Improved robustness and safety for autonomous vehicle control with adversarial reinforcement learning. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1665–1671. IEEE, 2018.

A. Mandlekar, Y. Zhu, A. Garg, L. Fei-Fei, and S. Savarese. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3932–3939. IEEE, 2017.

X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

X. Pan, D. Seita, Y. Gao, and J. Canny. Risk averse robust adversarial reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8522–8528. IEEE, 2019.

L. D. Persio and M. Garbelli. Deep learning and mean-field games: A stochastic optimal control perspective. *Symmetry*, 13(1):14, 2021.

I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, 1986.

L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

L. S. Pontryagin. The mathematical theory of optimal processes and differential games. *Trudy Matematicheskogo Instituta imeni VA Steklova*, 169:119–158, 1985.

L. S. Pontryagin. *Mathematical theory of optimal processes*. CRC press, 1987.

S. Sonoda and N. Murata. Double continuum limit of deep neural networks. In *ICML Workshop Principled Approaches to Deep Learning*, volume 1740, 2017.

N. Subbotina. The method of characteristics for hamilton—jacobi equations and applications to dynamical optimization. *Journal of mathematical sciences*, 135(3):2955–3091, 2006.

W. Uther and M. Veloso. Adversarial reinforcement learning. Technical report, Tech. rep., Carnegie Mellon University. Unpublished, 1997.

Y. Zang, G. Bao, X. Ye, and H. Zhou. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics*, 411:109409, 2020.

K. Zhang, B. Hu, and T. Basar. On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems. *Advances in Neural Information Processing Systems*, 33, 2020.

A Appendix

Proof A.1 (Proof of Theorem 3.1) *As the ordinary differential equation equation 3 is subject to terminal constraint $g(x(t_f), t_f) = 0$, we transform it into an unconstrained problem by introducing Lagrangian multiplier vectors $\xi \in \mathbb{R}^p$, thus the objective functionals equation 4 transforms into*

$$J_2(\theta) = \mathbb{E}_{(x_0,y_0)\sim\mu}\left[\Phi(x(t_f), y_0) + \xi^T g(x(t_f)) + \int_0^{t_f} L(x(t), \theta_z(t), \theta_d(t))dt\right]. \tag{25}$$

Then by Mean-field Pontryagin's maximum principle (Theorem 3 in (E et al., 2019)), the conclusion can be proved.

Proof A.2 (Proof of Theorem 4.1) *When (θ_z^*, θ_d^*) exists, we can split problem equation 9 into the following two optimal control problems.*

Problem 1:

$$\begin{aligned} \dot{x} &= f(x, \theta_z, \theta_d^*), \quad x(0) = x_0, \\ \mathcal{S} &:= \{x \big| g(x(t_f)) = 0\}, \\ J(\theta_z, \theta_d^*) =& \mathbb{E}_{(x_0,y_0)\sim\mu}\left[\Phi(x(t_f), y_0) + \int_0^{t_f} L(x(t), \theta_z(t), \theta_d^*(t))dt\right], \end{aligned} \tag{26}$$

find $\theta_z^* = \arg\inf_{\theta_z \in \mathcal{U}_z} J(\theta_z, \theta_d^*)$.

The Hamilton function of Problem 1 is

$$H_1(x(t), \theta_z(t), \theta_d^*(t), \psi_1(t)) := -L(x(t), \theta_z(t), \theta_d^*(t)) + \psi_1^T(t) f(x(t), \theta_z(t), \theta_d^*(t)).$$

Problem 2:

$$\begin{aligned} &\dot{x} = f(x, \theta_z^*, \theta_d), \quad x(0) = x_0, \\ &\mathcal{S} := \{x_z \big| g(x(t_f)) = 0\}, \\ J(\theta_z^*, \theta_d) =& \mathbb{E}_{(x_0,y_0)\sim\mu}\left[\Phi(x(t_f), y_0) + \int_0^{t_f} L(x(t), \theta_z^*(t), \theta_d(t))dt\right], \end{aligned} \tag{27}$$

find $\theta_d^* = \arg\sup_{\theta_d \in \mathcal{U}_d} J(\theta_z^*, \theta_d)$.

The Hamilton function of Problem 2 is

$$H_2(x(t), \theta_z^*(t), \theta_d(t), \psi_2(t)) := -L(x(t), \theta_z^*(t), \theta_d(t)) + \psi_2^T(t) f(x(t), \theta_z^*(t), \theta_d(t)).$$

*(θ_z^*, θ_d^*) is the optimal strategy of **Problem 1** and **Problem 2**. From the definition of $H_1(x, \theta_z, \theta_d^*, \psi_1, t)$ and $H_2(x, \theta_z^*, \theta_d, \psi_2, t)$, by Theorem 3.1, there exists Lagrangian multiplier $\xi_1, \xi_2 \in \mathbb{R}^p$ such that*

$$\begin{aligned} \dot{\psi}_1(t) &= -\nabla_x H_1(x^*(t), \theta_z(t), \theta_d^*(t), \psi_1(t)), \\ \dot{\psi}_2(t) &= -\nabla_x H_2(x^*(t), \theta_z^*(t), \theta_d(t), \psi_2(t)), \\ \psi_1^*(t_f) &= -\nabla_x \Phi(x^*(t_f), y_0) - \xi_1^T \nabla_x g(x^*(t_f)), \\ \psi_2^*(t_f) &= -\nabla_x \Phi(x^*(t_f), y_0) - \xi_2^T \nabla_x g(x^*(t_f)). \end{aligned} \tag{28}$$

Here both ξ_1 and ξ_2 are determined by $g(x^*(t_f), t_f) = 0$, so $\xi_1 = \xi_2$. Then it can be obtained from the uniqueness of linear differential terminal value problem's solution that

$$\psi_1(t) = \psi_2(t), \qquad \forall t \in [0, t_f].$$

Now let $\psi(t) \equiv \psi_1(t)$, $\forall t \in [0, t_f]$ we have

$$\begin{aligned} H(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi(t)) &= H_1(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi(t)), \\ &= H_2(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi(t)), \\ H(x^*(t), \theta_z, \theta_d^*(t), \psi(t)) =& H_1(x^*(t), \theta_z, \theta_d^*(t), \psi(t)), \\ H(x^*(t), \theta_z^*(t), \theta_d, \psi(t)) =& H_2(x^*(t), \theta_z^*(t), \theta_d, \psi(t)). \end{aligned} \tag{29}$$

By Theorem 3.1,

$$\begin{aligned} &\mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi^*(t)) \\ =& \sup_{\theta_z \in \Theta_z} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d^*(t), \psi^*(t)) \\ =& \inf_{\theta_d \in \Theta_d} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z^*(t), \theta_d, \psi^*(t)), \quad a.e.\ t \in [0, t_f] \end{aligned} \tag{30}$$

Thus

$$\begin{aligned} &\inf_{\theta_d \in \Theta_d} \sup_{\theta_z \in \Theta_z} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t)) \\ \le& \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi^*(t)) \\ \le& \sup_{\theta_z \in \Theta_z} \inf_{\theta_d \in \Theta_d} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t)). \quad a.e.\ t \in [0, t_f] \end{aligned} \tag{31}$$

On the other hand, from the well known max-min inequality,

$$\begin{aligned} &\sup_{\theta_z \in \Theta_z} \inf_{\theta_d \in \Theta_d} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t)) \\ \le& \inf_{\theta_d \in \Theta_d} \sup_{\theta_z \in \Theta_z} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t)). \quad a.e.\ t \in [0, t_f] \end{aligned} \tag{32}$$

Consequently,

$$\begin{aligned} &\mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z^*(t), \theta_d^*(t), \psi^*(t)) \\ =& \inf_{\theta_d \in \Theta_d} \sup_{\theta_z \in \Theta_z} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t)) \\ =& \sup_{\theta_z \in \Theta_z} \inf_{\theta_d \in \Theta_d} \mathbb{E}_{(x_0,y_0)\sim\mu} H(x^*(t), \theta_z, \theta_d, \psi^*(t)), \quad a.e.\ t \in [0, t_f] \end{aligned} \tag{33}$$

we have finished the proof.

B APPENDIX

Before proving Theorem 4.2, we write the express in Theorem 4.1 more compactly. For each control process $\theta_z \in L^\infty([0,t_f],\Theta_z)$ and $\theta_d \in L^\infty([0,t_f],\Theta_d)$, we denote by $x^{\theta_z,\theta_d} := \{x_t^{\theta_z,\theta_d} : 0 \le t \le t_f\}$ and $\psi^{\theta_z,\theta_d} := \{\psi_t^{\theta_z,\theta_d} : 0 \le t \le t_f\}$ the solutions of Hamilton's Equation equation 11, i.e.

$$\begin{aligned}
&\dot{x}_t^{\theta_z,\theta_d} = f(x_t^{\theta_z(t),\theta_d(t)},\theta_z,\theta_d), && x_0^{\theta_z,\theta_d} = x_0,\\
&\dot{\psi}_t^{\theta_z,\theta_d} = -\nabla_x H(x^{\theta_z,\theta_d},\theta_z(t),\theta_d(t),\psi_t^{\theta_z,\theta_d}), && \psi_{t_f}^{\theta_z,\theta_d} = -\nabla_x \Phi(x_{t_f}^{\theta_z,\theta_d},y_0) - \xi\nabla_x g(x_{t_f}^{\theta_z,\theta_d}).
\end{aligned}$$

We have the following lemma, which provides an estimate of the difference between $x^{\theta_z^1,\theta_d^1}, \psi^{\theta_z^1,\theta_d^1}$ and $x^{\theta_z^2,\theta_d^2}, \psi^{\theta_z^2,\theta_d^2}$.

Lemma B.1 *Let $\theta_z^1,\theta_z^2 \in L^\infty([0,t_f],\Theta_z)$ and $\theta_d^1,\theta_d^2 \in L^\infty([0,t_f],\Theta_d)$. Then there exists a constant T_0 such that for all $t_f \in [0,T_0)$, it holds that:*

$$\|x^{\theta_z^1,\theta_d^1} - x^{\theta_z^2,\theta_d^2}\|_{L^\infty} + \|\psi^{\theta_z^1,\theta_d^1} - \psi^{\theta_z^2,\theta_d^2}\|_{L^\infty} \le C(t_f)(\|\theta_z^1-\theta_z^2\|_{L^\infty} + \|\theta_d^1-\theta_d^2\|_{L^\infty}),$$

where $C(t_f) > 0$ satisfies $C(t_f) \to 0$ as $t_f \to 0$.

Proof B.1 (Proof of Lemma B.1) *Denote $\delta\theta_z := \theta_z^1 - \theta_z^2$, $\delta\theta_d := \theta_d^1 - \theta_d^2$, $\delta x := x^{\theta_z^1,\theta_d^1} - x^{\theta_z^2,\theta_d^2}$ and $\delta\psi := \psi^{\theta_z^1,\theta_d^1} - \psi^{\theta_z^2,\theta_d^2}$. The first two assumptions of Theorem 4.2 leads to*

$$\begin{aligned}
\|\delta x_t\| &\le \int_0^t \|f(x_s^{\theta_z^1,\theta_d^1},\theta_z^1(s),\theta_d^1(s)) - f(x_s^{\theta_z^2,\theta_d^2},\theta_z^2(s),\theta_d^2(s))\|ds\\
&\le K\int_0^{t_f}\|\delta x_s\|ds + K\int_0^{t_f}\|\delta\theta_z(s)\|ds + K\int_0^{t_f}\|\delta\theta_d(s)\|ds,
\end{aligned}$$

and so

$$\|\delta x\|_{L^\infty} \le Kt_f\|\delta x\|_{L^\infty} + Kt_f\|\delta\theta_z\|_{L^\infty} + Kt_f\|\delta\theta_d\|_{L^\infty}.$$

If $t_f \le T_0 := 1/K$, we have

$$\|\delta x\|_{L^\infty} \le \frac{Kt_f}{1-Kt_f}(\|\delta\theta_z\|_{L^\infty} + \|\delta\theta_d\|_{L^\infty}). \tag{34}$$

Similarly,

$$\begin{aligned}
\|\delta\psi_t\| &\le K\|\delta x_{t_f}\| + K\int_t^{t_f}\|\delta x_s\| + \|\delta\psi_s\| + \|\delta\theta_z(s)\| + \|\delta\theta_d(s)\|ds,\\
\|\delta\psi\|_{L^\infty} &\le (K+Kt_f)\|\delta x\|_{L^\infty} + Kt_f(\|\delta\psi\|_{L^\infty} + \|\delta\theta_z\|_{L^\infty} + \|\delta\theta_d\|_{L^\infty}),
\end{aligned}$$

hence

$$\|\delta\psi\|_{L^\infty} \le \frac{K(1+t_f)}{1-Kt_f}\|\delta x\|_{L^\infty} + \frac{Kt_f}{1-Kt_f}(\|\delta\theta_z\|_{L^\infty} + \|\delta\theta_d\|_{L^\infty}),$$

which combined with equation 34 proves the lemma.

We can now prove Theorem 4.2.

Proof B.2 (Proof of Theorem 4.2) *By uniform strong concavity and the second assumption of Theorem 4.2, there exists a $\lambda_0 > 0$ such that*

$$\begin{aligned}
\lambda_0\|\theta_z^1(t)-\theta_z^2(t)\|^2 \le{}& \big[\mathbb{E}_{\mu_0}\nabla_{\theta_z}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^1,\theta_d^1})\\
&- \mathbb{E}_{\mu_0}\nabla_{\theta_z}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^1(t),\theta_d^1(t),\psi_t^{\theta_z^1,\theta_d^1})\big]\cdot(\theta_z^1(t)-\theta_z^2(t)),\\
\lambda_0\|\theta_d^1(t)-\theta_d^2(t)\|^2 \le{}& \big[\mathbb{E}_{\mu_0}\nabla_{\theta_d}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^1,\theta_d^1})\\
&- \mathbb{E}_{\mu_0}\nabla_{\theta_d}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^1(t),\theta_d^1(t),\psi_t^{\theta_z^1,\theta_d^1})\big]\cdot(\theta_d^2(t)-\theta_d^1(t)).
\end{aligned}$$

Note that $\mathbb{E}_{\mu_0}\nabla_{\theta_z,\theta_d}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^1(t),\theta_d^1(t),\psi_t^{\theta_z^1,\theta_d^1}) = \mathbb{E}_{\mu_0}\nabla_{\theta_z,\theta_d}H(x_t^{\theta_z^2,\theta_d^2},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^2,\theta_d^2}) = 0,\ \forall t\in[0,t_f]$ *due to the optimality and continuity, then combining the two inequalities above we have*

$$
\begin{aligned}
&\lambda_0(\|\theta_z^1(t)-\theta_z^2(t)\|^2+\|\theta_d^1(t)-\theta_d^2(t)\|^2)\\
\le&\big[\mathbb{E}_{\mu_0}\nabla_{\theta_z}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^1,\theta_d^1})\\
&-\mathbb{E}_{\mu_0}\nabla_{\theta_z}H(x_t^{\theta_z^2,\theta_d^2},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^2,\theta_d^2})\big]\cdot(\theta_z^1(t)-\theta_z^2(t))\\
+&\big[\mathbb{E}_{\mu_0}\nabla_{\theta_d}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^1,\theta_d^1})\\
&-\mathbb{E}_{\mu_0}\nabla_{\theta_d}H(x_t^{\theta_z^2,\theta_d^2},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^2,\theta_d^2})\big]\cdot(\theta_d^2(t)-\theta_d^1(t))\\
\le&\mathbb{E}_{\mu_0}\|\nabla_{\theta_z}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^1,\theta_d^1})\\
&-\nabla_{\theta_z}H(x_t^{\theta_z^2,\theta_d^2},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^2,\theta_d^2})\|\|\theta_z^1(t)-\theta_z^2(t)\|\\
+&\mathbb{E}_{\mu_0}\|\nabla_{\theta_d}H(x_t^{\theta_z^1,\theta_d^1},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^1,\theta_d^1})\\
&-\nabla_{\theta_d}H(x_t^{\theta_z^2,\theta_d^2},\theta_z^2(t),\theta_d^2(t),\psi_t^{\theta_z^2,\theta_d^2})\|\|\theta_d^1(t)-\theta_d^2(t)\|\\
\le&K(\|\delta x\|_{L^\infty}+\|\delta\psi\|_{L^\infty})(\|\delta\theta_z\|_{L^\infty}+\|\delta\theta_d\|_{L^\infty}).
\end{aligned}
$$

Combining the above with Lemma B.1, we have

$$
\|\delta\theta_z\|_{L^\infty}^2+\|\delta\theta_d\|_{L^\infty}^2\le\frac{KC(t_f)}{\lambda_0}(\|\delta\theta_z\|_{L^\infty}+\|\delta\theta_d\|_{L^\infty})^2\le\frac{2KC(t_f)}{\lambda_0}(\|\delta\theta_z\|_{L^\infty}^2+\|\delta\theta_d\|_{L^\infty}^2).
$$

$C(t_f)\to 0$ *as* $t_f\to 0$*, by taking* t_f *sufficiently small, so that* $2KC(t_f)<\lambda_0$*, which implies* $\|\delta\theta_z\|_{L^\infty}=\|\delta\theta_d\|_{L^\infty}=0$*.*

C APPENDIX

Now we introduce the definition of viscosity solution. Consider a function $v(t,\mathbb{P}_X):[0,t_f]\times\mathcal{P}_2(\mathbb{R}^{n+m})\to\mathbb{R}$, the Hamiltonian $H(X,\partial_{\mathbb{P}_X}v(t,\mathbb{P}_X)(X)):L^2(\Omega,\mathbb{R}^{n+m})\times L^2(\Omega,\mathbb{R}^{n+m})\to\mathbb{R}$ and $\Psi:L^2(\Omega,\mathbb{R}^{n+m})\to\mathbb{R}$, where v satisfies

$$
\begin{aligned}
&\frac{\partial v}{\partial t}+H(X,\partial_{\mathbb{P}_X}v(t,\mathbb{P}_X)(X))=0, && on\ [0,t_f)\times L^2(\Omega,\mathbb{R}^{n+m}),\\
&v(t_f,\mathbb{P}_X)=\Psi(X), && on\ L^2(\Omega,\mathbb{R}^{n+m}).
\end{aligned}
\tag{35}
$$

Then the lifted function $V(t,X)=v(t,\mathbb{P}_X)$ satisfies

$$
\begin{aligned}
&\frac{\partial V}{\partial t}+H(X,D_XV(t,X))=0, && on\ [0,t_f)\times L^2(\Omega,\mathbb{R}^{n+m}),\\
&V(T,X)=\Psi(X), && on\ L^2(\Omega,\mathbb{R}^{n+m}).
\end{aligned}
\tag{36}
$$

We say that a bounded, uniformly continuous function $u:[0,t_f]\times\mathcal{P}_2(\mathbb{R}^{n+m})\to\mathbb{R}$ is a viscosity solution to equation 35 if its lifted function $U:[0,t_f]\times L^2(\Omega,\mathbb{R}^{n+m})\to\mathbb{R}$ defined by

$$
U(t,X)=u(t,\mathbb{P}_X),
$$

is a viscosity solution to the lifted equation equation 36, namely:

i) $U(t_f,X)\le\Psi(X)$ and for any test function $\gamma\in C^{1,1}([0,t_f]\times L^2(\Omega,\mathbb{R}^{n+m}))$ such that the map $U-\gamma$ has a local maximum at $(t_0,X_0)\in[0,t_f)\times L^2(\Omega,\mathbb{R}^{n+m})$, one has

$$
\partial_t\gamma(t_0,X_0)+H(X_0,D\gamma(t_0,X_0))\ge 0.
$$

ii) $U(t_f,X)\ge\Psi(X)$ and for any test function $\gamma\in C^{1,1}([0,t_f]\times L^2(\Omega,\mathbb{R}^{n+m}))$ such that the map $U-\gamma$ has a local minimum at $(t_0,X_0)\in[0,t_f)\times L^2(\Omega,\mathbb{R}^{n+m})$, one has

$$
\partial_t\gamma(t_0,X_0)+H(X_0,D\gamma(t_0,X_0))\le 0.
$$

For further details we refer the interested readers to (E et al., 2019).

Proof C.1 (Proof of Theorem 4.3) *Suppose $v'(t,\mu)$ is a viscosity solution to equation 16 and (θ'_z, θ'_d) is the corresponding optimal strategy.*

We first fix θ'_z, consider

$$\begin{aligned}
&\partial_t v_1(t,\mu) + \sup_{\theta_d \in \Theta_d} \left\{ \int_{\mathbb{R}^{n+m}} [\partial_\mu v(t,\mu)(x,y)]^T [f(x,\theta'_z,\theta_d),0] + L(x,\theta'_z,\theta_d) d\mu(x,y) \right\} = 0, \\
&v_1(t_f,\mu) = \int_{\mathbb{R}^{n+m}} \Phi(x,y) d\mu(x,y).
\end{aligned} \tag{37}$$

By Theorem 1 and Theorem 2 in (E et al., 2019), $v'(t,\mu)$ is the unique viscosity solution to equation 37 satisfies

$$v'(t,\mu) = \sup_{\theta_d \in \mathcal{U}_d} \mathbb{E}_{(x,y)\sim\mu} \left[\int_t^{t_f} \Phi(x(t_f),y) + L(x(t),\theta'_z(t),\theta_d(t)) dt \right]. \tag{38}$$

Then fix θ'_d, similarly we have

$$v'(t,\mu) = \inf_{\theta_z \in \mathcal{U}_z} \mathbb{E}_{(x,y)\sim\mu} \left[\int_t^{t_f} \Phi(x(t_f),y) + L(x(t),\theta_z(t),\theta'_d(t)) dt \right]. \tag{39}$$

Now for (θ'_z, θ'_d), equation 15 is satisfied, thus $v'(t,\mu) = v^*(t,\mu)$.

D APPENDIX

Proof D.1 (Proof of Theorem 5.2) *We first give a lemma, then use it alternatively in the training process of the two networks to prove the theorem.*

Define the generalization bound taking expectation with respect to randomized algorithm

$$er(w) := \mathbb{E}_{\mathcal{A}}(\mathbb{E}l(w;z) - \hat{\mathbb{E}}_N l(w;z)),$$

where $\hat{\mathbb{E}}_N l(w;z)$ is the average loss at N sample points.

Lemma D.1 (Theorem 8 in (Mou et al., 2018)) *Consider n rounds of SGLD with parameters β and $\{\eta_i\}$. Suppose the loss function $l(w;z)$ is uniformly bounded by C, and $\forall z, z'$, there is $\|\nabla l(w;z) - \nabla l(w;z')\| \leq L$. By setting k_0 such that $\eta_{k_0} \leq \ln 2/\beta L^2$, then we have the following generalization bound in expectation*

$$\mathbb{E}[er(w_n)] \leq \frac{2k_0}{N} + \frac{\sqrt{\beta}LC}{N}\Big(\sum_{i=k_0+1}^{n} \eta_i\Big)^{1/2}.$$

Now let $l(w;z) = J^0(\theta_z,\theta_d;X)$. During the training process, θ_z and θ_d are updated alternately, and the generalization error is accumulating. Using Lemma D.1 alternatively, we can finish the proof.