Appendix

Appendix A Proof Details of the Theoretical Analysis

A.1 Generalization Error

In this section, we analyze the generalization error on the model learning task. We denote $\mathcal{F}$ and $\mathcal{G}$ as the function spaces of $\mathbf{f}$ and $\mathbf{g}$, respectively, and the $\mathcal{D}$ as the function space of the $\{D_t\}_{t=0}^T$, where T stands for the number of steps, and $\mathbf{g}^{\circ t}(x, \xi_t) = \underbrace{((I+\mathbf{g}) \circ (I+\mathbf{g}) \circ \ldots \circ (I+\mathbf{g}))}_{t}(x) + \xi_t$ with $\xi_t \sim \mathcal{N}(0, \Delta t)$. We define

$$\ell(\mathbf{f}, \mathbf{g}) = \mathbb{E}_{y_{0:T}, x_0 \sim p(x), \xi_{0:T}} \left[\sum_{t=0}^{T} \max_{D_t \in \mathcal{D}} \left[D_t(y_t) - D_t((\mathbf{f} \circ \mathbf{g}^{\circ t}(x_0, \xi_t))) \right] \right] := \ell_t(\mathbf{f}, \mathbf{g}),$$

where

$$\ell_t(\mathbf{f}, \mathbf{g}) = \mathbb{E}_{y_t, x_0, \xi_{0:T}} \left[\max_{D_t \in \mathcal{D}} \underbrace{\left[D_t(y_t) - D_t((f \circ g^{\circ t}(x_0, \xi_t))) \right]}_{\phi_t(\mathbf{f}, \mathbf{g}, D_t)} \right].$$

Without the loss of generality, we assume in each timestamp the number of the observations is N. Given the samples $\mathcal{Y} = \{(y_t^i)_{t=0}^T\}_{i=1}^N$, where $y_{0:T} = (y_t^i)_{t=0}^T$ are sampled *i.i.d.* from the underline stochastic processes, and $\mathcal{X} = \{x_0^i\}_{i=1}^N$, $\Xi = \{\xi_{0:T}^i\}_{i=1}^N$ are also *i.i.d.* sampled, we have the empirical loss function as

$$\hat{\ell}(\mathbf{f}, \mathbf{g}) = \hat{\mathbb{E}}_{\mathcal{Y}} \hat{\mathbb{E}}_{\mathcal{X}} \left[\sum_{t=0}^{T} \max_{D_t \in \mathcal{D}} \left[D_t(y_t) - D_t((\mathbf{f} \circ \mathbf{g}^{\circ t}(x_0, \xi_t))) \right] \right] = \sum_{t=0}^{T} \hat{\ell}_t(\mathbf{f}, \mathbf{g}).$$

With the notations defined above, we provide the proof for Theorem 1 as below.

Proof. Denote the $\hat{\mathbf{f}}$ and $\hat{\mathbf{g}}$ are the solutions provided by the algorithm, and $\mathbf{f}^*$ and $\mathbf{g}^*$ be the optimal solutions, we have

$$\begin{aligned} |\ell_t(\hat{\mathbf{f}}, \hat{\mathbf{g}}) - \ell_t(\mathbf{f}^*, \mathbf{g}^*)| &= \left| \mathbb{E}[\max_{D_t \in \mathcal{D}} \phi_t(\hat{\mathbf{f}}, \hat{\mathbf{g}}, D_t)] - \mathbb{E}[\max_{D_t \in \mathcal{D}} \phi_t(\mathbf{f}^*, \mathbf{g}^*, D_t)] \right| \\ &\leq \left| \max_{D_t \in \mathcal{D}} \mathbb{E}[\phi_t(\hat{\mathbf{f}}, \hat{\mathbf{g}}, D_t) - \phi_t(\mathbf{f}^*, \mathbf{g}^*, D_t)] \right| \\ &\leq 2 \sup_{\mathbf{f} \in \mathcal{F}, \mathbf{g} \in \mathcal{G}, D_t \in \mathcal{D}} |\hat{\Phi}_t(\hat{\mathbf{f}}, \hat{\mathbf{g}}, D_t) - \Phi_t(\mathbf{f}^*, \mathbf{g}^*, D_t)|, \end{aligned}$$

where

$$\begin{aligned} \hat{\Phi}_t(\hat{\mathbf{f}}, \hat{\mathbf{g}}, D_t) &= \hat{\mathbb{E}}_{y_t \in \mathcal{Y}_t} \hat{\mathbb{E}}_{x_0, \xi_t} \big[\phi_t(\hat{\mathbf{f}}, \hat{\mathbf{g}}, D_t) \big], \\ \Phi_t(\mathbf{f}^*, \mathbf{g}^*, D_t) &= \mathbb{E} \big[\phi_t(\mathbf{f}^*, \mathbf{g}^*, D_t) \big]. \end{aligned}$$

Assume $\mathcal{D} \in \mathcal{L}_k$, where $\mathcal{L}_k$ denotes the k-Lipschitz function space, and $|\mathcal{Y}|_\infty = C_{\mathcal{Y}}$, we have,

$$\begin{aligned} &\sup_{\mathbf{f} \in \mathcal{F}, \mathbf{g} \in \mathcal{G}, D_t \in \mathcal{D}} |\hat{\Phi}_t(\hat{\mathbf{f}}, \hat{\mathbf{g}}, D_t) - \Phi_t(\mathbf{f}^*, \mathbf{g}^*, D_t)| \leq 2\mathbb{E} \left[\sup_{\mathbf{f} \in \mathcal{F}, \mathbf{g} \in \mathcal{G}, D_t \in \mathcal{D}} \left| \frac{1}{N} \sum_{i=1}^{N} \tau_i \phi_t(\mathbf{f}, \mathbf{g}, D_t) \right| \right] \\ \leq \; & 2\mathbb{E} \left[\sup_{D_t \in \mathcal{D}} \left| \frac{1}{N} \sum_{i=1}^{N} \tau_i D_t(y_i) \right| \right] + 2\mathbb{E} \left[\sup_{\mathbf{f} \in \mathcal{F}, \mathbf{g} \in \mathcal{G}, D_t \in \mathcal{D}} \left| \frac{1}{N} \sum_{i=1}^{N} \tau_i D_t((\mathbf{f} \circ \mathbf{g}^{\circ t}(x_0, \xi_t))) \right| \right] \\ \leq \; & 2\frac{kC}{\sqrt{N}} + 2k\mathbb{E} \left| \frac{1}{N} \sum_{i=1}^{N} \tau_i \mathbf{f} \circ \mathbf{g}^{\circ t}(x_0, \xi_t)) \right| = 2\frac{kC}{\sqrt{N}} + 2k\mathfrak{R}(\mathcal{F} \circ \mathcal{G}^{\circ t}), \end{aligned}$$

where the $\mathfrak{R}(\mathcal{F} \circ \mathcal{G}^{\circ t})$ denotes the Rademacher complexity of the function space $\mathcal{F} \circ \mathcal{G}^{\circ t}$. Therefore, we have

$$\frac{1}{T} \ell(\mathbf{f}, \mathbf{g}) \leq \frac{1}{T} \hat{\ell}(\mathbf{f}, \mathbf{g}) + \frac{4kC}{\sqrt{N}} + 4\frac{k \sum_{i=1}^{T} \mathfrak{R}(\mathcal{F} \circ \mathcal{G}^{\circ t})}{T}.$$

□

A.2 Convergence Analysis

Inspired by (Dai et al., 2017), we can see that once we obtain the D_t^*, the Algorithm 1 can be understood as a special case of stochastic gradient descent for non-convex problem. We prove the Theorem 2 as below.

Proof. We compute the gradient of $\ell(\mathbf{f}, \mathbf{g})$ w.r.t. $\mathbf{f}$, the same argument is also for gradient w.r.t. gb.

$$\nabla_{\mathbf{f}}\ell(\mathbf{f},\mathbf{g}) = \nabla_{\mathbf{f}}\mathbb{E}\Big[\sum_{t=0}^{T}\phi_t(\mathbf{f},\mathbf{g},D_t^*)\Big] = \mathbb{E}\Big[\sum_{t=0}^{T}\nabla_{\mathbf{f}}\phi_t(\mathbf{f},\mathbf{g},D_t^*)\Big] \tag{34}$$

$$= \mathbb{E}\Big[\sum_{t=0}^{T}(\nabla_{\mathbf{f}}\phi_t(\mathbf{f},\mathbf{g},D_t^*) + \underbrace{\nabla_{D_t^*}\phi_t(\mathbf{f},\mathbf{g},D_t^*)\nabla_{\mathbf{f}}D_t^*(\mathbf{f}\circ\mathbf{g}^{\circ t})}_{0})\Big] \tag{35}$$

$$= -\mathbb{E}\Big[\sum_{t=0}^{T}\nabla_{\mathbf{f}}D_t^*(\mathbf{f}\circ\mathbf{g}^{\circ t})\Big] \tag{36}$$

The second term in the last second line is zero due to the optimality of D^*. Therefore, we achieve the unbiasedness of the gradient estimators.

As long as the gradient estimator for $\mathbf{f}$ and $\mathbf{g}$ are unbiased, the convergence rate in Theorem 2 will be automatically hold from (Ghadimi & Lan, 2013). □