PROVABLE CONVERGENCE AND GLOBAL OPTIMALITY OF GENERATIVE ADVERSARIAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative adversarial networks (GANs) train implicit generative models through solving minimax problems. Such minimax problems are known as nonconvex-nonconcave, for which the dynamics of first-order methods are not well understood. In this paper, we consider GANs in the type of the integral probability metrics (IPMs) with the generator represented by an overparametrized neural network. When the discriminator is solved to approximate optimality in each iteration, we prove that stochastic gradient descent on a regularized IPM objective converges globally to a stationary point with a sublinear rate. Moreover, we prove that when the width of the generator network is sufficiently large and the discriminator function class has enough discriminative ability, the obtained stationary point corresponds to a generator that yields a distribution that is close to the distribution of the observed data in terms of the total variation. To the best of our knowledge, we seem to first establish both the global convergence and global optimality of training GANs when the generator is parametrized by a neural network.

1 INTRODUCTION

Generative adversarial network (GANs) learn generative models from unlabeled data (Goodfellow et al., 2014). GANs aim to learn a generator, a transformation of a known distribution, which is close to the unknown distribution of the real data. To measure the proximity between the generator and the distribution of interest, another function, known as the discriminator, is learned to distinguish the real data from the artificial samples created by the generator. Recently, GANs have demonstrated tremendous empirical success in generating realistic data in various domains such as computer vision (Radford et al., 2016; Goodfellow, 2016) and natural language processing (Rajeswar et al., 2017; Wang et al., 2019).

Despite their wide popularity, theoretical justification of GANs is rather limited due to the following two reasons. First, when both the generator and discriminator are represented by neural networks, the optimization problem associated with GANs is nonconvex-nonconcave. Unlike convex-concave minimax problems that have been well studied in existing literature (Korpelevich, 1976; Nemirovski, 2004; Nedić & Ozdaglar, 2009; Nemirovski & Yudin, 1978), the behavior of gradient descent-ascent algorithm is much more complicated when the objective function is nonconvex-nonconcave – the sequence of iterates might either diverge, converge to a locally minimax solution, or possess a limiting cycle (Mescheder et al., 2018; Daskalakis & Panageas, 2018; Letcher et al., 2019; Zhu et al., 2019; Jin et al., 2019). Although convergence guarantees have been established for a few variants of gradient methods for training GANs (Heusel et al., 2017; Mescheder et al., 2017; Li et al., 2018), nearly all of these results are local and based on assumptions that are not easily satisfied or verifiable for GANs used in practice. Second, from a statistical perspective, the minimax formulation of GANs results in a trade-off between generalization and discriminative ability (Arora et al., 2017; Zhang et al., 2018c). In particular, when the discriminator function class $\mathcal{F}$ have perfect discriminative power, although it can uniquely identify the true distribution, the generalization error cannot be controlled due to the huge capacity of $\mathcal{F}$. Whereas when $\mathcal{F}$ is not sufficiently discriminative, it can fail to differentiate a generator distribution that is disparate from the underlying truth, which leads to the mode collapse phenomenon observed in practice (Arora & Zhang, 2018; Srivastava et al., 2017).

In this work, we address these two challenges for GANs where the performance of the generator is characterized by a integral probability metric (Müller, 1997) induced by the discriminator function

class $\mathcal{F}$. Specifically, an IPM-based GAN considers the the following optimization problem:

$$\min_{g\in\mathcal{G}} \max_{f\in\mathcal{F}} L(f,g) = \mathbb{E}\big[f\big(g(Z)\big) - f(Y)\big], \tag{1.1}$$

where g and f are the generator and discriminator, respectively. Besides, the real data consist of i.i.d. observations of the unknown random variable $Y \in \mathbb{R}^D$, $Z \in \mathbb{R}^d$ follows a fixed and known distribution, and $\max_{f\in\mathcal{F}} L(f,g)$ is the IPM between the distributions of $g(Z)$ and Y. In particular, when $\mathcal{F}$ is the family of 1-Lipschitz functions, the IPM induced by $\mathcal{F}$ is the Wasserstein distance and the formulation in (1.1) recovers the well-known Wasserstein GAN (Arjovsky et al., 2017). The IPM-based GANs offer a flexible framework that enables us to directly take the discriminative ability of function class $\mathcal{F}$ into consideration and includes various variants of GANs as special cases (Li et al., 2017; Mroueh et al., 2017; Unterthiner et al., 2017).

We provide a joint statistical and computational analysis for IPM-based GANs where the generator is represented by an overparametrized neural network g_θ where $\theta \in \Theta$ is the network weights. That is, the number of neurons in the generator network is extremely large and can even goes to infinity. Under this setting, in terms of computation, we propose to learn the generator by minimizing a regularized IPM, viewed as a function of θ, via stochastic gradient descent. Assuming having access to an oracle which approximately solves the maximization problem of the discriminator in (1.1) within some accuracy for any fixed generator, we prove that stochastic gradient descent converges globally to an ϵ-stationary point within $\mathcal{O}(\epsilon^{-2})$ iterations.

More importantly, through a statistical lens, we build a connection between the global optimality of the obtained stationary point and the discriminative ability of the discriminator function class $\mathcal{F}$. In particular, we prove that, when $\mathcal{F}$ is a sufficiently rich function class and the width of the generator network is large enough, the optimal discriminator $f^*_{\theta^*} \in \mathcal{F}$ corresponding to any stationary point θ^* satisfies that $\|\nabla f^*_{\theta^*}\|_2$ diminishes to zero on the support of $g_{\theta^*}(Z)$. In other words, when $\mathcal{F}$ is sufficiently discriminative, the optimal discriminator converges to a constant function in the overparametrization regime, which further implies that the distribution of $g_{\theta^*}(Z)$ matches that of Y.

Our theoretical results are based on a careful combination of the geometry of (1.1) and the technique of the neural tangent kernel (NTK) (Jacot et al., 2018). In specific, the NTK theory demonstrates that any overparametrized neural network determines a kernel function induced by the random initialization scheme, which governs the expressiveness of the neural network when its parameter is within a neighborhood of the initialization. As we will show in the proof, such a kernel enables us to link the updates in the parameter space to those in the function space. That is, as we update the parameters of the generator via gradient descent, we equivalently tune the generator in the function space. Moreover, when lifting the generator to the function space, the IPM has a unique minimizer which is the unknown distribution of the real data. Thus, combining this fact and the NTK theory enable us to characterize the global optimality of a stationary point.

Our Contributions. Our contributions are twofold. First, for IPM-based GANs where the generator is represented as a overparameterized neural network, we prove that gradient descent converges globally to a stationary point of the regularized IPM with a sublinear rate. Second, by relating the optimality of any stationary point to the expressiveness of the generator network and the discriminative ability of the discriminator function class $\mathcal{F}$, we prove that any stationary point of the regularized IPM yields a distribution that is close to the distribution of the observed real data in terms of total variation. To the best of our knowledge, our results are the first provide global convergence and optimality guarantees for training GANs via first-order methods.

Related Work. The vanilla GAN (Goodfellow et al., 2014) is known to suffer from issues such as unstable training, vanishing gradient (Arjovsky & Bottou, 2017), and mode collapse (Arjovsky et al., 2017; Arora & Zhang, 2018). Various modifications have been proposed to alleviate these problems by improving the performance of the learned distribution or the stability of training dynamics. See, e.g., Nowozin et al. (2016); Huang et al. (2017); Odena et al. (2017); Arjovsky et al. (2017); Gulrajani et al. (2017); Karras et al. (2017); Berthelot et al. (2017); Li et al. (2017); Mroueh et al. (2017); Unterthiner et al. (2017); Zhang et al. (2018a;b); Hsieh et al. (2019) and the references therein. Our IPM-based GANs originates from the the Wasserstein GAN (Arjovsky et al., 2017) and also include the models studied in Li et al. (2017); Mroueh et al. (2017); Unterthiner et al. (2017) as special cases. It is shown in (Arjovsky et al., 2017) that IPM-based GANs is free of the vanishing gradient problem and it is desirable to solve for the optimal discriminator for each iterate of the generator,

which motivates our usage of an oracle that solves the inner maximization problem to approximate optimality in each iteration of the stochastic gradient update.

Moreover, there is growing research interest in theoretical analysis of GANs. From a statistical perspective, Jiang et al. (2019); Zhang et al. (2018c); Zhu et al. (2019) establish the generalization errors of GANs, Liang (2018) study the statistical error of IPM-based GANs under the nonparametric setup, and Zhang et al. (2018c); Arora et al. (2017); Bai et al. (2018) study the tradeoff between the generalization error and discriminative power. In addition, there is a line of research on convergent algorithms for solving the minimax optimization problem associated with GANs. For example, Daskalakis et al. (2017); Feizi et al. (2017); Mescheder et al. (2017); Nagarajan & Kolter (2017); Hsieh et al. (2019); Liang & Stokes (2019) propose convergent first-order methods that update the generator and discriminator either simultaneously or alternatively. However, their convergence guarantees either rely on local stability assumptions on the iterates, or require the objective function to possess a convex-concave structure locally or globally. A more related line of work is on studying the optimization problem of GAN only from the perspective of the generator, assuming the subproblem of the discriminator can be solved approximately to the global optima. In particular, Li et al. (2018) consider the problem of learning a one-dimensional Gaussian mixture or a two-dimensional Gaussian distribution. They prove global convergence to optimality for gradient updates of the generator, assuming the discriminator problem is solved to global maximum at each step. In addition, Hsieh et al. (2019) utilize the two-timescale updating rule (Borkar, 1997) where the discriminator updates much faster than the generator, which essentially is equivalent to solving the discriminator problem completely for each iterate of the generator. Compared with our work, they require additional local stability assumptions and their convergence guarantee is only asymptotic. A more related work is Sanjabi et al. (2018), which propose to train Wasserstein GAN via regularized optimal transport. Although they also consider gradient descent for a regularized IPM, their algorithm relies on the dual formulation of the Wasserstein distance (Villani, 2008) and thus might not be applied to IPM-based GANs in general. Furthermore, they only establish global convergence to a stationary point whereas we also characterize the optimality of the obtained stationary point.

Furthermore, our work is also related to the recent literature on the training and generalization error of training overparametrized neural networks via (stochastic) gradient descent. See, e.g., (Daniely, 2017; Lee et al., 2017; Chizat & Bach, 2018; Du et al., 2018a;b; Li & Liang, 2018; Allen-Zhu et al., 2018b;a; Zou et al., 2018; Arora et al., 2019a;b; Lee et al., 2019) and the references therein. Compared with these work, the dynamics of our gradient descent algorithm are more complicated due to the minimax formulation in (1.1), which makes the convergence analysis more challenging.

Notation. We use $\|\cdot\|_2$ to represent the ℓ^2-norm of a vector or the spectral norm of a matrix or the L^2-norm of a function. For a subset of D-dimensional Euclidean space $S \subseteq \mathbb{R}^D$, we use $|S|$ to represent its Lebesgue measure. We denote by $B_r(x)$ a ball neighborhood of a point x with radius r.

2 Background

In this section, we present the background knowledge of generative adversarial networks and integral probability metrics.

2.1 Learning Implicit Generative Models

An implicit generative model uses a map $g_\theta(Z)$ to generate data that are similar to the real data $Y \sim \mathbb{Q}$. Here θ is the parameter of the map and Z is the random variable with a fixed and known distribution $\mathbb{P}_Z$. The map g_θ is usually referred as the generator and we denote the learned distribution by $\mathbb{P}_\theta$. When g_θ is a complicated function, the distribution of $g_\theta(Z)$ is unattainable in closed form even when $\mathbb{P}_Z$ is a simple distribution such as uniform or Gaussian. However, it is easier to generate samples $\{g_\theta(Z_i)\}_{i=1}^n$ where $\{Z_i\}_{i=1}^n$ are i.i.d. random variables draw from $\mathbb{P}_Z$. Generative adversarial networks provide an algorithm framework that learns an implicit generative model using samples. Specifically, a discriminator $\mathcal{D}$ is trained to distinguish the samples from the model and real data. Therefore, to train a good generative model, it suffice to learn to confuse the best possible discriminator, which leads to a minimax optimization problem

$$\min_{\theta} \max_{\mathcal{D}} \mathbb{E}\Big[c_1\Big(\mathcal{D}\big(g_\theta(Z)\big)\Big) + c_2\Big(\mathcal{D}(Y)\Big)\Big]. \tag{2.1}$$

Here $\mathcal{D}(\cdot)$ returns the discrimination results and c_1, c_2 are the corresponding loss functions with regard to two types of data.

2.2 Integral Probability Metrics

The vanilla GAN (Goodfellow et al., 2014) proposes to let $\mathcal{D}$ return a probability of classifying a data point to be obtained from $\mathbb{Q}$ or $\mathbb{P}_\theta$, and let $c_1(x) = \log(x), c_2(x) = \log(1-x)$. Then the objective in (2.1) can be interpreted as the log likelihood of correct classification over all data points. Furthermore, the solution to the inner maximization problem indeed provides the Jensen-Shannon divergence between $\mathbb{P}_\theta$ and $\mathbb{Q}$. Similarly, other distances can also serve as optimization objectives of GANs. For example, using the IPM induced by a function class $\mathcal{F}$ yields

$$\min_{\theta\in\Theta} \text{IPM}_{\mathcal{F}}(\mathbb{P}_\theta, \mathbb{Q}) := \sup_{f\in\mathcal{F}} \big|\mathbb{E}_{X\sim\mathbb{P}_\theta}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]\big|. \tag{2.2}$$

Such formulation covers several distance definitions between probabilities. For example, the Wasserstein distance is obtained by (2.2) when $\mathcal{F}$ is all 1-Lipschitz continuous function on $\mathbb{R}^D$. Also, the maximum mean discrepancy (MMD) corresponds to $\mathcal{F}$ being a unit ball in a reproducing kernel Hilbert space, and the total variation (TV) distance corresponds to $\mathcal{F}$ being all indicator functions of Borel measurable sets in $\mathbb{R}^D$.

3 Stochastic Gradient Descent for GANs

In this section, we introduce the problem formulation and the optimization algorithm.

3.1 Generator as a Two-layer Neural Network

We parametrizethe generator using a two-layer neural network $g_\theta : \mathbb{R}^d \to \mathbb{R}^D$, which is given by

$$g_\theta(z) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \sigma([\theta]_r^\top z) \cdot a_r.$$

Such architecture has a hidden layer with m neurons where m is typically very large. Here $\sigma(\cdot)$ is an activation function, which is assumed to be 1-Lipschitz and 1-smooth. Vectors $\{[\theta]_r\}_{r=1}^m \subseteq \mathbb{R}^d$ are the weights of the hidden layer, and vectors $\{a_r\}_{r=1}^m \subset \mathbb{R}^D$ are the weights of the output layer. For all neuron index $r \in [m]$, we independently initialize the parameters $[\theta]_r$ and a_r by $[\theta]_r \sim N(0, I_d/d)$ and $a_r \sim \text{Unif}(\{-1/\sqrt{D}, 1/\sqrt{D}\}^D)$. During training, we fix the output layer weights $\{a_r\}_{r=1}^m$ as their initial values and only update the values of $\{[\theta]_r\}_{r=1}^m$, which is denote by a long vector $\theta \in \mathbb{R}^{md}$. The iterates of θ during the algorithm are then denoted by $\{\theta_t\}_{t\geq 0}$. Our generative model is then denoted by $g_\theta(Z)$ and we let $\mathbb{P}_Z$ be the uniform distribution over a compact set $\mathcal{Z} \subseteq \mathbb{R}^d$. We denote by p_z the costant probability density of $\mathbb{P}_Z$. Moreover, we focus on the case where both $\mathbb{P}_\theta$ and $\mathbb{Q}$ are continuous distributions in $\mathbb{R}^D$, which implies that the input dimension d of g_θ is no less than D. We denote their densities by $p_\theta(\cdot)$ and $q(\cdot)$, respectively.

In the following lemma, we show the neural network generator enjoys smoothness in its parameters.

Lemma 3.1. For any 1-Lipschitz and 1-smooth activation σ and input $z \in \mathbb{R}^d$ such that $\|z\|_2 \leq 1$, $g_\theta(z)$ is 1-Lipschitz and $(1/\sqrt{m})$-smooth in θ. In other words, we have $\|g_\theta(z) - g_{\theta'}(z)\|_2 \leq \|\theta - \theta'\|_2$ and $\|\nabla_\theta g_\theta(z) - \nabla_\theta g_{\theta'}(z)\|_{\text{F}} \leq 1/\sqrt{m} \cdot \|\theta - \theta'\|_2$ for any $\theta, \theta' \in \mathbb{R}^{md}$.

3.2 Minimax Formulation of a Regularized IPM

In the sequel, we focus on the following minimax formulation of a regularized IPM:

$$\min_{\theta\in\Theta} \max_{f\in\mathcal{F}} L(\theta, f) := \mathbb{E}_{X\sim\mathbb{P}_\theta}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)] - \gamma \cdot \psi(f). \tag{3.1}$$

Here we assume $\mathcal{F}$ is a convex and compact function class and satisfy $0 \in \mathcal{F}$, which enables us to remove the absolute value operation in (2.2). Besides, for any $f \in \mathcal{F}$, we further assume that f is

ν_1-Lipschitz continuous and $f(0) = 0$. Moreover, $\psi(\cdot)$ in (3.1) is a 1-strongly convex regularizer on f such that, for any $\theta \in \Theta$ and $f_1, f_2 \in \mathcal{F}$, we have

$$L(\theta, f_1) \leq L(\theta, f_2) + \mathrm{D}(f_1 - f_2, f_2) - \gamma/2 \cdot \|f_2 - f_1\|^2, \tag{3.2}$$

where $\mathrm{D}(\cdot, f_1)$ is a linear functional and $\mathrm{D}(f_1 - f_2, \cdot)$ is a continuous functional. The norm in (3.2) may depend on the choices of the class $\mathcal{F}$ and the regularizer ψ. For example, when $\mathcal{F} \subseteq L^2(\mathbb{R}^D)$, we may use the L^2-norm $\|f\|_2^2 = \int_{\mathbb{R}^D} |f(x)|^2 \mathrm{d}x$ and let $\psi(f) = \|f\|_2^2/2$. It follows that

$$\mathrm{D}(f_1 - f_2, f_2) = \langle f_1 - f_2, p_\theta - q - \gamma \cdot f_2 \rangle.$$

Moreover, we assume $\psi(f)$ takes its minimum zero at $f = 0$.

Since $\mathcal{F}$ is convex and compact and $\psi(\cdot)$ is strongly convex, there is a unique optimal discriminator for any generator θ, which is denoted by $f_\theta^* = \operatorname{argmax}_{f \in \mathcal{F}} L(\theta, f)$. Note that a Lipschitz continuous function is almost surely differentiable. Since both $\mathbb{P}_\theta$ and $\mathbb{Q}$ have probability densities, $\nabla f(X)$ almost surely exists under $\mathbb{P}_\theta$ and $\mathbb{Q}$, which further implies that $L(\theta, f)$ is differentiable for any θ. Then, by Danskin's Theorem (Rockafellar, 1970), the regularized IPM $\Phi(\theta) := \max_{f \in \mathcal{F}} L(\theta, f) = L(\theta, f_\theta^*)$ is differentiable and its gradient is given by

$$\nabla \Phi(\theta) = \nabla_\theta L(\theta, f)|_{f = f_\theta^*} = \mathbb{E}_{Z \sim \mathbb{P}_Z} \Big[\nabla_\theta \Big(f\big(g_\theta(Z)\big) \Big) \Big] \Big|_{f = f_\theta^*}. \tag{3.3}$$

In particular, when $\mathbb{P}_\theta = \mathbb{Q}$, we have $f_\theta^* = 0$ and $\nabla \Phi(\theta) = 0$.

3.3 Gradient Algorithm

The gradient in (3.3) can be further written as $\nabla \Phi(\theta) = \mathbb{E}_{Z \sim \mathbb{P}_Z} \big[\nabla_\theta g_\theta(Z) (\nabla f_\theta^*) \big(g_\theta(Z) \big) \big]$ where $\nabla_\theta g_\theta(z)$ is the Jacobian matrix of $g_\theta(z)$ with respect to θ.

According to Arjovsky et al. (2017), the discriminator problem is preferably to be exactly solved in each iteration so that we obtain accurate gradient information to decrease IPM. Thanks to the recent breakthroughs in deep learning, especially the optimization of overparametrized neural networks (Arora et al., 2019a), it is shown that stochastic gradient descent can obtain an arbitrarily accurate solution to problems that are strongly convex to the network outputs, which implies the global optimality of the inner maximization problem in (3.1) is attainable. Therefore, for simplicity, we assume having access to a maximization oracle which returns an approximately optimal discriminator for any g_θ up to an error $\varepsilon_f > 0$. This approach is also adopted in Sanjabi et al. (2018).

Assumption 3.2 (max-oracle)**.** We assume that we have an oracle such that for any $\theta \in \Theta$, it returns a discriminator $f_\theta^o \in \mathcal{F}$ satisfying $L(\theta, f_\theta^o) \geq L(\theta, f_\theta^*) - \varepsilon_f$.

With the max-oracle specified in Assumption 3.2, the stochastic gradient descent algorithm for updating the generator parameter is given by

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \big(\nabla_\theta L(\theta_t, f)|_{f = f_{\theta_t}^o} + \xi_t \big), \tag{3.4}$$

where $\eta > 0$ is the stepsize, $f_{\theta_t}^o$ is provided by the max-oracle, and $\{\xi_t\}_{t=0}^T$ are independent and mean-zero noises. Here T is the maximum of the iteration numbers. The noise arises due to estimating $\nabla_\theta L(\theta, f)$ using a batch of samples from $\mathbb{P}_0$ with size n, that is,

$$\widehat{\nabla_\theta L}(\theta, f) = \frac{1}{n} \sum_{i=1}^n \nabla_\theta g_\theta(Z_i) (\nabla f) \big(g_\theta(Z_i) \big).$$

Such an estimator is unbiased and we assume its variance is bounded by σ^2/n for a constant $\sigma > 0$. We assume the feasible set Θ is large enough such that any update in (3.4) satisfies $\theta_{t+1} \in \Theta$ for all iteration $t \leq T$, for example, letting $\Theta = \mathbb{R}^{md}$.

4 Main Results

In this section we present our main results on the provable convergence. We first prove that the updates given in (3.4) converge to a stationary point of the regularized IPM.

4.1 Convergence to a Stationary Point

For the simplicity of presentation, from now on we take $\psi(f) = \|f\|_2^2/2$ as the functional regularizer, while our results can be extended to any strongly convex regularizer. Based on Lemma 3.1 which shows that the generator $g_\theta(z)$ is smooth in θ, we assume the density p_θ satisfies the following regularity conditions.

Assumption 4.1. We assume the following conditions hold.

1. The probability density $p_\theta(\cdot)$ is almost everywhere differentiable in θ and the L^2-norm of the gradient $\|\nabla_\theta p_\theta\|_2$ is upper bounded by a constant $P_1 > 0$. Also, there exists a constant $F_1 > 0$ such that for any $\theta \in \Theta$, the optimal discriminator f_θ^* is upper bounded as $\|f_\theta^*\|_2 \leq F_1/\gamma$.
2. Moreover, p_θ and $\nabla_\theta p_\theta$ are Lipschitz continuous. In specific, there exists constants $\iota_1, \iota_2 > 0$, such that $\|p_{\theta_1} - p_{\theta_2}\|_2 \leq \iota_1 \cdot \|\theta_1 - \theta_2\|$ and $\|\nabla_\theta p_{\theta_1} - \nabla_\theta p_{\theta_2}\|_2 \leq \iota_2 \cdot \|\theta_1 - \theta_2\|_2$ for any $\theta_1, \theta_2 \in \Theta$.

Here $\|\nabla_\theta p_\theta\|_2$ and $\|\nabla_\theta p_{\theta_1} - \nabla_\theta p_{\theta_2}\|_2$ are defined respectively as

$$\|\nabla_\theta p_\theta\|_2^2 = \int_{\mathbb{R}^D} \|\nabla p_\theta(x)\|_2^2 \cdot \mathrm{d}x, \quad \|\nabla_\theta p_{\theta_1} - \nabla_\theta p_{\theta_2}\|_2^2 = \int_{\mathbb{R}^D} \|\nabla_\theta p_{\theta_1}(x) - \nabla_\theta p_{\theta_2}(x)\|_2^2 \cdot \mathrm{d}x.$$

The above assumption essentially implies that the optimal discriminator f_θ^* is Lipschitz in θ and the regularized IPM $\Phi(\theta)$ is smooth in θ, which is shown in the following lemma.

Lemma 4.2. Under Assumption 4.1, the optimal discriminator f_θ^* is $(\iota_1/2\gamma)$-Lipschitz continuous and $\Phi(\theta)$ is $(\iota_2 F_1 + \iota_1 P_1)/\gamma$-smooth, that is to say, for any $\theta_1, \theta_2 \in \Theta$ we have

$$\|f_{\theta_1}^* - f_{\theta_2}^*\|_2 \leq (\iota_1/2\gamma) \cdot \|\theta_1 - \theta_2\|_2, \ \ \|\nabla\Phi(\theta_1) - \nabla\Phi(\theta_2)\|_2 \leq (\iota_2 F_1 + \iota_1 P_1)/\gamma \cdot \|\theta_1 - \theta_2\|_2.$$

With the above conditions, we show that the gradient algorithm (3.4) converges to an ε-stationary point of the regularized $\Phi(\theta)$ in sublinear rate. By equation (3.3), the update approximately follows the negative gradient direction. When the gradient of the regularized IPM is accurate, such convergence result is standard in nonconvex optimization. For the case when the oracle has an error and updates are stochastic, the convergence result is discussed in Lin et al. (2019), which is concluded in the following Lemma 4.3. For notational simplicity, we denote by $\iota = (\iota_2 F_1 + \iota_1 P_1)/\gamma$.

Lemma 4.3. Under Assumption 4.1, by performing the update (3.4) for T times and letting the step size η satisfy $\eta \leq 1/4\iota$, we have that the average of the gradient $\|\nabla_\theta \Phi(\theta_t)\|_2$ satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_\theta \Phi(\theta_t)\|_2^2] \leq (4/\eta T) \cdot \Phi(\theta_0) + 2\iota\eta\sigma^2/n + 6P_1^2\varepsilon_f/\gamma,$$

where the expectation is taken on the noises $\{\xi_t\}_{t=0}^{T-2}$.

Lemma 4.3 shows that when the stepsize η is properly chosen as $\eta \leq 1/4\iota$, then with at most $O(\varepsilon^{-2})$ iterations, $O(\varepsilon^{-2})$ batch size, and $O(\varepsilon^2)$ max-oracle error, we can obtain an ε-stationary point θ^*, which is defined by $\|\nabla_\theta \Phi(\theta^*)\|_2 \leq \varepsilon$. In the following, we demonstrate that when the generator is a wide neural network, the generative distribution $\mathbb{P}_{\theta^*}$ at such stationary point is indeed globally optimal for approximating the real data distribution $\mathbb{Q}$.

4.2 Global Optimality of the Stationary Point

Although TV distance is a strong notion of distances in probability measures, as the discriminator f is an indicator, no gradient information of $\mathrm{TV}(\mathbb{P}_\theta, \mathbb{Q})$ can we get from the optimal discriminator. In this paper, we show that, under certain regularity conditions, $\mathbb{P}_\theta$ can actually converge to $\mathbb{Q}$ in the sense of TV distance when $\mathcal{F}$ has similar discriminative ability as all 1-Lipschitz functions (see Assumption 4.5 for a concrete description). In the following, we denote by θ^* an ε-stationary point of $\Phi(\theta)$ and denote by f^* the optimal discriminator $f_{\theta^*}^*$.

To establish the global optimality of θ^*, we discuss $\mathbb{P}_{\theta^*}$ and $\mathbb{Q}$ in a mild case when their densities are upper and lower bounded on their supports, which is concluded in the following condition.

Condition 4.4 . Distributions $\mathbb{P}_{\theta^*}$ and $\mathbb{Q}$ have continuous probability densities $p_*(\cdot)$ and $q(\cdot)$ in $\mathbb{R}^D$ with supports S_P and S_Q. Also, there exist constants $\underline{p}, \overline{p}, \overline{R} > 0$ such that S_P and S_Q are both bounded by $\overline{R}$ and $p_*(\cdot)$ and $q(\cdot)$ are both lower bounded by $\underline{p}$ and upper bounded by $\overline{p}$.

In practice, maximization over the class of all 1-Lipschitz function is implemented through maximization over parametrized functions. Therefore, as a more general case, we consider $\mathcal{F}$ to be a convex class of functions that have similar discriminative ability as all 1-Lipschitz functions. We summaries such discriminative ability in the following assumption on $\mathcal{F}$, which also allows an error term.

Assumption 4.5 (Discriminative Ability)**.** We assume there exist constants $\bar{\varepsilon}, \beta, \varepsilon_D > 0$, such that for any continuous function $h(x)$, any set $S \subset \mathbb{R}^D$ such that any neighborhood of a point in S has a nonzero Lebesgue measure, and any $f \in \mathcal{F}$, it holds that as long as $\sup_{x\in S} \|\nabla f(x)\|_2 \le \bar{\varepsilon}$, there exists a discriminator $\widetilde{f} \in \mathcal{F}$ such that (1) $\widetilde{f}(x) = f(x)$ for all $x \in S$ and (2) it holds that $\int_S |\widetilde{f}(x) - f(x)|\mathrm{d}x > 0$ and

$$\int_S \widetilde{f}(x)h(x)\mathrm{d}x - \int_S f(x)h(x)\mathrm{d}x \ge \big(\beta \cdot \sup_{x\in S} |h(x)| - \varepsilon_D\big) \cdot \int_S |\widetilde{f}(x) - f(x)|\mathrm{d}x.$$

It is easy to see such assumption holds when $\mathcal{F}$ is the class of all functions that are 1-Lipschitz on $\mathbb{R}^D$ and take value 0 at $x = 0$, which is concluded in the following lemma.

Lemma 4.6. Suppose $\mathcal{F}$ is the class of all functions that are 1-Lipschitz on $\mathbb{R}^D$ and take value 0 at $x = 0$, then $\mathcal{F}$ satisfies Assumption 4.5 with $\bar{\varepsilon} = 1/2$, $\beta = 1/2$, and $\varepsilon_D = 0$.

When we take $h(\cdot) = p_*(\cdot) - q(\cdot)$ and $S = S_P$, Assumption 4.5 states that as long as f's gradient on S_P is uniformly less then $\bar{\epsilon}$, the constraint $f \in \mathcal{F}$ is slack and we are able to find another discriminator $\widetilde{f} \in \mathcal{F}$ that is only different on S_P and increases the difference of expectations by a volume proportional to $\sup_{x\in S_P} |p_*(x) - q(x)|$ with an error up to ε_D.

To formally state our results, we define the following kernel function $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$,

$$K(z_1, z_2) := \mathbb{E}_{w\sim N(0, I_d/d)}[\sigma'(w^\top z_1)\sigma'(w^\top z_2)]. \tag{4.1}$$

Such $K(\cdot, \cdot)$ is a positive definite kernel when the activation function $\sigma(\cdot)$ provides good nonlinearity, which we conclude in the following lemma.

Lemma 4.7. For any function $u \in L^2(\mathcal{Z})$ and Z_1, Z_2 independently follows $\mathbb{P}_Z$, it holds that

$$\mathbb{E}[K(Z_1, Z_2)u(Z_1)u(Z_2)] \ge 0. \tag{4.2}$$

Moreover, if $\{v(\cdot) \,|\, v(z) = \sigma'(w^\top z), w \in \mathbb{R}^D\}$ contains a complete basis of $L^2(\mathcal{Z})$, then the equality in (4.2) holds if and only if $u = 0$.

With the kernel $K(\cdot, \cdot)$ defined in (4.1), there exists a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ induced by this kernel, which represents a class of functions on $\mathcal{Z}$. Such function class is a rich class and is dense in the class of continuous functions. We denote the the RKHS-norm in $\mathcal{H}$ by $\|\cdot\|_{\mathcal{H}}$. Now we are ready to establish the optimality of the obtained stationary point.

Theorem 4.8 (Main Theorem)**.** For any ε-stationary point θ^* and its optimal discriminator f^*, under Condition 4.4 and Assumption 4.5, assuming $\partial_{x_j} f^* \circ g_{\theta^*} \in \mathcal{H}$ for all $j \in [D]$, we have

$$\|\mathbb{P}_{\theta^*} - \mathbb{Q}\|_{\mathrm{TV}} \le (2\gamma\nu_1\overline{R} + \varepsilon_D)/\beta \cdot |S_P| + \overline{p}/\underline{p} \cdot \|\nabla f^* \circ g_{\theta^*}\|_{\mathcal{H}} \cdot \sqrt{D(\varepsilon^2 + \varepsilon_K^* \nu_1^2)/\bar{\varepsilon}^4},$$

for probability at least $1 - \delta$. Here $\varepsilon_K^* = [2\log(4D^2/\delta) + 4D\log(2\sqrt{m}) + 32 + 8 \cdot \|\theta^* - \theta_0\|_2^2]/m$ and $\|\nabla f^* \circ g_{\theta^*}\|_{\mathcal{H}}$ is defined by $\|\nabla f^* \circ g_{\theta^*}\|_{\mathcal{H}}^2 = \sum_{i=1}^D \|\partial_{x_i} f^* \circ g_{\theta^*}\|_{\mathcal{H}}^2$.

Here the first term is the bias incurred by the regularizer. The second term is the training error which comes from two sources. The first source is ε, which is the optimization error when searching for a stationary point and is controlled by the iteration number T, the batch size n, and the accuracy ε_f of the max-oracle, as we have discussed in Lemma 4.3. The second source is ε_K, which is due to the expressiveness power of the generator network and is controlled by the number of hidden neurons m.

5 PROOF SKETCH

Convergence to a Constant Discriminator One key observation in our proof is that, under the neural network parameterization of the generator g_θ, when θ approaches a stationary point of $\Phi(\theta)$, the composition $\nabla f_\theta^* \circ g_\theta$ converges to zero. Recalling the property of the kernel $K(\cdot,\cdot)$ in Lemma 4.7, it is natural to define the kernel norm

$$\|u\|_K^2 := \mathbb{E}[K(Z_1,Z_2)u(Z_1)u(Z_2)] = \int_{\mathcal{Z}\times\mathcal{Z}} K(z_1,z_2)\cdot u(z_1)^\top u(z_2)\cdot p_Z^2\cdot \mathrm{d}z_1\mathrm{d}z_2.$$

The following lemma shows that when the hidden layer in the generator model is sufficiently wide, for any discriminator f and generator g_θ, the kernel norm of the composition $\nabla f\circ g_\theta : \mathcal{Z}\to\mathbb{R}$ can be approximated by the L^2-norm of the gradient of $L(\theta,f)$ with respect to the generator parameter θ.

Lemma 5.1. With probability at least $1-\delta$, for any $f\in\mathcal{F}$ and any $\theta\in\Theta$, it holds that

$$\big\|\nabla f\big(g_\theta(\cdot)\big)\big\|_K^2/D \le \|\nabla_\theta L(\theta,f)\|_2^2 + \varepsilon_K(\theta)\cdot\nu_1^2,$$

where $\varepsilon_K(\theta)$ is defined by $\varepsilon_K(\theta) = (2\log(4D^2/\delta) + 4D\log(2\sqrt{m}) + 32 + 8\cdot\|\theta-\theta_0\|_2^2)/m$.

We note that by the definition of $\varepsilon_K(\theta)$, it dacays to zero as $m\to\infty$ if $\|\theta-\theta_0\|_2$ is upper bounded by $O(m^{1/4})$. Therefore, when we find a first-order stationary point with respect to a infinitely wide neural network generator, Lemma 5.1 shows that the kernel norm $\|\nabla f^*\circ g_{\theta^*}\|_K$ goes to zero as ε and ε_K go to zero. That is to say, the optimal discriminator f^* tends to be constant over all possible value of $g_{\theta^*}(Z)$, which is the support of $\mathbb{P}_{\theta^*}$.

The next lemma shows that when $\|\nabla f^*\circ g_{\theta^*}\|_{\mathcal{H}}$ is upper bounded, then the convergence in the kernel norm $\|\nabla f^*\circ g_{\theta^*}\|_K\to 0$ can be extended to the convergence in the L^2-norm $\|\nabla f^*\circ g_{\theta^*}\|_2\to 0$.

Lemma 5.2. If f^* and g_{θ^*} corresponding to the ε-stationary point θ^* satisfy $\partial_{x_j} f^*\circ g_{\theta^*}\in\mathcal{H}$ for all $j\in[D]$, then with probability at least $1-\delta$, it holds that

$$\mathbb{E}_{\mathbb{P}_{\theta^*}}[\|\nabla f^*(X)\|_2^2] \le \|\nabla f^*\circ g_{\theta^*}\|_{\mathcal{H}}\cdot\sqrt{D(\varepsilon^2+\varepsilon_K^*\nu_1^2)},$$

where the $\mathcal{H}$-norm $\|\nabla f^*\circ g_{\theta^*}\|_{\mathcal{H}}$ and the error term ε_K^* are defined in Theorem 4.8.

Convergence of the Distributions Lemma 5.2 shows that when the generator network is infinitely wide and θ^* is a stationary point of $\Phi(\theta)$, $\|\nabla f^*(x)\|_2$ goes to zero for most $x\in S_P$. Moreover, if we have $S_Q\subseteq S_P$ and S_P is a connected set, then $f^*(X)$ takes a constant value for any sample of $X = g_{\theta^*}(Z)$ and Y. The difference of expectations in (3.1) is naturally canceled to zero and we have $\Phi(\theta^*) = \gamma\cdot\psi(f^*)$, which is globally optimal if $\psi(f^*)$ is upper bounded and γ is small.

However, the condition $S_P\subseteq S_Q$ is usually unavailable in practice since matching the supports of distributions is just as difficult as matching the whole distributions in learning implicit generative models. Therefore, the discriminator should also be sufficiently powerful, as shown in Assumption 4.5, so that we can still guarantee the convergence of distributions even we only have $f^*(X)$ constant on S_P but not S_Q. To briefly summarize the proof of Theorem 4.8, we lay out the key elements of the proof in the following. Note that under Condition 4.4, for any $\theta\in\Theta$ and $f, f'\in\mathcal{F}$, it holds that

$$L(\theta,f) - L(\theta,f') = \int_{S_P\cup S_Q}\big(f(x)-f'(x)\big)\big(p_*(x)-q(x)\big)\mathrm{d}x - \gamma\cdot\big(\psi(f)-\psi(f')\big).$$

Thus, if we can find $\widetilde{f}$ satisfying the conditions defined in Assumption 4.5 and let $f=\widetilde{f}$, $f'=f^*$ in (5.1), then by the optimality of f^* we have

$$\Big(\beta\cdot\sup_{x\in S_P}\big|p_*(x)-q(x)\big| - \varepsilon_D\Big)\cdot\int_{S_P}|\widetilde{f}(x)-f^*(x)|\mathrm{d}x \le \gamma\cdot\big(\psi(\widetilde{f})-\psi(f^*)\big), \tag{5.1}$$

which implies that $\sup_{x\in S_P}|p_*(x)-q(x)|$ on S_P is $O(\gamma)$. Thus, we can further upper bounds TV as

$$\mathrm{TV}(\mathbb{P}_{\theta^*},\mathbb{Q}) = \frac{1}{2}\int_{S_P\cup S_Q}\big(p_*(x)-q(x)\big)\mathrm{d}x \le \int_{S_P}\big(p_*(x)-q(x)\big)\mathrm{d}x \le |S_P|\sup_{x\in S_P}\big|p_*(x)-q(x)\big|.$$

Then, to prove Theorem 4.8, it suffices to verify that $\sup_{x\in S_P}\|\nabla f^*(x)\|_2\le\bar{\varepsilon}$. This is true upon we drop a small subset of S_P, since Lemma 5.2 already shows that $\mathbb{E}_{\mathbb{P}_{\theta^*}}[\|\nabla f^*(X)\|_2^2]$ is small and the density $p_*(\cdot)$ of $\mathbb{P}_{\theta^*}$ is lower bounded on S_P. See Appendix B.5 for a detailed proof.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Sanjeev Arora and Yi Zhang. Do GANs actually learn the distribution? an empirical study. In *International Conference on Learning Representations*, 2018.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pp. 224–232. JMLR. org, 2017.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 2019a.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhangdong Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *ArXiv*, abs/1904.11955, 2019b.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

V.S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294, 1997.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2017.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pp. 9236–9246, 2018.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding GANs: the LQG setting. *arXiv preprint arXiv:1710.10793*, 2017.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pp. 2810–2819, 2019.

Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5077–5086, 2017.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pp. 8571–8580, 2018.

Haoming Jiang, Zhehui Chen, Minshuo Chen, Feng Liu, Dingding Wang, and Tuo Zhao. On computation and generalization of generative adversarial networks under spectrum control. In *International Conference on Learning Representations*, 2019.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Alistair Letcher, David Balduzzi, Sébastien Racanière, James Martens, Jakob N Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *Journal of Machine Learning Research*, 20 (84):1–40, 2019.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pp. 2203–2213, 2017.

Jerry Li, Aleksander Madry, John Peebles, and Ludwig Schmidt. On the limitations of first-order approximation in gan dynamics. In *International Conference on Machine Learning*, pp. 3011–3019, 2018.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 907–915, 2019.

Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of GANs. In *Advances in Neural Information Processing Systems*, pp. 1825–1835, 2017.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

Youssef Mroueh, Tom Sercu, and Vaibhava Goel. McGAN: mean and covariance feature matching gan. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2527–2535. JMLR. org, 2017.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems*, pp. 5585–5595, 2017.

Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

Arkadi Nemirovski. Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Arkadi S Nemirovski and David Berkovich Yudin. Cesari convergence of the gradient method of approximating saddle points of convex-concave functions. *Doklady Akademii Nauk*, 239(5): 1056–1059, 1978.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2642–2651. JMLR. org, 2017.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

Sai Rajeswar, Sandeep Subramanian, Francis Dutil, Christopher Pal, and Aaron Courville. Adversarial generation of natural language. *arXiv preprint arXiv:1705.10929*, 2017.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1970.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training GANs with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7091–7101, 2018.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. VeeGAN: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pp. 3308–3318, 2017.

Thomas Unterthiner, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb GANs: Provably optimal nash equilibria via potential fields. *arXiv preprint arXiv:1708.08819*, 2017.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

William Yang Wang, Sameer Singh, and Jiwei Li. Deep adversarial learning for nlp. In *Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 1–5, 2019.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018a.

Hongyang Zhang, Susu Xu, Jiantao Jiao, Pengtao Xie, Ruslan Salakhutdinov, and Eric P Xing. Stackelberg gan: Towards provable minimax equilibrium via multi-generator architectures. *arXiv preprint arXiv:1811.08010*, 2018b.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. In *International Conference on Learning Representations*, 2018c.

Banghua Zhu, Jiantao Jiao, and David Tse. Deconstructing generative adversarial networks. *arXiv preprint arXiv:1901.09465*, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

A PROOFS FOR SECTION 3

A.1 PROOF OF LEMMA 3.1

Proof. **Lipschitz:** We compute the gradient in $[\theta]_r$ of the j-th component of $g_\theta(z)$,

$$\nabla_{[\theta]_r}\big(g_\theta(z)\big)_j = \frac{1}{\sqrt{m}}\sigma'([\theta]_r^\top z)\cdot a_{r,j}\cdot z,$$

which yields

$$\big\|\nabla_\theta\big(g_\theta(z)\big)_j\big\|_2^2 = \sum_{r=1}^m\big\|\nabla_{[\theta]_r}\big(g_\theta(z)\big)_j\big\|_2^2 \le \sum_{r=1}^m \frac{1}{m}\cdot\frac{1}{D}\cdot\|z\|_2^2 = \frac{1}{D}.$$

Thus, each component of $g_\theta(z)$ is $(1/\sqrt{D})$-Lipschitz, and we have

$$\|g_\theta(z)-g_{\theta'}(z)\|_2^2 = \sum_{j=1}^D\big\|\big(g_\theta(z)\big)_j-\big(g_{\theta'}(z)\big)_j\big\|_2^2 \le \sum_{j=1}^D\frac{1}{D}\|\theta-\theta'\|_2^2 = \|\theta-\theta'\|_2^2,$$

for any $\theta,\theta'\in\mathbb{R}^{md}$.

Smooth: When σ is smooth, we can compute the Hessian

$$\nabla^2_{[\theta]_r}\big(g_\theta(z)\big)_j = \frac{1}{\sqrt{m}}\sigma''([\theta]_r^\top z)\cdot a_{r,j}\cdot zz^\top,\quad \nabla^2_{[\theta]_r,\theta_s}\big(g_\theta(z)\big)_j = 0\ (r\neq s).$$

By the fact that σ is 1-smooth, it holds that $|\sigma''([\theta]_r^\top z)|\le 1$. Combining with the fact $|a_{r,j}| = 1/\sqrt{D}$ and $\|zz^\top\|_2\le 1$, we have $\|\nabla^2_{[\theta]_r}(g_\theta(z))_j\|_2\le 1/\sqrt{mD}$. Then for any $v\in\mathbb{R}^{mD}$ we have

$$\big\|\nabla^2_\theta\big(g_\theta(z)\big)_j v\big\|_2^2 = \sum_{r=1}^m\big\|\nabla^2_{[\theta]_r}\big(g_\theta(z)\big)_j v_r\big\|_2^2 \le \sum_{r=1}^m\frac{1}{m}\cdot\frac{1}{D}\cdot 1\cdot\|v_r\|_2^2 = \frac{1}{mD}\|v\|_2^2.$$

By the definition $\|A\|_2 = \sup_{\|v\|_2=1}\|Av\|_2$ we obtain $\|\nabla^2_\theta(g_\theta(z))_j\|_2\le 1/\sqrt{mD}$, which implies $(g_\theta(z))_j$ is $(1/\sqrt{mD})$-smooth. Thus, we have

$$\begin{aligned}\big\|\nabla_\theta g_\theta(z)-\nabla_\theta g_{\theta'}(z)\big\|_F^2 &= \sum_{j=1}^D\big\|\nabla_\theta\big(g_\theta(z)\big)_j-\nabla_\theta\big(g_{\theta'}(z)\big)_j\big\|_2^2\\ &\le \sum_{j=1}^D\frac{1}{mD}\|\theta-\theta'\|_2^2 = \frac{1}{m}\|\theta-\theta'\|_2^2,\end{aligned}$$

for any $\theta,\theta'\in\mathbb{R}^{md}$. Thus, we conclude the proof of this lemma. □

B PROOFS FOR SECTION 4

B.1 PROOF OF LEMMA 4.2

Proof. We first prove that f^*_θ is Lipschitz in function space with respect to the L^2-nrom. For any $\theta_1,\theta_2\in\Theta$, since $(p_\theta - q - 2\gamma\cdot f)$ is the functional gradient of $L(\theta,f)$ in f, by the optimality of $f^*_{\theta_1}$ and $f^*_{\theta_2}$ we have

$$\langle f-f^*_{\theta_1}, p_{\theta_1}-q-2\gamma\cdot f^*_{\theta_1}\rangle\le 0,\quad \langle f-f^*_{\theta_2}, p_{\theta_2}-q-2\gamma\cdot f^*_{\theta_2}\rangle\le 0, \tag{B.1}$$

for any $f\in\mathcal{F}$. Leting $f=f^*_{\theta_2}$ and $f=f^*_{\theta_1}$ in (B.1) respectively and summing up yields

$$2\gamma\cdot\|f^*_{\theta_1}-f^*_{\theta_2}\|_2^2 + \langle f^*_{\theta_1}-f^*_{\theta_2}, p_{\theta_2}-p_{\theta_1}\rangle\le 0. \tag{B.2}$$

Rearranging terms in (B.2) and by Assumption 4.1 we have

$$\|f^*_{\theta_1}-f^*_{\theta_2}\|_2^2\le(1/2\gamma)\cdot\langle f^*_{\theta_1}-f^*_{\theta_2}, p_{\theta_1}-p_{\theta_2}\rangle\le(\iota_1/2\gamma)\cdot\|f^*_{\theta_1}-f^*_{\theta_2}\|_2\cdot\|\theta_1-\theta_2\|_2.$$

Cancelling out $\|f^*_{\theta_1} - f^*_{\theta_2}\|_2$ on both sides we obtain that f^*_θ is $(\iota_1/2\gamma)$-Lipschitz continuous. Next, we prove that $\Phi(\theta) = L(\theta, f^*_\theta)$ is smooth. Indeed, for any $\theta_1, \theta_2 \in \Theta$ we have

$$\begin{aligned}\|\nabla\Phi(\theta_1) - \nabla\Phi(\theta_2)\|_2 &= \|\nabla_\theta L(\theta_1, f^*_{\theta_1}) - \nabla_\theta L(\theta_2, f^*_{\theta_2})\|_2 \\ &\le \|\nabla_\theta L(\theta_1, f^*_{\theta_1}) - \nabla_\theta L(\theta_2, f^*_{\theta_1})\|_2 + \|\nabla_\theta L(\theta_2, f^*_{\theta_1}) - \nabla_\theta L(\theta_2, f^*_{\theta_2})\|_2.\end{aligned} \tag{B.3}$$

Note that by Assumption 4.1 we have $\nabla_\theta \mathbb{E}[f(g_\theta(Z))] = \nabla_\theta(\langle f, p_\theta\rangle) = \langle f, \nabla p_\theta\rangle$. Therefore, we can upper bound the difference in (B.3) as

$$\begin{aligned}&\|\nabla_\theta L(\theta_1, f^*_{\theta_1}) - \nabla_\theta L(\theta_2, f^*_{\theta_1})\|_2 \\ &\quad = \|\langle f^*_{\theta_1}, \nabla p_{\theta_1} - \nabla p_{\theta_2}\rangle\|_2 \le \|f^*_{\theta_1}\|_2 \cdot \|\nabla p_{\theta_1} - \nabla p_{\theta_2}\|_2 \le \iota_2 F_1/\gamma \cdot \|\theta_1 - \theta_2\|_2,\end{aligned} \tag{B.4}$$

where the last inequality is because Assumption 4.1. Also, we have

$$\begin{aligned}&\|\nabla_\theta L(\theta_2, f^*_{\theta_1}) - \nabla_\theta L(\theta_2, f^*_{\theta_2})\|_2 \\ &\quad = \|\langle f^*_{\theta_1} - f^*_{\theta_2}, \nabla_\theta p_{\theta_2}\rangle\|_2 \le \|f^*_{\theta_1} - f^*_{\theta_2}\|_2 \cdot \|\nabla_\theta p_{\theta_2}\|_2 \le \iota_1 P_1/2\gamma \cdot \|\theta_1 - \theta_2\|_2.\end{aligned} \tag{B.5}$$

Plugging (B.4) and (B.5) into (B.3) we obtain

$$\|\nabla\Phi(\theta_1) - \nabla\Phi(\theta_2)\|_2 \le (\iota_2 F_1/\gamma + \iota_1 P_1/2\gamma) \cdot \|\theta_1 - \theta_2\|_2,$$

which implies $\Phi(\theta)$ is $(\iota_2 F_1/\gamma + \iota_1 P_1/2\gamma)$-smooth. Thus, we conclude the proof. □

B.2 Proof of Lemma 4.3

Proof. We first prove that when ε_f is small, $f^o_{\theta_t}$ is close to $f^*_{\theta_t}$. For any $f \in \mathcal{F}$, since $L(\theta, f) \le L(\theta_t, f^*_{\theta_t})$ for all $f \in \mathcal{F}$, using the strong convexity (3.2) we have

$$L(\theta_t, f) - L(\theta_t, f^*_{\theta_t}) \le \langle f - f^*_{\theta_t}, p_{\theta_t} - q - \gamma \cdot f^*_{\theta_t}\rangle - \gamma/2 \cdot \|f - f^*_{\theta_t}\|_2^2, \quad \text{for all } \ f \in \mathcal{F}. \tag{B.6}$$

By the optimality of $f^*_{\theta_t}$, we have $\langle f - f^*_{\theta_t}, p_{\theta_t} - q - \gamma \cdot f^*_{\theta_t}\rangle \le 0$. Thus, we can rewrite (B.6) as

$$L(\theta_t, f) - L(\theta_t, f^*_{\theta_t}) \le -\gamma/2 \cdot \|f - f^*_{\theta_t}\|_2^2, \quad \text{for all } \ f \in \mathcal{F}. \tag{B.7}$$

Recalling the definition of $f^o_{\theta_t}$ we have

$$L(\theta_t, f^*_{\theta_t}) - L(\theta_t, f^o_{\theta_t}) \le \varepsilon_f. \tag{B.8}$$

Letting $f = f^o_{\theta_t}$ in (B.7) and summing it up with (B.8) we obtain

$$\gamma/2 \cdot \|f^*_{\theta_t} - f^o_{\theta_t}\|_2^2 \le \varepsilon_f. \tag{B.9}$$

Now we characterize the objective's change in each iteration. By Lemma 4.2, it holds that

$$\Phi(\theta_{t+1}) \le \Phi(\theta_t) + \nabla_\theta\Phi(\theta_t)^\top(\theta_{t+1} - \theta_t) + \iota/2 \cdot \|\theta_{t+1} - \theta_t\|_2^2. \tag{B.10}$$

Since $\theta_{t+1} = \theta_t - \eta \cdot (\nabla_\theta L(\theta_t, f^o_{\theta_t}) + \xi_t)$ and $\nabla_\theta\Phi(\theta_t) = \nabla_\theta L(\theta_t, f^*_{\theta_t})$, we have

$$\theta_{t+1} - \theta_t = -\eta \cdot \big(\nabla_\theta\Phi(\theta_t) + \nabla_\theta L(\theta_t, f^o_{\theta_t}) - \nabla_\theta L(\theta_t, f^*_{\theta_t}) + \xi_t\big),$$

plugging which into (B.10) and taking conditional expectations, we obtain

$$\begin{aligned}&\eta \cdot \|\nabla_\theta\Phi(\theta_t)\|_2^2 \\ &\quad \le \Phi(\theta_t) - \mathbb{E}_t[\Phi(\theta_{t+1})] + \eta \cdot \nabla_\theta\Phi(\theta_t)^\top\big(\nabla_\theta L(\theta_t, f^*_{\theta_t}) - \nabla_\theta L(\theta_t, f^o_{\theta_t})\big) \\ &\qquad + \iota\eta^2/2 \cdot \big(\|\nabla_\theta\Phi(\theta_t) + \nabla_\theta L(\theta_t, f^o_{\theta_t}) - \nabla_\theta L(\theta_t, f^*_{\theta_t})\|_2^2 + \mathbb{E}_t[\|\xi_t\|_2^2]\big),\end{aligned} \tag{B.11}$$

where $\mathbb{E}_t[\cdot]$ represents the expectation conditional on $\{\xi_k\}_{k=0}^{t-1}$. Using the Cauchy-Schwarz inequality we have

$$\begin{aligned}&\eta \cdot \nabla_\theta\Phi(\theta_t)^\top\big(\nabla_\theta L(\theta_t, f^*_{\theta_t}) - \nabla_\theta L(\theta_t, f^o_{\theta_t})\big) \\ &\quad \le \eta \cdot \|\nabla_\theta\Phi(\theta_t)\|_2 \cdot \|\nabla_\theta L(\theta_t, f^*_{\theta_t}) - \nabla_\theta L(\theta_t, f^o_{\theta_t})\|_2 \\ &\quad \le \eta/2 \cdot \|\nabla_\theta\Phi(\theta_t)\|_2^2 + \eta/2 \cdot \|\nabla_\theta L(\theta_t, f^*_{\theta_t}) - \nabla_\theta L(\theta_t, f^o_{\theta_t})\|_2^2,\end{aligned} \tag{B.12}$$

and also,

$$\begin{aligned}&\iota\eta^2/2\cdot\|\nabla_\theta\Phi(\theta_t)+\nabla_\theta L(\theta_t,f^o_{\theta_t})-\nabla_\theta L(\theta_t,f^*_{\theta_t})\|_2^2\\&\quad\le\iota\eta^2\cdot\|\nabla_\theta\Phi(\theta_t)\|_2^2+\iota\eta^2\cdot\|\nabla_\theta L(\theta_t,f^*_{\theta_t})-\nabla_\theta L(\theta_t,f^o_{\theta_t})\|_2^2.\end{aligned}\tag{B.13}$$

Plugging (B.12) and (B.13) into (B.11) we obtain

$$\begin{aligned}(\eta/2-\iota\eta^2)\cdot\|\nabla_\theta\Phi(\theta_t)\|_2^2\le\Phi(\theta_t)-\mathbb{E}_t[\Phi(\theta_{t+1})]+\iota\eta^2/2\cdot\mathbb{E}_t[\|\xi_t\|^2]\\+(\eta/2+\iota\eta^2)\cdot\|\nabla_\theta L(\theta_t,f^*_{\theta_t})-\nabla_\theta L(\theta_t,f^o_{\theta_t})\|_2^2.\end{aligned}\tag{B.14}$$

By our assumption on the step size $\eta\le 1/4\iota$, it holds that $\eta/2-\iota\eta^2\ge\eta/4$ and $\eta/2+\iota\eta^2\le 3\eta/4$. Thus, rearranging terms in (B.14) yields,

$$\|\nabla_\theta\Phi(\theta_t)\|_2^2\le\frac{\Phi(\theta_t)-\Phi(\theta_{t+1})}{\eta/4}+2\iota\eta\cdot\mathbb{E}_t[\|\xi_t\|^2]+3\cdot\|\nabla_\theta L(\theta_t,f^*_{\theta_t})-\nabla_\theta L(\theta_t,f^o_{\theta_t})\|_2^2.\tag{B.15}$$

Note that by (B.9) we have

$$\|\nabla_\theta L(\theta_t,f^*_{\theta_t})-\nabla_\theta L(\theta_t,f^o_{\theta_t})\|_2^2=\|\langle f^*_{\theta_t}-f^o_{\theta_t},\nabla_\theta p_{\theta_t}\rangle\|_2^2\le P_1^2\cdot\|f^*_{\theta_t}-f^o_{\theta_t}\|_2^2\le 2P_1^2\varepsilon_f/\gamma,$$

plugging which into (B.15) and recalling $\mathbb{E}_t[\|\xi_t\|_2^2]=\sigma^2/n$ we obtain

$$\|\nabla_\theta\Phi(\theta_t)\|_2^2\le(4/\eta)\cdot\big(\Phi(\theta_t)-\mathbb{E}_t[\Phi(\theta_{t+1})]\big)+2\iota\eta\sigma^2/n+6P_1^2\varepsilon_f/\gamma.\tag{B.16}$$

Since $0\in\mathcal{F}$, we know that $\Phi(\theta)\ge 0$ for any $\theta\in\Theta$. Telescoping (B.16) for $t=0,\ldots,T-1$ and taking expectation on $\{\xi_t\}_{t=0}^{T-2}$ we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_\theta\Phi(\theta_t)\|_2^2]\le(4/\eta T)\cdot\Phi(\theta_0)+2\iota\eta\sigma^2/n+6P_1^2\varepsilon_f/\gamma,$$

which concludes the proof. □

B.3 Proof of Lemma 4.7

Proof. For any function $u\in L^2(\mathcal{Z})$, we have

$$\begin{aligned}&\int_{\mathcal{Z}\times\mathcal{Z}}K(z_1,z_2)u(z_1)u(z_2)\mathrm{d}z_1\mathrm{d}z_2\\&=\int_{\mathcal{Z}\times\mathcal{Z}}\mathbb{E}_{w\sim N(0,I_d/d)}[\sigma'(w^\top z_1)\sigma'(w^\top z_2)]u(z_1)u(z_2)\mathrm{d}z_1\mathrm{d}z_2\\&=\mathbb{E}_{w\sim N(0,I_d/d)}\bigg[\int_{\mathcal{Z}\times\mathcal{Z}}\sigma'(w^\top z_1)\sigma'(w^\top z_2)u(z_1)u(z_2)\mathrm{d}z_1\mathrm{d}z_2\bigg]\\&=\mathbb{E}_{w\sim N(0,I_d/d)}\bigg[\int_{\mathcal{Z}}\sigma'(w^\top z_1)u(z_1)\mathrm{d}z_1\int_{\mathcal{Z}}\sigma'(w^\top z_2)u(z_2)\mathrm{d}z_2\bigg]\\&=\mathbb{E}_{w\sim N(0,I_d/d)}\bigg[\bigg(\int_{\mathcal{Z}}\sigma'(w^\top z)u(z)\mathrm{d}z\bigg)^2\bigg]\ge 0,\end{aligned}$$

where the equality holds if and only if $\int_{\mathcal{Z}}\sigma'(w^\top z)u(z)\mathrm{d}z=0$ for all $w\in\mathbb{R}^D$. Such condition can be strengthen to $\|u\|_2=0$ if $\{v_w\in L^2(\mathcal{Z})\,|\,v_w(z)=\sigma'(w^\top z),w\in\mathbb{R}^D\}$ contains a complete basis of $L^2(\mathcal{Z})$. □

B.4 Proof of Lemma 4.6

Proof. We denote $\overline{h}=\sup_{x\in S}|h(x)|>0$. Without loss of generality, we assume $h(x')=\overline{h}$ for some $x'\in S$. Since $h(\cdot)$ is continuous, we know there exists a subset $S'\subset S$ such that $h(x)\ge\overline{h}/2$ for any $x\in S'$ and $|S'|>0$ and $0\notin S'$. Thus, we can find $r>0$ and $x''\in S'$ such that $B_r(x'')\subset S'$. Then, we define

$$v(x)=1/2\cdot(r-\|x-x''\|_2)\cdot\mathbb{1}\{x\in B_r(x'')\}.$$

It is easy to see such $v(\cdot)$ is $1/2$-Lipschitz on S and constantly zero for any x not in S. Thus, if we have $\sup_{x\in S}\|\nabla f(x)\|_2 \le 1/2$, by triangle inequality it holds that $f + v\cdot \mathbb{1}_S$ is 1-Lipschitz. Combined with the fact $f(0) = v(0) = 0$, it implies that $\widetilde{f} := f + v\cdot\mathbb{1}_S$ belongs to $\mathcal{F}$. At last, since $h(x) \ge \overline{h}/2$ on the support of $v(x)$ and $v(x) \ge 0$, we have

$$\int_S v(x)h(x)\mathrm{d}x \ge \overline{h}/2 \cdot \int_S |v(x)|\mathrm{d}x,$$

which coincides with Assumption 4.5 if we let $\overline{\varepsilon} = 1/2$, $\beta = 1/2$, and $\varepsilon_D = 0$. □

B.5 Proof of Theorem 4.8

Proof. Let $S_1' = \{x \in S_P \,|\, \|\nabla f(x)\|_2 \le \overline{\varepsilon}\}$ and $S_2 = \{x \in S_P \,|\, \|\nabla f(x)\|_2 \in (\overline{\varepsilon}, \nu_1]\}$. Moreover, let

$$S_1 = S_1'/S_1'', \quad \text{where } S_1'' = \{x \in S_1' \,|\, \text{there exists } r > 0 \text{ such that } |B_r(x)\cap S_1'| = 0\}. \tag{B.17}$$

It is easy to see that $|S_1''| = 0$. Otherwise, suppose that $|S_1''| > 0$, then there exists $x_0 \in S_1''$ and $r > 0$ such that $B_r(x_0) \subset S_1'' \subset S_1'$, which implies $|B_{r'}(x_0)\cap S_1'| \ge |B_{r'\wedge r}(x_0)| > 0$ for all $r' > 0$. This contradicts to the fact that $x_0 \in S_1''$. Then, since f^* is ν_1-Lipschitz on S_P, we have $\|\nabla f^*(x)\|_2 \le \nu_1$ almost everywhere on S_P and $S_P = S_1 \cup S_2$ except a set with zero measure. By Lemma 5.2 we have

$$\begin{aligned}\|\nabla f^*\circ g_{\theta^*}\|_{\mathcal{H}}\cdot\sqrt{D(\varepsilon^2 + \varepsilon_K^*\nu_1^2)} &= \mathbb{E}_{\mathbb{P}_{\theta^*}}[\|\nabla f^*(X)\|_2^2] \\ &\ge \int_{S_P} \underline{p}\cdot\|\nabla f^*(X)\|_2^2\cdot\mathrm{d}x \ge \int_{S_2}\underline{p}\overline{\varepsilon}^2\mathrm{d}x = \underline{p}\overline{\varepsilon}^2\cdot|S_2|,\end{aligned} \tag{B.18}$$

which provides an upper bound on $|S_2|$ as $|S_2| \le \|\nabla f^*\circ g_{\theta^*}\|_{\mathcal{H}}\cdot\sqrt{D(\varepsilon^2+\varepsilon_K^*\nu_1^2)}/\underline{p}\overline{\varepsilon}^2$. Therefore, we only need to care about $|p_*(x) - q(x)|$ on S_1, where the gradient is upper bounded by ε. By our definition of S_1 in (B.17), Condition 4.4, and Assumption 4.5, letting $h(\cdot) = p_*(\cdot) - q(\cdot)$ and $S = S_1$, there exists function $v^*(\cdot)$ such that $f^\dagger = f^* + v^*\cdot\mathbb{1}_{S_1} \in \mathcal{F}$ and

$$\int_{S_1} v^*(x)\big(p_*(x) - q(x)\big)\mathrm{d}x \ge \big(\beta\cdot\sup_{x\in S_1}|p_*(x) - q(x)| - \varepsilon_D\big)\cdot\int_{S_1}|v^*(x)|\mathrm{d}x. \tag{B.19}$$

By the definition of f^*, we have

$$\mathbb{E}_{X\sim\mathbb{P}_{\theta^*},Y\sim\mathbb{Q}}[f^*(X) - f^\dagger(X) - f^*(Y) + f^\dagger(Y)] \ge \gamma\cdot\big(\psi(f^*) - \psi(f^\dagger)\big). \tag{B.20}$$

Note that $f^* \equiv f^\dagger$ on $\mathbb{R}^D\setminus S_1$, which implies

$$\mathbb{E}_{X\sim\mathbb{P}_{\theta^*},Y\sim\mathbb{Q}}[f^*(X) - f^\dagger(X) - f^*(Y) + f^\dagger(Y)] = \int_{S_1} -v^*(x)\big(p_*(x) - q(x)\big)\mathrm{d}x. \tag{B.21}$$

Since $f^*, f^\dagger \in \mathcal{F}$ and $f^*(0) = f^\dagger(0) = 0$, we have $\max\{|f^*(x)|, |f^\dagger(x)|\} \le \nu_1\overline{R}$ for all $x \in S_P$. Thus, by the definition of ψ we have

$$\psi(f^\dagger) - \psi(f^*) = \int_{S_1} v^*(x)\big(f^*(x) + f^\dagger(x)\big)\mathrm{d}x \le 2\nu_1\overline{R}\cdot\int_{S_1}|v^*(x)|\mathrm{d}x. \tag{B.22}$$

Thus, combining (B.19), (B.20), (B.21), and (B.22) we have

$$\sup_{x\in S_1}|p_*(x) - q(x)| \le (2\gamma\nu_1\overline{R} + \varepsilon_D)/\beta. \tag{B.23}$$

Thus, we have

$$\begin{aligned}\int_{S_P}|p_*(x) - q(x)|\mathrm{d}x &= \int_{S_1}|p_*(x) - q(x)|\mathrm{d}x + \int_{S_2}|p_*(x) - q(x)|\mathrm{d}x \\ &\le \int_{S_1}(2\gamma\nu_1\overline{R} + \varepsilon_D)/\beta\cdot\mathrm{d}x + \overline{p}\cdot|S_2| \\ &\le (2\gamma\nu_1\overline{R} + \varepsilon_D)/\beta\cdot|S_P| + \overline{p}/\underline{p}\cdot\|\nabla f^*\circ g_{\theta^*}\|_{\mathcal{H}}\cdot\sqrt{D(\varepsilon^2 + \varepsilon_K^*\nu_1^2)/\overline{\varepsilon}^4},\end{aligned} \tag{B.24}$$

where the first inequality uses the upper bound in (B.23) and the second inequality is becasue $|p_{\theta^*}(x) - q(x)| \leq \overline{p}$ and the upper bound of $|S_2|$ in (B.18). The total variation distance between $\mathbb{P}_\theta$ and $\mathbb{Q}$ is then upper bounded as

$$\|\mathbb{P}_{\theta^*} - \mathbb{Q}\|_{\mathrm{TV}} = 1/2 \cdot \int_{S_P \cup S_Q} |p_*(x) - q(x)| \mathrm{d}x \leq \int_{S_P} |p_*(x) - q(x)| \mathrm{d}x. \tag{B.25}$$

Here the first inequality is because

$$\begin{aligned}
&\int_{S_P \cup S_Q} |p_*(x) - q(x)| \mathrm{d}x \\
&= \int_{S_P} |p_*(x) - q(x)| \mathrm{d}x + \int_{S_Q \setminus S_P} q(x) \mathrm{d}x \\
&= \int_{S_P} |p_*(x) - q(x)| \mathrm{d}x + 1 - \int_{S_P} q(x) \mathrm{d}x \\
&= \int_{S_P} |p_*(x) - q(x)| \mathrm{d}x + 1 - \int_{S_P} \big(p_*(x) + q(x) - p_*(x)\big) \mathrm{d}x \\
&\leq \int_{S_P} |p_*(x) - q(x)| \mathrm{d}x + 1 - \int_{S_P} \big(p_*(x) - |q(x) - p_*(x)|\big) \mathrm{d}x \\
&= \int_{S_P} |p_*(x) - q(x)| \mathrm{d}x + 1 - \int_{S_P} p_*(x) \mathrm{d}x + \int_{S_P} |q(x) - p_*(x)| \mathrm{d}x \\
&= 2 \int_{S_P \cup S_Q} |p_*(x) - q(x)| \mathrm{d}x.
\end{aligned}$$

Combining (B.24) and (B.25) we finish the proof of Theorem 4.8. □

C Proofs for Section 5

C.1 Proof of Lemma 5.1

Proof. Recall that the gradient of $L(\theta, f)$ takes the form

$$\nabla_\theta L(\theta, f) = \nabla_\theta \mathbb{E}\big[f\big(g_\theta(Z)\big)\big] = \mathbb{E}\big[\nabla_\theta g_\theta(Z) \nabla f\big(g_\theta(Z)\big)\big].$$

Therefore, for two independent $Z_1, Z_2 \sim \mathbb{P}_0$, we can write

$$\begin{aligned}
\|\nabla_\theta L(\theta, f)\|_2^2 &= \mathbb{E}\big[\nabla_\theta g_\theta(Z_1) \nabla f\big(g_\theta(Z_1)\big)\big]^\top \mathbb{E}\big[\nabla_\theta g_\theta(Z_2) \nabla f\big(g_\theta(Z_2)\big)\big] \\
&= \mathbb{E}\big[\nabla f\big(g_\theta(Z_1)\big)^\top \big(\nabla_\theta g_\theta(Z_1)^\top \nabla_\theta g_\theta(Z_2)\big) \nabla f\big(g_\theta(Z_2)\big)\big].
\end{aligned} \tag{C.1}$$

Thus, to prove Lemma 5.1, it remains to connect $\nabla_\theta g_\theta(z_1)^\top \nabla_\theta g_\theta(z_2)$ with $K(z_1, z_2)$. We have the following lemma.

Lemma C.1. With probability at least $1 - \delta$ over the initial parameter θ_0 and $\{a_r\}_{r=1}^m$, for any $z_1, z_2 \in \mathcal{Z}$ and any $\theta \in \Theta$, it holds that

$$\begin{aligned}
&\big\|K(z_1, z_2)/D \cdot I_D - \nabla_\theta g_\theta(z_1)^\top \nabla_\theta g_\theta(z_2)\big\|_2 \\
&\leq \sqrt{\frac{2\log(4D^2/\delta) + 4D\log(2\sqrt{m}) + 32 + 8 \cdot \|\theta - \theta_0\|_2^2}{m}}.
\end{aligned}$$

Proof. See Appendix D.1 for a detailed proof. □

We denote the upper bound in Lemma C.1 by $\varepsilon_K(\theta)$. Then combining Lemma C.1 with (C.1) and the fact that $\|\nabla f(g_\theta(Z))\|_2 \leq \nu_1$ we have

$$\begin{aligned}
\|\nabla_\theta L(\theta, f)\|_2^2 &= \mathbb{E}\big[K(z_1, z_2) \nabla f\big(g_\theta(Z_1)\big)^\top \nabla f\big(g_\theta(Z_2)\big)\big]/D \\
&\quad + \mathbb{E}\big[\nabla f\big(g_\theta(Z_1)\big)^\top \big(\nabla_\theta g_\theta(Z_1)^\top \nabla_\theta g_\theta(Z_2) - K(z_1, z_2)/D \cdot I_D\big) \nabla f\big(g_\theta(Z_2)\big)\big] \\
&\geq \big\|\nabla f\big(g_\theta(\cdot)\big)\big\|_K^2/D - \varepsilon_K(\theta) \cdot \nu_1^2.
\end{aligned}$$

Rearranging terms we finish the proof of Lemma 5.1. □

C.2 Proof of Lemma 5.2

Proof. For a positive definite kernel $K(\cdot,\cdot)$ we have its eigenvalue decomposition

$$K(z_1,z_2)=\sum_{i=1}^{\infty}\lambda_i\phi_i(z_1)\phi_i(z_2),$$

where $\{\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of $L^2(\mathcal{Z})$ under the inner product

$$\langle\phi_i,\phi_j\rangle=\mathbb{E}[\phi_i(Z)\phi_j(Z)]=\int_{\mathcal{Z}}\phi_i(z)\phi_j(z)p_z\mathrm{d}z.$$

Then for any $f(\cdot)=\sum_{i=1}^{\infty}a_i\cdot\phi_i(\cdot)$, the operator K maps it as

$$(K\circ f)(z_2)=\mathbb{E}[K(Z_1,z_2)f(Z_1)]=\sum_{i=1}^{\infty}\lambda_i\cdot a_i\cdot\phi_i(z_2).$$

Also, we know that $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ is the basis of the RKHS $\mathcal{H}$ induced by this kernel. Recall that this construction provides the reproducing property

$$\begin{aligned}\langle K(\cdot,z_1),K(\cdot,z_2)\rangle_{\mathcal{H}}&=\left\langle\sum_{i=1}^{\infty}\lambda_i\cdot\phi_i(z_1)\cdot\phi_i,\sum_{i=1}^{\infty}\lambda_i\cdot\phi_i(z_2)\cdot\phi_i\right\rangle_{\mathcal{H}}\\&=\sum_{i=1}^{\infty}\big(\sqrt{\lambda_i}\cdot\phi_i(z_1)\big)\cdot\big(\sqrt{\lambda_i}\cdot\phi_i(z_2)\big)=K(z_1,z_2).\end{aligned}$$

Thus for a vector function $f(\cdot)=\sum_{i=1}^{\infty}\alpha_i\cdot\sqrt{\lambda_i}\phi_i(\cdot)$ where $\alpha_i\in\mathbb{R}^D$ for all i, its K-norm is

$$\|f\|_K^2=\sum_{j}^{D}\langle f_j,K\circ f_j\rangle=\sum_{j=1}^{D}\left\langle\sum_{i=1}^{\infty}\alpha_{i,j}\cdot\sqrt{\lambda_i}\cdot\phi_i,\ \sum_{i=1}^{\infty}\lambda_i\cdot\alpha_{i,j}\cdot\sqrt{\lambda_i}\cdot\phi_i\right\rangle=\sum_{i=1}^{\infty}\lambda_i^2\cdot\|\alpha_i\|^2.$$

Also, the $\mathcal{H}$-norm and L^2-norm of f is given by

$$\|f\|_{\mathcal{H}}^2=\sum_{i=1}^{\infty}\|\alpha_i\|^2,\quad \|f\|_2^2=\sum_{i=1}^{\infty}\left\langle\sum_{i=1}^{\infty}\alpha_{i,j}\cdot\sqrt{\lambda}\cdot\phi_i,\ \sum_{i=1}^{\infty}\cdot\alpha_{i,j}\cdot\sqrt{\lambda}\cdot\phi_i\right\rangle=\sum_{i=1}^{\infty}\lambda_i\cdot\|\alpha_i\|^2.$$

Thus, applying the Cauchy Schwarz inequality we obtain

$$\|f\|_2^2\le\|f\|_K\cdot\|f\|_{\mathcal{H}}. \tag{C.2}$$

In our context, we let $f=\nabla f^*\circ g_{\theta^*}$ and have

$$\|\nabla f^*\circ g_{\theta^*}\|_2^2=\mathbb{E}\big[\big\|\nabla f^*\big(g_{\theta^*}(Z)\big)\big\|_2^2\big]=\mathbb{E}_{\mathbb{P}_{\theta^*}}[\|\nabla f^*(X)\|_2^2]. \tag{C.3}$$

Then plugging (C.3) into (C.2) and using the upper bound of $\|\nabla f^*\circ g_{\theta^*}\|_K$ in Lemma 5.1 we finish the proof of Lemma 5.2. □

D Proofs for Supporting Lemmas

D.1 Proof of Lemma C.1

Proof. The first part of the proof is simply using concentration. We note that for any $z_1,z_2\in\mathcal{Z}$ and any $i,j\in[D](i\neq j)$, by the definition of g_θ in Section 3.1, we have

$$\nabla_\theta[g_{\theta_0}(z_1)]_i^\top\nabla_\theta[g_{\theta_0}(z_2)]_j=\frac{1}{m}\sum_{r=1}^{m}a_{r,i}a_{r,j}\cdot\sigma'([\theta_0]_r^\top z_1)\sigma'([\theta_0]_r^\top z_2). \tag{D.1}$$

Then it holds that

$$\mathbb{E}\big[\nabla_\theta[g_{\theta_0}(z_1)]_i^\top\nabla_\theta[g_{\theta_0}(z_2)]_i\big]=K(z_1,z_2)/D,\quad \mathbb{E}\big[\nabla_\theta[g_{\theta_0}(z_1)]_i^\top\nabla_\theta[g_{\theta_0}(z_2)]_j\big]=0, \tag{D.2}$$

where the expectation is taken on the initialization of θ_0 and $\{a_r\}_{r=1^m}$ in $g_{\theta_0}(\cdot)$. Since we need concentration over all $(z_1, z_2) \in \mathcal{Z} \times \mathcal{Z}$, we employ an ϵ-net on $\mathcal{Z} \times \mathcal{Z}$. By the covering number of a Euclidean ball (Vershynin, 2010) we know there exists an ϵ-net $\mathcal{Z}_\epsilon$ of $\mathcal{Z} \times \mathcal{Z}$ with $|\mathcal{Z}_\epsilon| \leq (1 + 1/\sqrt{2}\epsilon)^{2D}$. Note that for each term in the summation in (D.1), we have

$$a_{r,i}a_{r,j} \cdot \sigma'([\theta_0]_r^\top z_1)\sigma'([\theta_0]_r^\top z_2) \in [-1/D, 1/D].$$

Then, we apply Hoeffding's inequality on all points $(\widehat{z}_1, \widehat{z}_2)$ in $\mathcal{Z}_\epsilon$ and obtain

$$\big|\nabla_\theta [g_{\theta_0}(\widehat{z}_1)]_i^\top \nabla_\theta [g_{\theta_0}(\widehat{z}_2)]_j - \mathbb{E}\big[\nabla_\theta [g_{\theta_0}(\widehat{z}_1)]_i^\top \nabla_\theta [g_{\theta_0}(\widehat{z}_2)]_j\big]\big| \leq \sqrt{\frac{\log(2D^2|\mathcal{Z}_\epsilon|/\delta)}{2mD^2}} \tag{D.3}$$

for all $i, j \in [D]$, with probability at least $1 - \delta$.

By the definition of an ϵ-net, for any $z = (z_1, z_2) \in \mathcal{Z} \times \mathcal{Z}$, there exists $\widehat{z} = (\widehat{z}_1, \widehat{z}_2) \in \mathcal{Z}_\epsilon$ such that $\|z - \widehat{z}\|_2 \leq \epsilon$. Thus, it holds that

$$\begin{aligned}
&\big\|K(z_1, z_2)/D \cdot I_D - \nabla_\theta g_{\theta_0}(z_1)^\top \nabla_\theta g_{\theta_0}(z_2)\big\|_2 \\
&\quad \leq \big\|K(\widehat{z}_1, \widehat{z}_2)/D \cdot I_D - \nabla_\theta g_{\theta_0}(\widehat{z}_1)^\top \nabla_\theta g_{\theta_0}(\widehat{z}_2)\big\|_2 + \Delta_1 + \Delta_2,
\end{aligned} \tag{D.4}$$

where $\Delta_1 = \|K(z_1, z_2)/D \cdot I_D - K(\widehat{z}_1, \widehat{z}_2)/D \cdot I_D\|_2$ and $\Delta_2 = \|\nabla_\theta g_{\theta_0}(z_1)^\top \nabla_\theta g_{\theta_0}(z_2) - \nabla_\theta g_{\theta_0}(\widehat{z}_1)^\top \nabla_\theta g_{\theta_0}(\widehat{z}_2)\|_2$. By (D.2) and (D.3) we have

$$\begin{aligned}
&\big\|K(\widehat{z}_1, \widehat{z}_2)/D \cdot I_D - \nabla_\theta g_{\theta_0}(\widehat{z}_1)^\top \nabla_\theta g_{\theta_0}(\widehat{z}_2)\big\|_2 \\
&\quad \leq \big\|K(\widehat{z}_1, \widehat{z}_2)/D \cdot I_D - \nabla_\theta g_{\theta_0}(\widehat{z}_1)^\top \nabla_\theta g_{\theta_0}(\widehat{z}_2)\big\|_{\mathrm{F}} \leq \sqrt{\frac{\log(2D^2|\mathcal{Z}_\epsilon|/\delta)}{2m}}.
\end{aligned} \tag{D.5}$$

Also, note that $K(z_1, z_2)$ is $\sqrt{2}$-Lipschitz since

$$\begin{aligned}
\|\nabla K\|_2^2 &= \|\nabla_{z_1} K\|_2^2 + \|\nabla_{z_2} K\|_2^2 \\
&= \|\mathbb{E}_w[\sigma''(w^\top z_1)\sigma'(w^\top z_2) \cdot w]\|_2^2 + \|\mathbb{E}_w[\sigma'(w^\top z_1)\sigma''(w^\top z_2) \cdot w]\|_2^2 \leq 2 \cdot \mathbb{E}_w[\|w\|_2^2] = 2.
\end{aligned}$$

And similarly,

$$\begin{aligned}
&\big\|\nabla_z \big(\nabla_\theta [g_{\theta_0}(z_1)]_i^\top \nabla_\theta [g_{\theta_0}(z_2)]_j\big)\big\|_2^2 \\
&= \big\|\nabla_{z_1} \big(\nabla_\theta [g_{\theta_0}(z_1)]_i^\top \nabla_\theta [g_{\theta_0}(z_2)]_j\big)\big\|_2^2 + \big\|\nabla_{z_2} \big(\nabla_\theta [g_{\theta_0}(z_1)]_i^\top \nabla_\theta [g_{\theta_0}(z_2)]_j\big)\big\|_2^2 \\
&= \big\|\frac{1}{m}\sum_{r=1}^m a_{r,i}a_{r,j} \cdot \sigma''([\theta_0]_r^\top z_1)\sigma'([\theta_0]_r^\top z_2) \cdot [\theta_0]_r\big\|_2^2 \\
&\quad + \big\|\frac{1}{m}\sum_{r=1}^m a_{r,i}a_{r,j} \cdot \sigma'([\theta_0]_r^\top z_1)\sigma''([\theta_0]_r^\top z_2) \cdot [\theta_0]_r\big\|_2^2 \leq 2/D^2 \cdot \Big(\frac{1}{m}\sum_{r=1}^m \|[\theta_0]_r\|_2\Big)^2.
\end{aligned}$$

Note that $\mathbb{E}[\|[\theta_0]_r\|_2^2] = 1$. By the Chernoff bound we obtain

$$\frac{1}{m}\sum_{r=1}^m \|[\theta_0]_r\|_2 \leq 1 + \sqrt{2/m \cdot \log(1/\delta)} \tag{D.6}$$

with probability at least $1 - \delta$. Without loss of generality we assume m is sufficiently large such that the right-hand side of (D.6) is less than $\sqrt{2}$. Then $\nabla_\theta [g_{\theta_0}(z_1)]_i^\top \nabla_\theta [g_{\theta_0}(z_2)]_j$ is $2/D$-Lipschitz in (z_1, z_2) for any $i, j \in [D]$. Thus, we have

$$\Delta_1 \leq \|K(z_1, z_2)/D \cdot I_D - K(\widehat{z}_1, \widehat{z}_2)/D \cdot I_D\|_{\mathrm{F}} \leq \sqrt{2/D}\epsilon \leq 2\epsilon, \tag{D.7}$$

$$\Delta_2 \leq \|\nabla_\theta g_{\theta_0}(z_1)^\top \nabla_\theta g_{\theta_0}(z_2) - \nabla_\theta g_{\theta_0}(\widehat{z}_1)^\top \nabla_\theta g_{\theta_0}(\widehat{z}_2)\|_{\mathrm{F}} \leq 2\epsilon. \tag{D.8}$$

Letting $\epsilon = 1/\sqrt{2m}$ we have $|\mathcal{Z}_\epsilon| \leq (2/\sqrt{2}\epsilon)^{2D} = (2\sqrt{m})^{2D}$, plugging which into (D.5) we obtain

$$\big\|K(\widehat{z}_1, \widehat{z}_2)/D \cdot I_D - \nabla_\theta g_{\theta_0}(\widehat{z}_1)^\top \nabla_\theta g_{\theta_0}(\widehat{z}_2)\big\|_2 \leq \sqrt{\frac{\log(2D^2/\delta) + 2D\log(2\sqrt{m})}{2m}}. \tag{D.9}$$

Combining (D.4), (D.7), (D.8), and (D.9) we obtain

$$\begin{aligned}\left\|K(z_1,z_2)/D\cdot I_D-\nabla_\theta g_{\theta_0}(z_1)^\top\nabla_\theta g_{\theta_0}(z_2)\right\|_2 &\le \sqrt{\frac{\log(2D^2/\delta)+2D\log(2\sqrt{m})}{2m}}+\sqrt{8/m}\\ &\le \sqrt{\frac{\log(2D^2/\delta)+2D\log(2\sqrt{m})+16}{m}} \qquad \text{(D.10)}\end{aligned}$$

with probability at least $1-2\delta$, where the last inequality uses the basic inequality $\sqrt{x}+\sqrt{y}\le\sqrt{2(x+y)}$ for any $x,y\ge 0$. To simplify the notation, we rewrite $\delta\leftarrow 2\delta$ and (D.10) is now

$$\left\|K(z_1,z_2)/D\cdot I_D-\nabla_\theta g_{\theta_0}(z_1)^\top\nabla_\theta g_{\theta_0}(z_2)\right\|_2 \le \sqrt{\frac{\log(4D^2/\delta)+2D\log(2\sqrt{m})+16}{m}} \qquad \text{(D.11)}$$

with probability at least $1-\delta$.

In the sequel we characterize the change of $\nabla_\theta g_\theta(z_1)^\top\nabla_\theta g_\theta(z_2)$ over θ. By triangle inequality we have

$$\begin{aligned}&\|\nabla_\theta g_{\theta_0}(z_1)^\top\nabla_\theta g_{\theta_0}(z_2)-\nabla_\theta g_\theta(z_1)^\top\nabla_\theta g_\theta(z_2)\|_2\\ &\quad\le \left\|\nabla_\theta g_{\theta_0}(z_1)^\top\left(\nabla_\theta g_{\theta_0}(z_2)-\nabla_\theta g_\theta(z_2)\right)\right\|_2+\left\|\left(\nabla_\theta g_{\theta_0}(z_1)^\top-\nabla_\theta g_\theta(z_1)\right)^\top\nabla_\theta g_\theta(z_2)\right\|_2\\ &\quad\le \|\nabla_\theta g_{\theta_0}(z_1)\|_2\cdot\left\|\nabla_\theta g_{\theta_0}(z_2)-\nabla_\theta g_\theta(z_2)\right\|_2+\|\nabla_\theta g_{\theta_0}(z_1)^\top-\nabla_\theta g_\theta(z_1)\|_2\cdot\|\nabla_\theta g_\theta(z_2)\|_2\\ &\quad\le 1\cdot 1/\sqrt{m}\cdot\|\theta-\theta_0\|_2+1\cdot 1/\sqrt{m}\cdot\|\theta-\theta_0\|_2=2/\sqrt{m}\cdot\|\theta-\theta_0\|_2, \qquad \text{(D.12)}\end{aligned}$$

where the last inequality is from Lemma 3.1.

Combining (D.11) and (D.12), we finish the proof of Lemma C.1. □