Non-asymptotic Error Bounds for Bidirectional GANs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We derive nearly sharp non-asymptotic error bounds for the bidirectional GAN estimators under the Dudley distance between the latent joint distribution and the data joint distribution. To the best of our knowledge, this is the first theoretical guarantee for the bidirectional GAN learning approach. An appealing feature of our results is that the latent and the data distributions are not assumed to have the same dimension or bounded support. These assumptions are commonly assumed in the existing convergence analysis of the unidirectional GANs but may not be satisfied in practice. We show that the prefactors in the error bounds depend on the square root of the dimension of the target distribution. This is a significant improvement over the exponential dependence on the dimension in the existing results on the error bound for unidirectional GANs. Our results are also applicable to the Wasserstein bidirectional GAN if the target distribution is assumed to have a bounded support. To prove these results, we construct neural network functions that push forward an empirical distribution to another arbitrary empirical distribution on a possibly different-dimensional space. We also develop a novel decomposition of the integral probability metric for the error analysis of bidirectional GANs. These basic theoretical results are of independent interest and can be applied to other related learning problems.

1 Introduction

Generative adversarial networks (GAN) (Goodfellow et al., 2014) is an important approach to implicitly learning and sampling from high-dimensional complex distributions. GANs have been shown to achieve impressive performance in many machine learning tasks (Radford et al., 2016; Reed et al., 2016; Zhu et al., 2017; Karras et al., 2018, 2019; Brock et al., 2019). There has been an intensive effort devoted to developing various extensions and alternative formulations of the original GAN (Li et al., 2015; Nowozin et al., 2016; Sutherland et al., 2017; Arjovsky et al., 2017). In particular, several recent studies have generalized GANs to bidirectional generative learning, which also learns an encoder function mapping the data distribution to the latent distribution (also called reference distribution) together with the generator function. These studies include the adversarial autoencoder (AAE) (Makhzani et al., 2016), bidirectional GAN (BiGAN) (Donahue et al., 2017), adversarially learned inference (ALI) (Dumoulin et al., 2017), and bidirectional generative modeling using adversarial gradient estimation (AGE) (Shen et al., 2020). A common feature of these methods is that they generalize the basic adversarial training framework of the original GAN from unidirectional to bidirectional. Comparing with the unidirectional GANs, bidirectional generative modeling is capable of learning representations. Moreover, join distribution matching in the train-

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

ing of bidirectional GANs alleviates mode dropping and encourages cycle consistency (Shen et al., 2020).

Several elegant and stimulating papers have analyzed the theoretical properties of unidirectional GANs. Arora et al. (2017) considered the generalization error of GANs under the neural net distance. Zhang et al. (2018) improved the generalization error bound in Arora et al. (2017). Liang (2020) studied the rates of convergence for learning distributions implicitly with GAN. Bai et al. (2019) analyzed the estimation error of GANs under the Wasserstein distance for a special class of distributions implemented by a generator, while the discriminator is designed to guarantee zero bias. Chen et al. (2020) studies the convergence properties of GANs when both the evaluation class and the target density class are Hölder classes. While impressive progresses have been made on the theoretical understanding of GANs, there are still some drawbacks in the existing results. For example,

(a) The latent distribution and the data distribution are assumed to have the same dimension, which is not the actual setting for GAN training in practice.

(b) The latent and the data distributions are assumed to be supported on bounded sets.

(c) The prefactors in the convergence rates may depend on d exponentially.

In practice, GANs are usually trained using a latent distribution with a lower dimension than that of the data distribution. Indeed, an important strength of GANs is that they can model low-dimensional latent structures via using a low-dimensional latent distribution. The bounded support assumption excludes some commonly used distributions such as Gaussian as the latent distribution. Therefore, strictly speaking, the existing convergence analysis results do not apply to what have been done in practice. Also, there have been no theoretical analyses of bidirectional GANs in the literature.

1.1 Contributions

We derive nearly sharp non-asymptotic bounds for the GAN estimation error under the Dudley distance between the latent joint distribution and the data joint distribution.To the best of our knowledge, this is the first result providing theoretical guarantees for bidirectional GAN estimation error rate. We do not assume that the latent and the data distributions have the same dimension or these distributions have bounded support. Also, our results are applicable to the Wasserstein distance if the target distribution is assumed to have a bounded support.

The main novel aspects of our work are as follows.

(1) We allow the dimension of the latent distribution to be different from the dimension of the data distribution, in particular, it can be much lower than that of the data distribution.

(2) We allow unbounded support for the latent distribution and the data distribution under mild conditions on the tail probabilities of the data and the latent distributions.

(3) We explicitly establish that the prefactors in the error bounds depend on the square root of the dimension of the target distribution. This is a significant improvement over the exponential dependence on d in the existing works.

Moreover, we develop a novel decomposition of integral probability metric for the error analysis of bidirectional GANs. We also show that the pushforward distribution of an empirical distribution based on neural networks can perfectly approximate another arbitrary empirical distribution.

Notation We use σ to denote the ReLU activation function in neural networks, which is $\sigma(x) = \max\{x, 0\}$. Without further indication, $\|\cdot\|$ represents the L_2 norm. For any function g, let $\|g\|_\infty = \sup_x \|g(x)\|$. We use notation $O(\cdot)$ and $\tilde{O}(\cdot)$ to express the order of function slightly differently, where $O(\cdot)$ omits the universal constant not relying on d while $\tilde{O}(\cdot)$ omits the constant related with d. We use $B_2^d(a)$ to denote L_2 ball in $\mathbb{R}^d$ with center at $\mathbf{0}$ and radius a. Let $g_\#\nu$ be the pushforward distribution of ν by function g in the sense that $g_\#\nu(A) = \nu(g^{-1}(A))$ for any measurable set A.

2 Bidirectional generative learning

We describe the setup of the bidirectional GAN estimation problem and present the assumptions in our analysis.

2.1 Bidirectional GAN estimators

Let μ be the data distribution supported on $\mathbb{R}^d, d \geq 1$. Let ν be a latent distribution which is easy to sample from. We first consider the case when ν is supported on $\mathbb{R}$, and then extend it to $\mathbb{R}^k$, where $k \geq 1$ can be different from d. Usually, $k \ll d$ in practical machine learning tasks such as image generation. The goal is to learn functions $g : \mathbb{R} \to \mathbb{R}^d$ and $e : \mathbb{R}^d \to \mathbb{R}$ such that $\tilde{g}_{\#}\nu = \tilde{e}_{\#}\mu$, where $\tilde{g} := (g, I)$ and $\tilde{e} := (I, e)$. Here $\tilde{g}_{\#}\nu$ is the pushforward distribution of ν under $\tilde{g}$ and $\tilde{e}_{\#}\mu$ is the pushforward distribution of μ under $\tilde{e}$. At the population level, the bidirectional GAN solves the minimax problem:

$$(g^*, e^*, f^*) \in \arg \min_{g \in \mathcal{G}, e \in \mathcal{E}} \max_{f \in \mathcal{F}} \mathbb{E}_{Z \sim \nu}[f(g(Z), Z)] - \mathbb{E}_{x \sim \mu}[f(X, e(X))],$$

where $\mathcal{G}, \mathcal{E}, \mathcal{F}$ are referred to as the generator class, the encoder class, and the discriminator class, respectively. Suppose we have two independent random samples $Z_1, \ldots, Z_n$ i.i.d. ν and $X_1, \ldots, X_n$ i.i.d. μ. At the sample level, the bidirectional GAN solves the empirical version of the above minimax problem:

$$(\hat{g}_\theta, \hat{e}_\varphi, \hat{f}_\omega) = \arg \min_{g_\theta \in \mathcal{G}_{NN}, e_\varphi \in \mathcal{E}_{NN}} \max_{f_\omega \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^{n} f_\omega(g_\theta(Z_i), Z_i) - \frac{1}{n} \sum_{j=1}^{n} f_\omega(X_j, e_\varphi(X_j)), \tag{2.1}$$

where $\mathcal{G}_{NN}$ and $\mathcal{E}_{NN}$ are two classes of neural networks approximating the generator class $\mathcal{G}$ and the encoder class $\mathcal{E}$ respectively, and $\mathcal{F}_{NN}$ is a class of neural networks approximating the discriminator class $\mathcal{F}$.

2.2 Assumptions

We assume the data distribution μ and the latent distribution ν satisfy the following assumptions.

Assumption 1 (Thinner than exponential tail)**.** *For a large n, the data distribution μ on $\mathbb{R}^d$ and the latent distribution ν on $\mathbb{R}$ satisfy the first moment tail condition for some $\delta > 0$,*

$$\max\{\mathbb{E}_\nu \|Z\| \mathbb{1}_{\{\|Z\| > \log n\}}, \mathbb{E}_\mu \|X\| \mathbb{1}_{\{\|X\| > \log n\}}\} = O(n^{-\frac{(\log n)^\delta}{d}}).$$

Assumption 2 (Absolute continuity)**.** *The data distribution μ on $\mathbb{R}^d$ and the latent distribution ν on $\mathbb{R}$ are absolutely continuous with respect to the Lebesgue measure λ.*

Assumption 1 is a technical condition for dealing with the case when μ and ν are supported on $\mathbb{R}^d$ and $\mathbb{R}$ instead of just some compact subsets. For distributions with bounded supports, this assumption is automatically satisfied. Here the factor $(\log n)^\delta$ ensures that the tails of μ and ν are thinner than the exponential tail, which is satisfied if X is sub-Gaussian. For the latent distribution, Assumptions 1 and 2 can be easily satisfied by specifying ν as some common distribution with easy-to-sample density such as Gaussian or uniform, which is usually done in the applications of GANs. For the target distribution, Assumptions 1 and 2 specifies the type of distributions that are learnable by bidirectional GAN with our theoretical guarantees.

2.3 Generator, encoder and discriminator classes

Let $\mathcal{F}_{NN} := \mathcal{NN}(W_1, L_1)$ be the discriminator class consisting of the feedforward ReLU neural networks $f_\omega : \mathbb{R}^{d+1} \mapsto \mathbb{R}$ with width W_1 and depth L_1. Similarly, let $\mathcal{G}_{NN} := \mathcal{NN}(W_2, L_2)$ be the generator class consisting of the feedforward ReLU neural networks $g_\theta : \mathbb{R} \mapsto \mathbb{R}^d$ with width W_2 and depth L_2, and $\mathcal{E}_{NN} := \mathcal{NN}(W_3, L_3)$ the encoder class consisting of the feedforward ReLU neural networks $e_\varphi : \mathbb{R}^d \mapsto \mathbb{R}$ with width W_3 and depth L_3.

The functions $f_\omega \in \mathcal{F}_{NN}$ have the following form:

$$f_\omega(x) = A_{L_1} \cdot \sigma(A_{L_1-1} \cdots \sigma(A_1 x + b_1) \cdots + b_{L_1-1}) + b_{L_1}$$

where A_i are the weight matrices with number of rows and columns no larger than the width W_1, b_i are the bias vector with compatible dimensions, and σ is the ReLU activation function $\sigma(x) = x \vee 0$. Similarly, functions $g_\theta \in \mathcal{G}_{NN}$ and $e_\varphi \in \mathcal{E}_{NN}$ have the following form:

$$g_\theta(x) = A'_{L_2} \cdot \sigma(A'_{L_2-1} \cdots \sigma(A'_1 x + b'_1) \cdots + b'_{L_2-1}) + b'_{L_2}$$
$$e_\varphi(x) = A''_{L_3} \cdot \sigma(A''_{L_3-1} \cdots \sigma(A''_1 x + b''_1) \cdots + b''_{L_3-1}) + b''_{L_3}$$

where A'_i and A''_i are the weight matrices with number of rows and columns no larger than W_2 and W_3, respectively, and b'_i and b''_i are the bias vector with compatible dimensions.

We impose the following conditions on $\mathcal{G}_{NN}$, $\mathcal{E}_{NN}$, and $\mathcal{F}_{NN}$.

Condition 1. *For any $g_\theta \in \mathcal{G}_{NN}$ and $e_\varphi \in \mathcal{E}_{NN}$, we have* $\max\{\|g_\theta\|_\infty, \|e_\varphi\|_\infty\} \leq \log n$.

Condition 1 on $\mathcal{G}_{NN}$ can be easily satisfied by adding an additional clipping layer ℓ after the original output layer, with $c_{n,d} \equiv (\log n)/\sqrt{d}$,

$$\ell(a) = a \wedge c_{n,d} \vee (-c_{n,d}) = \sigma(a + c_{n,d}) - \sigma(a - c_{n,d}) - c_{n,d}. \tag{2.2}$$

We truncate the output of $\|g_\theta\|$ to an increasing interval $[-\log n, \log n]$ to include the whole $\mathbb{R}^d$ support for the evaluation function class. Condition 1 on $\mathcal{E}_{NN}$ can be satisfied in the same manner.

3 Non-asymptotic error bounds

We characterize the bidirectional GAN solutions based on minimizing the integral probability metric (IPM, Müller (1997)) between two distributions μ and ν with respect to a symmetric evaluation function class $\mathcal{F}$, defined by

$$d_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} [\mathbb{E}_\mu f - \mathbb{E}_\nu f]. \tag{3.1}$$

By specifying the evaluation function class $\mathcal{F}$ differently, we can obtain many commonly-used metrics (Liu et al., 2017). Here we focus on the following two

- $\mathcal{F}$ = bounded Lipschitz function class $\Rightarrow$ $d_{\mathcal{F}} = d_{BL}$, (bounded Lipschitz (or Dudley) metric: metricizing weak convergence, Dudley (2018)),
- $\mathcal{F}$ = 1-Lipschitz function class $\Rightarrow$ $d_{\mathcal{F}} = d_{W_1}$ (Wasserstein GAN, Arjovsky et al. (2017)).

We consider the estimation error under the Dudley metric d_{BL}. We note that in the case when μ and ν have bounded supports, the Dudley metric d_{BL} is equivalent to the 1-Wasserstein metric d_{W_1}. Therefore, under the bounded support condition for μ and ν, all our convergence results also hold under the 1-Wasserstein metric d_{W_1}. Even if the support of μ and ν are unbounded, we can still apply the result of Lu and Lu (2020) to avoid empirical process theory and obtain an error bound under the 1-Wasserstein metric d_{W_1}. However, the result of Lu and Lu (2020) did not clearly defined the prefactor in the error bound. Since making the prefactors explicit is one of the main goals in this work, we only consider the Dudley metric d_{BL}.

The bidirectional GAN solution $(\hat{g}_\theta, \hat{e}_\varphi)$ in (2.1) also minimizes the distance between $(\tilde{g}_\theta)_\# \hat{\nu}_n$ and $(\tilde{e}_\varphi)_\# \hat{\mu}_n$ under $d_{\mathcal{F}_{NN}}$

$$\min_{g_\theta \in \mathcal{G}_{NN}, e_\varphi \in \mathcal{E}_{NN}} d_{\mathcal{F}_{NN}}((\tilde{g}_\theta)_\# \hat{\nu}_n, (\tilde{e}_\varphi)_\# \hat{\mu}_n).$$

However, even if two distributions are close with respect to $d_{\mathcal{F}_{NN}}$, there is no automatic guarantee that they will still be close under other metrics, for example, the Dudley metric or the Wasserstein metric (Arora et al., 2017). Therefore, it is natural to ask the question:

- How close are the two bidirectional GAN estimators

$$\hat{\boldsymbol{\nu}} := (\hat{g}_\theta, I)_{\#}\nu \text{ and } \hat{\boldsymbol{\mu}} := (I, \hat{e}_\varphi)_{\#}\mu$$

  under some other stronger metrics?

We consider the IPM with the uniformly bounded 1-Lipschitz function class on $\mathbb{R}^{d+1}$, as the evaluation class, which is defined as, for some finite $B > 0$,

$$\mathcal{F}^1 := \left\{ f : \mathbb{R}^{d+1} \mapsto \mathbb{R} \middle| \, |f(x) - f(y)| \leq \|x - y\|, x, y \in \mathbb{R}^{d+1} \text{ and } \|f\|_\infty \leq B \right\} \tag{3.2}$$

We first present a result when μ is supported on a compact subset $[-M, M]^d \subset \mathbb{R}^d$ and ν is supported on $[-M, M] \subset \mathbb{R}$ for a finite $M > 0$.

Theorem 3.1. *Suppose that the data distribution μ is supported on $[-M, M]^d \subset \mathbb{R}^d$ and the latent distribution ν is supported on $[-M, M] \subset \mathbb{R}$ for a finite $M > 0$, and Assumption 2 holds. Let the outputs of g_θ and e_φ be within $[-M, M]^d$ and $[-M, M]$ for $g_\theta \in \mathcal{G}_{NN}$ and $e_\varphi \in \mathcal{E}_{NN}$, respectively.*

By specifying the three network structures properly as $W_1 L_1 = \lceil \sqrt{n} \rceil$, $W_2^2 L_2 = C_1 dn$, and $W_3^2 L_3 = C_2 n$ for some constants $12 \leq C_1, C_2 \leq 384$, we have

$$\mathbb{E} d_{\mathcal{F}^1}(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\mu}}) \leq C_0 \sqrt{d} n^{-\frac{1}{d+1}} (\log n)^{\frac{1}{d+1}},$$

where $C_0 > 0$ is a constant independent of d and n.

The prefactor $C_0\sqrt{d}$ in the error bound depends on d through $\sqrt{d}$. This is drastically different from the existing works where the dependence of the prefactor on d is either not clearly described or is exponential. In high-dimensional settings with large d, this makes a substantial difference in the quality of the error bounds. These remarks apply to all the results stated below.

The next theorem deals with the case of unbounded support.

Theorem 3.2. *Suppose Assumptions 1 and 2 hold, and Conditions 1 is satisfied. By specifying the structures of the three network classes as $W_1 L_1 = \lceil \sqrt{n} \rceil$, $W_2^2 L_2 = C_1 dn$, and $W_3^2 L_3 = C_2 n$ for some constants $12 \leq C_1, C_2 \leq 384$, we have*

$$\mathbb{E} d_{\mathcal{F}^1}(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\mu}}) \leq \min \left\{ C_0 \sqrt{d} n^{-\frac{1}{d+1}} (\log n)^{1+\frac{1}{d+1}}, C_d n^{-\frac{1}{d+1}} \log n \right\},$$

where C_0 is a constant independent of d and n, but C_d depends on d.

Our next result generalizes the results to the case when the latent distribution ν is supported on $\mathbb{R}^k$ for $k \in \mathbb{N}_+$.

Assumption 3. *Both the data distribution μ on $\mathbb{R}^d$ and the latent distribution ν on $\mathbb{R}^k$ are absolutely continuous with respect to the Lebesgue measure λ.*

With the above assumption, we have the following theorem providing theoretical guarantees for the validity of any dimensional latent distribution ν.

Theorem 3.3. *Suppose Assumptions 1 and 3 hold, and Conditions 1 is satisfied. By specifying generator and discriminator class structure as $W_1 L_1 = \lceil \sqrt{n} \rceil$, $W_2^2 L_2 = C_1 dn$, and $W_3^2 L_3 = C_2 kn$ for some constants $12 \leq C_1, C_2 \leq 384$, we have*

$$\mathbb{E} d_{\mathcal{F}^1}(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\mu}}) \leq \min \left\{ C_0 \sqrt{d} n^{-\frac{1}{d+1}} (\log n)^{1+\frac{1}{d+1}}, C_d n^{-\frac{1}{d+1}} \log n \right\},$$

where C_0 is a constant independent of d and n, but C_d depends on d.

4 Approximation and stochastic errors

In this section we present a novel inequality for decomposing the total error into approximation and stochastic errors and establish bounds on these errors.

4.1 Decomposition of the estimation error

We first define the approximation error of a function class $\mathcal{F}$ to another function class $\mathcal{H}$ by

$$\mathcal{E}(\mathcal{H}, \mathcal{F}) := \sup_{h \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|h - f\|_\infty.$$

We decompose the Dudley distance $d_{\mathcal{F}^1}(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\mu}})$ between the latent joint distribution and the data joint distribution into four different error terms:

- the approximation error of the discriminator class $\mathcal{F}_{NN}$ to $\mathcal{F}^1$:

$$\mathcal{E}_1 = \mathcal{E}(\mathcal{F}^1, \mathcal{F}_{NN}),$$

- the approximation error of the generator and encoder classes:

$$\mathcal{E}_2 = \inf_{g_\theta \in \mathcal{G}_{NN}, e_\varphi \in \mathcal{E}_{NN}} \sup_{f_\omega \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^{n} \Big(f_\omega(g_\theta(z_i), z_i) - f_\omega(x_i, e_\varphi(x_i)) \Big),$$

- the stochastic error for the latent joint distribution $\hat{\boldsymbol{\nu}}$:

$$\mathcal{E}_3 = \sup_{f_\omega \in \mathcal{F}^1} \mathbb{E} f_\omega(g^*(z), z) - \hat{\mathbb{E}} f_\omega(g^*(z), z),$$

- the stochastic error for the data joint distribution $\hat{\boldsymbol{\mu}}$:

$$\mathcal{E}_4 = \sup_{f_\omega \in \mathcal{F}^1} \hat{\mathbb{E}} f_\omega(x, e^*(x)) - \mathbb{E} f_\omega(x, e^*(x)).$$

Lemma 4.1. *Let $(\hat{g}_\theta, \hat{e}_\varphi)$ be the bidirectional GAN solution in* (2.1) *and $\mathcal{F}^1$ be the uniformly bounded 1-Lipschitz function class defined in* (3.2)*. Then the Dudley distance between the latent joint distribution $\hat{\boldsymbol{\nu}} = (\hat{g}_\theta, I)_\# \nu$ and the data joint distribution $\hat{\boldsymbol{\mu}} = (I, \hat{e}_\varphi)_\# \mu$ can be decomposed as follows*

$$d_{\mathcal{F}^1}(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\mu}}) \quad \leq \quad 2\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4. \tag{4.1}$$

The novel decomposition (4.1) is fundamental to our error analysis. Based on (4.1), we bound each error term on the right side of (4.1) and balance the bounds to obtain an overall bound for the bidirectional GAN estimation.

For proving Lemma 4.1, we introduce the following useful inequality, which states that for any two probability distributions, the difference in IPMs with two distinct evaluation classes will not exceed 2 times the approximation error between the two evaluation classes, that is, for any probability distributions μ and ν and symmetric function classes $\mathcal{F}$ and $\mathcal{H}$,

$$d_{\mathcal{H}}(\mu, \nu) - d_{\mathcal{F}}(\mu, \nu) \leq 2\mathcal{E}(\mathcal{H}, \mathcal{F}). \tag{4.2}$$

It is easy to check that if we replace $d_{\mathcal{H}}(\mu, \nu)$ by $\hat{d}_{\mathcal{H}}(\mu, \nu) := \sup_{h \in \mathcal{H}} [\hat{\mathbb{E}}_\mu h - \hat{\mathbb{E}}_\nu h]$, (4.2) still holds. The proof of (4.2) is given in the appendix.

Proof of Lemma 4.1. We have

$$\begin{aligned}
d_{\mathcal{F}^1}(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\mu}}) = & \sup_{f_\omega \in \mathcal{F}^1} \mathbb{E} f_\omega(\hat{g}(z), z) - \mathbb{E} f_\omega(x, \hat{e}(x)) \\
\leq & \sup_{f_\omega \in \mathcal{F}^1} \mathbb{E} f_\omega(\hat{g}(z), z) - \hat{\mathbb{E}} f_\omega(\hat{g}(z), z) + \sup_{f_\omega \in \mathcal{F}^1} \hat{\mathbb{E}} f_\omega(\hat{g}(z), z) - \hat{\mathbb{E}} f_\omega(x, \hat{e}(x)) \\
& + \sup_{f_\omega \in \mathcal{F}^1} \hat{\mathbb{E}} f_\omega(x, \hat{e}(x)) - \mathbb{E} f_\omega(x, \hat{e}(x)) \\
= & \mathcal{E}_3 + \mathcal{E}_4 + \sup_{f_\omega \in \mathcal{F}^1} \hat{\mathbb{E}} f_\omega(\hat{g}(z), z) - \hat{\mathbb{E}} f_\omega(x, \hat{e}(x))
\end{aligned}$$

Denote $A := \sup_{f_\omega \in \mathcal{F}^1} \hat{\mathbb{E}} f_\omega(\hat{g}(z), z) - \hat{\mathbb{E}} f_\omega(x, \hat{e}(x))$. By (4.2) and the optimality of the bidirectional GAN solutions, A satisfies

$$
\begin{aligned}
A &= \sup_{f_\omega \in \mathcal{F}^1} \frac{1}{n} \sum_{i=1}^{n} \Big(f_\omega(\hat{g}(z_i), z_i) - f_\omega(x_i, \hat{e}(x_i)) \Big) \\
&\leq \sup_{f_\omega \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^{n} \Big(f_\omega(\hat{g}(z_i), z_i) - f_\omega(x_i, \hat{e}(x_i)) \Big) + 2\mathcal{E}(\mathcal{F}^1, \mathcal{F}_{NN}) \\
&= \inf_{g_\theta \in \mathcal{G}_{NN}, e_\varphi \in \mathcal{E}_{NN}} \sup_{f_\omega \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^{n} \Big(f_\omega(g_\theta(z_i), z_i) - f_\omega(x_i, e_\varphi(x_i)) \Big) + 2\mathcal{E}_1 \\
&= 2\mathcal{E}_1 + \mathcal{E}_2
\end{aligned}
$$

□

4.2 Approximation errors

We now discuss the errors due to the discriminator approximation and the generator and encoder approximation.

4.2.1 The discriminator approximation error $\mathcal{E}_1$

The discriminator approximation error $\mathcal{E}_1$ describes how well the discriminator neural network class approximates functions from the Lipschitz class $\mathcal{F}^1$. Lemma 4.2 below can be applied to obtain the neural network approximation error of Lipschitz functions. It leads to a quantitative and non-asymptotic approximation rate in terms of the width and depth of the neural networks when bounding $\mathcal{E}_1$.

Lemma 4.2 (Shen et al. (2021))**.** *Let f be a Lipschitz continuous function defined on $[-R, R]^d$. For arbitrary $W, L \in \mathbb{N}_+$, there exists a function ψ implemented by a ReLU feedforward neural network with width W and depth L such that*

$$||f - \psi||_\infty = O\big(\sqrt{d} R (WL)^{-\frac{2}{d}}\big).$$

By Lemma 4.2 and our choice of the architecture of discriminator class $\mathcal{F}_{NN}$, we have $\mathcal{E}_1 = O\big(\sqrt{d}(W_1 L_1)^{-\frac{2}{d+1}} \log n\big)$. Theorem 4.2 also informs about how to choose the architecture of the discriminator networks based on how small we want the approximation error $\mathcal{E}_1$ to be. We will set $(W_1 L_1)^2 = n$, in which case $\mathcal{E}_1$ is dominated by the stochastic terms $\mathcal{E}_3$ and $\mathcal{E}_4$.

4.2.2 The generator and encoder approximation error $\mathcal{E}_2$

The generator and encoder approximation error $\mathcal{E}_2$ describes how powerful the generator and encoder classes are in pushing the empirical distributions $\hat{\mu}_n$ and $\hat{\nu}_n$ to each other. A natural question is

- Can we find some generator and encoder neural network functions such that $\mathcal{E}_2 = 0$?

Most of the current literature concerning the error analysis of GANs applied the optimal transport theory (Villani, 2008) to minimize an error term similar to $\mathcal{E}_2$, see, for example, Chen et al. (2020). However, the existence of the optimal transport function from $\mathbb{R} \to \mathbb{R}^d$ is not guaranteed. Therefore, the existing analysis of GANs can only deal with the scenario when the latent and the data distribution are assumed to have the same dimension. This equal dimensionality assumption is not satisfied in the actual training of GANs or bidirectional GANs in many applications. Here, instead of using the optimal transport theory, we establish the following approximation results in Theorem 4.3, which enables us to forgo the equal dimensionality assumption.

Theorem 4.3. *Suppose that ν supported on $\mathbb{R}$ and μ supported on $\mathbb{R}^d$ are both absolutely continuous w.r.t. the Lebesgue measures, and $z_i's$ and $x_i's$ are i.i.d. samples from ν and μ, respectively for*

$1 \leq i \leq n$. *Then there exist generator and encoder neural network functions* $g : \mathbb{R} \mapsto \mathbb{R}^d$ *and* $e : \mathbb{R}^d \mapsto \mathbb{R}$ *such that* g *and* e *are inverse bijections of each other between* $\{z_i : 1 \leq i \leq n\}$ *and* $\{x_i : 1 \leq i \leq n\}$. *Moreover, such neural network functions* g *and* e *can be obtained by properly specifying* $W_2^2 L_2 = c_2 dn$ *and* $W_3^2 L_3 = c_3 n$ *for some constant* $12 \leq c_2, c_3 \leq 384$.

Proof. By the absolute continuity of ν and μ, all the $z_i's$ and $x_i's$ are distinct a.s.. We can reorder $z_i's$ from the smallest to the largest, so $z_1 < z_2 < \ldots < z_n$. Let $z_{i+1/2}$ be any point between z_i and z_{i+1} for $i \in \{1, 2, \ldots, n-1\}$. We define the continuous piece-wise linear function $g : \mathbb{R} \mapsto \mathbb{R}^d$ by

$$g(z) = \begin{cases} x_1 & z \leq z_1, \\ \frac{z - z_{i+1/2}}{z_i - z_{i+1/2}} x_i + \frac{z - z_i}{z_{i+\frac{1}{2}} - z_i} x_{i+1} & z = (z_i, z_{i+1/2}), \text{ for } i = 1, \ldots, n-1, \\ x_{i+1} & z \in [z_{i+1/2}, z_{i+1}], \text{ for } i = 1, \ldots, n-2, \\ x_n & z \geq z_{n-1+1/2}. \end{cases}$$

By Yang et al. (2021, Lemma 4.1) , $g \in \mathcal{NN}(W_2, L_2)$ if $n \leq (W_2 - d - 1)\lfloor \frac{W_2 - d - 1}{6d} \rfloor \lfloor \frac{L_2}{2} \rfloor$. Taking $n = (W_2 - d - 1)\lfloor \frac{W_2 - d - 1}{6d} \rfloor \lfloor \frac{L_2}{2} \rfloor$, a simple calculation shows $W_2^2 L_2 = cdn$ for some constant $12 \leq c \leq 384$. The existence of neural net function e can be constructed in the same way due to the fact that the first coordinate of $x_i's$ are distinct almost surely. □

When the number of point masses of the empirical distributions are relatively moderate compared with the structure of the neural nets, we can approximate empirical distributions arbitrarily well with any empirical distribution with the same number of point masses pushforwarded by the neural nets.

Theorem 4.3 provides an effective way to specify the architecture of generator and encoder classes. According to this lemma, we can take $n = \frac{W_2 - d}{2} \lfloor \frac{W_2 - d}{6d} \rfloor \lfloor \frac{L_2}{2} \rfloor + 2 = \frac{W_3 - 1}{2} \lfloor \frac{W_3 - 1}{6} \rfloor \lfloor \frac{L_3}{2} \rfloor + 2$, which gives rise to $W_2^2 L_2 / d \asymp W_3^2 L_3 \asymp n$. More importantly, Theorem 4.3 can be applied to bound $\mathcal{E}_2$ as follows.

$$\begin{aligned} \mathcal{E}_2 = & \inf_{g_\theta \in \mathcal{G}_{NN}, e_\varphi \in \mathcal{E}_{NN}} \sup_{f_\omega \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^n \Big(f_\omega(g_\theta(z_i), z_i) - f_\omega(x_i, e_\varphi(x_i)) \Big) \\ \leq & \inf_{g_\theta \in \mathcal{G}_{NN}} \sup_{f_\omega \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^n \Big(f_\omega(g_\theta(z_i), z_i) - f_\omega(x_i, z_i) \Big) \\ & + \inf_{e_\varphi \in \mathcal{E}_{NN}} \sup_{f_\omega \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^n \Big(f_\omega(x_i, z_i) - f_\omega(x_i, e_\varphi(x_i)) \Big) \\ = & 0. \end{aligned}$$

We simply reordered $z_i's$ and $x_i's$ as in the proof. Therefore, this error term can be perfectly controlled.

4.3 Stochastic errors

We apply the refined Dudley inequality (Schreuder, 2020) in Lemma 4.4 to bound the stochastic error terms $\mathcal{E}_3$ and $\mathcal{E}_4$.

Lemma 4.4 (Refined Dudley Inequality)**.** *For a symmetric function class* $\mathcal{F}$ *with* $\sup_{f \in \mathcal{F}} ||f||_\infty \leq M$*, we have*

$$\mathbb{E}[d_{\mathcal{F}}(\hat{\mu}_n, \mu)] \leq \inf_{0 < \delta < M} \left(4\delta + \frac{12}{\sqrt{n}} \int_\delta^M \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, ||\cdot||_\infty)} d\epsilon \right).$$

The original Dudley inequality (Dudley, 1967; Van der Vaart and Wellner, 1996) suffers from the problem that if the covering number $\mathcal{N}(\epsilon, \mathcal{F}, ||\cdot||_\infty)$ increases too fast as ϵ goes to 0, then the upper bound will be infinity, which is totally meaningless. The improved Dudley inequality circumvents such a problem by only allowing ϵ to integrate from $\delta > 0$ as is shown in Lemma 4.4, which also indicates that $\mathbb{E}\mathcal{E}_3$ scales with the covering number $\mathcal{N}(\epsilon, \mathcal{F}^1, ||\cdot||_\infty)$.

By calculating the cover numbers of $\mathcal{F}^1$ and utilizing the refined Dudley inequality, we can obtain the upper bound

$$\max\{\mathbb{E}\mathcal{E}_3, \mathbb{E}\mathcal{E}_4\} = O\left(C_d n^{-\frac{1}{d+1}} \log n \wedge \sqrt{d} n^{-\frac{1}{d+1}} (\log n)^{1+\frac{1}{d+1}}\right) \tag{4.3}$$

5 Related work

Recently, several impressive works have studied the challenging problem of the convergence properties of unidirectional GANs. Arora et al. (2017) noted that training of GAN may not have good generalization properties in the sense that even if training may appear successful but the trained distribution may be far from target distribution in standard metrics. On the other hand, Bai et al. (2019) showed that GANs can learn distributions in Wasserstein distance with polynomial sample complexity. Liang (2020) studied the rates of convergence of a class of GANs, including Wasserstein, Sobolev and MMD GANs. The results of Bai et al. (2019) and Liang (2020) require invertible generator networks. Chen et al. (2020) established an upper bound for the estimation error of GANs. They assumed that the latent distribution has the same dimension as the data distribution and applied the optimal transport theory to control the generator approximation error. However, how the prefactor depends in the error bounds on the dimension d in the existing results is either not clearly described or it depends on d exponentially (Liang, 2020; Chen et al., 2020). In high-dimensional settings with large d, this makes a substantial difference in the quality of the error bounds.

Singh et al. (2018) studied minimax convergence rates of nonparametric density estimation under a class of adversarial losses and investigated how the choice of loss and the assumed smoothness of the underlying density together determine the minimax rate; they also discussed connections to learning generative models in a minimax statistical sense. Uppal et al. (2019) generates the idea of Sobolev IPM to Besov IPM, where both target density and the evaluation classes are Besov classes. They also showed how their results imply bounds on the statistical error of a GAN.

These results provide important insights in the understanding of GANs. However, as we mentioned earlier, some of the assumptions made in these results, including equal dimension between the latent and the data distributions and bounded support of the distributions, are not satisfied in the training of GANs in practice. Our results avoid these assumptions. Moreover, the prefactors in our error bounds are clearly described as being dependent on the square root of the dimension d. Finally, the aforementioned results only dealt with unidirectional GANs. Our work is the first to address the convergence properties of bidirectional GANs.

6 Conclusion

This paper derives the error bounds for the bidirectional GANs under the Dudley distance between the latent joint distribution and the data joint distribution. The results are established without the two crucial conditions that are commonly assumed in the existing literature: equal dimensionality between the latent and the data distributions and bounded support for these distributions. A novel decomposition of integral probability metric is also developed for error analysis of bidirectional GANs, which can be useful in other generative learning problems.

A limitation of our results, as well as all the existing results on the convergence properties of GANs, is that they suffer from the curse of dimensionality, which cannot be circumvented by assuming sufficient smoothness assumptions. In many applications, high-dimensional complex data such as images, texts and natural languages, tend to be supported on approximate lower-dimensional manifolds. It is desirable to take into account such structure in the theoretical analysis. An important extension of the present results is to show that bidirectional GANs can circumvent the curse of dimensionality if the target distribution is assumed to be supported on a lower-dimensional manifold. This is a technically challenging problem and will be pursued in our future work. This work is basic research on the convergence properties of bidirectional GANs, we believe that it will not have potential negative societal impacts.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *ICML*.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. (2017). Generalization and equilibrium in generative adversarial nets (GANs). In *ICML*.

Bai, Y., Ma, T., and Risteski, A. (2019). Approximability of discriminators implies diversity in GANs. In *ICLR*.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. *arXiv*, abs/1809.11096.

Chen, M., Liao, W., Zha, H., and Zhao, T. (2020). Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv*, 2002.03938.

Donahue, J., Krähenbühl, P., and Darrell, T. (2017). Adversarial feature learning. In *ICLR*.

Dudley, R. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

Dudley, R. M. (2018). *Real Analysis and Probability*. CRC Press.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. (2017). Adversarially learned inference. In *ICLR*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS*.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *arXiv*, 1812.04948.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *ICML*.

Liang, T. (2020). How well generative adversarial networks learn distributions. *arXiv*, 1811.03179.

Liu, S., Bousquet, O., and Chaudhuri, K. (2017). Approximation and convergence properties of generative adversarial learning. In *NeurIPS*.

Lu, Y. and Lu, J. (2020). A universal approximation theorem of deep neural networks for expressing probability distributions. In *NeurIPS*.

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. (2016). Adversarial autoencoders. In *ICLR*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In *NeurIPS*.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. In *ICML*.

Schreuder, N. (2020). Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. *arXiv*, 2003.13530.

Shen, X., Zhang, T., and Chen, K. (2020). Bidirectional generative modeling using adversarial gradient estimation. *arXiv*, 2002.09161.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. (2018). Nonparametric density estimation with adversarial losses. In *NeurIPS*.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*.

Uppal, A., Singh, S., and Póczos, B. (2019). Nonparametric density estimation & convergence rates for GANs under Besov ipm losses. *arXiv*, 1902.03511.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer.

Villani, C. (2008). *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Yang, Y., Li, Z., and Wang, Y. (2021). On the capacity of deep generative networks for approximating distributions. *arXiv*, 2101.12353.

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. (2018). On the discrimination-generalization tradeoff in GANs. In *ICLR*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This work is basic theoretical research and there will not be potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 2 and 3
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix in the Supplementary Material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]