
Uncertainty Principles of Encoding GANs

Supplementary Material

Anonymous Authors¹

A. Potential Solutions

To help understand the implication of our theory for real problems, we discuss the connection between our theory and some potential empirical solutions in recent works.

All the difficulties of encoding GANs we mention in this paper are caused by two factors:

- the intrinsic dimensions of latent spaces and data manifolds are different;
- the neural networks to capture the underlying encoder and generator have to be smooth.

Removing either of them can free encoding GANs from the uncertainty principles uncovered in this paper. Based on that, we introduce two potential solutions to this issue: adaptive latents and noise disturbance techniques.

Adaptive latents Perhaps the most direct and reliable way is to let the network learn a latent distribution which can adapt to the intrinsic dimension of specific data distribution. To achieve this goal, one may train an extra network to produce latents for the generator and encoder. We are aware that the state-of-the-art autoencoder network, NVAE (Vahdat & Kautz, 2021), applies this design in its top-down model to learn an adaptive latent space for both encoder and generator (decoder), and achieves impressive performance. Our theory may explain the success of NVAE, and support the plausibility of this idea in encoding GANs.

Noise disturbance A potential way is to use noise disturbance in the encoder and generator networks. Noise disturbance to neural networks has been proved to improve robustness and generalization performance (Hayakawa et al., 1995; Baldi & Sadowski, 2013; He et al., 2019; Jenni & Favaro, 2019). Based on our theory, the introduction of noise disturbance can also enlarge the function space that neural network can approximate and express. With the help of noise disturbance, we may easily build neural networks that map a curve to a surface without the worry of gradient explosion. In (Arjovsky & Bottou, 2017), noise disturbance is also proved to help stabilize training of generators. Previous works have enabled noise disturbance in GANs such as StyleGAN (Karras et al., 2019; 2020), but very few works apply this technique to encoders. Although the theoretical property of noise disturbance is still unclear, it might potentially help encode GANs.

B. Preliminaries

We start by introducing our settings. See Tab. 1 for meanings and examples of notations used in this paper. The current framework of encoding GAN researches can be abstracted as follows. Let \mathcal{Z} and \mathcal{X} be the latent space and the data manifold, and $\mathbb{P}_{\mathcal{Z}}$ and $\mathbb{P}_{\mathcal{X}}$ be the latent distribution and the data distribution on \mathcal{Z} and \mathcal{X} , respectively. Encoding GANs introduces a bijection between the latent and the data: an underlying ‘perfect’ generator g transports the latent distribution into the data one,

$$\mathbb{P}_{g(\mathcal{Z})}(\mathcal{A}) = \int_{g^{-1}(\mathcal{A})} d\mathbb{P}_{\mathcal{Z}} = \mathbb{P}_{\mathcal{X}}(\mathcal{A}), \forall \mathcal{A} \subset \mathcal{F}_{\mathcal{X}}, \quad (1)$$

where $\mathcal{F}_{\mathcal{X}}$ is the collection of measurable sets in \mathcal{X} ; and an underlying ‘perfect’ encoder e inverts the generator,

$$e \circ g(z) = z, g \circ e(x) = x, \forall z \in \mathcal{Z}, x \in \mathcal{X}. \quad (2)$$

The training algorithms then aim at approximating the underlying ‘perfect’ encoder and generator with parameterized neural networks E_{θ} and G_{ϕ} , respectively.

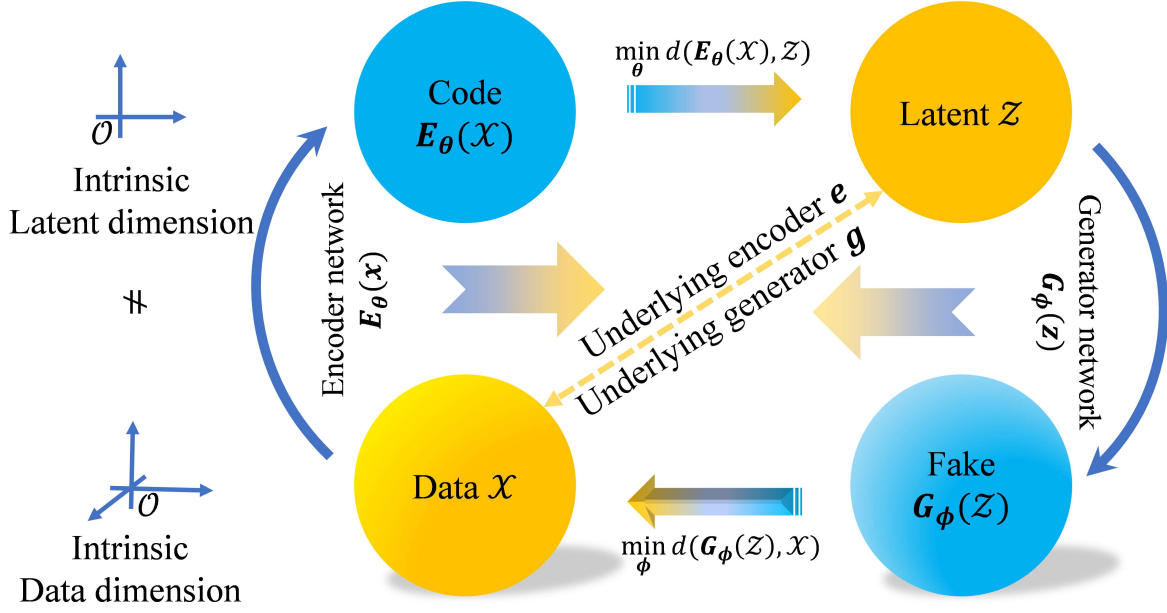


Figure 1. A typical training process. Training algorithms guide encoder network E_θ and generator network G_ϕ to the underlying e and g by minimizing divergences from $E_\theta(\mathcal{X})$ to \mathcal{Z} and from $G_\phi(\mathcal{Z})$ to \mathcal{X} . It is worthwhile to note that, as the latent distribution is pre-assigned and fixed, we usually have $\dim(\mathcal{Z}) \neq \dim(\mathcal{X})$.

Table 1. Examples of notation used in this paper.

d -dimensional volume	$m_d(\cdot)$
Scalars	$\sigma, \delta, \epsilon, m, n, d$
Scalar value functions	f, g
Vectors	\mathbf{x}, \mathbf{z}
Vector value functions	\mathbf{e}, \mathbf{g}
Sets & Manifolds	$\mathcal{X}, \mathcal{Z}, \mathcal{D}$
Neural networks	E_θ, G_ϕ, D_ψ
Distributions	$\mathbb{P}_\mathcal{X}(\mathbf{x}), \mathbb{P}_\mathcal{Z}(\mathbf{z})$
Induced Distributions	$\mathbb{P}_{\mathbf{g}(\mathcal{Z})}(\mathbf{x}), \mathbb{P}_{E_\theta(\mathcal{X})}(\mathbf{z})$
Intrinsic dimension of manifolds	$\dim(\mathcal{X}), \dim(\mathcal{Z})$

Fig. 1 illustrates a typical training process. Training algorithms guide the encoder network E_θ and the generator network G_ϕ to the underlying ‘perfect’ encoder e and generator g by minimizing divergence from $E_\theta(\mathcal{X})$ to \mathcal{Z} and from $G_\phi(\mathcal{Z})$ to \mathcal{X} . Popular divergences include the Jensen-Shannon divergence (Donahue et al., 2017; Dumoulin et al., 2017)

$$D_{JS}(\mathbb{P}_A, \mathbb{Q}_B) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_A} \left[\log \left(\frac{2\mathbb{P}_A(d\mathbf{x})}{\mathbb{P}_A(d\mathbf{x}) + \mathbb{Q}_B(d\mathbf{x})} \right) \right] + \mathbb{E}_{\mathbf{y} \sim \mathbb{Q}_B} \left[\log \left(\frac{2\mathbb{Q}_B(d\mathbf{y})}{\mathbb{P}_A(d\mathbf{y}) + \mathbb{Q}_B(d\mathbf{y})} \right) \right], \quad (3)$$

KL divergence (Makhzani et al., 2015), l_2 reconstruction loss (Choi et al., 2020), and Wasserstein divergence (Tolstikhin et al., 2017)

$$W_1(\mathbb{P}_A, \mathbb{Q}_B) = \inf_{\pi \in \Pi(\mathbb{P}_A, \mathbb{Q}_B)} \int_{\mathcal{A} \times \mathcal{B}} \|\mathbf{x} - \mathbf{y}\| d\pi(\mathbf{x}, \mathbf{y}), \quad (4)$$

where $\Pi(\mathbb{P}_A, \mathbb{Q}_B)$ is the collection of all joint distributions of $(\mathbf{x}, \mathbf{y}) \in \mathcal{A} \times \mathcal{B}$ which have marginal distribution \mathbb{P}_A for \mathbf{x} and \mathbb{Q}_B for \mathbf{y} .

Usually, latent space \mathcal{Z} and data space \mathcal{X} are treated as manifolds embedded in some Euclidean ambient spaces. We introduce the concept of manifolds and their *intrinsic dimensions* (Gallot et al., 1990) below, and give examples in Fig. 2. Note that throughout this paper, we use the word ‘dimension’ for the intrinsic dimension of manifold, not the dimension of its ambient space.

Definition 1 (Intrinsic Dimension and Manifold). *If for any point $\mathbf{x} \in \mathcal{A}$, it has a small open neighborhood \mathcal{U} and a continuous bijection \mathbf{b} (also called the **chart** at \mathbf{x}) that maps $\mathcal{U} \cap \mathcal{A}$ to an open set in \mathbb{R}^n , then n is the intrinsic dimension of \mathcal{A} . We denote it as $\dim(\mathcal{A}) = n$. Accordingly, \mathcal{A} is called a manifold.*

We introduce two specific examples of the training process. The concurrent training process in BiGAN (Donahue et al., 2017) solves a zero-sum game

$$\min_{\theta, \phi} \max_{\psi} V(\mathbf{E}_{\theta}, \mathbf{G}_{\phi}, \mathbf{D}_{\psi}), \quad (5)$$

where \mathbf{D}_{ψ} is the discriminator network for the (\mathbf{x}, \mathbf{z}) pair, and

$$V(\mathbf{E}_{\theta}, \mathbf{G}_{\phi}, \mathbf{D}_{\psi}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\log(D(\mathbf{x}, \mathbf{E}_{\theta}(\mathbf{x})))] + \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_{\mathcal{Z}}} [1 - \log(D(\mathbf{G}_{\phi}(\mathbf{z}), \mathbf{z}))]. \quad (6)$$

Another example is the two phase training process of LIA (Zhu et al., 2019), where a generator is trained by solving

$$\min_{\theta} \max_{\psi} V(\mathbf{G}_{\phi}, \mathbf{D}_{\psi}), \quad (7)$$

in which

$$V(\mathbf{G}_{\phi}, \mathbf{D}_{\psi}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_{\mathcal{Z}}} [1 - \log(D(\mathbf{G}_{\phi}(\mathbf{z})))], \quad (8)$$

and then an encoder is trained by optimizing

$$\min_{\theta} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\|\mathbf{G}_{\phi} \circ \mathbf{E}_{\theta}(\mathbf{x}) - \mathbf{x}\|_2^2] + d(\mathbb{P}_{\mathbf{E}_{\theta}(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}), \quad (9)$$

in which d is among the divergences of distributions introduced at the beginning of this section.

Current design of generative models assigns a fixed latent distribution to the generator, which also fixes the intrinsic dimension of latent distribution. Specifically, for the popular standard Gaussian latents, the intrinsic dimension is the number of variables (Goodfellow et al., 2014; Gallot et al., 1990). We disallow the networks to adjust the latent distribution during training, because we need each sample $\mathbf{z} \in \mathcal{Z}$ from the latent distribution to produce meaningful synthesis in \mathcal{X} through the generator. This is essentially different from auto-encoders (Hinton & Zemel, 1994; Ng et al., 2011) which are not designed for synthesis and allow self-adaptation in the latent distribution. As $\dim(\mathcal{X})$ is often unclear, and $\dim(\mathcal{Z})$ is manually assigned before training, we are safe to assume that the latent space \mathcal{Z} and domain of interest \mathcal{X} have different intrinsic dimensions, i.e. $\dim(\mathcal{X}) \neq \dim(\mathcal{Z})$.

To build the foundation of our theory, we make the following assumptions, which are almost the minimum requests for theoretical analysis.

Assumption 1. *Throughout this paper, we assume that:*

- *the data domain \mathcal{X} is a manifold with an intrinsic dimension n , where n is unknown;*
- *the neural networks $\mathbf{E}_{\theta}(\mathbf{x})$ and $\mathbf{G}_{\phi}(\mathbf{z})$ are continuous and piece-wise continuously differentiable with respect to inputs \mathbf{x} and \mathbf{z} ; we do not make any assumption on the training method or the loss function;*
- *the latent and the data distributions are absolutely continuous with respect to the Lebesgue measure on \mathcal{Z} and \mathcal{X} respectively, which are the minimum requirements for calculating the Jensen-Shannon and Wasserstein divergences.*

Remark 1. *Obviously, neural network components such as MLPs, CNNs, Relu, Tanh, LeakyRelu, Softmax, Sigmoid, and neural networks composed of them are all continuous and piece-wise continuously differentiable with respect to their inputs.*

C. Proof to Theorems

C.1. Proof to Theorem 1

Theorem 1. *When $\dim(\mathcal{Z}) \neq \dim(\mathcal{X})$, at least one of the underlying encoder and generator in Eq. (1) & (2) is discontinuous; and for any $\mathbf{x} \in \mathcal{X}, \delta > 0$, there is a point \mathbf{x}' in the geodesic ball centered at \mathbf{x} with radius δ , such that e is not continuous at \mathbf{x}' or g is not continuous at $e(\mathbf{x}')$. The same thing holds for \mathcal{Z} .*

Proof. We divide the proof into two parts. In the first part we prove the discontinuity of optimal encoder and generator; in the second part we prove that the discontinuous points are dense.

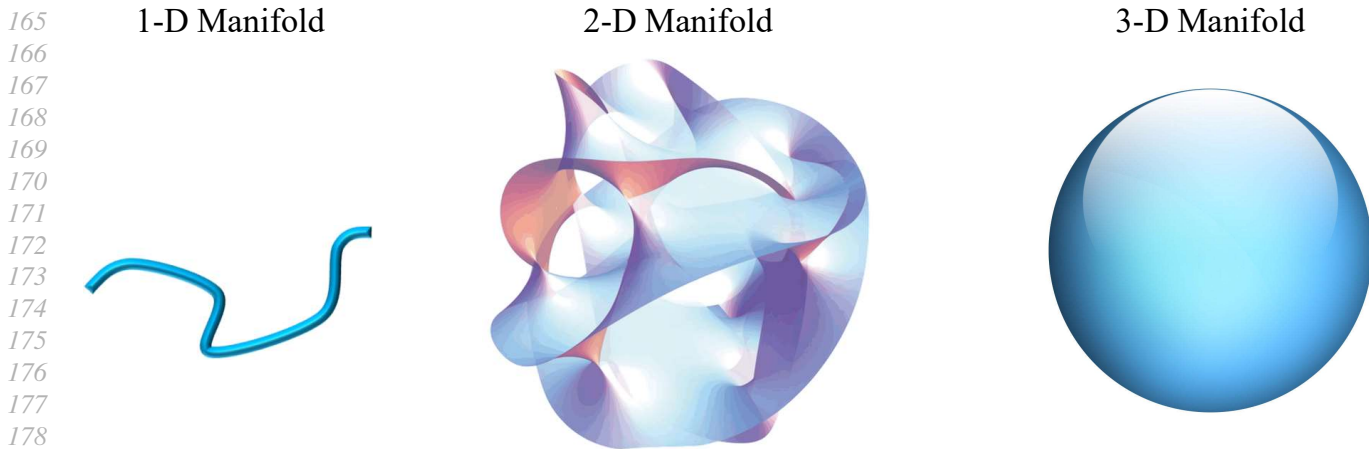


Figure 2. Intrinsic dimensions of manifolds in \mathbb{R}^3 . All the above sets have 3-D coordinates (x, y, z) in \mathbb{R}^3 , but their intrinsic dimensions are different.

C.1.1. PART I

The discontinuity is a deduction from the following invariance of domain theorem.

Proposition 1 (Invariance of Domain). *Let \mathcal{A} and \mathcal{B} be manifolds. Suppose that $\dim(\mathcal{A}) \neq \dim(\mathcal{B})$, then there is no diffeomorphism between \mathcal{A} and \mathcal{B} .*

It is easy to check that the underlying g and f introduce diffeomorphism between \mathcal{X} and \mathcal{Z} if they are continuous.

C.1.2. PART II

Suppose that for some $\mathbf{x} \in \mathcal{X}$, there is an open geodesic ball $B_{\mathcal{X}}(\mathbf{x}, r)$ with radius r centered at \mathbf{x} , and f is continuous on $B_{\mathcal{X}}(\mathbf{x}, r)$. If g is also continuous on $f(B_{\mathcal{X}}(\mathbf{x}, r))$, then it is easy to see that $f(B_{\mathcal{X}}(\mathbf{x}, r))$ is open subset of \mathcal{Z} . Thus f and g introduce diffeomorphism between them, which is contradictory to Proposition 1. \square

C.2. Proof to Theorem 2

Theorem 2. *When $\dim(\mathcal{Z}) \neq \dim(\mathcal{X})$, neural networks are not universal approximators to the underlying encoder and generator in Eq. (1) & (2). More specifically, we have:*

$$\inf_{\theta, \phi} \delta_e(\theta) + \delta_g(\phi) \geq D_e + D_g > 0, \quad (10)$$

where

$$D_e = \frac{1}{2} \sup_{\mathbf{x} \in \mathcal{X}} \limsup_{\mathbf{y} \rightarrow \mathbf{x}} \|e(\mathbf{y}) - e(\mathbf{x})\|, \quad (11)$$

$$D_g = \frac{1}{2} \sup_{\mathbf{z} \in \mathcal{Z}} \limsup_{\mathbf{w} \rightarrow \mathbf{z}} \|g(\mathbf{w}) - g(\mathbf{z})\|, \quad (12)$$

and

$$\delta_e(\theta) = \sup_{\mathbf{x} \in \mathcal{X}} \|E_{\theta}(\mathbf{x}) - e(\mathbf{x})\|, \quad (13)$$

$$\delta_g(\phi) = \sup_{\mathbf{z} \in \mathcal{Z}} \|G_{\phi}(\mathbf{z}) - g(\mathbf{z})\|. \quad (14)$$

Moreover, if $\dim(\mathcal{Z}) < \dim(\mathcal{X})$, we have

$$D_{JS}(\mathbb{P}_{G_{\phi}(\mathcal{Z})}, \mathbb{P}_{\mathcal{X}}) \geq \frac{\log 2}{2}, \quad (15)$$

and if $\dim(\mathcal{X}) < \dim(\mathcal{Z})$, we have

$$D_{JS}(\mathbb{P}_{\mathbf{E}_\theta(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}) \geq \frac{\log 2}{2}. \quad (16)$$

Proof. By Theorem 1, we know that at least one of \mathbf{f} and \mathbf{g} is not continuous. Assume that it is \mathbf{f} . It is then easy to see $D_{\mathbf{f}} > 0$. Thus $D_{\mathbf{f}} + D_{\mathbf{g}} > 0$.

Then we prove that

$$\inf_{\boldsymbol{\theta}} \delta_{\mathbf{f}}(\boldsymbol{\theta}) \geq D_{\mathbf{f}}, \quad (17)$$

$$\inf_{\phi} \delta_{\mathbf{g}}(\phi) \geq D_{\mathbf{g}}. \quad (18)$$

For any $\epsilon > 0$, we can pick some $\mathbf{x} \in \mathcal{X}$, such that

$$\limsup_{\mathbf{y} \rightarrow \mathbf{x}} \|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})\| > 2D_{\mathbf{f}} - \epsilon. \quad (19)$$

By definition, it means that we have some

$$\{\mathbf{x}_n\}_{n=1}^{\infty} \subset \mathcal{X}, \mathbf{x}_n \rightarrow \mathbf{x} \quad (20)$$

such that there is some positive $N \in \mathbb{N}$ to satisfy

$$\|\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\mathbf{x})\| > 2D_{\mathbf{f}} - 2\epsilon, \forall n > N. \quad (21)$$

For neural network \mathbf{F}_θ , if

$$\|\mathbf{F}_\theta(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| \geq D_{\mathbf{f}}, \quad (22)$$

then we have Eq. (17) already. Otherwise, we have

$$\|\mathbf{F}_\theta(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| < D_{\mathbf{f}}. \quad (23)$$

By triangular inequality, we have

$$\begin{aligned} & \|\mathbf{F}_\theta(\mathbf{x}_n) - \mathbf{f}(\mathbf{x}_n)\| + \|\mathbf{F}_\theta(\mathbf{x}_n) - \mathbf{F}_\theta(\mathbf{x})\| + \|\mathbf{F}_\theta(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| \geq \|\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\mathbf{x})\| \\ \Leftrightarrow & \|\mathbf{F}_\theta(\mathbf{x}_n) - \mathbf{f}(\mathbf{x}_n)\| \geq \|\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\mathbf{x})\| - \|\mathbf{F}_\theta(\mathbf{x}_n) - \mathbf{F}_\theta(\mathbf{x})\| - \|\mathbf{F}_\theta(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| \\ & \geq 2D_{\mathbf{f}} - 2\epsilon - D_{\mathbf{f}} - \|\mathbf{F}_\theta(\mathbf{x}_n) - \mathbf{F}_\theta(\mathbf{x})\|. \end{aligned} \quad (24)$$

Recall that \mathbf{F}_θ is differentiable and continuous. There exists $M \in \mathbb{N}$, such that $\|\mathbf{F}_\theta(\mathbf{x}_n) - \mathbf{F}_\theta(\mathbf{x})\| < \epsilon, \forall n > \max\{N, M\}$.

Then we have

$$\|\mathbf{F}_\theta(\mathbf{x}_n) - \mathbf{f}(\mathbf{x}_n)\| \geq D_{\mathbf{f}} - 3\epsilon, \forall n > \max\{N, M\}. \quad (25)$$

Recalling that ϵ can be arbitrarily small and $\mathbf{x}_n \in \mathcal{X}$, we then get

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{F}_\theta(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| \geq D_{\mathbf{f}}. \quad (26)$$

Applying the same argument to \mathbf{g} , we can prove Eq. (10).

At last we prove Eq. (16).

By definition, we have

$$2D_{JS}(\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}) = \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})}} \left[\log \frac{2d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})}}{d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})} + d\mathbb{P}_{\mathcal{Z}}} \right] + \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_{\mathcal{Z}}} \left[\log \frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})} + d\mathbb{P}_{\mathcal{Z}}} \right]. \quad (27)$$

Assume that $\dim(\mathcal{X}) < \dim(\mathcal{Z})$. We then have $\mathbf{F}_\theta(\mathcal{X}) \cap \mathcal{Z}$ is a zero measure set in \mathcal{Z} , as \mathbf{F}_θ is continuously differentiable. Recall we assume that all the probability distributions are absolutely continuous about the Lebesgue measure, which suggests

$$\begin{aligned} \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_{\mathcal{Z}}} \left[\log \frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})} + d\mathbb{P}_{\mathcal{Z}}} \right] &= \int_{\mathcal{Z} \setminus \mathbf{F}_\theta(\mathcal{X})} \log \frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})} + d\mathbb{P}_{\mathcal{Z}}} d\mathbb{P}_{\mathcal{Z}} + \int_{\mathbf{F}_\theta(\mathcal{X})} \log \frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})} + d\mathbb{P}_{\mathcal{Z}}} d\mathbb{P}_{\mathcal{Z}} \\ &= \int_{\mathcal{Z} \setminus \mathbf{F}_\theta(\mathcal{X})} \log \frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})} + d\mathbb{P}_{\mathcal{Z}}} d\mathbb{P}_{\mathcal{Z}} = \int_{\mathcal{Z} \setminus \mathbf{F}_\theta(\mathcal{X})} \log \frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\mathcal{Z}}} d\mathbb{P}_{\mathcal{Z}} = \int_{\mathcal{Z} \setminus \mathbf{F}_\theta(\mathcal{X})} \log(2) d\mathbb{P}_{\mathcal{Z}} = \log(2). \end{aligned} \quad (28)$$

As the KL divergence is non-negative, we have

$$\mathbb{E}_{\mathbf{z} \sim \mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})}} \left[\log \frac{2d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})}}{d\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})} + d\mathbb{P}_{\mathcal{Z}}} \right] \geq 0, \quad (29)$$

thus we get

$$2D_{JS}(\mathbb{P}_{\mathbf{F}_\theta(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}) \geq \log(2). \quad (30)$$

Applying the same argument to $\mathbb{P}_{\mathcal{X}}$ completes the proof. \square

C.3. Proof to Theorem 3

Theorem 3. Denote $n = \dim(\mathcal{X})$ and $d = \dim(\mathcal{Z})$. Let $m_d(\mathcal{Z})$ and $m_n(\mathcal{X})$ be the volumes of \mathcal{Z} and \mathcal{X} with respect to their intrinsic dimensions, respectively. Assume that \mathcal{Z} and \mathcal{X} are bounded manifolds embedded in high dimensional Euclidean spaces, but are almost everywhere diffeomorphism to open subsets in \mathbb{R}^d and \mathbb{R}^n , respectively. Denote $\text{diam}(\mathcal{Z}) = \sup_{\mathbf{z}, \mathbf{w} \in \mathcal{Z}} \|\mathbf{z} - \mathbf{w}\|$, $\text{diam}(\mathcal{X}) = \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$, and ω_i to be the volume of unit ball of dimension i . For simplicity, let $i, j \in \{d, n\}$ and

$$\begin{aligned} & \Gamma(\mathcal{A}, \mathcal{B}, i, j, a, b) \\ &= \left(\frac{\text{diam}(\mathcal{A})^{j-i} m_j(\mathcal{B})}{3a(2^j m_j(\mathcal{B}) + \omega_j \text{diam}(\mathcal{A})^j)} \right)^{\frac{1}{j-i}} \frac{b m_j(\mathcal{B})}{3}. \end{aligned} \quad (31)$$

Then there is a trade-off between the approximation error and the maximum gradient norm of networks if $\dim(\mathcal{Z}) \neq \dim(\mathcal{X})$. Specifically, if $\dim(\mathcal{Z}) < \dim(\mathcal{X})$, there exist constants $C_{\mathcal{X}} > 0$ that only depends on $\mathbb{P}_{\mathcal{X}}$ and $C_d > 0$ that only depends on d , such that

$$\begin{aligned} & W_1(\mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}, \mathbb{P}_{\mathcal{X}}) \left(\sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\| + 1 \right)^{\frac{n}{n-d}} \\ & \geq \Gamma(\mathcal{Z}, \mathcal{X}, d, n, C_d, C_{\mathcal{X}}); \end{aligned} \quad (32)$$

if $D_{JS}(\mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}, \mathbb{P}_{\mathcal{X}}) < \log 2$, then we further have

$$\begin{aligned} & \frac{D_{JS}(\mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}, \mathbb{P}_{\mathcal{X}}) (\sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\| + 1)^{\frac{2n}{n-d}}}{(\text{diam}(\mathcal{Z}) (\sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\|) + \text{diam}(\mathcal{X}))^2} \\ & \geq 4\Gamma(\mathcal{Z}, \mathcal{X}, d, n, C_d, C_{\mathcal{X}})^2. \end{aligned} \quad (33)$$

On the other hand, if $\dim(\mathcal{Z}) > \dim(\mathcal{X})$, there exist constants $C_{\mathcal{Z}} > 0$ that only depends on $\mathbb{P}_{\mathcal{Z}}$ and $C_n > 0$ that only depends on n , such that

$$\begin{aligned} & W_1(\mathbb{P}_{\mathbf{E}_\theta(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}) \left(\sup_{\mathbf{x} \in \mathcal{X}} \|\nabla \mathbf{E}_\theta\| + 1 \right)^{\frac{n}{d-n}} \\ & \geq \Gamma(\mathcal{X}, \mathcal{Z}, n, d, C_n, C_{\mathcal{Z}}); \end{aligned} \quad (34)$$

if $D_{JS}(\mathbb{P}_{\mathbf{E}_\theta(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}) < \log 2$, then we further have

$$\begin{aligned} & \frac{D_{JS}(\mathbb{P}_{\mathbf{E}_\theta(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}) (\sup_{\mathbf{x} \in \mathcal{X}} \|\nabla \mathbf{E}_\theta\| + 1)^{\frac{2n}{d-n}}}{(\text{diam}(\mathcal{X}) (\sup_{\mathbf{x} \in \mathcal{X}} \|\nabla \mathbf{E}_\theta\|) + \text{diam}(\mathcal{Z}))^2} \\ & \geq 4\Gamma(\mathcal{X}, \mathcal{Z}, n, d, C_n, C_{\mathcal{Z}})^2, \end{aligned} \quad (35)$$

where W_1 is the 1-Wasserstein distance (Villani, 2008).

Proof. Considering the Wasserstein divergence is integral over given domains, without loss of generality, we can assume that both \mathcal{X} and \mathcal{Z} are diffeomorphism to open subsets in \mathbb{R}^d and \mathbb{R}^n . Let $c_{\mathcal{X}}$ and $c_{\mathcal{Z}}$ be those diffeomorphisms. They are only decided by \mathcal{X} and \mathcal{Z} . Then $c_{\mathcal{X}} \circ \mathbf{G}_\phi \circ c_{\mathcal{Z}}^{-1}$ and $c_{\mathcal{Z}} \circ \mathbf{E}_\theta \circ c_{\mathcal{X}}^{-1}$ are networks for encoding GANs between open subsets of \mathbb{R}^d and \mathbb{R}^n .

For simplicity, we only prove the theorem for Euclidean case. For the manifold case, using the corresponding meaning of measure, distance, and gradient induced by diffeomorphism can yield the same result.

C.3.1. KANTOROVICH DUALITY

We first introduce the Kantorovich duality of W_1 divergence (Villani, 2008; Bottou et al., 2019)

Proposition 2. *When both \mathcal{A} and \mathcal{B} are bounded manifolds, let \mathcal{Q} be the set of all pairs $(f_{\mathcal{A}}, f_{\mathcal{B}})$ of \mathcal{A} and \mathcal{B} -integrable functions satisfying the property $\forall \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}, f_{\mathcal{A}}(\mathbf{x}) - f_{\mathcal{B}}(\mathbf{y}) \leq \|\mathbf{x} - \mathbf{y}\|$. We have*

$$W_1(\mathbb{P}_{\mathcal{A}}, \mathbb{P}_{\mathcal{B}}) = \sup_{(f_{\mathcal{A}}, f_{\mathcal{B}}) \in \mathcal{Q}} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{A}}} [f_{\mathcal{A}}(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{B}}} [f_{\mathcal{B}}(\mathbf{x})]. \quad (36)$$

It is then easy to see that

$$\hat{f}_{\mathcal{A}}(\mathbf{x}) = d(\mathbf{x}, \mathcal{B}) = \inf_{\mathbf{y} \in \mathcal{B}} \|\mathbf{x} - \mathbf{y}\|, \quad \hat{f}_{\mathcal{B}}(\mathbf{y}) = 0, \quad \forall \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B} \quad (37)$$

satisfies the condition of Proposition 2.

C.3.2. PROPERTY OF DISTANCE FUNCTION

Without loss of generality, assume that $\dim(\mathcal{Z}) < \dim(\mathcal{X})$. Then it is easy to see that

$$\overline{\mathbf{G}_{\phi}(\mathcal{Z})} = \mathbf{G}_{\phi}(\overline{\mathcal{Z}}) \quad (38)$$

is bounded zero measure set with respect to $m_n(\cdot)$. Let

$$\Omega_{\delta} = \{\mathbf{x} \in \mathcal{X} : d(\mathbf{x}, \mathbf{G}_{\phi}(\mathcal{Z})) \leq \delta\}. \quad (39)$$

We then have $m_n(\Omega_0) = 0$. On the other hand, it is easy to see $\lim_{\delta \rightarrow \infty} m_n(\Omega_{\delta}) \rightarrow \infty$, and $m_n(\Omega_{\delta})$ is continuous about δ , thus $m_n(\Omega_{\delta})$ can reach any value in \mathbb{R}_+ .

When $\delta = \frac{\text{diam}(\mathcal{Z})}{2}$, we have $B_n(\mathbf{x}, \delta) \subset \Omega_{\delta}$ for any $\mathbf{x} \in \mathbf{G}_{\phi}(\mathcal{Z})$. Thus we have

$$m_n(\Omega_{\delta}) \geq \omega_n \left(\frac{\text{diam}(\mathcal{Z})}{2} \right)^n \quad (40)$$

when $\delta = \frac{\text{diam}(\mathcal{Z})}{2}$, where ω_n is the volume of unit ball in \mathcal{X} . Note that

$$0 < \frac{w_n \text{diam}(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} m_n(\mathcal{X}) < \omega_n \left(\frac{\text{diam}(\mathcal{Z})}{2} \right)^n. \quad (41)$$

By the continuity of $m_n(\Omega_{\delta})$, we can conclude that there is some $\delta \in (0, \frac{\text{diam}(\mathcal{Z})}{2})$, such that

$$m_n(\Omega_{\delta}) = \frac{w_n \text{diam}(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} m_n(\mathcal{X}) < \frac{m_n(\mathcal{X})}{3}. \quad (42)$$

We give an estimation to the lower bound of δ below. Assume that there are N_{δ} balls $\{B_d(\mathbf{z}_k, \delta)\}_{k=1}^{N_{\delta}}$ in \mathbb{R}^d covering \mathcal{Z} . It then follows that

$$\mathbf{G}_{\phi}(\mathcal{Z}) \subset \cup_{k=1}^{N_{\delta}} B_n \left(\mathbf{G}_{\phi}(\mathbf{z}_k), \sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_{\phi}\| \delta \right), \quad (43)$$

where B_n are balls in \mathbb{R}^n . It is then easy to see that

$$\Omega_{\delta} \subset \left\{ \mathbf{x} \in \mathcal{X} : d \left(\mathbf{x}, \cup_{k=1}^{N_{\delta}} B_n \left(\mathbf{G}_{\phi}(\mathbf{z}_k), \sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_{\phi}\| \delta \right) \right) \leq \delta \right\} \subset \cup_{k=1}^{N_{\delta}} B_n \left(\mathbf{G}_{\phi}(\mathbf{z}_k), \sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_{\phi}\| \delta + \delta \right). \quad (44)$$

Thus we get

$$m_n(\Omega_{\delta}) = \frac{w_n \text{diam}(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} m_n(\mathcal{X}) \leq N_{\delta} \omega_n \left(\sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_{\phi}\| + 1 \right)^n \delta^n. \quad (45)$$

In (Rogers, 1957), for ball with radius R of \mathbb{R}^d , there is a constant C_d that only depends on d , such that $C_d(\frac{R}{\delta})^d$ is the upper bound for minimal number of balls with radius $\delta < R$ to cover the ball with radius R . Thus we can pick $\{B_d(\mathbf{z}_k, \delta)\}_{k=1}^{N_\delta}$ of \mathbb{R}^d such that

$$N_\delta \leq C_d \left(\frac{\text{diam}(\mathcal{Z})}{\delta} \right)^d. \quad (46)$$

We then have

$$\begin{aligned} \frac{w_n \text{diam}(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} m_n(\mathcal{X}) &\leq C_d \left(\frac{\text{diam}(\mathcal{Z})}{\delta} \right)^d \omega_n \left(\sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\| + 1 \right)^n \delta^n \\ \Leftrightarrow \delta &\geq \left(\frac{\text{diam}(\mathcal{Z})^{n-d} m_n(\mathcal{X})}{3C_d (\sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\| + 1)^n (2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} \right)^{\frac{1}{n-d}}. \end{aligned} \quad (47)$$

C.3.3. LOWER BOUND ESTIMATION

By the Littlewood principles of real analysis (Stein & Shakarchi, 2009), there exists compact set $\mathcal{X}' \subset \mathcal{X}$, such that

$$m_n(\mathcal{X}') \geq \frac{2}{3} m_n(\mathcal{X}) \quad (48)$$

and the density $p_{\mathcal{X}}$ of $\mathbb{P}_{\mathcal{X}}$ is continuous on it. As \mathcal{X} is the support of $p_{\mathcal{X}}$ and \mathcal{X}' is compact, there exists some positive real number $C_{\mathcal{X}}$ that is only depends on $\mathbb{P}_{\mathcal{X}}$, such that

$$p_{\mathcal{X}}(\mathbf{x}) \geq C_{\mathcal{X}}, \forall \mathbf{x} \in \mathcal{X}'. \quad (49)$$

Let $I_{\mathcal{A}}$ be identity function

$$I_{\mathcal{A}}(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in \mathcal{A} \\ 0, & \mathbf{x} \notin \mathcal{A} \end{cases}, \quad (50)$$

$\hat{f}_{\mathcal{X}}(\mathbf{x}) = d(\mathbf{x}, \mathbf{G}_\phi(\mathcal{Z}))$, and $\hat{f}_{\mathbf{G}_\phi(\mathcal{Z})} = 0$. We then have

$$\begin{aligned} W_1(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}) &= \sup_{(f_{\mathcal{X}}, f_{\mathbf{G}_\phi(\mathcal{Z})}) \in \mathcal{Q}} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [f_{\mathcal{X}}(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}} [f_{\mathbf{G}_\phi(\mathcal{Z})}(\mathbf{x})] \\ &\geq \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\hat{f}_{\mathcal{X}}(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}} [\hat{f}_{\mathbf{G}_\phi(\mathcal{Z})}(\mathbf{x})] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\hat{f}_{\mathcal{X}}(\mathbf{x})] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\hat{f}_{\mathcal{X}}(\mathbf{x}) I_{\mathcal{X} \setminus \mathbf{G}_\phi(\mathcal{Z})}] + \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\hat{f}_{\mathcal{X}}(\mathbf{x}) I_{\mathcal{X} \cap \mathbf{G}_\phi(\mathcal{Z})}] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\hat{f}_{\mathcal{X}}(\mathbf{x}) I_{\mathcal{X} \setminus \mathbf{G}_\phi(\mathcal{Z})}] \\ &= \int_{\mathcal{X} \setminus \mathbf{G}_\phi(\mathcal{Z})} \hat{f}_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x} \\ &\geq \int_{\mathcal{X}' \setminus \mathbf{G}_\phi(\mathcal{Z})} \hat{f}_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x} \\ &\geq \int_{\mathcal{X}' \setminus \Omega_\delta} \delta C_{\mathcal{X}} d\mathbf{x} \\ &\geq \delta C_{\mathcal{X}} \frac{m_n(\mathcal{X})}{3} \\ &\geq \left(\frac{\text{diam}(\mathcal{Z})^{n-d} m_n(\mathcal{X})}{3C_d (\sup_{\mathbf{z} \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\| + 1)^n (2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} \right)^{\frac{1}{n-d}} \frac{C_{\mathcal{X}} m_n(\mathcal{X})}{3}. \end{aligned} \quad (51)$$

Note that in the penultimate step, we use the estimation

$$m_n(\Omega_\delta) = \frac{w_n \text{diam}(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} m_n(\mathcal{X}) \approx \frac{m_n(\mathcal{X})}{3}. \quad (52)$$

This estimation is very loose when $\text{diam}(\mathcal{Z})$ is tiny. We may use

$$\int_{\mathcal{X}' \setminus \Omega_\delta} \delta C_{\mathcal{X}} dx \geq \delta C_{\mathcal{X}} \left(\frac{2}{3} - \frac{w_n \text{diam}(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n \text{diam}(\mathcal{Z})^n)} \right) m_n(\mathcal{X}) \quad (53)$$

directly in this case to get more accurate estimation. Another issue is Eq. (47). When $\text{diam}(\mathcal{Z})$ is tiny, the estimation for δ is also very loose. A detailed analysis for this case can offer more accurate estimation, but may break the current uniform format of the final inequality Eq. (51). As our inequality is already very complicated, we just use the current result of Eq. (51).

C.3.4. INEQUALITY FOR JENSEN-SHANNON DIVERGENCE

Recall the inequality (Arjovsky & Bottou, 2017; Villani, 2008) induced by the Kantorovich duality

$$W_1(\mathbb{P}, \mathbb{Q}) \leq C \cdot TV(\mathbb{P}, \mathbb{Q}), \quad (54)$$

where $TV(\mathbb{P}, \mathbb{Q})$ is the total variation between \mathbb{P} and \mathbb{Q} , C is the diameter of the ball that contains the supports of \mathbb{P} and \mathbb{Q} , and Pinsker's inequality of Kullback–Leibler divergence is

$$TV(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} D_{KL}(\mathbb{P}, \mathbb{Q})}. \quad (55)$$

We then have

$$\begin{aligned} D_{JS}(\mathbb{P}, \mathbb{Q}) &\geq \left(\frac{1}{2} \left(\sqrt{\frac{1}{2} D_{KL}(\mathbb{P}, \mathbb{M})} + \sqrt{\frac{1}{2} D_{KL}(\mathbb{M}, \mathbb{Q})} \right) \right)^2 \geq \frac{1}{4} (TV(\mathbb{P}, \mathbb{M}) + TV(\mathbb{M}, \mathbb{Q}))^2 \\ &\geq \frac{1}{4} TV(\mathbb{P}, \mathbb{Q})^2 \geq \frac{1}{4C^2} W_1(\mathbb{P}, \mathbb{Q})^2, \end{aligned} \quad (56)$$

where

$$\mathbb{M} = \frac{\mathbb{P} + \mathbb{Q}}{2}. \quad (57)$$

The first inequality follows the Jensen inequality, and the third inequality follows the triangle inequality of total variation.

Note that when

$$D_{JS}(\mathbb{P}, \mathbb{Q}) < \log 2, \quad (58)$$

the supports of \mathbb{P} and \mathbb{Q} must intersect each other, thus C is not larger than the sum of diameters of supports of \mathbb{P} and \mathbb{Q} . Substituting Eq. (56) into (51) and noting that when

$$D_{JS}(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}) < \log 2, \quad (59)$$

we can take C as

$$C = (\text{diam}(\mathcal{X}) + \text{diam}(\mathbf{G}_\phi(\mathcal{Z}))) \leq \left(\text{diam}(\mathcal{X}) + \left(\sup_{z \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\| \right) \text{diam}(\mathcal{Z}) \right). \quad (60)$$

Substituting Eq. (60) into (56), we then have the inequalities in Eq. (33) & (35) for Jensen-Shannon divergence.

Applying the same process to the case of $\text{dim}(\mathcal{Z}) > \text{dim}(\mathcal{X})$ completes the proof. \square

C.4. Proof to Corollary 1

Corollary 1. *Under the condition of Theorem 3, we let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be N independent samples from $\mathbb{P}_{\mathcal{X}}$. Let $\mathbb{Q}_N = \frac{1}{N} \sum_{i=1}^N \delta_{\mathbf{x}_i}$ be the empirical distribution of those samples, where $\delta_{\mathbf{x}}$ is the Dirac distribution for sample \mathbf{x} . Further assume that $\int_{\mathcal{X}} \|\mathbf{x} - \mathbf{y}\| \mathbb{P}_{\mathcal{X}}(d\mathbf{y}) < \infty, \forall \mathbf{x} \in \mathcal{X}$. If $d < n$, then there exists constant $C > 0$ such that for all generator network \mathbf{G}_ϕ*

$$\mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_N \sim \mathbb{P}_{\mathcal{X}}} [W_1(\mathbb{P}_{\mathbf{G}_\phi(\mathcal{Z})}, \mathbb{Q}_N)] \geq \frac{C}{(\sup_{z \in \mathcal{Z}} \|\nabla \mathbf{G}_\phi\| + 1)^{\frac{n}{n-d}}} - \mathcal{O}(N^{-\frac{1}{n}}). \quad (61)$$

Proof. Taking the following proposition (Bottou et al., 2019; Villani, 2008) and triangle inequality of Wasserstein distance into Theorem 3 yields the proof.

Proposition 3. Let \mathbb{Q}_N be defined in Corollary 1. We have

$$\mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_N} [W_1(\mathbb{Q}_N, \mathbb{P}_{\mathcal{X}})] = \mathcal{O}(N^{-\frac{1}{n}}). \quad (62)$$

□

C.5. Proof to Lemma 1

Lemma 1. If \mathbf{h} is continuous and \mathcal{D} is connected, then $\mathbf{h}(\mathcal{D})$ is also connected.

Proof. See (Rudin et al., 1964).

□

C.6. Proof to Lemma 2

Lemma 2 (Estimation of the JS Divergence). For a fixed generator \mathbf{G} , when the discriminator \mathbf{D} is optimal, we have

$$V(\mathbf{D}, \mathbf{G}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [\log(\mathbf{D}(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim \mathbb{P}_{\mathcal{Z}}} [\log(1 - \mathbf{D}(\mathbf{G}(\mathbf{z})))] \quad (63)$$

and

$$D_{JS}(\mathbb{P}_{\mathbf{G}}, \mathbb{P}_{data}) = \log 2 + \frac{1}{2} V(\mathbf{D}, \mathbf{G}). \quad (64)$$

Proof. See proof of Theorem 1 in (Goodfellow et al., 2014).

□

C.7. Proof to Lemma 3

Lemma 3. For differentiable map $\mathbf{h} : \mathcal{D} \rightarrow \mathbb{R}^d$, $\mathcal{D} \subset \mathbb{R}^d$, we have

$$m_d(\mathbf{h}(\mathcal{D})) \leq \left(\sup_{\mathbf{x} \in \mathcal{D}} \|\nabla \mathbf{h}\| \right)^d m_d(\mathcal{D}), \quad (65)$$

where $m_d(\cdot)$ is the volume of set of dimension d .

Proof. By definition, for any $\epsilon > 0$, there is balls B_k with radius r_k , $k \in \mathbb{N}$, such that

$$\mathcal{D} \subset \cup_k B_k, \quad m_d(\mathcal{D}) \leq \sum_k r_k^d \leq m_d(\mathcal{D}) + \epsilon. \quad (66)$$

It is easy to see that $\mathbf{h}(B_k)$ is contained in some ball with radius $(\sup_{\mathbf{x} \in \mathcal{D}} \|\nabla \mathbf{h}\|) r_k$, thus we have

$$m_d(\mathbf{h}(\mathcal{D})) \leq \left(\sup_{\mathbf{x} \in \mathcal{D}} \|\nabla \mathbf{h}\| \right)^d \sum_k r_k^d \leq \left(\sup_{\mathbf{x} \in \mathcal{D}} \|\nabla \mathbf{h}\| \right)^d m_d(\mathcal{D}) + \left(\sup_{\mathbf{x} \in \mathcal{D}} \|\nabla \mathbf{h}\| \right)^d \epsilon. \quad (67)$$

Let $\epsilon \rightarrow 0$. We then prove the lemma.

□

References

- Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.
- Baldi, P. and Sadowski, P. J. Understanding dropout. *Advances in neural information processing systems*, 26:2814–2822, 2013.
- Bottou, L., Arjovsky, M., Lopez-Paz, D., and Oquab, M. Geometrical insights for implicit generative modeling, 2019.

- 550 Choi, K., Hawthorne, C., Simon, I., Dinculescu, M., and Engel, J. Encoding musical style with transformer autoencoders. In
551 *International Conference on Machine Learning*, pp. 1899–1908. PMLR, 2020.
- 552 Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial Feature Learning, 2017.
- 554 Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. Adversarially Learned
555 Inference, 2017.
- 556 Gallot, S., Hulin, D., and Lafontaine, J. *Riemannian geometry*, volume 2. Springer, 1990.
- 557 Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative
558 adversarial networks, 2014.
- 559 Hayakawa, Y., Marumoto, A., and Sawada, Y. Effects of the chaotic noise on the performance of a neural network model for
560 optimization problems. *Physical review E*, 51(4):R2693, 1995.
- 561 He, Z., Rakin, A. S., and Fan, D. Parametric noise injection: Trainable randomness to improve deep neural network
562 robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
563 Recognition*, pp. 588–597, 2019.
- 564 Hinton, G. E. and Zemel, R. S. Autoencoders, minimum description length and helmholtz free energy. In *Advances in
565 neural information processing systems*, pp. 3–10, 1994.
- 566 Jenni, S. and Favaro, P. On stabilizing generative adversarial training with noise. In *Proceedings of the IEEE/CVF
567 Conference on Computer Vision and Pattern Recognition*, pp. 12145–12153, 2019.
- 568 Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings
569 of the IEEE conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.
- 570 Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of
571 StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119,
572 2020.
- 573 Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*,
574 2015.
- 575 Ng, A. et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.
- 576 Rogers, C. A note on coverings. *Mathematika*, 4(1):1–6, 1957.
- 577 Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.
- 578 Stein, E. M. and Shakarchi, R. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press,
579 2009.
- 580 Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*,
581 2017.
- 582 Vahdat, A. and Kautz, J. NVAE: A deep hierarchical variational autoencoder, 2021.
- 583 Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
- 584 Zhu, J., Zhao, D., Zhou, B., and Zhang, B. LIA: Latently invertible autoencoder with adversarial learning. *arXiv*, pp.
585 arXiv–1906, 2019.

586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604