Uncertainty Principles of Encoding GANs
Supplementary Material

Ruili Feng [1] Zhouchen Lin [‡ 2 3] Jiapeng Zhu [4] Deli Zhao [5] Jinren Zhou [5] Zheng-Jun Zha [† 1]

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. S1 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 $\boldsymbol{g}$ transports the latent distribution into the data one,

$$\mathbb{P}_{\boldsymbol{g}(\mathcal{Z})}(\mathcal{A}) = \int_{\boldsymbol{g}^{-1}(\mathcal{A})} d\mathbb{P}_{\mathcal{Z}} = \mathbb{P}_{\mathcal{X}}(\mathcal{A}), \forall \mathcal{A} \subset \mathcal{F}_{\mathcal{X}}, \tag{S1}$$

† Corresponding author, ‡ co-corresponding author. [1]University of Science and Technology of China, Hefei, China. [2]Key Lab. of Machine Perception (MoE), School of EECS, Peking University, Beijing, China. [3]Pazhou Lab, Guangzhou, China. [4]Hong Kong University of Science and Technology, Hong Kong, China. [5]Alibaba Group. Correspondence to: Ruili Feng <frl1996@mail.ustc.edu.cn>, Zhouchen Lin <zlin@pku.edu.cn>, Zheng-Jun Zha <zhazj@ustc.edu.cn>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

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

$$\boldsymbol{e} \circ \boldsymbol{g}(\boldsymbol{z}) = \boldsymbol{z}, \boldsymbol{g} \circ \boldsymbol{e}(\boldsymbol{x}) = \boldsymbol{x}, \forall \boldsymbol{z} \in \mathcal{Z}, \boldsymbol{x} \in \mathcal{X}. \tag{S2}$$

The training algorithms then aim at approximating the underlying 'perfect' encoder and generator with parameterized neural networks $\boldsymbol{E_\theta}$ and $\boldsymbol{G_\phi}$, respectively.

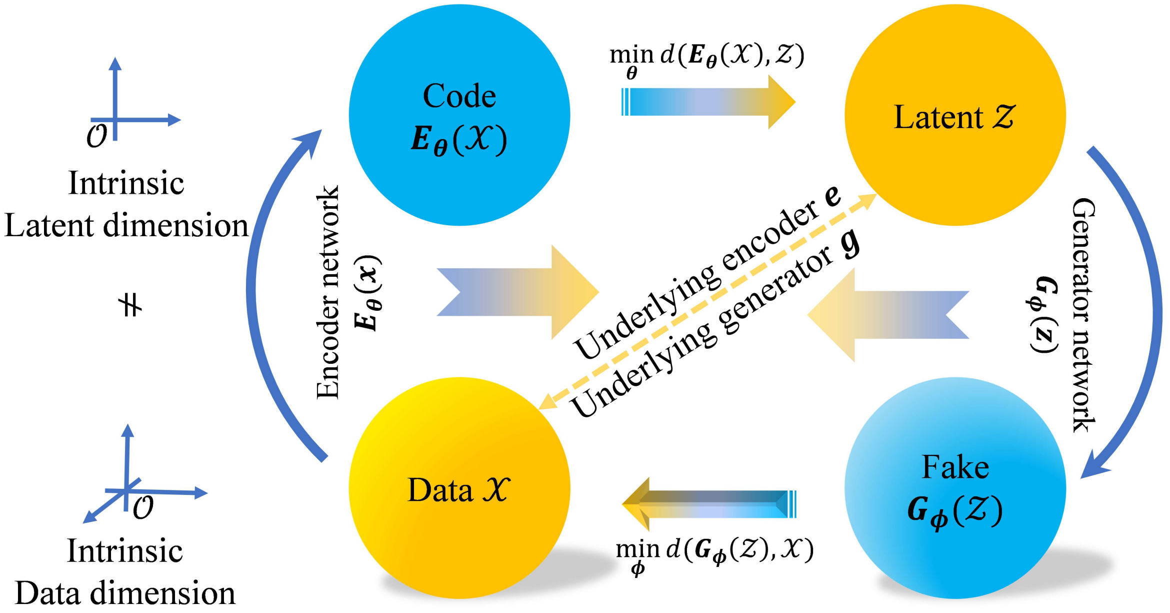
$\min_{\boldsymbol{\theta}} d(\boldsymbol{E_\theta}(\mathcal{X}), \mathcal{Z})$
Code
$\boldsymbol{E_\theta}(\mathcal{X})$
Latent $\mathcal{Z}$
$\mathcal{O}$
Intrinsic
Latent dimension
$\neq$
Encoder network
$\boldsymbol{E_\theta}(\boldsymbol{x})$
Underlying encoder $\boldsymbol{e}$
Underlying generator $\boldsymbol{g}$
Generator network
$\boldsymbol{G_\phi}(\boldsymbol{z})$
$\mathcal{O}$
Intrinsic
Data dimension
Data $\mathcal{X}$
Fake
$\boldsymbol{G_\phi}(\mathcal{Z})$
$\min_{\boldsymbol{\phi}} d(\boldsymbol{G_\phi}(\mathcal{Z}), \mathcal{X})$

Figure S1. A typical training process. Training algorithms guide encoder network $\boldsymbol{E_\theta}$ and generator network $\boldsymbol{G_\phi}$ to the underlying $\boldsymbol{e}$ and $\boldsymbol{g}$ by minimizing divergences from $\boldsymbol{E_\theta}(\mathcal{X})$ to $\mathcal{Z}$ and from $\boldsymbol{G_\phi}(\mathcal{Z})$ 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 S1. 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	$\boldsymbol{x}, \boldsymbol{z}$
Vector value functions	$\boldsymbol{e}, \boldsymbol{g}$
Sets & Manifolds	$\mathcal{X}, \mathcal{Z}, \mathcal{D}$
Neural networks	$\boldsymbol{E_\theta}, \boldsymbol{G_\phi}, \boldsymbol{D_\psi}$
Distributions	$\mathbb{P}_{\mathcal{X}}(\boldsymbol{x}), \mathbb{P}_{\mathcal{Z}}(\boldsymbol{z})$
Induced Distributions	$\mathbb{P}_{\boldsymbol{g}(\mathcal{Z})}(\boldsymbol{x}), \mathbb{P}_{\boldsymbol{E_\theta}(\mathcal{X})}(\boldsymbol{z})$
Intrinsic dimension of manifolds	$dim(\mathcal{X}), dim(\mathcal{Z})$

Fig. S1 illustrates a typical training process. Training algorithms guide the encoder network $\boldsymbol{E_\theta}$ and the generator network $\boldsymbol{G_\phi}$ to the underlying 'perfect' encoder $\boldsymbol{e}$ and generator $\boldsymbol{g}$ by minimizing divergence from $\boldsymbol{E_\theta}(\mathcal{X})$ to $\mathcal{Z}$ and from $\boldsymbol{G_\phi}(\mathcal{Z})$ to $\mathcal{X}$. Popular divergences include the Jensen-Shannon divergence (Donahue et al., 2017; Dumoulin et al., 2017)

$$D_{JS}(\mathbb{P}_{\mathcal{A}}, \mathbb{Q}_{\mathcal{B}}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{A}}}\left[\log\left(\frac{2\mathbb{P}_{\mathcal{A}}(d\boldsymbol{x})}{\mathbb{P}_{\mathcal{A}}(d\boldsymbol{x}) + \mathbb{Q}_{\mathcal{B}}(d\boldsymbol{x})}\right)\right] + \mathbb{E}_{\boldsymbol{y}\sim\mathbb{Q}_{\mathcal{B}}}\left[\log\left(\frac{2\mathbb{Q}_{\mathcal{B}}(d\boldsymbol{y})}{\mathbb{P}_{\mathcal{A}}(d\boldsymbol{y}) + \mathbb{Q}_{\mathcal{B}}(d\boldsymbol{y})}\right)\right], \tag{S3}$$

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

$$W_1(\mathbb{P}_{\mathcal{A}}, \mathbb{Q}_{\mathcal{B}}) = \inf_{\pi\in\Pi(\mathbb{P}_{\mathcal{A}}, \mathbb{Q}_{\mathcal{B}})} \int_{\mathcal{A}\times\mathcal{B}} \|\boldsymbol{x} - \boldsymbol{y}\| \, d\pi(\boldsymbol{x}, \boldsymbol{y}), \tag{S4}$$

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

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. S2. *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 $\boldsymbol{x} \in \mathcal{A}$, it has a small open neighborhood $\mathcal{U}$ and a continuous bijection $\boldsymbol{b}$ (also called the **chart** at $\boldsymbol{x}$) 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

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \max_{\boldsymbol{\psi}} V(\boldsymbol{E}_{\boldsymbol{\theta}}, \boldsymbol{G}_{\boldsymbol{\phi}}, \boldsymbol{D}_{\boldsymbol{\psi}}), \tag{S5}$$

where $\boldsymbol{D}_{\boldsymbol{\psi}}$ is the discriminator network for the $(\boldsymbol{x}, \boldsymbol{z})$ pair, and

$$V(\boldsymbol{E}_{\boldsymbol{\theta}}, \boldsymbol{G}_{\boldsymbol{\phi}}, \boldsymbol{D}_{\boldsymbol{\psi}}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}} \left[\log\left(D(\boldsymbol{x}, \boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}))\right)\right] + \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\mathcal{Z}}} \left[1 - \log(D(\boldsymbol{G}_{\boldsymbol{\phi}}(\boldsymbol{z}), \boldsymbol{z}))\right]. \tag{S6}$$

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

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\psi}} V(\boldsymbol{G}_{\boldsymbol{\phi}}, \boldsymbol{D}_{\boldsymbol{\psi}}), \tag{S7}$$

in which

$$V(\boldsymbol{G}_{\boldsymbol{\phi}}, \boldsymbol{D}_{\boldsymbol{\psi}}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}[\log(D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\mathcal{Z}}}[1 - \log(D(\boldsymbol{G}_{\boldsymbol{\phi}}(\boldsymbol{z})))], \tag{S8}$$

and then an encoder is trained by optimizing

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}[\|\boldsymbol{G}_{\boldsymbol{\phi}} \circ \boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{x}\|_2^2] + d(\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}), \tag{S9}$$

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 $\boldsymbol{z} \in \mathcal{Z}$ 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 $\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $\boldsymbol{G}_{\boldsymbol{\phi}}(\boldsymbol{z})$ are continuous and piece-wise continuously differentiable with respect to inputs $\boldsymbol{x}$ and $\boldsymbol{z}$; 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.*

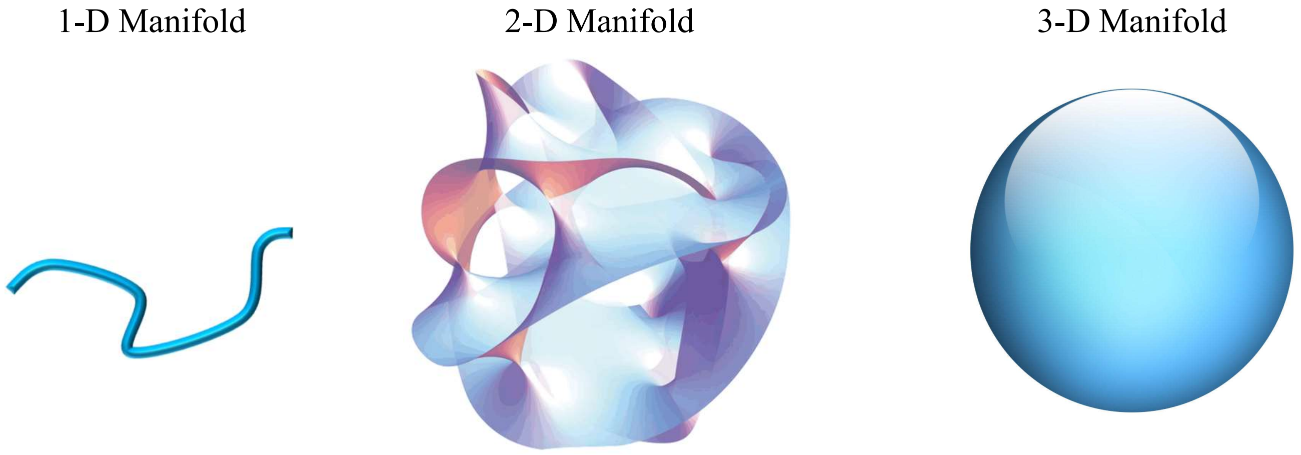

Figure S2. Intrinsic dimensions of manifolds in $\mathbb{R}^3$. All the above sets have 3-D coordinates (x, y, z) in $\mathbb{R}^3$, but their intrinsic dimensions are different.

C. 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. (S1) & (S2) is discontinuous; and for any $\boldsymbol{x} \in \mathcal{X}, \delta > 0$, there is a point $\boldsymbol{x}'$ in the geodesic ball centered at $\boldsymbol{x}$ with radius δ, such that $\boldsymbol{e}$ is not continuous at $\boldsymbol{x}'$ or $\boldsymbol{g}$ is not continuous at $\boldsymbol{e}(\boldsymbol{x}')$. 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.

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 $\boldsymbol{g}$ and $\boldsymbol{e}$ introduce diffeomorphism between $\mathcal{X}$ and $\mathcal{Z}$ if they are continuous.

C.1.2. Part II

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

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. (S1) & (S2). More specifically, we have:*

$$\inf_{\boldsymbol{\theta},\boldsymbol{\phi}} \delta_{\boldsymbol{e}}(\boldsymbol{\theta}) + \delta_{\boldsymbol{g}}(\boldsymbol{\phi}) \geq D_{\boldsymbol{e}} + D_{\boldsymbol{g}} > 0, \tag{S10}$$

where

$$D_{\boldsymbol{e}} = \frac{1}{2} \sup_{\boldsymbol{x} \in \mathcal{X}} \limsup_{\boldsymbol{y} \to \boldsymbol{x}} \|\boldsymbol{e}(\boldsymbol{y}) - \boldsymbol{e}(\boldsymbol{x})\|, \tag{S11}$$

$$D_{\boldsymbol{g}} = \frac{1}{2} \sup_{\boldsymbol{z} \in \mathcal{Z}} \limsup_{\boldsymbol{w} \to \boldsymbol{z}} \|\boldsymbol{g}(\boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{z})\|, \tag{S12}$$

and

$$\delta_{\boldsymbol{e}}(\boldsymbol{\theta}) = \sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{e}(\boldsymbol{x})\|, \tag{S13}$$

$$\delta_{\boldsymbol{g}}(\boldsymbol{\phi}) = \sup_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{G}_{\boldsymbol{\phi}}(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{z})\|. \tag{S14}$$

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

$$D_{JS}(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}, \mathbb{P}_{\mathcal{X}}) \geq \frac{\log 2}{2}, \tag{S15}$$

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

$$D_{JS}(\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}}) \geq \frac{\log 2}{2}. \tag{S16}$$

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

Then we prove that

$$\inf_{\boldsymbol{\theta}} \delta_{\boldsymbol{e}}(\boldsymbol{\theta}) \geq D_{\boldsymbol{e}}, \tag{S17}$$

$$\inf_{\boldsymbol{\phi}} \delta_{\boldsymbol{g}}(\boldsymbol{\phi}) \geq D_{\boldsymbol{g}}. \tag{S18}$$

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

$$\limsup_{\boldsymbol{y} \to \boldsymbol{x}} \|\boldsymbol{e}(\boldsymbol{y}) - \boldsymbol{e}(\boldsymbol{x})\| > 2D_{\boldsymbol{e}} - \epsilon. \tag{S19}$$

By definition, it means that we have some

$$\{\boldsymbol{x}_n\}_{n=1}^{\infty} \subset \mathcal{X}, \boldsymbol{x}_n \to \boldsymbol{x} \tag{S20}$$

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

$$\|\boldsymbol{e}(\boldsymbol{x}_n) - \boldsymbol{e}(\boldsymbol{x})\| > 2D_{\boldsymbol{e}} - 2\epsilon, \forall n > N. \tag{S21}$$

For neural network $\boldsymbol{E}_{\boldsymbol{\theta}}$, if

$$\|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{e}(\boldsymbol{x})\| \geq D_{\boldsymbol{e}}, \tag{S22}$$

then we have Eq. (S17) already. Otherwise, we have

$$\|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{e}(\boldsymbol{x})\| < D_{\boldsymbol{e}}. \tag{S23}$$

By triangular inequality, we have

$$\begin{aligned} &\|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - \boldsymbol{e}(\boldsymbol{x}_n)\| + \|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - \boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x})\| + \|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{e}(\boldsymbol{x})\| \geq \|\boldsymbol{e}(\boldsymbol{x}_n) - \boldsymbol{e}(\boldsymbol{x})\| \\ \Leftrightarrow &\|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - \boldsymbol{e}(\boldsymbol{x}_n)\| \geq \|\boldsymbol{e}(\boldsymbol{x}_n) - \boldsymbol{e}(\boldsymbol{x})\| - \|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - \boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x})\| - \|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{e}(\boldsymbol{x})\| \\ &\geq 2D_{\boldsymbol{e}} - 2\epsilon - D_{\boldsymbol{e}} - \|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - \boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x})\|. \end{aligned} \tag{S24}$$

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

$$\|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - \boldsymbol{e}(\boldsymbol{x}_n)\| \geq D_{\boldsymbol{e}} - 3\epsilon, \forall n > max\{N, M\}. \tag{S25}$$

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

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{e}(\boldsymbol{x})\| \geq D_{\boldsymbol{e}}. \tag{S26}$$

Applying the same argument to $\boldsymbol{g}$, we can prove Eq. (S10).

At last we prove Eq. (S16).

By definition, we have

$$2D_{JS}(\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})},\mathbb{P}_{\mathcal{Z}}) = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}}\left[\log\frac{2d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}}{d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}+d\mathbb{P}_{\mathcal{Z}}}\right] + \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\mathcal{Z}}}\left[\log\frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}+d\mathbb{P}_{\mathcal{Z}}}\right]. \tag{S27}$$

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

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\mathcal{Z}}}\left[\log\frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}+d\mathbb{P}_{\mathcal{Z}}}\right] &= \int_{\mathcal{Z}\setminus\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}\log\frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}+d\mathbb{P}_{\mathcal{Z}}}d\mathbb{P}_{\mathcal{Z}} + \int_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}\log\frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}+d\mathbb{P}_{\mathcal{Z}}}d\mathbb{P}_{\mathcal{Z}} \\ &= \int_{\mathcal{Z}\setminus\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}\log\frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}+d\mathbb{P}_{\mathcal{Z}}}d\mathbb{P}_{\mathcal{Z}} = \int_{\mathcal{Z}\setminus\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}\log\frac{2d\mathbb{P}_{\mathcal{Z}}}{d\mathbb{P}_{\mathcal{Z}}}d\mathbb{P}_{\mathcal{Z}} = \int_{\mathcal{Z}\setminus\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}\log(2)d\mathbb{P}_{\mathcal{Z}} = \log(2).\end{aligned} \tag{S28}$$

As the KL divergence is non-negative, we have

$$\mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}}\left[\log\frac{2d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}}{d\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}+d\mathbb{P}_{\mathcal{Z}}}\right] \geq 0, \tag{S29}$$

thus we get

$$2D_{JS}(\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})},\mathbb{P}_{\mathcal{Z}}) \geq \log(2). \tag{S30}$$

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

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 $diam(\mathcal{Z}) = \sup_{\boldsymbol{z},\boldsymbol{w}\in\mathcal{Z}}\|\boldsymbol{z}-\boldsymbol{w}\|$, $diam(\mathcal{X}) = \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}}\|\boldsymbol{x}-\boldsymbol{y}\|$, and ω_i to be the volume of unit ball of dimension i. For simplicity, let $i, j \in \{d, n\}$ and*

$$\begin{aligned}&\Gamma(\mathcal{A},\mathcal{B},i,j,a,b) \\ =&\left(\frac{diam(\mathcal{A})^{j-i}m_j(\mathcal{B})}{3a(2^j m_j(\mathcal{B}) + \omega_j diam(\mathcal{A})^j)}\right)^{\frac{1}{j-i}}\frac{bm_j(\mathcal{B})}{3}.\end{aligned} \tag{S31}$$

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

$$\begin{aligned}&W_1(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})},\mathbb{P}_{\mathcal{X}})\left(\sup_{\boldsymbol{z}\in\mathcal{Z}}\|\nabla\boldsymbol{G}_{\boldsymbol{\phi}}\| + 1\right)^{\frac{n}{n-d}} \\ \geq&\Gamma(\mathcal{Z},\mathcal{X},d,n,C_d,C_{\mathcal{X}});\end{aligned} \tag{S32}$$

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

$$\begin{aligned}&\frac{D_{JS}(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})},\mathbb{P}_{\mathcal{X}})(\sup_{\boldsymbol{z}\in\mathcal{Z}}\|\nabla\boldsymbol{G}_{\boldsymbol{\phi}}\| + 1)^{\frac{2n}{n-d}}}{(diam(\mathcal{Z})(\sup_{\boldsymbol{z}\in\mathcal{Z}}\|\nabla\boldsymbol{G}_{\boldsymbol{\phi}}\|) + diam(\mathcal{X}))^2} \\ \geq&4\Gamma(\mathcal{Z},\mathcal{X},d,n,C_d,C_{\mathcal{X}})^2.\end{aligned} \tag{S33}$$

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

$$\begin{aligned}&W_1(\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})},\mathbb{P}_{\mathcal{Z}})\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\|\nabla\boldsymbol{E}_{\boldsymbol{\theta}}\| + 1\right)^{\frac{n}{d-n}} \\ \geq&\Gamma(\mathcal{X},\mathcal{Z},n,d,C_n,C_{\mathcal{Z}});\end{aligned} \tag{S34}$$

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

$$\begin{aligned}&\frac{D_{JS}(\mathbb{P}_{\boldsymbol{E}_{\boldsymbol{\theta}}(\mathcal{X})}, \mathbb{P}_{\mathcal{Z}})(\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla \boldsymbol{E}_{\boldsymbol{\theta}}\| + 1)^{\frac{2n}{d-n}}}{(diam(\mathcal{X})(\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla \boldsymbol{E}_{\boldsymbol{\theta}}\|) + diam(\mathcal{Z}))^2}\\ \geq & 4\Gamma(\mathcal{X}, \mathcal{Z}, n, d, C_n, C_{\mathcal{Z}})^2,\end{aligned} \tag{S35}$$

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

Proof. Considering the Wasserstein divergence is integral over given domains, without loss of generalirity, we can assume that both $\mathcal{X}$ and $\mathcal{Z}$ are diffeomorphism to open subsets in $\mathbb{R}^d$ and $\mathbb{R}^n$. Let $\boldsymbol{c}_{\mathcal{X}}$ and $\boldsymbol{c}_{\mathcal{Z}}$ be those diffeomorphisms. They are only decided by $\mathcal{X}$ and $\mathcal{Z}$. Then $\boldsymbol{c}_{\mathcal{X}} \circ \boldsymbol{G}_{\boldsymbol{\phi}} \circ \boldsymbol{c}_{\mathcal{Z}}^{-1}$ and $\boldsymbol{c}_{\mathcal{Z}} \circ \boldsymbol{E}_{\boldsymbol{\theta}} \circ \boldsymbol{c}_{\mathcal{X}}^{-1}$ 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* $\forall \boldsymbol{x} \in \mathcal{A}, \boldsymbol{y} \in \mathcal{B}, f_{\mathcal{A}}(\boldsymbol{x}) - f_{\mathcal{B}}(\boldsymbol{y}) \leq \|\boldsymbol{x} - \boldsymbol{y}\|$*. We have*

$$W_1(\mathbb{P}_{\mathcal{A}}, \mathbb{P}_{\mathcal{B}}) = \sup_{(f_{\mathcal{A}}, f_{\mathcal{B}})\in\mathcal{Q}} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{A}}}[f_{\mathcal{A}}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{B}}}[f_{\mathcal{B}}(\boldsymbol{x})]. \tag{S36}$$

It is then easy to see that

$$\hat{f}_{\mathcal{A}}(\boldsymbol{x}) = d(\boldsymbol{x}, \mathcal{B}) = \inf_{\boldsymbol{y}\in\mathcal{B}} \|\boldsymbol{x} - \boldsymbol{y}\|,\ \hat{f}_{\mathcal{B}}(\boldsymbol{y}) = 0,\ \forall \boldsymbol{x} \in \mathcal{A},\ \boldsymbol{y} \in \mathcal{B} \tag{S37}$$

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

$$\overline{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})} = \boldsymbol{G}_{\boldsymbol{\phi}}(\overline{\mathcal{Z}}) \tag{S38}$$

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

$$\Omega_\delta = \{\boldsymbol{x} \in \mathcal{X} : d(\boldsymbol{x}, \boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})) \leq \delta\}. \tag{S39}$$

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

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

$$m_n(\Omega_\delta) \geq \omega_n \left(\frac{diam(\mathcal{Z})}{2}\right)^n \tag{S40}$$

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

$$0 < \frac{w_n diam(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)} m_n(\mathcal{X}) < \omega_n \left(\frac{diam(\mathcal{Z})}{2}\right)^n. \tag{S41}$$

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

$$m_n(\Omega_\delta) = \frac{w_n diam(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)} m_n(\mathcal{X}) < \frac{m_n(\mathcal{X})}{3}. \tag{S42}$$

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

$$\boldsymbol{G_\phi}(\mathcal{Z}) \subset \cup_{k=1}^{N_\delta} B_n\left(\boldsymbol{G_\phi}(\boldsymbol{z}_k), \sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G_\phi}\|\delta\right), \tag{S43}$$

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

$$\Omega_\delta \subset \left\{\boldsymbol{x} \in \mathcal{X} : d\left(\boldsymbol{x}, \cup_{k=1}^{N_\delta} B_n\left(\boldsymbol{G_\phi}(\boldsymbol{z}_k), \sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G_\phi}\|\delta\right)\right) \leq \delta\right\} \subset \cup_{k=1}^{N_\delta} B_n\left(\boldsymbol{G_\phi}(\boldsymbol{z}_k), \sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G_\phi}\|\delta + \delta\right). \tag{S44}$$

Thus we get

$$m_n(\Omega_\delta) = \frac{w_n diam(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)} m_n(\mathcal{X}) \leq N_\delta \omega_n \left(\sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G_\phi}\| + 1\right)^n \delta^n. \tag{S45}$$

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 $\{B_d(\boldsymbol{z}_k, \delta)\}_{k=1}^{N_\delta}$ of $\mathbb{R}^d$ such that

$$N_\delta \leq C_d \left(\frac{diam(\mathcal{Z})}{\delta}\right)^d. \tag{S46}$$

We then have

$$\begin{aligned} \frac{w_n diam(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)} m_n(\mathcal{X}) \leq C_d \left(\frac{diam(\mathcal{Z})}{\delta}\right)^d \omega_n \left(\sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G_\phi}\| + 1\right)^n \delta^n \\ \Leftrightarrow \delta \geq \left(\frac{diam(\mathcal{Z})^{n-d} m_n(\mathcal{X})}{3C_d(\sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G_\phi}\| + 1)^n (2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)}\right)^{\frac{1}{n-d}}. \end{aligned} \tag{S47}$$

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

$$m_n(\mathcal{X}') \geq \frac{2}{3} m_n(\mathcal{X}) \tag{S48}$$

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

$$p_\mathcal{X}(\boldsymbol{x}) \geq C_\mathcal{X}, \forall \boldsymbol{x} \in \mathcal{X}'. \tag{S49}$$

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

$$I_\mathcal{A}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in \mathcal{A} \\ 0. & \boldsymbol{x} \notin \mathcal{A} \end{cases}, \tag{S50}$$

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

$$\begin{aligned}
W_1(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}) = &\sup_{(f_{\mathcal{X}}, f_{\mathcal{G}_{\boldsymbol{\phi}}(\mathcal{Z})}) \in \mathcal{Q}} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}[f_{\mathcal{X}}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}}\left[f_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}(\boldsymbol{x})\right] \\
&\geq \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}\left[\hat{f}_{\mathcal{X}}(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}}\left[\hat{f}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}(\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}\left[\hat{f}_{\mathcal{X}}(\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}\left[\hat{f}_{\mathcal{X}}(\boldsymbol{x}) I_{\mathcal{X}\setminus\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}\right] + \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}\left[\hat{f}_{\mathcal{X}}(\boldsymbol{x}) I_{\mathcal{X}\cap\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}\left[\hat{f}_{\mathcal{X}}(\boldsymbol{x}) I_{\mathcal{X}\setminus\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}\right] \\
&= \int_{\mathcal{X}\setminus\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})} \hat{f}_{\mathcal{X}}(\boldsymbol{x}) p_{\mathcal{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&\geq \int_{\mathcal{X}'\setminus\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})} \hat{f}_{\mathcal{X}}(\boldsymbol{x}) p_{\mathcal{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&\geq \int_{\mathcal{X}'\setminus\Omega_\delta} \delta C_{\mathcal{X}} d\boldsymbol{x} \\
&\geq \delta C_{\mathcal{X}} \frac{m_n(\mathcal{X})}{3} \\
&\geq \left(\frac{diam(\mathcal{Z})^{n-d} m_n(\mathcal{X})}{3C_d(\sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G}_{\boldsymbol{\phi}}\| + 1)^n (2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)}\right)^{\frac{1}{n-d}} \frac{C_{\mathcal{X}} m_n(\mathcal{X})}{3}.
\end{aligned} \tag{S51}$$

Note that in the penultimate step, we use the estimation

$$m_n(\Omega_\delta) = \frac{w_n diam(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)} m_n(\mathcal{X}) \approx \frac{m_n(\mathcal{X})}{3}. \tag{S52}$$

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

$$\int_{\mathcal{X}'\setminus\Omega_\delta} \delta C_{\mathcal{X}} dx \geq \delta C_{\mathcal{X}} \left(\frac{2}{3} - \frac{w_n diam(\mathcal{Z})^n}{3(2^n m_n(\mathcal{X}) + \omega_n diam(\mathcal{Z})^n)}\right) m_n(\mathcal{X}) \tag{S53}$$

directly in this case to get more accurate estimation. Another issue is Eq. (S47). When $diam(\mathcal{Z})$ 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. (S51). As our inequality is already very complicated, we just use the current result of Eq. (S51).

C.3.4. Inequality for Jensen-Shannon divergence

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

$$W_1(\mathbb{P}, \mathbb{Q}) \leq C \cdot TV(\mathbb{P}, \mathbb{Q}), \tag{S54}$$

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

$$TV(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} D_{KL}(\mathbb{P}, \mathbb{Q})}. \tag{S55}$$

We then have

$$\begin{aligned}
D_{JS}(\mathbb{P}, \mathbb{Q}) \geq &\left(\frac{1}{2}\left(\sqrt{\frac{1}{2} D_{KL}(\mathbb{P}, \mathbb{M})} + \sqrt{\frac{1}{2} D_{KL}(\mathbb{M}, \mathbb{Q})}\right)\right)^2 \geq \frac{1}{4}(TV(\mathbb{P}, \mathbb{M}) + TV(\mathbb{M}, \mathbb{Q}))^2 \\
&\geq \frac{1}{4} TV(\mathbb{P}, \mathbb{Q})^2 \geq \frac{1}{4C^2} W_1(\mathbb{P}, \mathbb{Q})^2,
\end{aligned} \tag{S56}$$

where

$$\mathbb{M} = \frac{\mathbb{P}+\mathbb{Q}}{2}. \tag{S57}$$

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

Note that when

$$D_{JS}(\mathbb{P},\mathbb{Q}) < \log 2, \tag{S58}$$

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. (S56) into (S51) and noting that when

$$D_{JS}(\mathbb{P}_{\mathcal{X}},\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})}) < \log 2, \tag{S59}$$

we can take C as

$$C = (diam(\mathcal{X}) + diam(\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})) \leq \left(diam(\mathcal{X}) + \left(\sup_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla \boldsymbol{G}_{\boldsymbol{\phi}}\|\right) diam(\mathcal{Z})\right). \tag{S60}$$

Substituting Eq. (S60) into (S56), we then have the inequalities in Eq. (S33) & (S35) for Jensen-Shannon divergence.

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

C.4. Proof to Corollary 1

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

$$\mathbb{E}_{\boldsymbol{x}_1,...,\boldsymbol{x}_N\sim\mathbb{P}_{\mathcal{X}}}[W_1(\mathbb{P}_{\boldsymbol{G}_{\boldsymbol{\phi}}(\mathcal{Z})},\mathbb{Q}_N)] \geq \frac{C}{(\sup_{\boldsymbol{z}\in\mathcal{Z}}\|\nabla \boldsymbol{G}_{\boldsymbol{\phi}}\|+1)^{\frac{n}{n-d}}} - \mathcal{O}(N^{-\frac{1}{n}}). \tag{S61}$$

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*

$$\mathbb{E}_{\boldsymbol{x}_1,...,\boldsymbol{x}_N}[W_1(\mathbb{Q}_N,\mathbb{P}_{\mathcal{X}})] = \mathcal{O}(N^{-\frac{1}{n}}). \tag{S62}$$

□

C.5. Proof to Lemma 1

Lemma 1. *If $\boldsymbol{h}$ is continuous and $\mathcal{D}$ is connected, then $\boldsymbol{h}(\mathcal{D})$ 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 $\boldsymbol{G}$, when the discriminator $\boldsymbol{D}$ is optimal, we have*

$$V(\boldsymbol{D},\boldsymbol{G}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}[\log(\boldsymbol{D}(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\mathcal{Z}}}[\log(1-\boldsymbol{D}(\boldsymbol{G}(\boldsymbol{z})))] \tag{S63}$$

and

$$D_{JS}(\mathbb{P}_{\boldsymbol{G}},\mathbb{P}_{data}) = \log 2 + \frac{1}{2}V(\boldsymbol{D},\boldsymbol{G}). \tag{S64}$$

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

C.7. Proof to Lemma 3

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

$$m_d(\boldsymbol{h}(\mathcal{D})) \leq \left(\sup_{\boldsymbol{x}\in\mathcal{D}} \|\nabla \boldsymbol{h}\|\right)^d m_d(\mathcal{D}), \tag{S65}$$

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

$$\mathcal{D} \subset \cup_k B_k,\ m_d(\mathcal{D}) \leq \sum_k r_k^d \leq m_d(\mathcal{D}) + \epsilon. \tag{S66}$$

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

$$m_d(\boldsymbol{h}(\mathcal{D})) \leq \left(\sup_{\boldsymbol{x}\in\mathcal{D}} \|\nabla \boldsymbol{h}\|\right)^d \sum_k r_k^d \leq \left(\sup_{\boldsymbol{x}\in\mathcal{D}} \|\nabla \boldsymbol{h}\|\right)^d m_d(\mathcal{D}) + \left(\sup_{\boldsymbol{x}\in\mathcal{D}} \|\nabla \boldsymbol{h}\|\right)^d \epsilon. \tag{S67}$$

Let $\epsilon \to 0$. 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.

Choi, K., Hawthorne, C., Simon, I., Dinculescu, M., and Engel, J. Encoding musical style with transformer autoencoders. In *International Conference on Machine Learning*, pp. 1899–1908. PMLR, 2020.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial Feature Learning, 2017.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. Adversarially Learned Inference, 2017.

Gallot, S., Hulin, D., and Lafontaine, J. *Riemannian geometry*, volume 2. Springer, 1990.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Hayakawa, Y., Marumoto, A., and Sawada, Y. Effects of the chaotic noise on the performance of a neural network model for optimization problems. *Physical review E*, 51(4):R2693, 1995.

He, Z., Rakin, A. S., and Fan, D. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 588–597, 2019.

Hinton, G. E. and Zemel, R. S. Autoencoders, minimum description length and Helmholtz free energy. In *Advances in neural information processing systems*, pp. 3–10, 1994.

Jenni, S. and Favaro, P. On stabilizing generative adversarial training with noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12145–12153, 2019.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Ng, A. et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Rogers, C. A note on coverings. *Mathematika*, 4(1):1–6, 1957.

Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Stein, E. M. and Shakarchi, R. *Real analysis: Measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Vahdat, A. and Kautz, J. NVAE: A deep hierarchical variational autoencoder, 2021.

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

Zhu, J., Zhao, D., Zhang, B., and Zhou, B. Disentangled inference for GANs with latently invertible autoencoder. *arXiv:1906.08090v3*, 2019.