Engineering 6 (2020) 361–374

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Article

A Geometric Understanding of Deep Learning

Na Lei [a,#], Dongsheng An [b,#], Yang Guo [b], Kehua Su [c], Shixia Liu [d], Zhongxuan Luo [a], Shing-Tung Yau [e], Xianfeng Gu [b,e,*]

[a] DUT-RU Co-Research Center of Advanced ICT for Active Life, Dalian University of Technology, Dalian 116620, China
[b] Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-2424, USA
[c] School of Computer Science, Wuhan University, Wuhan 430072, China
[d] School of Software, Tsinghua University, Beijing 100084, China
[e] Center of Mathematical Sciences and Applications, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 2 March 2019
Revised 31 August 2019
Accepted 11 September 2019
Available online 11 January 2020

Keywords:
Generative
Adversarial
Deep learning
Optimal transportation
Mode collapse

ABSTRACT

This work introduces an optimal transportation (OT) view of generative adversarial networks (GANs). Natural datasets have intrinsic patterns, which can be summarized as the manifold distribution principle: the distribution of a class of data is close to a low-dimensional manifold. GANs mainly accomplish two tasks: manifold learning and probability distribution transformation. The latter can be carried out using the classical OT method. From the OT perspective, the generator computes the OT map, while the discriminator computes the Wasserstein distance between the generated data distribution and the real data distribution; both can be reduced to a convex geometric optimization process. Furthermore, OT theory discovers the intrinsic collaborative—instead of competitive—relation between the generator and the discriminator, and the fundamental reason for mode collapse. We also propose a novel generative model, which uses an autoencoder (AE) for manifold learning and OT map for probability distribution transformation. This AE–OT model improves the theoretical rigor and transparency, as well as the computational stability and efficiency; in particular, it eliminates the mode collapse. The experimental results validate our hypothesis, and demonstrate the advantages of our proposed model.

1. Introduction

Generative adversarial networks (GANs) have emerged as one of the dominant approaches for unconditional image generation. When trained on several datasets, GANs are able to produce realistic and visually appealing samples. GAN methods train an unconditional generator that regresses real images from random noises and a discriminator that measures the difference between the generated samples and real images. GANs have received various improvements. One breakthrough was achieved by combing optimal transportation (OT) theory with GANs, such as the Wasserstein GAN (WGAN) [1]. In the WGAN framework, the generator computes the OT map from the white noise to the data distribution, and the discriminator computes the Wasserstein distance between the generated data distribution and the real data distribution.

1.1. Manifold distribution hypothesis

The great success of GANs can be explained by the fact that GANs effectively discover the intrinsic structures of real datasets, which can be formulated as the manifold distribution hypothesis: A specific class of natural data is concentrated on a low-dimensional manifold embedded in the high-dimensional background space [2].

Fig. 1 shows the manifold structure of the MNIST database. Each handwritten digit image has the dimensions 28 × 28, and is treated as a point in the image space $\mathbb{R}^{784}$. The MNIST database is concentrated close to a low-dimensional manifold. By using the t-SNE manifold embedding algorithm [3], the MNIST database is mapped onto a planar domain, and each image is mapped onto a single point. The images representing the same digit are mapped onto one cluster, and 10 clusters are color encoded. This demonstrates that the MNIST database is distributed close to a two-dimensional (2D) surface embedded in the unit cube in $\mathbb{R}^{784}$.

* Corresponding author.
E-mail address: gu@cs.stonybrook.edu (X. Gu).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.eng.2019.09.010

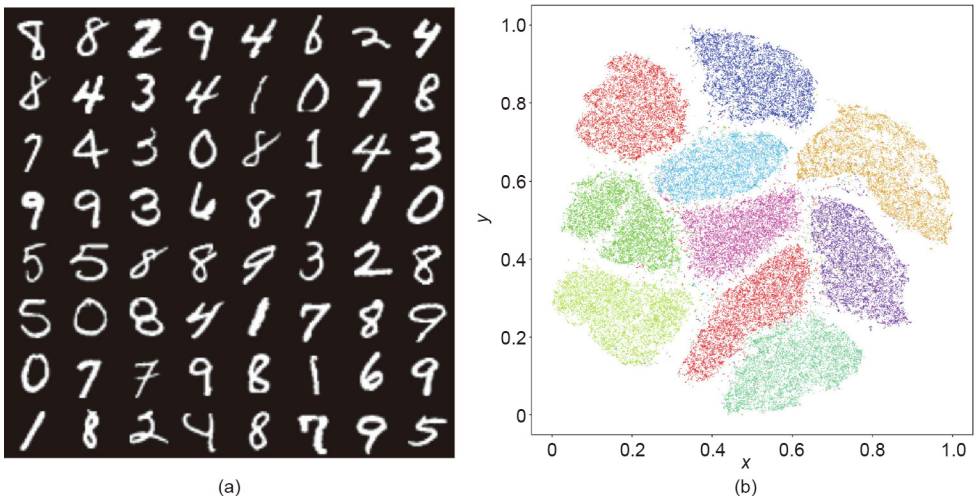

Fig. 1. Manifold distribution of the MNIST database. (a) Some handwritten digitals in MNIST database; (b) the embedded result of the digitals in two-dimensional (2D) plane by t-SNE algorithm. The x and y relative coordinates are normalized.

1.2. Theoretic model of GANs

Fig. 2 illustrates the theoretic model of GANs. The real data distribution v is concentrated on a manifold Σ embedded in the ambient space χ. (Σ, v) together show the intrinsic structure of the real datasets. A GAN model computes a generator map g_θ from the latent space Z to the manifold Σ, where θ represents the parameter of a deep neural network (DNN). ζ is a Gaussian distribution in the latent space, and g_θ pushes forward ζ to μ_θ. The discriminator calculates a distance between the real data distribution v and the generated distribution μ_θ, such as the Wasserstein distance $W_c(\mu_\theta, v)$, which is equivalent to the Kontarovich's potential φ_ξ (ξ: the parameter of the discriminator).

Despite GANs' advantages, they have critical drawbacks. In theory, the understanding of the fundamental principles of deep learning remains primitive. In practice, the training of GANs is tricky and sensitive to hyperparameters; GANs suffer from mode collapsing. Recently, Mescheder et al. [4] studied nine different GAN models and variants showing that gradient-descent-based GAN optimization is not always locally convergent.

According to the manifold distribution hypothesis, a natural dataset can be represented as a probability distribution on a manifold. Therefore, GANs mainly accomplish two tasks: ① manifold learning—namely, computing the decoding/encoding maps between the latent space and the ambient space; and ② probability distribution transformation, either in the latent or image space, which involves transformation between the given white noise and the data distribution.

Fig. 3 shows the decomposition of the generator map $g_\theta = h \circ T$, where h: $Z \to \Sigma$ is the decoding map from the latent space to the data manifold Σ in the ambient space, the probability distribution

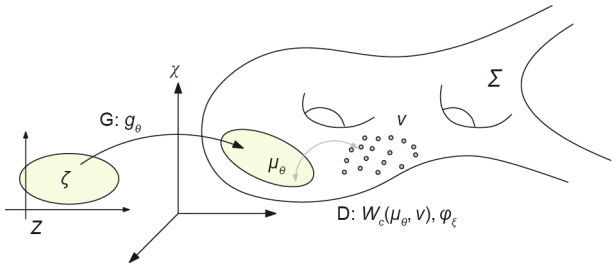

Fig. 2. The theoretic model of GANs. G: generator; D: discriminator.

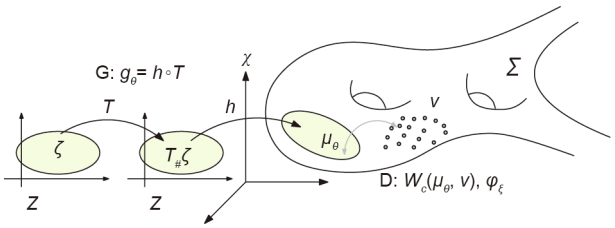

Fig. 3. The generator map is decomposed into a decoding map h and a transportation map T. $T_\#\zeta$ is the push-forward measure induced by T.

transformation map T: $Z \to Z$. The decoding map h is for manifold learning, and the map T is for measure transportation.

1.3. Optimal transportation view

OT theory [5] studies the problem of transforming one probability distribution into another distribution in the most economical way. OT provides rigorous and powerful ways to compute the optimal mapping to transform one probability distribution into another distribution, and to determine the distance between them [6].

As mentioned before, GANs accomplish two major tasks: manifold learning and probability distribution transformation. The latter task can be fully carried out by OT methods directly. In detail, in Fig. 3, the probability distribution transformation map T can be computed using OT theory. The discriminator computes the Wasserstein distance $W_c(\mu_\theta, v)$ between the generated data distribution and the real data distribution, which can be calculated directly using the OT method.

From the theoretical point of view, the OT interpretation of GANs makes part of the black box transparent, the probability distribution transformation is reduced to a convex optimization process using OT theory, the existence and uniqueness of the solution have theoretical guarantees, and the convergence rate and approximation accuracy are fully analyzed.

The OT interpretation also explains the fundamental reason for mode collapse. According to the regularity theory of the Monge–Ampère equation, the transportation map is discontinuous on some singular sets. However, DNN can only model continuous functions/mappings. Therefore, the target transportation mapping is outside of the functional space representable by GANs. This intrinsic conflict makes mode collapses unavoidable.

The OT interpretation also reveals a more complicated relation between the generator and the discriminator. In the current GAN models, the generator and the discriminator compete with each other without sharing the intermediate computational results. The OT theory shows that under the L^2 cost function, the optimal solution of the generator and the optimal solution of the discriminator can be expressed by each other in a closed form. Therefore, the competition between the generator and the discriminator should be replaced by collaboration, and the intermediate computation results should be shared to improve the efficiency.

1.4. Autoencoder–optimal transportation model

In order to reduce the training difficulty of GANs and, in particular, to avoid mode collapses, we propose a simpler generative model based on OT theory: an autoencoder (AE)–OT model, as shown in Fig. 4.

As mentioned before, the two major tasks for generative models are manifold learning and probability distribution transformation. The AE computes the encoding map, f_θ: $Z \to \Sigma$, and the decoding map g_ξ: $\Sigma \to Z$, for the purpose of manifold learning. The OT map, T: $Z \to Z$, transforms the white noise ζ to the data distribution push-forwarded by the encoding map, $(f_\theta)_\# v$.

The AE–OT model has many merits. From a theoretical standpoint, the OT theory has been well established and is fully understood. By decoupling the decoding map and the OT map, it is possible to improve the theoretical rigor of generative models and make part of the black box transparent. In practice, the OT map is reduced to a convex optimization problem, the existence and the uniqueness of the solution are guaranteed, and the training process will not be trapped in the local optimum. The convex energy associated with the OT map has an explicit Hessian matrix; hence, the optimization can be performed using Newton's method with second-order convergence, or using quasi-Newton's method with superlinear convergence. In contrast, current generative models are based on the gradient descent method with linear convergence; the number of unknowns is equal to that of the training samples, in order to avoid the over-paramerization problem; the error bound of the OT map can be fully controlled by the sampling density in the Monte Carlo method; the hierarchical algorithm with self-adaptivity further improves the efficiency; and the parallel OT map algorithm can be implemented using a graphics processing unit (GPU). Most importantly, the AE–OT model can eliminate mode collapse.

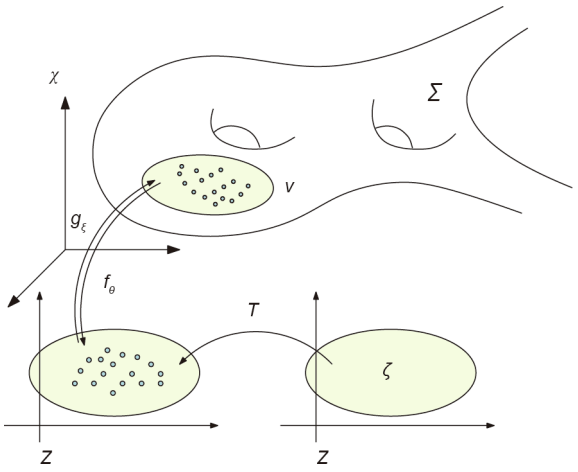

Fig. 4. A generative model, the AE–OT model, combining an AE and an OT map.

1.5. Contributions

This work interprets the GAN model using OT theory. GANs can accomplish two major tasks: manifold learning and probability distribution transformation. The latter task can be carried out using OT methods. The generator computes an OT map, and the discriminator calculates the Wasserstein distance between the generated data distribution and the real data distribution. Using Brenier's theory, the competition between the generator and the discriminator can be replaced by collaboration; according to the regularity theory of the Monge–Ampère equation, the discontinuity of the transportation map causes mode collapse. We further propose to decouple manifold learning and probability distribution transformation by means of an AE–OT model, which makes part of the black box transparent, improves the training efficiency, and prevents mode collapse. The experimental results demonstrate the efficiency and efficacy of our method.

This paper is organized as follows: Section 2 briefly reviews the most related works in OT and GANs; Section 3 briefly introduces the fundamental theories in OT and the regularity theory of the Monge–Ampère equation; Section 4 introduces a variational framework for computing OT, which is suitable for a deep learning setting; Section 5 analyzes the GAN model from the OT perspective, explains the collaborative (instead of competitive) relation between the generator and the discriminator, and reveals the intrinsic reason for mode collapse; Section 6 reports the experimental results; and the paper concludes in Section 7.

2. Previous works

2.1. Optimal transportation

The OT problem plays an important role in various kinds of fields. For detailed overviews, we refer readers to Refs. [7] and [8].

When both the input and output domains are Dirac masses, the OT problem can be treated as a standard linear programming (LP) task. In order to extend the problem to a large dataset, the authors of Ref. [9] added an entropic regularizer into the original LP problem; as a result, the regularized problem could be quickly computed with the Sinkhorn algorithm. Solomon et al. [10] then improved the computational efficiency by the introduction of fast convolution.

The second type of method for solving the OT problem computes the OT map between continuous and point-wise measures by minimizing a convex energy [6] through the connection between the OT problem and convex geometry. In Ref. [11], the authors then linked the convex-geometry-viewed OT to the Kantorovich duality by means of the Legendre dual theory. The proposed method is an extension of this method in high dimension. If both the input and output are continuous densities, solving the OT problem is equivalent to solving the famous Monge–Ampère equation, which is a highly nonlinear elliptic partial differential equation (PDE). With an additional virtual time dimension, this problem can be relaxed through computational fluid dynamics [12–14].

2.2. Generative models

In the machine learning field, generative models, which are capable of generating complex and high-dimensional data, are recently becoming increasingly important and popular. To be specific, generative models are largely utilized to generate new images from given image datasets. Several methods, including deep belief networks [15] and deep Boltzmann machines [16], have been introduced in early stages. However, the training in these

methods is generally tricky and inefficient. Later, a huge breakthrough was achieved from the scheme of variational AEs (VAEs) [17], where the decoders approximate real data distributions from a Gaussian distribution using a variational approach [17,18]. Various recent works following this scheme have been proposed, including adversarial AEs (AAEs) [19] and Wasserstein AEs (WAEs) [20]. Although VAEs are relatively simple to train, the images they generate look blurry. To some extent, this is because the explicitly expressed density functions may fail to represent the complexity of a real data distribution and learn the high-dimensional data distribution [21,22]. Other non-adversarial training models have been proposed, including PixelCNN [23], PixelRNN [24], and WaveNet [25]. However, due to their auto-regressive nature, the generation of new samples cannot be paralleled.

2.3. Adversarial generative models

GANs [26] were proposed to solve the disadvantages of the above models. Although they are a powerful tool for generating realistic-looking samples, GANs can be difficult to train and suffer from mode collapsing. Various improvements have been proposed for better GAN training, including changing the loss function (e.g., the WGAN [1]) and regularizing the discriminators to be Lipschitz by clipping [1], gradient regularization [4,27], or spectral normalization [28]. However, the training of GANs is still tricky and requires careful hyperparameter selection.

2.4. Evaluation of generative models

The evaluation of generative models remains challenging. Early works include probabilistic criteria [29]. However, recent generative models (particularly GANs) are not amenable to such evaluation. Traditionally, the evaluation of GANs relies on visual inspection of a handful of examples or a user study. Recently, several quantitative evaluation criteria were proposed. The inception score (IS) [30] measures both diversity and image quality. However, it is not a distance metric. To overcome the shortcomings of the IS, the Fréchet inception distance (FID) was introduced in Ref. [31]. The FID has been shown to be robust to image corruption, and correlates well with visual fidelity. In a more recent work [32], precision and recall for distributions (PRD) was introduced to measure both precision and recall between generated data distribution and real data distribution. In order to fairly compare the GANs, a large-scale comparison was performed in Ref. [33], where seven different GANs and VAEs were compared under a uniform network architecture, and a common baseline for evaluation was established.

2.5. Non-adversarial models

Various non-adversarial models have also been proposed recently. Generative latent optimization (GLO) [34] employs an "encoder-less AE" approach in which a generative model is trained with a non-adversarial loss function, achieving better results than VAEs. Implicit maximum likelihood estimation (IMLE) [35] proposed an iterative closest points (ICP)-related generative model training approach. Later, Hoshen and Malik [36] proposed generative latent nearest neighbors (GLANN), which combines the advantages of GLO and GLANN, in which an embedding from the image space to latent space was first found using GLO, and then a transformation between an arbitrary distribution and latent code was computed using IMLE.

Other methods directly approximate the distribution transformation map from the noise space to the image space by means of DNNs with a controllable Jacobian matrix [37–39]. Recently, the energy-based models [40–42] have been chosen to model the image distribution through the Gibbs distribution by representing the energy function with DNNs. These methods alternatively generate fake samples using the current models, and then optimize the model parameters with the generated fake samples and real samples.

3. Optimal transportation theory

In this section, we introduce basic concepts and theorems in classic OT theory, with a focus on Brenier's approach and their generalization to the discrete setting. Details can be found in Villani's book [5].

3.1. Monge's problem

Suppose $X \subset \mathbb{R}^d, Y \subset \mathbb{R}^d$ are two subsets of d-dimensional Euclidean space $\mathbb{R}^d$, and μ and ν are two probability measures defined on X and Y, respectively, with the following density functions:

$$\mu(\boldsymbol{x}) = f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$\nu(\boldsymbol{y}) = g(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$$

Suppose the total measures are equal, $\mu(X) = \nu(Y)$; that is

$$\int_X f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_Y g(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \tag{1}$$

We only consider maps that preserve the measures.

Definition 3.1 (measure-preserving map). A map $T: X \to Y$ is measure preserving if for any measurable set $B \subset Y$, the set $T^{-1}(B)$ is μ-measurable and $\mu[T^{-1}(B)] = \nu(B)$, that is,

$$\int_{T^{-1}(B)} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_B g(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \tag{2}$$

The measure-preserving condition is denoted as $T_{\#}\mu = \nu$, where $T_{\#}\mu$ is the push-forward measure induced by T.

Given a cost function $c(\boldsymbol{x},\boldsymbol{y})$: $X \times Y \to \mathbb{R}_{\geq 0}$, which indicates the cost of moving each unit mass from the source to the target, the total transport cost (C_t) of the map T: $X \to Y$ is defined to be

$$C_t = \int_X c[\boldsymbol{x}, T(\boldsymbol{x})]\mathrm{d}\mu(\boldsymbol{x}) \tag{3}$$

Monge's problem of OT arises from finding the measure-preserving map that minimizes the total transport cost.

Problem 3.2 (Monge's [43]; MP). Given a transport cost function $c(\boldsymbol{x},\boldsymbol{y})$: $X \times Y \to \mathbb{R}_{\geq 0}$, find the measure-preserving map T: $X \to Y$ that minimizes the total transport cost:

$$(\mathrm{MP}) \quad \min_{T_{\#}\mu=\nu} \int_X c[\boldsymbol{x}, T(\boldsymbol{x})]\mathrm{d}\mu(\boldsymbol{x}) \tag{4}$$

Definition 3.3 (OT map). The solutions to Monge's problem are called the OT maps. The total transportation cost of an OT map is called the Wasserstein distance between μ and ν, denoted as $W_c(\mu, \nu)$.

$$W_c(\mu, \nu) = \min_{T_{\#}\mu=\nu} \int_X c[\boldsymbol{x}, T(\boldsymbol{x})]\mathrm{d}\mu(\boldsymbol{x}) \tag{5}$$

3.2. Kontarovich's approach

Depending on the cost function and measures, an OT map between (X, μ) and (Y, ν) may not exist. Kontarovich relaxed the transportation maps to transportation plans, and defined the joint probability measure $\rho(\boldsymbol{x}, \boldsymbol{y})$: $X \times Y \to \mathbb{R}_{\geq 0}$, such that the marginal

probability of ρ is equal to μ and v, respectively. Let the projection maps formally be $\pi_x(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}$, $\pi_y(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y}$, then define the joint measure class as follows:

$$\Pi(\mu, v) = \left\{\rho(\boldsymbol{x},\boldsymbol{y}) : X \times Y \to \mathbb{R} : (\pi_x)_{\#}\rho = \mu, (\pi_y)_{\#}\rho = v\right\} \tag{6}$$

Problem 3.4 (Kontarovich's; KP). Given a transport cost function $c(\boldsymbol{x},\boldsymbol{y})$: $X \times Y \to \mathbb{R}_{\geq 0}$, find the joint probability measure $\rho(\boldsymbol{x}, \boldsymbol{y})$: $X \times Y \to \mathbb{R}_{\geq 0}$ that minimizes the total transport cost.

$$\text{(KP) } W_c(\mu, v) = \min_{\rho\in\Pi(\mu,v)} \int_{X\times Y} c(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\rho(\boldsymbol{x},\boldsymbol{y}) \tag{7}$$

KP can be solved using the LP method. Due to the duality of LP, Eq. (7) (the KP equation) can be reformulated as the duality problem (DP) as follows:

Problem 3.5 (duality; DP). Given a transport cost function $c(\boldsymbol{x},\boldsymbol{y})$: $X \times Y \to \mathbb{R}_{\geq 0}$, find the real functions φ: $X \to \mathbb{R}$ and ψ: $Y \to \mathbb{R}$, such that

$$\text{(DP) } \max_{\varphi,\psi}\left[\int_X \varphi(\boldsymbol{x})\mathrm{d}\mu + \int_Y \psi(\boldsymbol{y})\mathrm{d}v : \varphi(\boldsymbol{x}) + \psi(\boldsymbol{y}) \leq c(\boldsymbol{x},\boldsymbol{y})\right] \tag{8}$$

The maximum value of Eq. (8) gives the Wasserstein distance. Most existing WGAN models are based on the duality formulation under the L^1 cost function.

Definition 3.6 (*c*-transformation). The *c*-transformation of φ: $X \to \mathbb{R}$ is defined as φ^c: $Y \to \mathbb{R}$:

$$\varphi^c(\boldsymbol{y}) = \inf_{x\in X}[c(\boldsymbol{x},\boldsymbol{y}) - \varphi(\boldsymbol{x})] \tag{9}$$

The DP can then be rewritten as follows:

$$\text{(DP) } W_c(\mu, v) = \max_{\varphi} \int_X \varphi(\boldsymbol{x})\mathrm{d}\mu + \int_Y \varphi^c(\boldsymbol{y})\mathrm{d}v \tag{10}$$

3.3. Brenier's approach

For the quadratic Euclidean distance cost, the existence, uniqueness, and intrinsic structure of the OT map were proven by Brenier [44].

Theorem 3.7 (Brenier's [44]). Suppose X and Y are subsets of the Euclidean space $\mathbb{R}^d$ and the transportation cost is the quadratic Euclidean distance $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\| \boldsymbol{x} - \boldsymbol{y} \|^2$. Furthermore, μ is absolutely continuous and μ and v have finite second-order moments

$$\int_X \| \boldsymbol{x} \|^2 \mathrm{d}\mu(\boldsymbol{x}) + \int_Y \| \boldsymbol{y} \|^2 \mathrm{d}v(\boldsymbol{y}) < \infty \tag{11}$$

then there exists a convex function u: $X \to \mathbb{R}$, the so-called Brenier's potential, whose gradient map ∇u gives the solution to MP:

$$(\nabla u)_{\#}\mu = v \tag{12}$$

The Brenier's potential is unique up to a constant; hence, the optimal mass transportation map is unique.

Assuming that the Briener potential is C^2 smooth, then it is the solution to the following Monge–Ampère equation:

$$\det\left(\frac{\partial^2 u(\boldsymbol{x})}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_j}\right) = \frac{f(\boldsymbol{x})}{g\circ\nabla u(\boldsymbol{x})} \tag{13}$$

For the L^2 transportation cost $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\| \boldsymbol{x} - \boldsymbol{y} \|^2$ in $\mathbb{R}^d$, the *c*-transform and the classical Legendre transform have special relations.

Definition 3.8 (Legendre transform). Given a function φ: $\mathbb{R}^n \to \mathbb{R}$, its Legendre transform is defined as follows:

$$\varphi^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}}[\langle \boldsymbol{x},\boldsymbol{y}\rangle - \varphi(\boldsymbol{x})] \tag{14}$$

It can be shown that the following relation holds when $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\| \boldsymbol{x} - \boldsymbol{y} \|^2$:

$$\frac{1}{2}\| \boldsymbol{y} \|^2 - \varphi^c(\boldsymbol{y}) = \left[\frac{1}{2}\| \boldsymbol{x} \|^2 - \varphi(\boldsymbol{x})\right]^* \tag{15}$$

Theorem 3.9 (Brenier's polar factorization [44]). Suppose X and Y are the Euclidean space $\mathbb{R}^d$, μ is absolutely continuous with respect to the Lebesgue measure, and a mapping φ: $X \to Y$ pushes μ forward to v, $\varphi_{\#}\mu = v$, then there exists a convex function u: $X \to \mathbb{R}$, such that $\varphi = \nabla u \circ s$, where s: $X \to X$ is measure preserving, $s_{\#}\mu = \mu$. Furthermore, this factorization is unique.

The following theorem is well known in OT theory:

Theorem 3.10 (Villani [5]). Given μ and v on a compact convex domain $\Omega \subset \mathbb{R}^d$, there exists an OT plan ρ for the cost $c(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x} - \boldsymbol{y})$, with h strictly convex. It is unique and of the form $(\mathrm{id}, T_{\#})\mu$ (id: identity map), provided that μ is absolutely continuous and $\partial\Omega$ is negligible. Moreover, there exists a Kantorovich's potential φ, and T can be represented as follows:

$$T(\boldsymbol{x}) = \boldsymbol{x} - (\nabla \boldsymbol{h})^{-1}[\nabla\varphi(\boldsymbol{x})]$$

When $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\| \boldsymbol{x} - \boldsymbol{y} \|^2$, we have

$$T(\boldsymbol{x}) = \boldsymbol{x} - \nabla\varphi(\boldsymbol{x}) = \nabla\left[\frac{1}{2}\| \boldsymbol{x} \|^2 - \varphi(\boldsymbol{x})\right] = \nabla u(\boldsymbol{x})$$

In this case, the Brenier's potential u and the Kantorovich's potential φ are related by the following:

$$u(\boldsymbol{x}) = \frac{1}{2}\| \boldsymbol{x} \|^2 - \varphi(\boldsymbol{x}) \tag{16}$$

3.4. Regularity of OT maps

Let Ω and Λ be two bounded smooth open sets in $\mathbb{R}^d$, and let $\mu = f\mathrm{d}\boldsymbol{x}$ and $v = g\mathrm{d}\boldsymbol{y}$ be two probability measures on $\mathbb{R}^d$ such that $f|_{\mathbb{R}^d\setminus\Omega} = 0$ and $g|_{\mathbb{R}^d\setminus\Lambda} = 0$. Assume that f and g are bounded away from zero and infinity on Ω and Λ, respectively.

3.4.1. Convex target domain

Definition 3.11 (Hölder continuous). A real or complex-valued function f on a d-dimensional Euclidean space satisfies a Hölder condition, or is Hölder continuous, when there are nonnegative real constants C, $\alpha > 0$, such that $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq C\| \boldsymbol{x} - \boldsymbol{y} \|^{\alpha}$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in the domain of f.

Definition 3.12 (Hölder space). The Hölder space $C^{k,\alpha}(\Omega)$, where Ω is an open subset of some Euclidean space and $k \geq 0$ is an integer, consists of those functions on Ω having continuous derivatives up to order k and such that the kth partial derivatives are Hölder continuous with exponent α, where $0 < \alpha \leq 1$. $C^{k,\alpha}_{\mathrm{loc}}(\Omega)$ means the above conditions hold on any compact subset of Ω.

Theorem 3.13 (Caffarelli [45]). If Λ is convex, then the Brenier's potential u is strictly convex; furthermore,

(1) If $\lambda \leq f, g \leq 1/\lambda$ for some $\lambda > 0$, then $u \in C^{1,\alpha}_{\mathrm{loc}}(\Omega)$.

(2) If $f \in C^{k,\alpha}_{\mathrm{loc}}(\Omega)$ and $g \in C^{k,\alpha}_{\mathrm{loc}}(\Lambda)$, with $f, g > 0$, then $u \in C^{k+2,\alpha}_{\mathrm{loc}}(\Omega)$ and $(k \geq 0, \alpha \in (0.1))$.

3.4.2. Non-convex target domain

If Λ is not convex and there exist f and g that are smooth such that $u \notin C^1(\Omega)$, then the OT map ∇u is discontinuous at singularities.

Definition 3.14 (subgradient). Given an open set $\Omega \subset \mathbb{R}^d$ and a convex function u: $\Omega \to \mathbb{R}$, for $\boldsymbol{x} \in \Omega$, the subgradient (subdifferential) of u at $\boldsymbol{x}$ is defined as follows:

$$\partial u(\boldsymbol{x}) = \{\boldsymbol{p} \in \mathbb{R}^n : u(\boldsymbol{z}) \geq u(\boldsymbol{x}) + \langle \boldsymbol{p}, \boldsymbol{z} - \boldsymbol{x} \rangle, \ \forall \boldsymbol{z} \in \Omega\}$$

It is obvious that $u(\boldsymbol{x})$ is a closed convex set. Geometrically, if $\boldsymbol{p} \in u(\boldsymbol{x})$, then the hyperplane $l_{\boldsymbol{x},\boldsymbol{p}}(\boldsymbol{z}) = u(\boldsymbol{x}) + \langle \boldsymbol{p}, \boldsymbol{z} - \boldsymbol{x} \rangle$ touches u from below at $\boldsymbol{x}$; that is, $l_{\boldsymbol{x},\boldsymbol{p}} \leq u$ in Ω and $l_{\boldsymbol{x},\boldsymbol{p}}(\boldsymbol{x}) = u(\boldsymbol{x})$, where $l_{\boldsymbol{x},\boldsymbol{p}}$ is a supporting plane to u at $\boldsymbol{x}$.

The Brenier's potential u is differentiable at $\boldsymbol{x}$ if its subgradient $\partial u(\boldsymbol{x})$ is a singleton. We classify the points according to the dimensions of their subgradients, and define the sets $\Sigma_k(u) = \{\boldsymbol{x} \in \mathbb{R}^d | \dim[\partial u(\boldsymbol{x})] = k\}, \ k = 0, 1, 2, ..., d$

It is obvious that $\Sigma_0(u)$ is the set of regular points, and $\Sigma_k(u)$, where $k > 0$, is the set of singular points. We also define the reachable subgradients at $\boldsymbol{x}$ as follows:

$$\nabla_* u(\boldsymbol{x}) = \left\{ \lim_{k \to \infty} \nabla u(\boldsymbol{x}_k) | \boldsymbol{x}_k \in \Sigma_0, \boldsymbol{x}_k \to \boldsymbol{x} \right\}$$

It is well known that the subgradient is equal to the convex hull of the reachable subgradient:

$$\partial u(\boldsymbol{x}) = \text{Convex hull } \nabla_* u(\boldsymbol{x})$$

Theorem 3.15 (regularity). Let Ω, $\Lambda \subset \mathbb{R}^d$ be two bounded open sets, let f, g: $\mathbb{R}^d \to \mathbb{R}^+$ be two probability densities that are zero outside Ω and Λ, and are bounded away from zero and infinity on Ω and Λ, respectively. Denote by $T = \nabla u$: $\Omega \to \Lambda$ the OT map provided by Theorem 3.7. Then there exist two relatively closed sets $\Sigma_\Omega \subset \Omega$ and $\Sigma_\Lambda \subset \Lambda$ with $\Sigma_\Omega = \Sigma_\Lambda = 0$ such that T: $\Omega \backslash \Sigma_\Omega \to \Lambda \backslash \Sigma_\Lambda$ is a homeomorphism of class $C^{0,\alpha}_{\text{loc}}$ for some $\alpha > 0$.

We call Σ_Ω the singular set of the OT map ∇u: $\Omega \to \Lambda$. Fig. 5 illustrates the singularity set structure, computed using the algorithm based on Theorem 4.2. We obtain the following:

$$\Sigma_0 = \boldsymbol{\Omega}\{\{\Sigma_1 \cup \Sigma_2\}, \ \Sigma_1 = \bigcup_{k=0}^{3} \gamma_k, \ \Sigma_2 = \{\boldsymbol{x}_0, \boldsymbol{x}_1\}$$

The subgradient of $\boldsymbol{x}_0$, $\partial u(\boldsymbol{x}_0)$, is the entire inner hole of Λ, while $\partial u(\boldsymbol{x}_1)$ is the shaded triangle. For each point on $\gamma_k(t)$, $\partial u[\gamma_k(t)]$ is a line segment outside Λ. $\boldsymbol{x}_1$ is the bifurcation point of γ_1, γ_2, and γ_3. The Brenier's potential on Σ_1 and Σ_2 is not differentiable, and the OT map ∇u on them is discontinuous.

4. Computational algorithm

Brenier's theorem can be directly generalized to the discrete situation. In GAN models, the source measure μ is given as a uniform (or Gaussian) distribution defined on a compact convex domain Ω; the target measure ν is represented as the empirical measure, which is the sum of the Dirac measures:

$$\nu = \sum_{i=1}^{n} \nu_i \delta(\boldsymbol{y} - \boldsymbol{y}_i) \tag{17}$$

where $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n\}$ are training samples, with the weights $\sum_{i=1}^{n} \nu_i = \mu(\boldsymbol{\Omega})$; δ is the characteristic function.

Each training sample $\boldsymbol{y}_i$ corresponds to a supporting plane of the Brenier's potential, denoted as follows:

$$\pi_{\boldsymbol{h},i}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y}_i \rangle + \boldsymbol{h}_i \tag{18}$$

where the height h_i is an unknown variable. We represent all the height variables as $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n)$.

An envelope of a family of hyper-planes in the Euclidean space is a hypersurface that is tangential to each member of the family at some point, and these points of tangency together form the whole

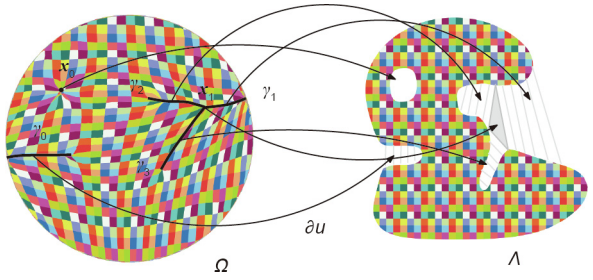

Fig. 5. Singularity structure of an OT map.

envelope. As shown in Fig. 6, the Brenier's potential $u_{\boldsymbol{h}}$: $\Omega \to \mathbb{R}$ is a piecewise linear convex function determined by $\boldsymbol{h}$, which is the upper envelope of all its supporting planes:

$$u_{\boldsymbol{h}}(\boldsymbol{x}) = \max_{i=1}^{n} [\pi_{\boldsymbol{h},i}(\boldsymbol{x})] = \max_{i=1}^{n} [\langle \boldsymbol{x}, \boldsymbol{y}_i \rangle + \boldsymbol{h}_i] \tag{19}$$

The graph of the Brenier's potential is a convex polytope. Each supporting plane $\pi_{\boldsymbol{h},i}$ corresponds to a facet of the polytope. The projection of the polytope induces a cell decomposition of Ω, where each supporting plane $\pi_i(\boldsymbol{x})$ projects onto a cell $W_i(\boldsymbol{h})$, $\boldsymbol{p}$ is any point in $\mathbb{R}^d$:

$$\boldsymbol{\Omega} = \bigcup_{i=1}^{n} W_i(\boldsymbol{h}) \cap \boldsymbol{\Omega}, \ W_i(\boldsymbol{h}) = \{\boldsymbol{p} \in \mathbb{R}^d | \nabla u_{\boldsymbol{h}}(\boldsymbol{p}) = \boldsymbol{y}_i\} \tag{20}$$

The cell decomposition is a power diagram. The μ-measure of $W_i \cap \Omega$ is denoted as $w_i(\boldsymbol{h})$:

$$w_i(\boldsymbol{h}) = \mu[W_i(\boldsymbol{h}) \cap \Omega] = \int_{W_i(\boldsymbol{h}) \cap \Omega} \mathrm{d}\mu \tag{21}$$

The gradient map $\nabla u_{\boldsymbol{h}}$: $\Omega \to Y$ maps each cell $W_i(\boldsymbol{h})$ to a single point $\boldsymbol{y}_i$:

$$\nabla u_{\boldsymbol{h}}: W_i(\boldsymbol{h}) \to \boldsymbol{y}_i, \ i = 1, 2, ..., n. \tag{22}$$

Given the target measure ν in Eq. (17), there exists a discrete Brenier's potential in Eq. (19) whose projected μ volume of each facet $w_i(\boldsymbol{h})$ is equal to the given target measure ν_i. This was proved by Alexandrov [46] in convex geometry.

Theorem 4.1 (Alexandrov [46]). Suppose Ω is a compact convex polytope with a non-empty interior in $\mathbb{R}^n$, $\boldsymbol{n}_1$, ..., $\boldsymbol{n}_k \subset \mathbb{R}^{n+1}$ are distinct k unit vectors, the $(n+1)$th coordinates are negative, and ν_1, ..., $\nu_k > 0$ so that $\sum_{i=1}^{k} \nu_i = \text{vol}(\boldsymbol{\Omega})$. Then there exists a convex polytope $P \subset \mathbb{R}^{n+1}$ with the exact k codimension-1 faces F_1, ..., F_k so that $\boldsymbol{n}_i$ is the normal vector to F_i and the intersection between Ω and the projection of F_i has the volume ν_i. Furthermore, such P is unique up to vertical translation.

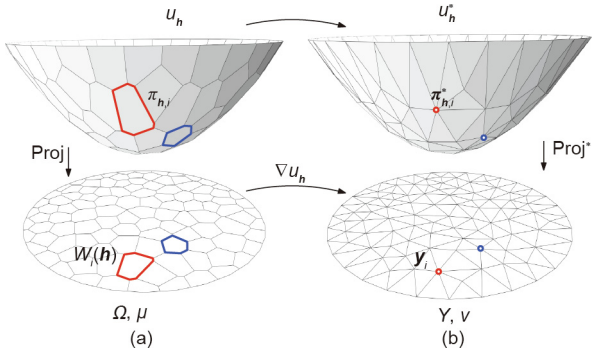

Fig. 6. (a) Piecewise linear Brenier's potential ($u_{\boldsymbol{h}}$) and (b) its Legendre transformation $u^*_{\boldsymbol{h}}$. $\pi^*_{\boldsymbol{h},i}$: the Legendre dual of $\pi_{\boldsymbol{h},i}$; $\nabla u_{\boldsymbol{h}}$: the gradient of $u_{\boldsymbol{h}}$; Proj: project map; Proj*: the projection map in the Legendre dual space.

Alexandrov's proof for the existence of the solution is based on algebraic topology, which is not constructive. Recently, Gu et al. [6] provided a constructive proof based on the variational approach.

Theorem 4.2 (Ref. [6]). Let μ be a probability measure defined on a compact convex domain Ω in $\mathbb{R}^d$, and let $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n\}$ be a set of distinct points in $\mathbb{R}^d$. Then for any $v_1, v_2, ..., v_n > 0$ with $\sum_{i=1}^{n} v_i = \mu(\Omega)$, there exists $\boldsymbol{h} = (h_1, h_2, ..., h_n) \in \mathbb{R}^n$, which is unique up to adding a constant $(c, c, ..., c)$, so that $w_i(\boldsymbol{h}) = v_i$, for all i. The vector $\boldsymbol{h}$ is the unique minimum argument of the following convex energy:

$$E(\boldsymbol{h}) = \int_0^{\boldsymbol{h}} \sum_{i=1}^{n} w_i(\eta) \mathrm{d}\eta_i - \sum_{i=1}^{n} \boldsymbol{h}_i v_i \tag{23}$$

defined on an open convex set

$$\boldsymbol{h} = \{\boldsymbol{h} \in \mathbb{R}^n : w_i(\boldsymbol{h}) > 0,\ i = 1, 2, ..., n\} \tag{24}$$

Furthermore, $\nabla u_{\boldsymbol{h}}$ minimizes the quadratic cost

$$\frac{1}{2} \int_{\Omega} \| \boldsymbol{x} - T(\boldsymbol{x}) \|^2 \mathrm{d}\mu(\boldsymbol{x}) \tag{25}$$

among all transport maps $T_{\#}\mu = \nu$.

The gradient of the above convex energy in Eq. (23) is given by the following:

$$\nabla E(\boldsymbol{h}) = [w_1(\boldsymbol{h}) - v_1, w_2(\boldsymbol{h}) - v_2, ..., w_n(\boldsymbol{h}) - v_n]^{\mathrm{T}} \tag{26}$$

The ith row and jth column element of the Hessian of the energy is given by the following:

$$\frac{\partial w_i}{\partial \boldsymbol{h}_j} = -\frac{\mu(W_i \cap W_j \cap \Omega)}{\| \boldsymbol{y}_i - \boldsymbol{y}_j \|}, \quad \frac{\partial w_i}{\partial \boldsymbol{h}_i} = \sum_{j \neq i} \frac{\partial w_i}{\partial \boldsymbol{h}_j} \tag{27}$$

As shown in Fig. 6, the Hessian matrix has an explicit geometric interpretation. Fig. 6(a) shows the discrete Brenier's potential $u_{\boldsymbol{h}}$, while Fig. 6(b) shows its Legendre transformation $u_{\boldsymbol{h}}^*$ using Definition 3.8. The Legendre transformation can be constructed geometrically: For each supporting plane $\pi_{\boldsymbol{h},i}$, we construct the dual point $\pi_{\boldsymbol{h},i}^* = (\boldsymbol{y}_i, \boldsymbol{h}_i)$; the convex hull of the dual points $\left\{\pi_{\boldsymbol{h},1}^*, \pi_{\boldsymbol{h},2}^*, ..., \pi_{\boldsymbol{h},n}^*\right\}$ is the graph of the Legendre transformation $u_{\boldsymbol{h}}^*$.

The projection of $u_{\boldsymbol{h}}^*$ induces a triangulation of $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n\}$, which is the weighted Delaunay triangulation. As shown in Fig. 7, the power diagram in Eq. (20) and the weighted Delaunay triangulation are Poincaré dual to each other: If, in the power diagram, $W_i(\boldsymbol{h})$ and $W_j(\boldsymbol{h})$ intersect at a $(d-1)$-dimensional cell, then in the weighted Delaunay triangulation, $\boldsymbol{y}_i$ connects with $\boldsymbol{y}_j$. The element of the Hessian matrix in Eq. (27) is the ratio between the μ volume of the $(d-1)$ cell in the power diagram and the length of the dual edge in the weighted Delaunay triangulation.

The conventional power diagram can be closely related to the above theorem.

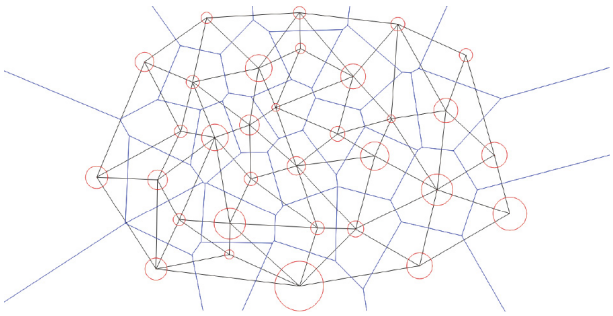

Fig. 7. Power diagram (blue) and its dual-weighted Delaunay triangulation (black).

Definition 4.3 (power distance). Given a point $\boldsymbol{y}_i \in \mathbb{R}^d$ with a power weight ψ_i, the power distance is given by the following:

$$\mathrm{pow}(\boldsymbol{x}, \boldsymbol{y}_i) = \| \boldsymbol{x} - \boldsymbol{y}_i \|^2 - \psi_i \tag{28}$$

Definition 4.4 (power diagram). Given the weighted points $(\boldsymbol{y}_1, \psi_1), ..., (\boldsymbol{y}_k, \psi_k)$, the power diagram is the cell decomposition of $\mathbb{R}^d$:

$$\mathbb{R}^d = \bigcup_{i=1}^{k} W_i(\psi) \tag{29}$$

where each cell is a convex polytope:

$$W_i(\psi) = \{\boldsymbol{x} \in \mathbb{R}^d | \mathrm{pow}(\boldsymbol{x}, \boldsymbol{y}_i) \leq \mathrm{pow}(\boldsymbol{x}, \boldsymbol{y}_j)\} \tag{30}$$

The weighted Delaunay triangulation, denoted as $T(\psi)$, is the Poincaré dual to the power diagram; if $W_i(\psi) \cap W_j(\psi) \neq \phi$, then there is an edge connecting $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ in the weighted Delaunay triangulation. Note that $\mathrm{pow}(\boldsymbol{x}, \boldsymbol{y}_i) \leq \mathrm{pow}(\boldsymbol{x}, \boldsymbol{y}_j)$ is equivalent to

$$\langle \boldsymbol{x}, \boldsymbol{y}_i \rangle + \frac{1}{2}\left(\psi_i - \| \boldsymbol{y}_i \|^2\right) \geq \langle \boldsymbol{x}, \boldsymbol{y}_j \rangle + \frac{1}{2}\left(\psi_j - \| \boldsymbol{y}_j \|^2\right) \tag{31}$$

Let $h_i = 1/2\left(\psi_i - \| \boldsymbol{y}_i \|^2\right)$; then we rewrite the definition of $W_i(\psi)$ as follows:

$$W_i(\psi) = \{\boldsymbol{x} \in \mathbb{R}^d | \langle \boldsymbol{x}, \boldsymbol{y}_i \rangle + \boldsymbol{h}_i \geq \langle \boldsymbol{x}, \boldsymbol{y}_j \rangle + h_j,\ \forall j\} \tag{32}$$

In practice, our goal is to compute the discrete Brenier's potential Eq. (19) by optimizing the convex energy Eq. (23). For low-dimensional cases, we can directly use Newton's method by computing the gradient Eq. (26) and the Hessian matrix Eq. (27). For deep learning applications, direct computation of the Hessian matrix is unfeasible; instead, we can use the gradient descend method or quasi-Newton's method with superlinear convergence. The key of the gradient is to estimate the μ volume $w_i(\boldsymbol{h})$. This can be done using the Monte Carlo method: We draw n random samples from the distribution μ, and count the number of samples falling within $W_i(\boldsymbol{h})$, which is the ratio converging to the μ volume. This method is purely parallel and can be implemented using a GPU. Moreover, we can use a hierarchical method to further improve the efficiency: First, we classify the target samples to clusters, and compute the OT map to the mass centers of the clusters; second, for each cluster, we compute the OT map from the corresponding cell to the original target samples within the cluster.

In order to avoid mode collapse, we need to find the singularity sets in Ω. As shown in Fig. 8, the target Dirac measure has two clusters; the source is the uniform distribution on the unit planar disk. The graph of the Brenier's potential function is a convex polyhedron with a ridge in the middle. The projection of the ridge on the disk is the singularity set $\Sigma_1(u)$, and the optimal mapping is discontinuous on Σ_1. In general cases, if two cells $W_i(\boldsymbol{h})$ and $W_j(\boldsymbol{h})$ are adjacent, then we compute the angle between the normals to the corresponding support planes:

$$\theta_{i,j} = \frac{\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle}{\| \boldsymbol{y}_i \| \| \boldsymbol{y}_j \|}$$

If $\theta_{i,j}$ is greater than a threshold, then the common facet $W_i(\boldsymbol{h}) \cap W_j(\boldsymbol{h})$ is in the discontinuity singular set.

5. GANs and optimal transportation

OT theory lays down the theoretical foundation for GANs. Many recent works, such as the WGAN [1], gradient penalty WGAN (WGAN-GP) [27], and relaxed Wasserstein with applications to GAN (RW-GAN) [47], use the Wasserstein distance to measure

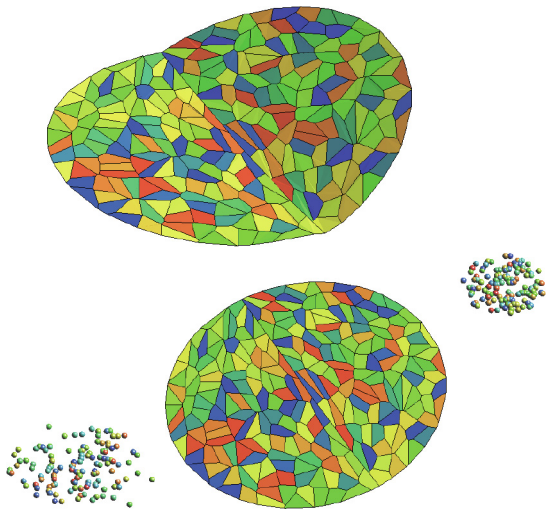

Fig. 8. Singularity set of the Brenier's potential function and discontinuity set of the OT map.

the deviation between the generated data distribution and the real data distribution.

From the OT perspective, the optimal solutions for the generator and discriminator are related by a closed form; hence, the generator and the discriminator should collaborate instead of compete. More details can be found in Ref. [11]. Furthermore, the regularity theory of the solutions to the Monge–Ampère equation can explain the mode collapse in GANs [48].

5.1. Competition versus collaboration

The OT view of the WGAN [1] is illustrated in Fig. 2. According to the manifold distribution hypothesis, the real data distribution ν is close to a manifold Σ embedded in the ambient space χ. The generator computes the decoding map g_θ from the latent space Z to the ambient space, and transforms the white noise ζ (i.e., the Gaussian distribution) to the generated distribution, μ_θ. The discriminator computes the Wasserstein distance between μ_θ and the real data distribution ν, $W_c(\mu_\theta, \nu)$, by computing the Kantorovich's potential φ_ξ. Both g_θ and φ_ξ are realized by DNNs.

In the training process, the generator improves g_θ in order to better approximate ν by $(g_\theta)_\#\zeta$; the discriminator refines the Kantorovich's potential φ_ξ to improve the estimation of the Wasserstein distance. The generator and the discriminator compete with each other, without sharing intermediate computational results. Under the L^1 cost function, the alternative training process of the WGAN can be formulated as the min–max optimization of expectations:

$$\min_\theta \max_\xi E_{z\sim\zeta}\{\varphi_\xi[g_\theta(\boldsymbol{z})]\} + E_{y\sim\nu}[\varphi^c_\xi(\boldsymbol{y})]$$

But if we change the cost function to be L^2 distance, then according to Theorem 3.10, at the optimum, the Brenier's potential u and the Kontarovich's potential φ are related by a closed form of Eq. (16), $u(\boldsymbol{x}) = 1/2\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x})$. The generator pursues the OT map ∇u; the discriminator computes φ. Hence, once the generator reaches the optimum, the optimal solution for the discriminator can be obtained without any training, and vice versa.

In more detail, suppose at the kth iteration, the generator map is g^k_θ. The discriminator computes the Kontarovich's potential φ_ξ, which gives the Wasserstein distance between the current generated data distribution $(g^k_\theta)_\#\zeta$ and the real data distribution ν; ∇u gives the OT map from $(g^k_\theta)_\#\zeta$ to ν. Therefore, we obtain the following:

$$\nu = (\nabla u)_\#\left[(g^k_\theta)_\#\zeta\right] = (\nabla u \circ g^k_\theta)_\#\zeta = \left[\left(\mathrm{id} - \nabla_{\varphi_\xi}\right)\circ g^k_\theta\right]_\#\zeta$$

This means that the generator map can be updated by the following:

$$g^{k+1}_\theta = (\mathrm{id} - \nabla\varphi_\xi)\circ g^k_\theta \tag{33}$$

This conclusion shows that, in principle, the training process of the generator can be skipped; in practice, the efficiency can be improved greatly by sharing the intermediate computational results. Therefore, in designing the architectures of GANs, collaboration is better than competition.

5.2. Mode collapse and regularity

Although GANs are powerful for many applications, they have critical drawbacks: First, the training of GANs is tricky, sensitive to hyperparameters, and difficult to converge; second, GANs suffer from mode collapsing; and third, GANs may generate unrealistic samples. The difficulties of convergence, mode collapse, and the generation of unrealistic samples can be explained by the regularity Theorem 3.15 of the OT map.

According to Brenier's polar factorization, Theorem 3.9, any measure-preserving map can be decomposed into two maps, one of which is an OT map, which is a solution to the Monge–Ampère equation. According to the regularity Theorem 3.15, if the support Λ of the target measure ν has multiple connected components—that is, if ν has multiple modes, or Λ is non-convex—then the OT map T: $\Omega \to \Lambda$ is discontinuous on the singular set Σ_Ω.

Fig. 9 shows the multi-cluster case: Λ has two connected components, where the OT map T is discontinuous along Σ_1. Fig. 10 shows that even Λ is connected, albeit non-convex. Ω is a rectangle, Λ is a dumbbell shape, the density functions are constants, the OT map is discontinuous, and the singularity set $\Sigma_1 = \gamma_1 \cup \gamma_2$.

Fig. 11 shows an OT map between two probability measures in $\mathbb{R}^3$. Both the source measure μ and the target measure ν are uniform distributions; the support of Ω is the unit solid ball, and the support of Λ is the solid Stanford bunny. We compute the Brenier's potential u: $\Omega \to \mathbb{R}$ based on Theorem 4.2. In order to visualize the mapping, we interpolate the probability measure as follows:

$$\rho_t := [(1-t)\mathrm{id} + t\nabla u]_\#\mu,\ 0 \le t \le 1$$

Fig. 11 shows the support of the interpolated measure ρ_t. The foldings on the surface are the singularity sets, where the OT map is discontinuous.

In a general situation, due to the complexity of the real data distributions, the embedding manifold Σ, and the encoding/decoding maps, the supports of the target measures are rarely convex; therefore, the transportation mapping cannot be globally continuous.

On the other hand, general DNNs, such as rectified linear unit (ReLU) DNNs, can only approximate continuous mappings. The functional space represented by ReLU DNNs does not contain the desired discontinuous transportation mapping. The training process or, equivalently, the searching process will lead to three alternative situations:

(1) The training process is unstable, and does not converge.

(2) The searching converges to one of the multiple connected components of Λ, and the mapping converges to one continuous branch of the desired transportation mapping. This means that a mode collapse is encountered.

(3) The training process leads to a transportation map, which covers all the modes successfully, but also covers the regions

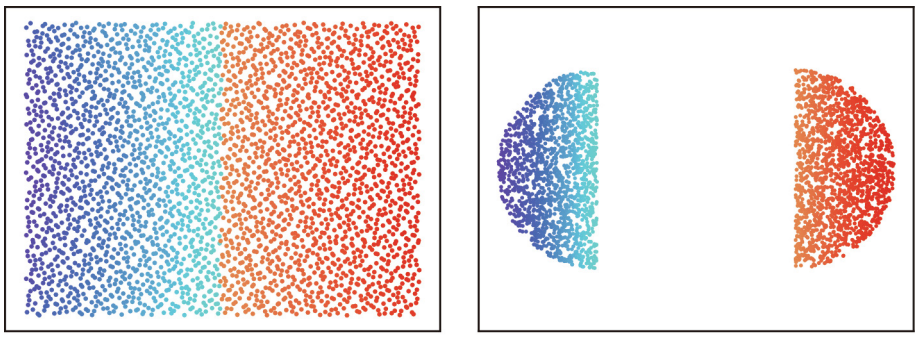

Fig. 9. Discontinuous OT map, produced by a GPU implementation of an algorithm based on Theorem 4.2: (a) is the source domain and (b) is the target domain. The middle line in (a) is the singularity set Σ_1.

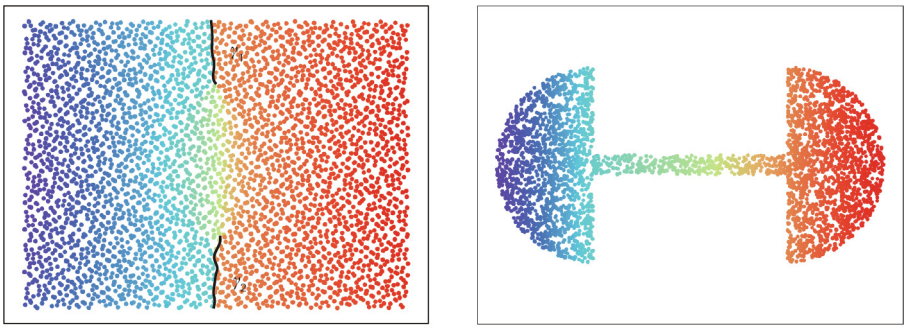

Fig. 10. Discontinuous OT map, produced by a GPU implementation of an algorithm based on Theorem 4.2: (a) is the source domain and (b) is the target domain. γ_1 and γ_2 in (a) are two singularity sets.

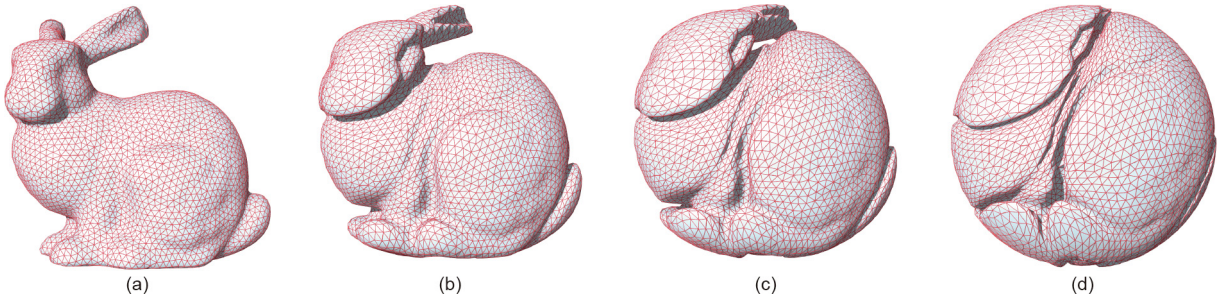

Fig. 11. OT from the Stanford bunny to a solid ball. The singular sets are the foldings on the boundary surface. (a–d) show the deformation procedure.

outside Λ. In practice, this will induce the phenomenon of generating unrealistic samples, as shown in the middle frame of Fig. 12.

Therefore, in theory, it is impossible to approximate OT maps directly using DNNs.

5.3. AE–OT model

As shown in Fig. 4, we separate the two main tasks of GANs: manifold learning and probability distribution transformation. The first task is carried out by an AE to compute the encoding/decoding maps f_θ, g_ξ; the second task is accomplished using the explicit variational method to compute the OT map T in the latent space. The real data distribution ν is pushed forward by the encoding map f_θ, inducing $(f_\theta)_\# \nu$. In the latent space, T maps the uniform distribution μ to $(f_\theta)_\# \nu$.

The AE–OT model has many advantages. In essence, finding the OT map is a convex optimization problem; the existence and the uniqueness of the solution are guaranteed. The training process is stable and has superlinear convergence by using quasi-Newton's method. The number of unknowns is equal to that of the training samples, avoiding over-paramerization. The parallel OT map algorithm can be implemented using a GPU. The error bound of the OT map can be controlled by the sampling density in the Monte Carlo method. The hierarchical algorithm with self-adaptivity further improves the efficiency. In particular, the AE–OT model can eliminate mode collapse.

6. Experimental results

In this section, we report our experimental results.

6.1. Training process

The training of the AE–OT model mainly includes two steps: training the AE and finding the OT map. The OT step is accomplished using a GPU implementation of the algorithm, as described

(a) (b)

Fig. 12. Facial images generated by an AE–OT model. (a) Generated realistic facial images; (b) a path through a singularity. The image in the center of (b) shows that the transportation map is discontinuous.

in Section 4. In the AE step, during the training process, we adopt the Adam algorithm [49] to optimize the parameters of the neutral network, with a learning rate of 0.003, β_1 = 0.5, and β_2 = 0.999. When the L^2 loss stops descending, which means that the network has found a good encoding map, we freeze the encoder part and continue to train the network for the decoding map. The training loss before and after the freezing of the encoder is shown in Table 1. Next, in order to find the OT map from the given distribution (here, we use uniform distribution) to the distribution of latent features, we randomly sample 100N random points from the uniform distribution to compute the gradient of the energy. Here, N is the number of latent features of the dataset. Also, in the experiment, $\theta_{i,j}$ is set to be different for different datasets. To be specific, for the MNIST and Fashion-MNIST datasets, $\theta_{i,j}$ is set to be 0.75, while for the CIFAR-10 and CelebA datasets, it is set to be 0.68 and 0.75, respectively.

Our AE–OT model was implemented using PyTorch on a Linux platform. All the experiments were conducted on a GTX1080Ti.

6.2. Transportation map discontinuity test

In this experiment, we want to test our hypothesis: In most real applications, the support of the target measure is non-convex, the singularity set is non-empty, and the probability distribution map is discontinuous along the singularity set.

As shown in Fig. 12, we use an AE to compute the encoding/decoding maps from the CelebA dataset (Σ, ν) to the latent space Z; the encoding map f_θ: $\Sigma \to Z$ pushes forward ν to $(f_\theta)_\#\nu$ on the latent space. In the latent space, we compute the OT map based on the algorithm described in Section 4, T: $Z \to Z$, where T maps the uniform distribution in a unit cube ζ to $(f_\theta)_\#\nu$. Then we randomly draw a sample z from the distribution ζ and use the decoding map g_ξ: $Z \to \Sigma$ to map $T(z)$ to a generated human facial image $g_\xi \circ T(z)$. Fig. 12(a) demonstrates the realistic facial images generated by this AE–OT framework.

Table 1
The L^2 loss of the AEs before and after the freezing of the encoder.

Situation	Dataset			
	MNIST	Fashion-MNIST	CIFAR-10	CelebA
Before	0.0013	0.0026	0.0023	0.0077
After	0.0005	0.0011	0.0018	0.0074

If the support of the push-forward measure $(f_\theta)_\#\nu$ in the latent space is non-convex, there will be a singularity set Σ_k, where $k > 0$. We would like to detect the existence of Σ_k. We randomly draw line segments in the unit cube in the latent space, and then densely interpolate along this line segment to generate facial images. As shown in Fig. 12(b), we find a line segment γ, and generate a morphing sequence between a boy with a pair of brown eyes and a girl with a pair of blue eyes. In the middle, we generate a face with one blue eye and one brown eye, which is definitely unrealistic and outside Σ. This result means that the line segment γ goes through a singularity set Σ_k, where the transportation map T is discontinuous. This also shows that our hypothesis is correct: The support of the encoded human facial image measure on the latent space is non-convex.

As a byproduct, we find that this AE–OT framework improves the training speed by a factor of five and increases the convergence stability, since the OT step is a convex optimization. Thus, it provides a promising way to improve existing GANs.

6.3. Mode collapse comparison

Since the synthetic dataset consists of explicit distributions and known modes, mode collapse can be accurately measured. We chose two synthetic datasets that have been studied or proposed in prior works [50,51]: a 2D grid dataset.

For a choice of the measurement metric of mode collapse, we adopted three previously used metrics [50,51]. Number of modes counts the quantity of modes captured by the samples produced by a generative model. In this metric, a mode is considered as lost if no sample is generated within three standard deviations of that mode. Percentage of high-quality samples measures the proportion of samples that are generated within three standard deviations of the nearest mode. The third metric, used in Ref. [51], is the reverse Kullback–Leibler (KL) divergence. In this metric, each generated sample is assigned to its nearest mode, and we count the histogram of samples assigned on each mode. This histogram then forms a discrete distribution, whose KL divergence with the histogram formed by real data is then calculated. Intuitively, this measures how well the generated samples balance among all modes regarding the real distribution.

In Ref. [51], the authors evaluated GAN [26], adversarially learned inference (ALI) [52], minibatch discriminati (MD) [30], and PacGAN [51] on synthetic datasets with the above three metrics. Each experiment was trained under the same generator

architecture with a total of approximately 4×10^5 training parameters. The networks were trained on 1×10^5 samples for 400 epochs. For the AE–OT experiment, since the source space and target space are both 2D, there is no need to train an AE. We directly compute a semi-discrete OT that maps between the uniform distribution on the unit square and the empirical real data distribution. Theoretically, the minimum amount of real sample needed for OT to recover all modes is one sample per mode. However, this may lead to the generation of low-quality samples during the interpolation process. Therefore, for OT computation, we take 512 real samples, and new samples are generated based on this map. We note that, in this case, there are only 512 parameters to optimize in OT computing, and the optimization process is stable due to the existence of the convex positive-definite Hessian. Our results are provided in Table 2, and benchmarks of previous methods are copied from Ref. [51]. For illustration purposes, we plotted our results on synthetic datasets along with those of GAN and PacGAN in Fig. 13.

6.4. Comparison with the state of the art

We designed experiments to compare our proposed AE–OT model with state-of-the-art generative models, including the adversarial models evaluated by Lucic et al. in Ref. [33], and the non-adversarial models studied by Hoshen and Malik in Ref. [36].

For the purpose of fair comparison, we used the same testing datasets and network architecture. The datasets included MNIST [53], Fashion-MNIST [54], CIFAR-10 [55], and CelebA [56], similar to those tested in Refs. [31,36]. The network architecture was similar to that used by Lucic et al. in Ref. [33]. In particular, in our AE–OT model, the network architecture of the decoder was the same as that of the generators of GANs in Ref. [33], and the encoder was symmetric to the decoder.

We compared our model with state-of-the-art generative models using the FID score [31] and PRD curve as the evaluation criteria. The FID score measures the visual fidelity of the generated results and is robust to image corruption. However, the FID score is sensitive to mode addition and dropping [33]. Hence, we also used the PRD curve, which can quantify the degree of mode dropping and mode inventing on real datasets [32].

Table 2
Mode collapse comparison for the 2D grid dataset.

Method	Modes	Samples	Reverse KL
GAN	17.3 ± 0.8	94.8 ± 0.7%	0.70 ± 0.07
ALI	24.1 ± 0.4	95.7 ± 0.6%	0.14 ± 0.03
MD	23.8 ± 0.5	79.9 ± 3.2%	0.17 ± 0.03
PacGAN2	23.8 ± 0.7	91.3 ± 0.8%	0.13 ± 0.04
PacGAN3	24.6 ± 0.4	94.2 ± 0.4%	0.06 ± 0.02
PacGAN4	24.8 ± 0.2	93.6 ± 0.6%	0.04 ± 0.01
AE–OT	25.0 ± 0.0	99.8 ± 0.2%	0.007 ± 0.002

6.4.1. Comparison with FID score

The FID score is computed as follows: ① Extract the visually meaningful features of both the generated and real images by running the inception network [30], ② fit the real and generated feature distributions with Gaussian distributions; and ③ compute the distance between the two Gaussian distributions using the following formula:

$$\mathrm{FID} = \| \mu_r - \mu_g \|_2^2 + T_r\left[\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2}\right] \tag{34}$$

where μ_r and μ_g represent the means of the real and generated distributions, respectively; and Σ_r and Σ_g represent the variances of these distributions.

The comparison results are summarized in Tables 3 and 4. The statistics of various GANs come from Lucic et al. [33], and those of the non-adversarial generative models come from Hoshen and Malik [36]. In general, our proposed model achieves better FID scores than the other state-of-the-art generative models.

Theoretically, the FID scores of our AE–OT model should be close to those of the pre-trained AEs; this is also validated by our experiments.

The fixed network architecture of our AE was adopted from Lucic et al. [33]; its capacity is not large enough to encode CIFAR-10 or CelebA, so we had to down-sample these datasets. We randomly selected 2.5×10^4 images from CIFAR-10 and 1×10^4 images from CelebA to train our model. Even so, our model obtained the best FID score in CIFAR-10. Dut to the limited capacity of the InfoGAN model, the performance of the AE of CelebA, whose FID of 67.5 is not ideal, further caused the FID of the generated dataset to be 68.4. By adding two more convolutional layers to the AE architecture, the L^2 loss in CelebA was less than 0.03, and the FID score beat all other models (28.6, as shown in the bracket of Table 4).

6.4.2. Comparison with the PRD curve

The FID score is an effective method to measure the difference between the generated distribution and the real data distribution, but it mainly focuses on precision, and cannot accurately capture what portion of real data a generative model can cover. The method proposed in Ref. [32] disentangles the divergence between distributions into two components: precision and recall.

Given a reference distribution P and a learned distribution Q, the precision intuitively measures the quality of samples from Q, while the recall measures the proportion of P that is covered by Q.

We used the concept of $(F_8, F_{1/8})$ introduced by Sajjadi et al. in Ref. [32] to quantify the relative importance of precision and recall. Fig. 14 summarizes the comparison results. Each dot represents a specific model with a set of hyperparameters. The closer a dot is

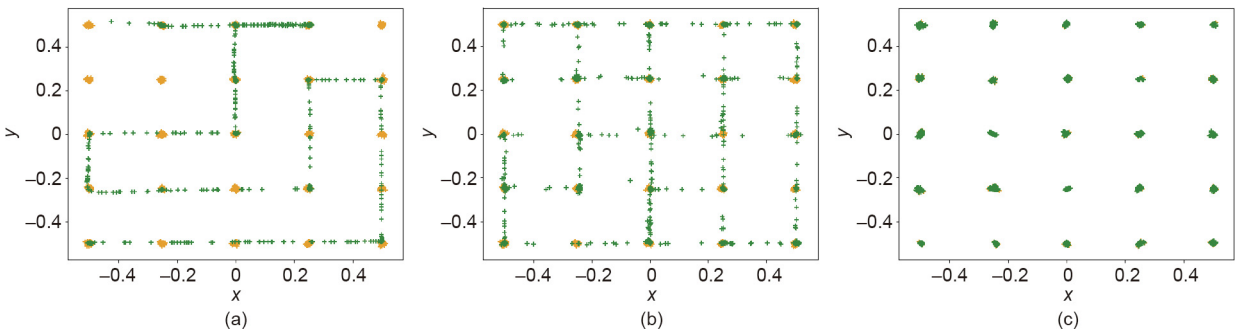

Fig. 13. Mode collapse comparison on a 2D grid dataset. (a) GAN; (b) PacGAN4; (c) AE–OT. Orange marks are real samples and green marks are generated ones.

Table 3
Quantitative comparison with FID-I.

Dataset	Adversarial				
	MM GAN	NS GAN	LSGAN	WGAN	BEGAN
MNIST	9.8	6.8	7.8	6.7	13.1
Fashion-MNIST	29.6	26.5	30.7	21.5	22.9
CIFAR-10	72.7	58.5	87.1	55.2	71.4
CelebA	65.6	55.0	53.9	41.3	**38.9**

The best result is shown in bold. MM: manifold matching; NS: non-saturating; LSGAN: least squares GAN; BEGAN: boundary equilibrium GAN.

Table 4
Quantitative comparison with FID-II.

Dataset	Non-adversarial			Reference	
	VAE	GLO	GLANN	AE	AE–OT
MNIST	23.8	49.6	8.6	5.5	**6.4**
Fashion-MNIST	58.7	57.7	13.0	4.7	**10.2**
CIFAR-10	155.7	65.4	46.5	28.2	**38.1**
CelebA	85.7	52.4	46.3	67.5	68.4 (**28.6**)

The best result is shown in bold.

to the upper-right corner, the better the performance of the model is. The blue and green dots show the GANs and VAEs evaluated in Ref. [32], the khaki dot represents the GLANN model in Ref. [36], and the red dot is our AE–OT model.

It is clear that our proposed model outperforms others for MNIST and Fashion-MNIST. For the CIFAR-10 dataset, the precision of our model is slightly lower than those of GANs and GLANN, but the recall is the highest. For the CelebA dataset, due to the limited capacity of the AE, the performance of our model is not impressive. However, after adding two more convolutional layers to the AE, our model achieves the best score.

6.4.3. Visual comparison

Fig. 15 shows a visual comparison between the images generated by our proposed method and those generated by the GANs studied by Lucic et al. in Ref. [33] and the non-adversarial models

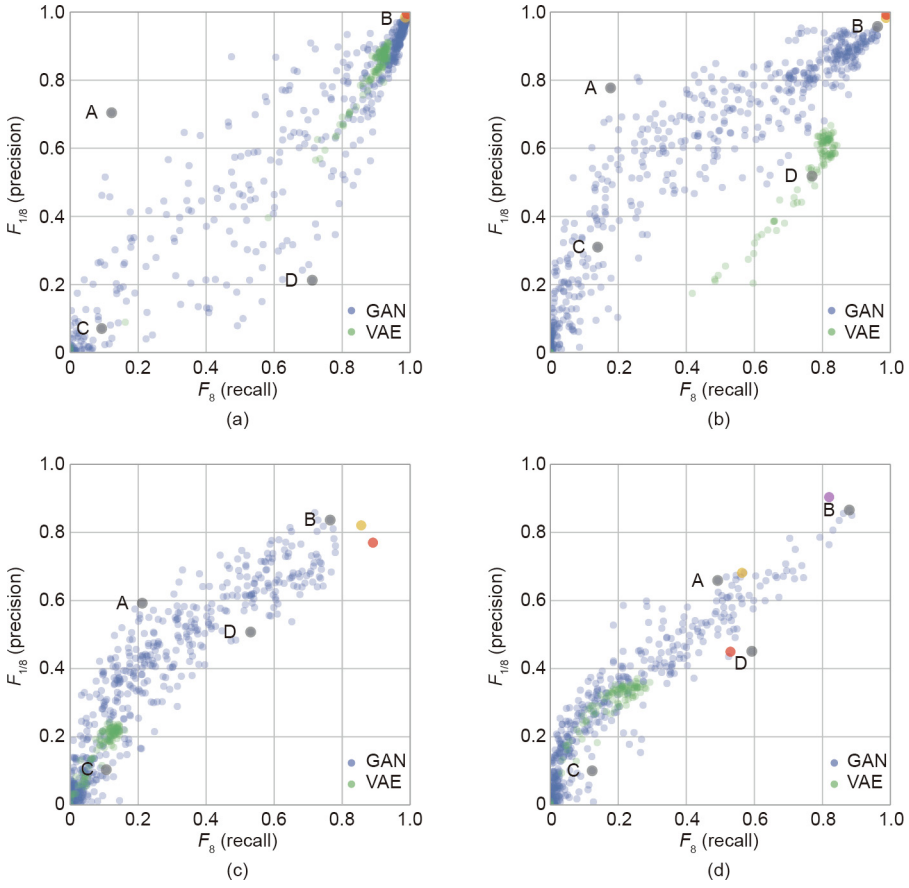

Fig. 14. A comparison of the precision–recall pair in (F_8, $F_{1/8}$) in the four datasets. (a) MNIST; (b) Fashion-MNIST; (c) CIFAR-10; (d) CelebA. The khaki dots are the results of Ref. [36]. The red dots are the results of the proposed method. The purple dot in the fourth subfigure corresponds to the results of the architecture with two more convolutional layers.

Fig. 15. A visual comparison of the four datasets. The first column (a) shows the real data; the second column (b) is generated by an AE; the third column (c) illustrates the generating results of the GANs [33] with the highest precision-recall scores of (F_8, $F_{1/8}$), corresponding to the B dots in Fig. 14; the fourth column (d) gives the results of Ref. [36]; and the last column (e) shows the results of the proposed method.

studied by Hoshen and Malik in Ref. [36]. The first column shows the original images, the second column shows the results generated by the AE, the third column shows the best generating results of the GANs in Lucic et al. [33], the fourth column displays the results generated by the models of Hoshen and Malik [36], and the fifth column displays the results from our method. It is clear that our method generates high-quality images and covers all modes.

7. Conclusion

This work uses OT theory to interpret GANs. According to the data manifold distribution hypothesis, GANs mainly accomplish two tasks: manifold learning and probability distribution transformation. The latter task can be carried out using the OT method directly. This theoretical understanding explains the fundamental reason for mode collapse, and shows that the intrinsic relation between the generator and the discriminator should be collaboration instead of competition. Furthermore, we propose an AE–OT model, which improves the theoretical rigor, training stability, and efficiency, and eliminates mode collapse.

Our experiment validates our assumption that if the distribution transportation map is discontinuous, then the existence of the singularity set leads to mode collapse. Furthermore, when our proposed model is compared with the state of the art, our method eliminates the mode collapse and outperforms the other models in terms of the FID score and PRD curve.

In the future, we will explore the theoretical understanding of the manifold learning stage, and use a rigorous method to make this part of the black box transparent.

Acknowledgements

The project is partially supported by the National Natural Science Foundation of China (61936002, 61772105, 61432003, 61720106005, and 61772379), US National Science Foundation (NSF) CMMI-1762287 collaborative research "computational framework for designing conformal stretchable electronics, Ford URP topology optimization of cellular mesostructures' nonlinear behaviors for crash safety," and NSF DMS-1737812 collaborative research "ATD: theory and algorithms for discrete curvatures on network data from human mobility and monitoring."

Compliance with ethics guidelines

Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu declare that they have no conflicts of interest or financial conflicts to disclose.

References

[1] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia; 2017. p. 214–23.
[2] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23.

[3] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11):2579–605.
[4] Mescheder L, Geiger A, Nowozin S. Which training methods for GANs do actually converge? In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholmsmässan, Sweden; 2018. p. 3478–87.
[5] Villani C. Optimal transport: old and new. Berlin: Springer Science & Business Media; 2008.
[6] Gu DX, Luo F, Sun J, Yau ST. Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge–Ampère equations. Asian J Math 2016;20(2):383–98.
[7] Peyré G, Cuturi M. Computational optimal transport. Found Trends Mach Learn 2019;11(5–6):355–607.
[8] Solomon J. Optimal transport on discrete domains. 2018. arXiv:1801.07745.
[9] Cuturi M. Sinkhorn distances: lightspeed computation of optimal transportation distances. Adv Neural Inf Process Syst 2013;26:2292–300.
[10] Solomon J, de Goes F, Peyré G, Cuturi M, Butscher A, Nguyen A, et al. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. ACM Trans Graph 2015;34(4):66.
[11] Lei N, Su K, Cui L, Yau ST, Gu XD. A geometric view of optimal transportation and generative model. Comput Aided Geom Des 2019;68:1–21.
[12] Benamou JD, Brenier Y, Guittet K. The Monge–Kantorovitch mass transfer and its computational fluid mechanics formulation. Int J Numer Methods Fluids 2002;40(1–2):21–30.
[13] Jean-David Benamou BDF, Oberman AM. Numerical solution of the optimal transportation problem using the Monge–Ampère equation. J Comput Phys 2014;260:107–26.
[14] Nicolas P, Gabriel P, Oudet E. Optimal transport with proximal splitting. SIAM J Imaging Sci 2014;7(1):212–38.
[15] Bengio Y, Mesnil G, Dauphin Y, Rifai S. Better mixing via deep representations. In: Proceedings of the 30th International Conference on Machine Learning; 2013 Jun 16–21; Atlanta, GA, USA; 2013. p. 552–60.
[16] Salakhutdinov R, Larochelle H. Efficient learning of deep Boltzmann machines. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics; 2010 May 13–15; Chia Laguna Resort, Italy; 2010. p. 693–700.
[17] Kingma DP, Welling M. Auto-encoding variational Bayes. 2013. arXiv:1312.6114.
[18] Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. 2014. arXiv:1401.4082.
[19] Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. 2015. arXiv:1511.05644.
[20] Tolstikhin I, Bousquet O, Gelly S, Schoelkopf B. Wasserstein auto-encoders. 2017. arXiv:1711.01558.
[21] He X, Yan S, Hu Y, Niyogi P, Zhang HJ. Face recognition using laplacianfaces. IEEE Trans Pattern Anal Mach Intell 2005;27(3):328–40.
[22] Arandjelović O. Unfolding a face: from singular to manifold. In: Proceedings of the 9th Asian Conference on Computer Vision; 2009 Sep 23–27; Xi'an, China; 2009. p. 203–13.
[23] Salimans T, Karpathy A, Chen X, Kingma DP. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. 2017. arXiv:1701.05517.
[24] Oord Ad, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. 2016. arXiv:1601.06759.
[25] Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. WaveNet: a generative model for raw audio. 2016. arXiv:1609.03499.
[26] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. 2014. arXiv:1406.2661.
[27] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein GANs. 2017. arXiv:1704.00028.
[28] Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. 2018. arXiv:1802.05957.
[29] Zoran D, Weiss Y. From learning models of natural image patches to whole image restoration. In: Proceedings of the 2011 International Conference on Computer Vision; 2011 Jun 6–11; Barcelona, Spain; 2011. p. 479–86.
[30] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. 2016. arXiv:1606.03498.
[31] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Klambauer G, Hochreiter S. GANs trained by a two time-scale update rule converge to a Nash equilibrium. 2017. arXiv:1706.08500.
[32] Sajjadi MS, Bachem O, Lucic M, Bousquet O, Gelly S. Assessing generative models via precision and recall. 2018. arXiv:1806.00035.
[33] Lucic M, Kurach K, Michalski M, Gelly S, Bousquet O. Are GANs created equal? A large-scale study. 2018. arXiv:1711.10337.
[34] Bojanowski P, Joulin A, Lopez-Paz D, Szlam A. Optimizing the latent space of generative networks. 2017. arXiv:1707.05776.
[35] Li K, Malik J. Implicit maximum likelihood estimation. 2018. arXiv:1809.09087.
[36] Hoshen Y, Malik J. Non-adversarial image synthesis with generative latent nearest neighbors. 2018. arXiv:1812.08985.
[37] Dinh L, Krueger D, Bengio Y. NICE: non-linear independent components estimation. 2014. arXiv:1410.8516.
[38] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. 2017. arXiv:1605.08803.
[39] Kingma DP, Dhariwal P. Glow: generative flow with invertible 1×1 convolutions. 2018. arXiv:1807.03039.
[40] LeCun Y, Chopra S, Hadsell R, Ranzota MA, Huang FJ. A tutorial on energy-based learning. In: Bakir G, Hofman T, Schölkopf T, Smola A, Taskar B, editors. Predicting structured data. Cambridge: The MIT Press; 2006.
[41] Dai J, Lu Y, Wu Y. Generative modeling of convolutional neural networks. In: Proceedings of the 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, CA, USA; 2015.
[42] Nijkamp E, Hill M, Zhu S, Wu Y. On learning non-convergent non-persistent short-run MCMC toward energy-based model. 2019. arXiv:1904.09770.
[43] Bonnotte N. From Knothe's rearrangement to Brenier's optimal transport map. SIAM J Math Anal 2013;45(1):64–87.
[44] Brenier Y. Polar factorization and monotone rearrangement of vector-valued functions. Commun Pure Appl Math 1991;44(4):375–417.
[45] Caffarelli L. Some regularity properties of solutions of Monge–Ampère equation. Commun Pure Appl Math 1991;44(8–9):965–9.
[46] Alexandrov AD. Convex polyhedra. New York: Springer; 2005.
[47] Guo X, Hong J, Lin T, Yang N. Relaxed wasserstein with applications to GANs. 2017. arXiv:1705.07164.
[48] Lei N, Guo Y, An D, Qi X, Luo Z, Gu X, et al. Mode collapse and regularity of optimal transportation maps. 2019. arXiv:1902.02934.
[49] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. arXiv:1412.6980.
[50] Srivastava A, Valkov L, Russell C, Gutmann MU, Sutton C. VeeGAN: reducing mode collapse in GANs using implicit variational learning. 2017. arXiv:1705.17761.
[51] Lin Z, Khetan A, Fanti G, Oh S. PacGAN: the power of two samples in generative adversarial networks. 2017. arXiv:1712.04086.
[52] Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky M, et al. Adversarially learned inference. 2016. arXiv:1606.00704.
[53] LeCun Y, Cortes C, Burges CJC. The MNIST database of handwritten digits. Available from: http://yann.lecun.com/exdb/mnist/.
[54] Xiao H, Rasul F, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. 2017. arXiv:1708.07747.
[55] Krizhevsky A. Learning multiple layers of features from tiny images. Technical report. Toronto: University of Toronto; 2009.
[56] Zhang Z, Luo P, Loy CC, Tang X. From facial expression recognition to interpersonal relation prediction. Int J Comput Vis 2018;126(5):550–69.