Computer Aided Geometric Design 68 (2019) 1–21

Contents lists available at ScienceDirect

Computer Aided Geometric Design

www.elsevier.com/locate/cagd

A geometric view of optimal transportation and generative model

Na Lei [a,f], Kehua Su [b], Li Cui [c], Shing-Tung Yau [d], Xianfeng David Gu [e,*]

[a] DUT-RU ISE, Dalian University of Technology, Dalian, China
[b] School of Compute Science, Wuhan university, Wuhan, China
[c] Mathematics Department, Beijing Normal University, Beijing, China
[d] Mathematics Department, Harvard University, Cambridge, USA
[e] State University of New York at Stony Brook, Stony Brook, USA
[f] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, China

ARTICLE INFO

Article history:
Available online 9 November 2018

Keywords:
Optimal Mass Transportation
Monge–Ampere
GAN
Wasserstein distance

ABSTRACT

In this work, we give a geometric interpretation to the Generative Adversarial Networks (GANs). The geometric view is based on the intrinsic relation between Optimal Mass Transportation (OMT) theory and convex geometry, and leads to a variational approach to solve the Alexandrov problem: constructing a convex polytope with prescribed face normals and volumes.

By using the optimal transportation view of GAN model, we show that the discriminator computes the Wasserstein distance via the Kantorovich potential, the generator calculates the transportation map. For a large class of transportation costs, the Kantorovich potential can give the optimal transportation map by a close-form formula. Therefore, it is sufficient to solely optimize the discriminator. This shows the adversarial competition can be avoided, and the computational architecture can be simplified.

Preliminary experimental results show the geometric method outperforms the traditional Wasserstein GAN for approximating probability measures with multiple clusters in low dimensional space.

1. Introduction

GAN model Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) aim at learning a mapping from a simple distribution to a given distribution. A GAN model consists of a generator G and a discriminator D, both are represented as deep neural networks (DNNs). The generator captures the data distribution and generates samples, the discriminator estimates the probability that a sample came from the training data rather than the generator. Both the generator and the discriminator are trained simultaneously. The competition drives both of them to improve their performance until the generated samples are indistinguishable from the genuine data samples. At the Nash equilibrium (Zhao et al., 2016), the distribution generated by G equals to the real data distribution. GANs have several advantages: they can automatically generate samples and reduce the amount of real data samples; furthermore, GANs do not need the explicit expression of the distribution of given data.

* Corresponding author.
E-mail addresses: nalei@dlut.edu.cn (N. Lei), skh@whu.edu.cn (K. Su), licui@bnu.edu.cn (L. Cui), yau@math.harvard.edu (S.-T. Yau), gu@cs.stonybrook.edu (X.D. Gu).

https://doi.org/10.1016/j.cagd.2018.10.005
0167-8396/

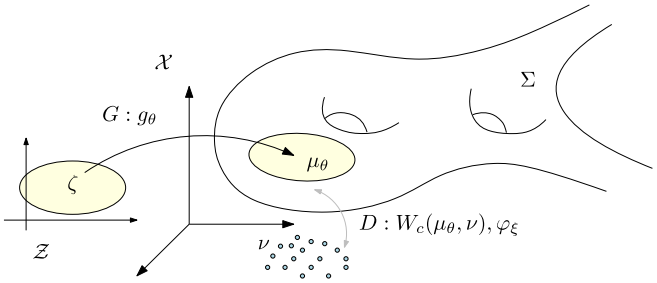


Fig. 1. Wasserstein Generative Adversarial Networks (W-GAN) framework.

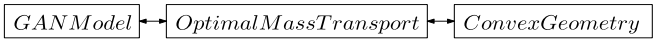


Fig. 2. The GAN model, OMT theory and convex geometry has intrinsic relations.

Recently, GANs receive an exploding amount of attention. For examples, GANs have been widely applied to numerous computer vision tasks such as image inpainting (Pathak et al., 2016; Yeh et al., 2017; Li et al., 2017b), image super resolution (Ledig et al., 2016; Iizuka et al., 2017), semantic segmentation (Zhu and Xie, 2016; Luc et al., 2016), object detection (Radford et al., 2015; Li et al., 2017a; Wang et al., 2017), video prediction (Mathieu et al., 2015; Vondrick et al., 2016), image translation (Isola et al., 2016; Zhu et al., 2017; Dong et al., 2017; Liu et al., 2017), 3D vision (Wu et al., 2016; Park et al., 2017), face editing (Larsen et al., 2015; Liu and Tuzel, 2016; Perarnau et al., 2016; Shen and Liu, 2017; Brock et al., 2017; Shu et al., 2017; Huang et al., 2017), etc. Also, in machine learning field, GANs have been applied to semi-supervised learning (Odena, 2016; Kumar et al., 2017; Salimans et al., 2016), clustering (Springenberg, 2016), cross domain learning (Taigman et al., 2016; Kim et al., 2017), and ensemble learning (Tolstikhin et al., 2017).

Optimal transportation view In deep learning, the "data distribution hypothesis" is well accepted: natural data sets distribute close to low dimensional manifolds. Therefore, the central goal of deep learning is to learn these manifolds and the distributions on them. Generative models, such as Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE), achieve this by mapping a data manifold embedded in the ambient space to the low dimensional latent space and manipulating the mapping to adjust the push forward distributions on the latent space.

Recently, Optimal Mass Transportation (OMT) theory has been applied to improve VAEs and GANs. The Wasserstein distance has been adapted by GANs as the loss function as the discriminator, such as WGAN (Arjovsky et al., 2017), WGAN-GP (Gulrajani et al., 2017) and RWGAN (Guo et al., 2017). When the supports of two distributions have no overlap, Wasserstein distance still provides a suitable gradient for the generator to update. The Wasserstein distance in VAE is calculated using linear programming method in Liu et al. (2018), which gives more transparent and accurate results.

Fig. 1 shows the optimal mass transportation point of view of WGAN (Arjovsky et al., 2017). The ambient image space is $\mathcal{X}$, with the real data distribution ν. The latent space is $\mathcal{Z}$ with much lower dimension. The generator G can be treated as a "decoding map" from the latent space to the sample space, $g_\theta : \mathcal{Z} \to \mathcal{X}$, realized by a deep neural network with parameter θ. Let ζ be a fixed distribution in the latent space, such as uniform distribution or Gaussian distribution. The generator G pushes forward ζ to a distribution $\mu_\theta = g_{\theta\#}\zeta$ in the ambient space $\mathcal{X}$. The discriminator D uses the power of Euclidean distance as the cost function and computes the Wasserstein distance between μ_θ and ν, $W_c(\mu_\theta, \nu)$, realized by another deep neural network with parameter ξ. Calculating the Wasserstein distance $W_c(\mu_\theta, \nu)$ is equivalent to finding the so-called Kantorovich potential φ_ξ. Therefore, G improves the decoding map g_θ to approximate ν by $g_{\theta\#}\zeta$; D improves the Kantorovich potential φ_ξ to increase the approximation accuracy to the Wasserstein distance. The generator G and the discriminator D are trained alternatively, until the competition reaches an equilibrium.

In summary, the Generative Adversarial Network model (GAN) has natural connection with the Optimal Mass Transportation (OMT) theory:

1. In generator G, the generating map g_θ in GAN is equivalent to the optimal transportation map in OMT;
2. In discriminator D, the Wasserstein distance between distributions is equivalent to the Kantorovich potential φ_ξ.
3. The alternative training process of W-GAN is the min-max optimization of expectations:

$$\min_\theta \max_\xi \mathbb{E}_{z\sim\zeta}(\varphi_\xi(g_\theta(z))) + \mathbb{E}_{y\sim\nu}(\varphi_\xi^c(y)).$$

The deep neural network of D computes the Wasserstein distance by the maximization process to approximate Kantorovich potentials φ_ξ, parameterized by ξ; the network G calculates the optimal transportation map g_θ by the minimization process, parameterized by θ (see Fig. 1).

The GAN model and the convex geometry are connected by the optimal transportation theory (see Fig. 2).

Geometric interpretation The Optimal Mass Transportation theory has intrinsic connections with the convex geometry. A special type of OMT problem is equivalent to the Alexandrov problem in convex geometry, specifically, finding the optimal transportation map with L^2 cost is equivalent to constructing a convex polytope with user prescribed normals and face volumes. The geometric view leads to a practical algorithm, which finds the generating map g_θ by a convex optimization. Furthermore, the optimization can be carried out using Newton's method with explicit geometric meaning. The geometric interpretation also gives the direct relation between the transportation map g_θ for G and the Kantorovich potential φ_ξ for D.

These concepts can be explained using the plain language in computational geometry (Edelsbrunner, 1987),

1. the Kantorovich potential φ_ξ corresponds to the power distance;
2. the optimal transportation map g_θ represents the mapping from the power diagram to the power centers, each power cell is mapped to the corresponding site.

Imaginary adversary In the current work, we use Optimal Mass Transportation theory to show the fact that: by carefully designing the model and choosing special distance functions c, the generator map g_θ and the discriminator function (Kantorovich potential) φ_ξ are equivalent, one can be deduced from the other by a simple closed formula. Therefore, once the Kantorovich potential reaches the optimum, the generator map can be obtained directly without training. One of the deep neural networks for G or D is redundant, one of the training processes is wasteful. The competition between the generator G and the discriminator D is unnecessary, and imaginary.

Contributions The major contributions of the current work are as follows:

1. Based on the connection between convex geometry and Optimal Transportation, develop an explicit geometric construction for optimal transportation map for the purpose of Generative Adversarial Networks;
2. Demonstrate in Theorem 3.7 that if the cost function $c(x, y) = h(x - y)$, where h is a strictly convex function, then once the optimal discriminator is obtained, the generator can be written down in an explicit formula. In this situation, the competition between the discriminator and the generator is unnecessary and the computational architecture can be simplified;
3. Propose a novel framework for generative model, which uses geometric construction of the optimal mass transportation map;
4. Conduct preliminary experiments for the proof of concepts.

Organization The article is organized as follows: section 2 explains the Optimal Mass Transportation view of WGAN in details; section 3 lists the main theory of OMT; section 4 gives the detailed exposition of Minkowski and Alexandrov theorems in convex geometry, and its close relation with power diagram theory in computational geometry, an explicit computational algorithm is given to solve Alexandrov's problem; section 5 analyzes semi-discrete optimal transportation problem, and connects Alexandrov problem with the optimal transportation map; preliminary experiments are conducted for proof of concept, which are reported in section 6. The work concludes in the section 7.

2. Optimal transportation view of GAN

In this section, the GAN model is interpreted from the optimal transportation point of view. We show that the discriminator mainly looks for the Kantorovich potential.

Let $\mathcal{X} \subset \mathbb{R}^n$ be the (ambient) image space, $\mathcal{P}(\mathcal{X})$ be the Wasserstein space of all probability measures on $\mathcal{X}$. Assume the *real data distribution* is $\nu \in \mathcal{P}(\mathcal{X})$, in practice approximated by an *empirical distribution*

$$\nu := \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j}, \tag{1}$$

where $y_j \in \mathcal{X}$, $j = 1, \ldots, n$ are data samples, δ_{y_j} is the Dirac function. A *generative model* produces a parametric family of probability distributions μ_θ, $\theta \in \Theta$, a Minimum Kantorovitch Estimator for θ is defined as any solution to the problem

$$\min_\theta W_c(\mu_\theta, \nu),$$

where W_c is the Kantorovich cost on $\mathcal{P}(\mathcal{X})$ and the ground cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. When c is a power of the Euclidean distance, W_c is the Wasserstein distance between μ and ν,

$$W_c(\mu, \nu) = \min_{\rho \in \mathcal{P}(\mathcal{X}\times\mathcal{X})} \left\{ \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\rho(x, y) | \pi_{x\#}\rho = \mu, \pi_{y\#}\rho = \nu \right\} \tag{2}$$

where π_x and π_y are projectors, $\pi_{x\#}$ and $\pi_{y\#}$ are marginalization operators. In a generative model, the image samples are encoded to a low dimensional latent space (or a feature space) $\mathcal{Z} \subset \mathbb{R}^m$, $m \ll n$. Let ζ be a fixed distribution supported on $\mathcal{Z}$. A Wasserstein GAN (WGAN) produces a parametric mapping $g_\theta : \mathcal{Z} \to \mathcal{X}$, which is treated as a "decoding map" from the latent space $\mathcal{Z}$ to the original image space $\mathcal{X}$. g_θ pushes ζ forward to $\mu_\theta \in \mathcal{P}(\mathcal{X})$, $\mu_\theta = g_{\theta\#}\zeta$. The minimal Kantorovich estimator in WGAN is formulated as

$$\min_\theta E(\theta) := \min_\theta W_c(g_{\theta\#}\zeta, \nu).$$

According to the optimal transportation theory (Villani, 2003), the Kantorovich problem has a dual formulation

$$E(\theta) = \max_{\varphi,\psi} \left\{ \int_{\mathcal{Z}} \varphi(g_\theta(z)) d\zeta(z) + \int_{\mathcal{X}} \psi(y) d\nu(y) : \varphi(x) + \psi(y) \le c(x, y) \right\} \tag{3}$$

The gradient of the dual energy with respect to θ can be written as

$$\nabla E(\theta) = \int_{\mathcal{Z}} [\partial_\theta g_\theta(z)]^T \nabla \varphi^\star(g_\theta(z)) d\zeta(z),$$

where $\varphi^\star$ is the optimal Kantorovich potential. In practice, ψ can be replaced by the c-transform of $\varphi^\star$, defined as

$$\varphi^c(y) := \inf_x c(x, y) - \varphi(x).$$

The function φ is called the *Kantorovich potential*. According to the optimal transportation theory (Villani, 2003, 2008), $\psi = \varphi^c$ and symmetrically $\varphi = \psi^c$. Since ν is discrete, ψ is just defined on the support $Y := \{y_i\}$ of ν, and $\psi^c = \varphi$. Let $\psi(y_i) = \psi_i$, the optimization over $\{\psi_i\}$ can then be achieved using stochastic gradient descent, as in Genevay et al. (2016).

In WGAN (Arjovsky et al., 2017), the dual problem Eqn. (3) is solved by approximating the Kantorovich potential φ by the so-called "adversarial" map, $\varphi_\xi : \mathcal{X} \to \mathbb{R}$, where ξ is represented by a discriminative deep network. This leads to the Wasserstein-GAN problem

$$\min_\theta \max_\xi \int_{\mathcal{Z}} \varphi_\xi \circ g_\theta(z) d\zeta(z) + \frac{1}{n} \sum_{j=1}^n \varphi_\xi^c(y_j). \tag{4}$$

The generator produces g_θ, the discriminator estimates φ_ξ, by simultaneous training, the competition reaches the equilibrium. In WGAN (Arjovsky et al., 2017), $c(x, y) = |x - y|$, then the c-transform of φ_ξ equals to $-\varphi_\xi$, subject to φ_ξ being a 1-Lipschitz function. This is used to replace φ_ξ^c by $-\varphi_\xi$ in Eqn. (4) and use deep network made of ReLu units whose Lipschitz constant is upper-bounded by 1.

3. Optimal mass transport theory

In this section, we give a brief review for the classical optimal mass transportation theory for engineering purposes, neglecting the technical and delicate aspects of the theory, such as the conditions for the existence of a feasible plan, the existence of optimal Kantorovich potentials and their regularities. For more rigorous and thorough treatments, we refer readers to Villani's books (Villani, 2003, 2008). Theorem 3.7 shows the intrinsic relation between the Wasserstein distance (equivalent to the Kantorovich potential) and the optimal transportation map (equivalent to the Brenier potential), this demonstrates that once the optimal discriminator is known, the generator is automatically obtained. The game between the discriminator and the generator is unnecessary.

The problem of finding a map that minimizes the inter-domain transportation cost while preserves measure quantities was first studied by Bonnotte (2012) in the 18th century. Let X and Y be two metric spaces with probability measures μ and ν respectively. Assume X and Y have equal total measure

$$\int_X d\mu = \int_Y d\nu.$$

Definition 3.1 *(Push-forward measure).* A map $T : X \to Y$ is given, if for any measurable set $B \subset Y$,

$$\mu(T^{-1}(B)) = \nu(B), \tag{5}$$

then ν is said to be the push-forward of μ by T, and we write $\nu = T_\#\mu$.

If the mapping $T : X \to Y$ is differentiable, X and Y are the same Euclidean space $\mathbb{R}^d$, μ and ν have Lebesgue densities, which are identified with the measures themselves, then Eqn. (5) can be formulated as the following Jacobian equation, $\mu(x)dx = \nu(T(x))dT(x)$,

$$det(DT(x)) = \frac{\mu(x)}{\nu \circ T(x)}. \tag{6}$$

Let us denote the transportation cost for sending $x \in X$ to $y \in Y$ by $c(x, y)$, then the total *transportation cost* is given by

$$\mathcal{C}(T) := \int_X c(x, T(x))d\mu(x). \tag{7}$$

Problem 3.2 *(Monge's Optimal Mass Transportation, Bonnotte, 2012).* Given measures μ and ν, and a transportation cost function $c : X \times Y \to \mathbb{R}$, find the transportation map $T : X \to Y$ that minimizes the total transportation cost

$$(MP) \qquad W_c(\mu, \nu) = \inf_{T:X\to Y} \left\{ \int_X c(x, T(x))d\mu(x) : T_{\#}\mu = \nu \right\}. \tag{8}$$

If c is the power of the Euclidean distance, the total transportation cost $W_c(\mu, \nu)$ is called the *Wasserstein distance* between the two measures μ and ν.

3.1. Kantorovich's approach

In the Monge formulation the infimum is not attained in general. In the 1940s, Kantorovich introduced the relaxation of Monge's problem (Kantorovich, 1948). Any strategy sending μ onto ν can be represented by a joint measure ρ on $X \times Y$, such that for every A, B Borel subsets of X, Y respectively,

$$\rho(A \times Y) = \mu(A), \rho(X \times B) = \nu(B), \tag{9}$$

$\rho(A \times B)$ is called a *transportation plan*, which represents the share to be moved from A to B. We denote the projection to X and Y as π_x and π_y respectively, then $\pi_{x\#}\rho = \mu$ and $\pi_{y\#}\rho = \nu$. The total cost of the transportation plan ρ is

$$\mathcal{C}(\rho) := \int_{X\times Y} c(x, y)d\rho(x, y). \tag{10}$$

The Monge–Kantorovich problem consists in finding the ρ, among all the suitable transportation plans with marginals μ and ν, minimizing $\mathcal{C}(\rho)$ in Eqn. (10),

$$(KP) \qquad W_c(\mu, \nu) := \min_{\rho} \left\{ \int_{X\times Y} c(x, y)d\rho(x, y) : \pi_{x\#}\rho = \mu, \pi_{y\#}\rho = \nu \right\} \tag{11}$$

When μ is a diffuse measure (i.e. $\mu(\{x\}) = 0$ for every $x \in X$) and c is continuous, the infimum of (MP) is equals to the minimum of (KP).

3.2. Kantorovich dual formulation

Because Eqn. (11) is a linear program, it has a dual formulation, known as the Kantorovich problem (Villani, 2008):

$$(DP) \qquad W_c(\mu, \nu) := \max_{\varphi,\psi} \left\{ \int_X \varphi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) : \varphi(x) + \psi(y) \le c(x, y) \right\} \tag{12}$$

where $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ are real functions defined on X and Y respectively. Equivalently, we can replace ψ by the c-transform of φ.

Definition 3.3 *(c-transform).* Given a real function $\varphi : X \to \mathbb{R}$, the c-transform of φ is defined by

$$\varphi^c(y) = \inf_{x\in X} \left(c(x, y) - \varphi(x)\right).$$

Then the Kantorovich problem can be reformulated as the following dual problem:

$$(DP) \qquad W_c(\mu,\nu) := \max_{\varphi}\left\{\int_X \varphi(x)d\mu(x) + \int_Y \varphi^c(y)d\nu(y)\right\}, \tag{13}$$

any optimal φ where the maximum is attained in Eqn. (13) is called a *Kantorovich potential*. φ and ψ plays a symmetric role in Eqn. (12), ψ can be treated as the Kantorovich potential as well. Computing the Wasserstein distance is equivalent to finding a Kantorovich potential.

When $X = Y$, for L^1 transportation cost $c(x,y) = |x-y|$ in $\mathbb{R}^n$, if the Kantorovich potential φ is 1-Lipschitz, then its c-transform has a special relation $\varphi^c = -\varphi$. The Wasserstein distance is given by

$$W_c(\mu,\nu) := \max_{\varphi}\left\{\int_X \varphi(x)d\mu(x) - \int_Y \varphi(y)d\nu(y)\right\}. \tag{14}$$

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

Definition 3.4. Given a function $\varphi : \mathbb{R}^n \to \mathbb{R}$, its Legendre transform is defined as

$$\varphi^*(y) := \sup_x \left(\langle x, y\rangle - \varphi(x)\right). \tag{15}$$

We can show the following relation holds when $c = 1/2|x-y|^2$,

$$\frac{1}{2}|y|^2 - \varphi^c = \left(\frac{1}{2}|x|^2 - \varphi\right)^*. \tag{16}$$

3.3. Brenier's approach

At the end of 1980's, Brenier (1991) discovered the intrinsic connection between optimal mass transport map and convex geometry (see also for instance Villani, 2003, Theorem 2.12(ii), and Theorem 2.32).

Assume $X = Y = \mathbb{R}^n$, function $u : X \to \mathbb{R}$ is a C^2 continuous convex function, namely its Hessian matrix is semi-positive definite. Its gradient map $\nabla u : X \to Y$ is defined as $x \mapsto \nabla u(x)$.

Theorem 3.5 *(Brenier, 1991). Suppose X and Y are the Euclidean space $\mathbb{R}^n$, and the transportation cost is the quadratic Euclidean distance $c(x,y) = |x-y|^2$. If μ is absolutely continuous and μ and ν have finite second order moments, then there exists a convex function $u : X \to \mathbb{R}$, such that the gradient map ∇u gives the unique solution to the Monge's problem, where u is called Brenier's potential, ∇u is called Brenier map or the optimal mass transportation map. In general, u is not unique.*

This theorem converts the Monge's problem to solving the following Monge–Ampère partial differential equation:

$$det\left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right)(x) = \frac{\mu(x)}{\nu \circ \nabla u(x)}. \tag{17}$$

The function $u : X \to \mathbb{R}$ is called the *Brenier potential*. Brenier proved the polar factorization theorem.

Theorem 3.6 *(Brenier factorization, Brenier, 1991). Suppose X and Y are the Euclidean space $\mathbb{R}^n$, μ is absolutely continuous with respect to Lebesgue measure, a mapping $\varphi : X \to Y$ pushes μ forward to ν, $\varphi_{\#}\mu = \nu$. 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 optimal transportation theory, the proof can be found in Villani's book (Villani, 2003) and in the book of Ambrosio et al. (2008). We apply this theorem to Deep Learning and show that the generator and the discriminator in WGAN model with L^2 cost are equivalent. For the completeness, we give the detailed proof here.

Theorem 3.7 *(Generator–discriminator equivalence). Given μ and ν on a compact domain $\Omega \subset \mathbb{R}^n$ there exists an optimal transport plan ρ for the cost $c(x, y) = h(x-y)$ with h strictly convex. It is unique and of the form $(id, T_{\#})\mu$, provided μ is absolutely continuous and $\partial\Omega$ is negligible. More over, there exists a Kantorovich potential φ, and T can be represented as*

$$T(x) = x - (\nabla h)^{-1}(\nabla\varphi(x)).$$

Proof. Assume ρ is the joint probability, satisfying the conditions $\pi_{x\#}\rho = \mu$, $\pi_{y\#}\rho = \nu$, (x_0, y_0) is a point in the support of ρ, by definition $\varphi^c(y_0) = \inf_x(c(x, y_0) - \varphi(x))$, hence $\nabla_x(c(x, y_0) - \varphi(x))|_{x=x_0} = 0$,

$$\nabla\varphi(x_0) = \nabla_x c(x_0, y_0) = \nabla h(x_0 - y_0).$$

Because h is strictly convex, therefore ∇h is invertible,

$$x_0 - y_0 = (\nabla h)^{-1}(\nabla\varphi(x_0)),$$

hence $y_0 = x_0 - (\nabla h)^{-1}(\nabla\varphi(x_0))$. □

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

$$T(x) = x - \nabla\varphi(x) = \nabla\left(\frac{x^2}{2} - \varphi(x)\right) = \nabla u(x).$$

In this case, the Brenier's potential u and the Kantorovich's potential φ is related by

$$u(x) = \frac{x^2}{2} - \varphi(x). \tag{18}$$

As discussed in section 2, in Wasserstein GAN framework, the discriminator D computes the Wasserstein distance, which is equivalent to find the Kantorovich potential φ; the generator G computes the optimal transportation map ∇u, which is equivalent to find the Brenier potential u. This shows the computational results of the discriminator and the generator are closely related, the result obtained by the discriminator can be used directly by the generator, and vice versa.

Corollary 3.8. *Under the conditions of Brenier theorem, the Monge–Kantorovich problem*

$$\min_{\rho}\left\{\int_{X\times Y} c(x, y)d\rho(x, y) : (\pi_x)_{\#}\rho = \mu, (\pi_y)_{\#}\rho = \nu\right\},$$

is equivalent to the Kantorovich dual problem,

$$\max_{\varphi,\psi}\left\{\int_X \varphi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) : \varphi(x) + \psi(y) \le c(x, y)\right\}.$$

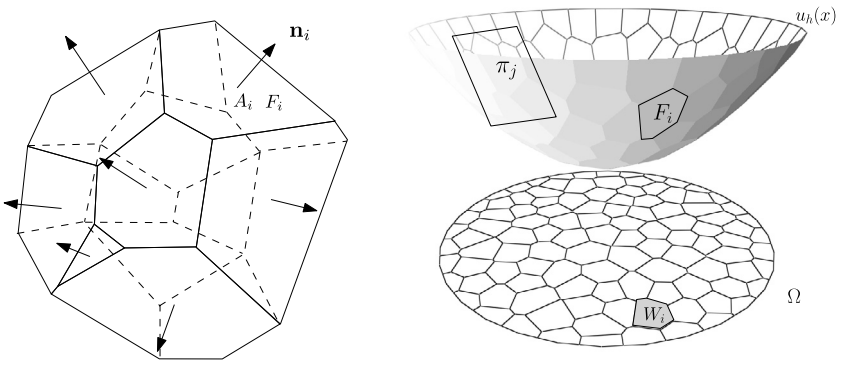
Proof. The quadratic cost of a general transport plan can be written as

$$\begin{aligned}\frac{1}{2}\int_{X\times Y}|x-y|^2d\rho &= \frac{1}{2}\int_{X\times Y}|x|^2d\rho + \frac{1}{2}\int_{X\times Y}|y|^2d\rho - \int_{X\times Y}\langle x, y\rangle d\rho\\ &= \frac{1}{2}\int_X |x|^2d\mu(x) + \frac{1}{2}\int_Y |y|^2d\nu(y) - \int_{X\times Y}\langle x, y\rangle d\rho\end{aligned}$$

up to substracting the quadratic moments of μ and ν, it is equivalent to minimize the cost $c(x, y) = -\langle x, y\rangle$. The inverse of the optimal transportation map $T : X \to Y$, $T^{-1} : Y \to X$ is also optimal, by Briener theorem, there exists a convex function $v : Y \to \mathbb{R}$, such that $\nabla v = T^{-1}$. Furthermore, the following relations hold

$$u(x) = \frac{1}{2}|x|^2 - \varphi(x),\ v(y) = \frac{1}{2}|y|^2 - \psi(y).$$

A similar decomposition holds for the dual problem:



(a) Minkowski theorem (b) Alexandrov theorem

Fig. 3. Minkowski and Alexandrov theorems for convex polytopes with prescribed normals and areas.

$$\int_X \varphi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) = \int_X \left(\frac{1}{2}|x|^2 - u(x)\right)d\mu(x) + \int_Y \left(\frac{1}{2}|y|^2 - v(y)\right)d\nu(y)$$

$$= \frac{1}{2}\int_X |x|^2 d\mu(x) + \frac{1}{2}\int_Y |y|^2 d\nu(y) - \left(\int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)\right)$$

$$= \frac{1}{2}\int_X |x|^2 d\mu(x) + \frac{1}{2}\int_Y |y|^2 d\nu(y) - \int_{X\times Y} (u(x) + v(y))d\rho$$

from the condition

$$\varphi(x) + \psi(y) \le \frac{1}{2}|x - y|^2$$

we obtain

$$u(x) + v(y) = \frac{1}{2}|x|^2 - \varphi(x) + \frac{1}{2}|y|^2 - \psi(y)$$
$$\ge \langle x, y\rangle$$

Hence

$$\int_X \varphi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) \le \frac{1}{2}\int_X |x|^2 d\mu(x) + \frac{1}{2}\int_Y |y|^2 d\nu(y) - \int_{X\times Y} \langle x, y\rangle d\rho \quad \square$$

4. Convex geometry

This section introduces Minkowski and Alexandrov problems in convex geometry, which can be described by Monge–Ampère equation as well. This intrinsic connection gives a geometric interpretation to optimal mass transportation map with L^2 transportation cost.

4.1. Alexandrov's theorem

Minkowski proved the existence and the uniqueness of convex polytope with user prescribed face normals and areas (see Fig. 3 left frame).

Theorem 4.1 *(Minkowski). Suppose $n_1, \ldots, n_k$ are unit vectors which span $\mathbb{R}^n$ and $\nu_1, \ldots, \nu_k > 0$ so that $\sum_{i=1}^k \nu_i n_i = 0$. There exists a compact convex polytope $P \subset \mathbb{R}^n$ with exactly k codimension-1 faces $F_1, \ldots, F_k$ so that n_i is the outward normal vector to F_i and the volume of F_i is ν_i. Furthermore, such P is unique up to parallel translation.*

Minkowski's proof is variational and suggests an algorithm to find the polytope. Minkowski theorem for unbounded convex polytopes was considered and solved by A.D. Alexandrov and his student A. Pogorelov. In his book on convex polyhedra (Alexandrov, 2005), Alexandrov proved the following fundamental theorem (Theorem 7.3.2 and theorem 6.4.2 in Alexandrov, 2005):

Theorem 4.2 *(Alexandrov, 2005). Suppose Ω is a compact convex polytope with non-empty interior in $\mathbb{R}^n$, $n_1, ..., n_k \subset \mathbb{R}^{n+1}$ are distinct k unit vectors, the $(n+1)$-th coordinates are negative, and $\nu_1, ..., \nu_k > 0$ so that $\sum_{i=1}^k \nu_i = vol(\Omega)$. Then there exists convex polytope $P \subset \mathbb{R}^{n+1}$ with exact k codimension-1 faces $F_1, \ldots, F_k$, so that n_i is the normal vector to F_i and the intersection between Ω and the projection of F_i is with volume ν_i. Furthermore, such P is unique up to vertical translation (see Fig. 3 right frame).*

Alexandrov's proof is based on algebraic topology and non-constructive. Gu et al. (2016) gave a variational proof for the generalized Alexandrov theorem stated in terms of convex functions.

Given $y_1, \ldots, y_k \in \mathbb{R}^n$ and $h = (h_1, \ldots, h_k) \in \mathbb{R}^k$, we define a piecewise linear convex function $u_h(x)$ as

$$u_h(x) = \max_{i=1}^{k} \{\langle x, y_i \rangle + h_i\}.$$

The graph of u_h is a convex polytope in $\mathbb{R}^{n+1}$, the projection induces a cell decomposition of $\mathbb{R}^n$. Each cell is a closed convex polytope,

$$W_i(h) = \left\{x \in \mathbb{R}^n | \nabla u_h(x) = y_i\right\}.$$

Some cells may be empty or unbounded. Given a probability measure μ defined on Ω, the μ-volume of $W_i(h)$ is defined as

$$w_i(h) := \mu(W_i(h) \cap \Omega) = \int_{W_i(h)\cap\Omega} d\mu.$$

Theorem 4.3 *(Gu et al., 2016). Let Ω be a compact convex domain in $\mathbb{R}^n$, $\{y_1, ..., y_k\}$ be a set of distinct points in $\mathbb{R}^n$ and μ a probability measure on Ω. Then for any $\nu_1, ..., \nu_k > 0$ with $\sum_{i=1}^k \nu_i = \mu(\Omega)$, there exists $h = (h_1, ..., h_k) \in \mathbb{R}^k$, unique up to adding a constant $(c, ..., c)$, so that $w_i(h) = \nu_i$, for all i. The vectors h are exactly maximum points of the concave function*

$$E(h) = \sum_{i=1}^{k} h_i \nu_i - \int_0^h \sum_{i=1}^{k} w_i(\eta) d\eta_i \tag{19}$$

on the open convex set

$$H = \{h \in \mathbb{R}^k | w_i(h) > 0, \forall i\}.$$

Furthermore, ∇u_h minimizes the quadratic cost

$$\int_\Omega |x - T(x)|^2 d\mu(x)$$

among all transport maps $T_\# \mu = \nu$, where the Dirac measure $\nu = \sum_{i=1}^k \nu_i \delta(y - y_i)$.

For the convenience of discussion, we define the Alexandrov's potential as follows:

Definition 4.4 *(Alexandrov potential).* Under the above condition, the convex function

$$\mathcal{A}(h) = \int^h \sum_{i=1}^{k} w_i(\eta) d\eta_i \tag{20}$$

is called the Alexandrov potential.

We define the admissible space for the height vector $\mathbf{h}$

$$\mathcal{H} := \left\{\mathbf{h} \in \mathbb{R}^k | w_i(\mathbf{h}) > 0\right\} \cap \left\{\sum_{i=1}^{k} h_k = 0\right\}$$

By Brunn–Minkowski inequality, we can show that $\mathcal{H}$ is a convex open set in $\mathbb{R}^k$.

From Eqn. (24), it is easy to show the following symmetric relation:

$$\frac{\partial w_i(h)}{\partial h_j} = \frac{\partial w_j(h)}{\partial h_i},$$

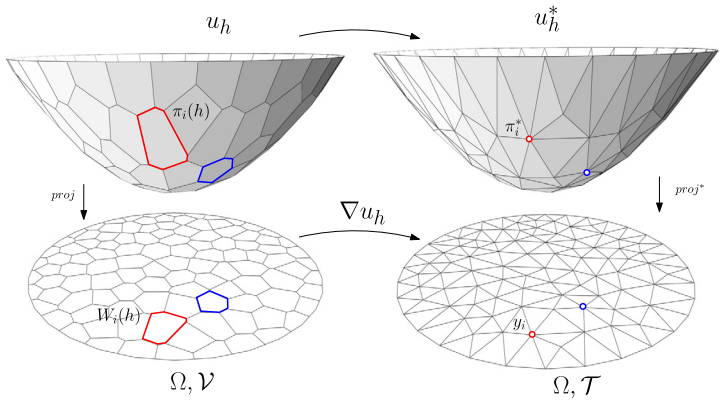


Fig. 4. Geometric Interpretation to Optimal Transport Map: Brenier potential $u_h : \Omega \to \mathbb{R}$, Legendre dual u_h^*, optimal transportation map $\nabla u_h : W_i(h) \to y_i$, power diagram $\mathcal{V}$, weighted Delaunay triangulation $\mathcal{T}$.

therefore the differential form

$$\omega := \sum_{i=1}^{k} w_i(\mathbf{h}) dh_i$$

is a closed one form. Because $\mathcal{H}$ is simply connected, ω is exact, hence the Alexandrov potential Eqn. (20) is well-defined. The geometric meaning of $\mathcal{A}(\mathbf{h})$ is the volume under the upper envelope. Details can be found in Gu et al. (2016).

4.2. Power diagram

Alexandrov's theorem has close relation with the conventional power diagram. We can use power diagram algorithm to solve the Alexandrov's problem.

Definition 4.5 *(power distance).* Given a point $y_i \in \mathbb{R}^n$ with a power weight ψ_i, the power distance is given by

$$\mathrm{pow}(x, y_i) = \frac{1}{2}|x - y_i|^2 - \psi_i.$$

Definition 4.6 *(power diagram).* Given weighted points $\{(y_1, \psi_1), (y_2, \psi_2), \ldots, (y_k, \psi_k)\}$, the power diagram is the cell decomposition of $\mathbb{R}^n$, denoted as $\mathcal{V}(\psi)$,

$$\mathbb{R}^n = \bigcup_{i=1}^{k} W_i(\psi),$$

where each cell is a convex polytope

$$W_i(\psi) = \{x \in \mathbb{R}^n | \mathrm{pow}(x, y_i) \leq \mathrm{pow}(x, y_j), \forall j\}.$$

The weighted Delaunay triangulation, denoted as $\mathcal{T}(\psi)$, is the Poincaré dual to the power diagram, if $W_i(\psi) \cap W_j(\psi) \neq \emptyset$ then there is an edge connecting y_i and y_j in the weighted Delaunay triangulation (see Fig. 5).

Note that $\mathrm{pow}(x, y_i) \leq \mathrm{pow}(x, y_j)$ is equivalent to

$$\langle x, y_i \rangle + \frac{1}{2}(\psi_i - |y_i|^2) \geq \langle x, y_j \rangle + \frac{1}{2}(\psi_j - |y_j|^2),$$

let

$$h_i = 1/2(\psi_i - |y_i|^2), \tag{21}$$

then $\mathrm{pow}(x, y_i) \leq \mathrm{pow}(x, y_j)$ is equivalent to $\langle x, y_i \rangle + h_i \geq \langle x, y_j \rangle + h_j$. The *upper envelope* of the planes $\{\langle x, y_i \rangle + h_i = 0\}$ is the graph of the convex function

$$u_h(x) = \max_i \{\langle x, y_i \rangle + h_i\}. \tag{22}$$

The projection of the graph of u_h gives the power diagram $\mathcal{V}(\psi)$.

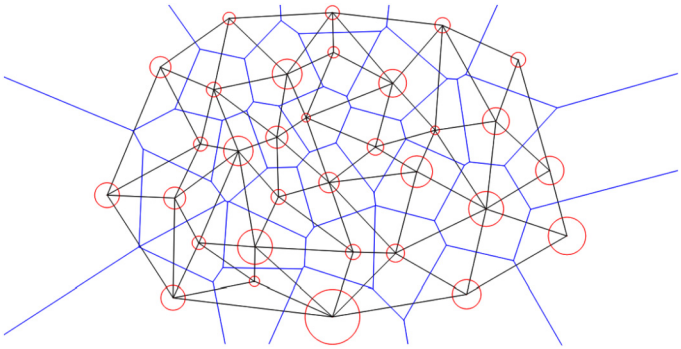


Fig. 5. Power diagram (blue) and its dual weighted Delaunay triangulation (black), the power weight ψ_i equal to the square of radius r_i (red circle). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

4.3. Convex optimization

Now, we can use the power diagram to explain the gradient and the Hessian of the energy Eqn. (19), by definition

$$\nabla E(h) = (\nu_1 - w_1(h), \nu_2 - w_2(h), \cdots, \nu_k - w_k(h))^T. \tag{23}$$

The Hessian matrix is given by power diagram – weighted Delaunay triangulation, for adjacent cells in the power diagram,

$$\frac{\partial^2 E(h)}{\partial h_i \partial h_j} = \frac{\partial w_i(h)}{\partial h_j} = -\frac{\mu(W_i(h) \cap W_j(h) \cap \Omega)}{|y_j - y_i|} \tag{24}$$

Suppose edge e_{ij} is in the weighted Delaunay triangulation, connecting y_i and y_j. It has a unique dual cell D_{ij} in the power diagram, then

$$\frac{\partial w_i(h)}{\partial h_j} = -\frac{\mu(D_{ij})}{|e_{ij}|},$$

the volume ratio between the dual cells. The diagonal element in the Hessian is

$$\frac{\partial^2 E(h)}{\partial h_i^2} = \frac{\partial w_i(h)}{\partial h_i} = -\sum_{j \neq i} \frac{\partial w_i(h)}{\partial h_j}. \tag{25}$$

Therefore, in order to solve Alexandrov's problem to construct the convex polytope with user prescribed normal and face volume, we can optimize the energy in Eqn. (19) using classical Newton's method directly.

Let's observe the convex function u_h^*, its graph is the convex hull $\mathcal{C}(h)$. Then the discrete Hessian determinant of u_h^* assigns each vertex v of $\mathcal{C}(h)$ the volume of the convex hull of the gradients of u_h^* at top-dimensional cells adjacent to v. Therefore, solving Alexandrov's problem is equivalent to solve a discrete Monge–Ampère equation.

5. Semi-discrete optimal mass transport

In this section, we solve the semi-discrete optimal transportation problem from geometric point of view. This special case is useful in practice.

Suppose the closure of μ, Ω is a compact convex subset of the Euclidean space X,

$$\Omega = \overline{\text{supp}\ \mu} = \overline{\{x \in X | \mu(x) > 0\}}.$$

The space Y is discretized to $Y = \{y_1, y_2, \cdots, y_k\}$ with Dirac measure $\nu = \sum_{j=1}^k \nu_j \delta_{y_j}$. The total mass is equal

$$\int_\Omega d\mu(x) = \sum_{i=1}^k \nu_i.$$

5.1. Kantorovich dual approach

We define the discrete Kantorovich potential $\psi : Y \to \mathbb{R}$, $\psi_j := \psi(y_j)$, here Y is a discrete set in $X = \mathbb{R}^n$, then

$$\int_Y \psi d\nu = \sum_{j=1}^k \psi_j \nu_j. \tag{26}$$

The c-transformation of ψ is given by

$$\psi^c(x) = \min_{1\le j\le k}\{c(x, y_j) - \psi_j\}. \tag{27}$$

This induces a cell decomposition of X,

$$X = \bigcup_{i=1}^{k} W_i(\psi),$$

where each cell is given by

$$W_i(\psi) = \left\{x \in X | c(x, y_i) - \psi_i \le c(x, y_j) - \psi_j, \forall 1 \le j \le k\right\}.$$

According to the dual formulation of the Wasserstein distance Eqn. (13) and integration Eqn. (26), we define the energy

$$E(\psi) = \int_X \psi^c d\mu + \int_Y \psi d\nu$$

then obtain the formula

$$E(\psi) = \sum_{i=1}^{k} \psi_i (\nu_i - w_i(\psi)) + \sum_{j=1}^{k} \int_{W_j(\psi)} c(x, y_j) d\mu, \tag{28}$$

where $w_i(\psi)$ is the measure of the cell $W_i(\psi)$,

$$w_i(\psi) = \mu(W_i(\psi)) = \int_{W_i(\psi)} d\mu(x). \tag{29}$$

Then the Wasserstein distance between μ and ν equals to

$$W_c(\mu, \nu) = \max_{\psi} E(\psi).$$

5.2. Brenier's approach

Kantorovich's dual approach is for general cost functions. When the cost function is the L^2 distance $c(x, y) = 1/2|x - y|^2$, we can apply Brenier's approach directly.

We define a *height vector* $h = (h_1, h_2, \cdots, h_k) \in \mathbb{R}^n$, consisting of k real numbers. For each $y_i \in Y$, we construct a hyperplane defined on X, $\pi_i(h) : \langle x, y_i\rangle + h_i = 0$. We define the Brenier potential function as

$$u_h(x) = \max_{i=1}^{k}\{\langle x, y_i\rangle + h_i\}, \tag{30}$$

then $u_h(x)$ is a convex function. The graph of $u_h(x)$ is an infinite convex polyhedron with supporting planes $\pi_i(h)$. The projection of the graph induces a polygonal partition of Ω,

$$\Omega = \bigcup_{i=1}^{k} W_i(h), \tag{31}$$

where each cell $W_i(h)$ is the projection of a facet of the graph of u_h onto Ω,

$$W_i(h) = \{x \in X | \nabla u_h(x) = y_i\} \cap \Omega. \tag{32}$$

The measure of $W_i(h)$ is given by

$$w_i(h) = \int_{W_i(h)} d\mu. \tag{33}$$

The convex function u_h on each cell $W_i(h)$ is a linear function $\pi_i(h)$, therefore, the gradient map

$$\nabla u_h : W_i(h) \to y_i, i = 1, 2, \cdots, k, \tag{34}$$

maps each $W_i(h)$ to a single point y_i. According to Alexandrov's theorem, and the Gu–Luo–Yau theorem, we obtain the following corollary:

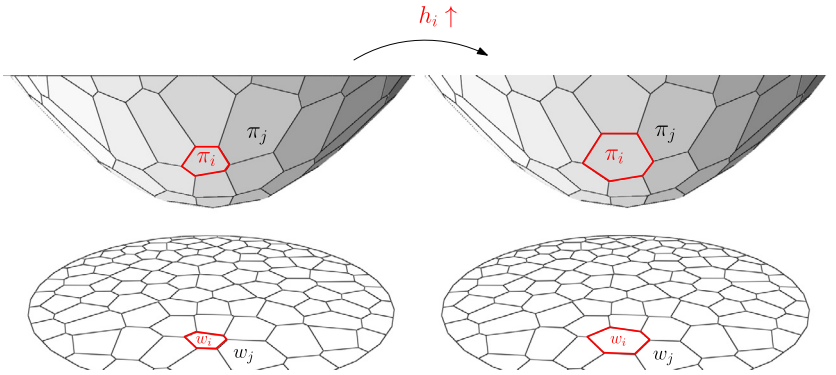


Fig. 6. Variation of the volume of top-dimensional cells.

Corollary 5.1. *Let Ω be a compact convex domain in $\mathbb{R}^n$, $\{y_1, ..., y_k\}$ be a set of distinct points in $\mathbb{R}^n$ and μ a probability measure on Ω. Then for any $\nu = \sum_{i=1}^k \nu_i \delta_{y_i}$, with $\sum_{i=1}^k \nu_i = \mu(\Omega)$, there exists $h = (h_1, ..., h_k) \in \mathbb{R}^k$, unique up to adding a constant $(c, ..., c)$, so that $w_i(h) = \nu_i$, for all i. The vectors h are exactly maximum points of the concave function*

$$E_B(h) = \sum_{i=1}^k h_i \nu_i - \int_0^h \sum_{i=1}^k w_i(\eta) d\eta_i \tag{35}$$

Furthermore, $T = \nabla u_h$ minimizes the quadratic cost

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

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

Fig. 6 shows the top-dimensional cell volumes depend on the heights smoothly.

5.3. Equivalence

For $c(x, y) = 1/2|x - y|^2$ cost case, we have introduced two approaches: Kantorovich's dual approach and Brenier's approach. In the following, we show these two approaches are equivalent.

In Kantorovich's dual approach, finding the optimal mass transportation is equivalent to maximize the following energy:

$$E_D(\psi) = \sum_{i=1}^k \psi_i(\nu_i - w_i(\psi)) + \sum_{j=1}^k \int_{W_j(\psi)} c(x, y_j) d\mu. \tag{36}$$

In Brenier's approach, finding the optimal transportation map boils down to maximize

$$E_B(h) = \sum_{i=1}^k h_i \nu_i - \int_0^h \sum_{i=1}^k w_i(\eta) d\eta. \tag{37}$$

Lemma 5.2. *Let Ω be a compact convex domain in $\mathbb{R}^n$, $\{y_1, ..., y_k\}$ be a set of distinct points in $\mathbb{R}^n$. Given μ a probability measure on Ω, $\nu = \sum_{i=1}^k \nu_i \delta_{y_i}$, with $\sum_{i=1}^k \nu_i = \mu(\Omega)$. If $c(x, y) = 1/2|x - y|^2$, then*

$$h_i = \psi_i - \frac{1}{2}|y_i|^2, \quad \forall i$$

and

$$E_D(\psi) - E_B(h) = Const$$

Here, E_D is the energy in Kantorovich's dual problem, E_B is the volume under the upper envelope (the graph of u_h).

Proof. Consider the power cell

$$c(x, y_i) - \psi_i \leq c(x, y_j) - \psi_j$$

is equivalent to

$$\langle x, y_i \rangle + \left(\psi_i - \frac{1}{2}|y_i|^2 \right) \geq \langle x, y_j \rangle + \left(\psi_j - \frac{1}{2}|y_j|^2 \right)$$

therefore

$$h_i = \psi_i - 1/2|y_i|^2. \tag{38}$$

Let the transportation cost to be defined as

$$\mathcal{C}(\psi) = \sum_{j=1}^{k} \int_{W_j(\psi)} c(x, y_j) d\mu,$$

then from Eqn. (36), we obtain

$$E_D(\psi) = \sum_{i=1}^{k} \psi_i (\nu_i - w_i(\psi)) + \mathcal{C}(\psi). \tag{39}$$

Suppose we infinitesimally change h to $h + dh$, then we define

$$D_{ij} = W_j(h) \cap W_i(h + dh) \cap \Omega.$$

Then $\mu(D_{ij}) = dw_i$, also $\mu(D_{ij}) = -dw_j$. For each $x \in D_{ij}$, $c(x, y_i) - \psi_i = c(x, y_j) - \psi_j$, then $c(x, y_i) - c(x, y_j) = \psi_i - \psi_j$, hence

$$\int_{D_{ij}} (c(x, y_i) - c(x, y_j)) d\mu = \psi_i dw_i + \psi_j dw_j.$$

This shows $d\mathcal{C} = \sum_{i=1}^{k} \psi_i dw_i$. Because the mapping $\psi \mapsto w$ is bijective, we change the parameter ψ to w and change $\mathcal{C}(\psi)$ to $\mathcal{C}(w)$, hence

$$\mathcal{C}(w) = \int^{w} \sum_{i=1}^{k} \psi_i dw_i.$$

Therefore, from Eqn. (39) we obtain

$$E_D(w) = \sum_{i=1}^{k} \psi_i(w)(\nu_i - w_i) + \mathcal{C}(w). \tag{40}$$

The Legendre dual of $\mathcal{C}(w)$ is

$$\mathcal{C}^*(\psi) = \int^{\psi} \sum_{i=1}^{k} w_i d\psi_i.$$

Hence by the definition of Legendre dual, we have

$$\mathcal{C}(w) + \mathcal{C}^*(\psi) = \sum_{i=1}^{k} w_i \psi_i. \tag{41}$$

From Eqn. (38), we change parameter h to ψ. By Eqn. (37), we get

$$E_B(\psi) = \sum_{i=1}^{k} h_i \nu_i - \mathcal{C}^*(\psi). \tag{42}$$

We put everything together, from Eqn. (40) and Eqn. (42),

$$\begin{aligned} E_D(w) - E_B(h) &= \left(\sum_{i=1}^{k} \psi_i(v_i - w_i) + \mathcal{C}(w)\right) - \left(\sum_{i=1}^{k} h_i v_i - \mathcal{C}^*(\psi)\right) \\ &= \sum_{i=1}^{k} (\psi_i - h_i) v_i - \left(\sum_{i=1}^{k} \psi_i w_i - \mathcal{C}(w) - \mathcal{C}^*(\psi)\right) \\ &= \frac{1}{2} \sum_{i=1}^{k} |y_i|^2 v_i, \end{aligned}$$

which is a constant. □

This shows Kantorovich's dual approach and Brienier's approach are equivalent. At the optimal point, $v_i = w_i(\psi)$, therefore $E_D(\psi)$ equals to the transportation cost $\mathcal{C}(\psi)$. Furthermore, the Brenier's potential is

$$u_h(x) = \max_{i=1}^{k} \{\langle x, p_i \rangle + h_i\},$$

where h_i is given by the power weight ψ_i. The Kantorovich's potential is the power distance

$$\varphi(x) = \psi^c(x) = \min_j \{c(x, y_j) - \psi_j\} = \min_j \{\mathrm{pow}(x, y_j)\} = \frac{1}{2}|x|^2 - \max_j \{\langle x, y_j \rangle + (\psi_j - \frac{1}{2}|y_j|^2)\}$$

hence at the optimum, the Brenier potential and the Kantorovich potential are related by

$$u_h(x) = \frac{1}{2}|x|^2 - \varphi(x). \tag{43}$$

The optimal transportation map is ∇u_h, which maps each power cell $W_j(\psi)$ to y_j.

6. Experiments

In order to demonstrate in principle the potential of our proposed method, we have designed and conducted the preliminary experiments.

6.1. Comparison with WGAN

In the first experiment, we use Wasserstein Generative Adversarial Networks (WGANs) (Arjovsky et al., 2017) to learn the mixed Gaussian distribution as shown in Fig. 7.

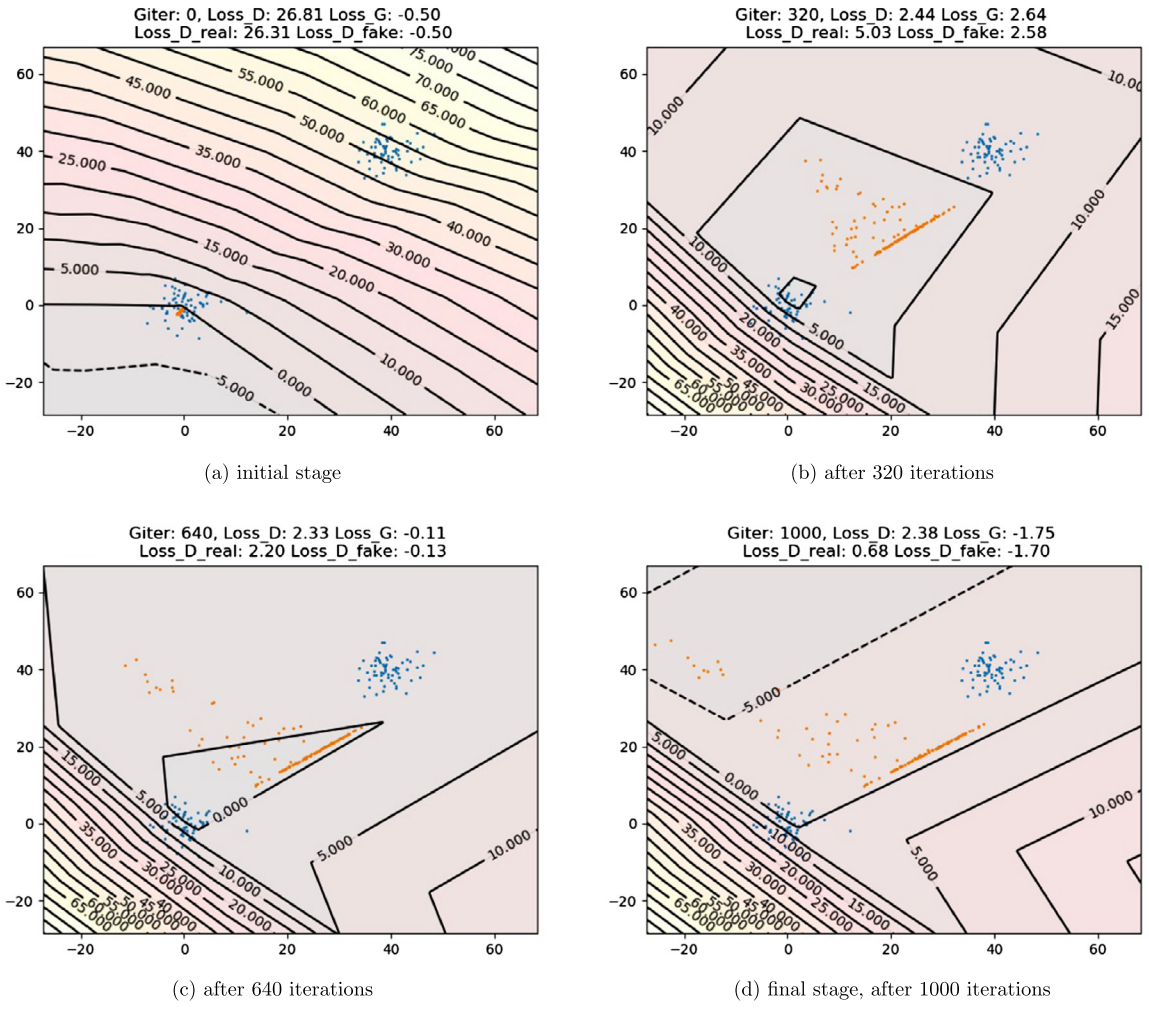
Dataset The distribution of data ν is described by a point cloud on a $2d$ plane. We sample 128 data points as real data from two Gaussian distributions, $\mathcal{N}(p_k, \sigma_k^2)$, $k = 1, 2$, where $p_1 = (0, 0)$ and $\sigma_1 = 3$, $p_2 = (40, 40)$ and $\sigma_2 = 3$. The latent space $\mathcal{Z}$ is a square on the $2d$ plane $[1k, 3k] \times [1k, 3k]$, the input distribution ζ is the uniform distribution on $\mathcal{Z}$. We generate 128 samples from ζ to approximate the data distribution ν.

Network structure The structure of the discriminator is 2-layer (2×10 FC)-ReLU-(10×1 FC) network, where FC denotes the fully connected layer. The number of inputs is 2 and the number of outputs is 1. The number of nodes of the hidden layer is 10.

The structure of the generator is a 6-layer (2×10 FC)-ReLU-(10×10 FC)-ReLU-(10×10 FC)-ReLU-(10×10 FC)-ReLU-(10×10 FC)-ReLU-(10×2 FC) network. The number of inputs is 2 and the number of outputs is 2. The number of nodes of all the hidden layer is 10.

Parameter setting For WGAN, we clip all the weights to $[-0.5, 0.5]$. We use the RMSprop (Hinton et al., 2012) as the optimizer for both discriminator and generator. The learning rate of both the discriminator and the generator are set to $1e-3$.

Deep learning framework and hardware We use the PyTorch (http://pytorch.org/) as our deep learning tool. Since the toy dataset is small, we do experiments on CPU. We perform experiments on a cluster with 48 cores and $193GB$ RAM. However, for this toy data, the running code only consumes 1 core with less than $500MB$ RAM, which means that it can run on a personal computer.



(a) initial stage

(b) after 320 iterations

(c) after 640 iterations

(d) final stage, after 1000 iterations

Fig. 7. WGAN learns the Gaussian mixture distribution.

Results analysis In Fig. 7, the blue points represent the real data distribution and the orange points represent the generated distribution. The left upper frame shows the initial stage, the right lower frame illustrates the stage after 1000 iterations. It seems that WGAN cannot capture the Gaussian mixture distribution. Generated data tend to lie in the middle of the two Gaussians. One reason is the well known mode collapse problem in GAN, meaning that if the data distribution has multiple clusters or data is distributed in multiple isolated manifolds, then the generator is hard to learn multiple modes well. Although there are efforts to deal with this problem (Gurumurthy et al., 2017; Hoang et al., 2017), it still remains open in machine learning community.

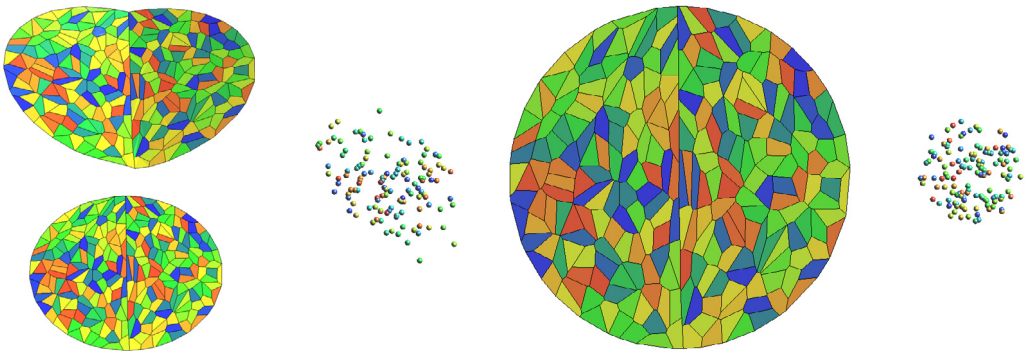
Geometric OMT Fig. 8 shows the geometric method to solve the same problem. The left frame shows the Brenier potential u_h, namely the upper envelope, which projects to the power diagram $\mathcal{V}$ on a unit disk $\mathbb{D} \subset \mathcal{Z}$, $\mathcal{V} = \bigcup_k W_i(h)$. The right frame shows the discrete optimal transportation map $T : \mathbb{D} \to \{y_i\}$, which maps each cell $W_i(h) \cap \mathbb{D}$ to a sample y_i, the cell $W_i(h)$ and the sample y_i have the same color. All the cells have the same area, this demonstrates that T pushes the uniform distribution ζ to the exact empirical distribution $T_{\#}\zeta = 1/n \sum_i \delta_{y_i}$.

The samples $\{y_i\}$ are generated according to the same Gauss mixture distribution, therefore there are two clusters. This doesn't cause any difficulty for the geometric method. In the left frame, we can see the upper envelope has a sharp ridge, the gradients point to the two clusters. Hence, the geometric method outperforms the WGAN model in the current experiment.

It is difficult for conventional deep learning methods to handle multiple modal distributions. This example demonstrates the geometric method has the potential to tackle this problem.

6.2. Geometric method

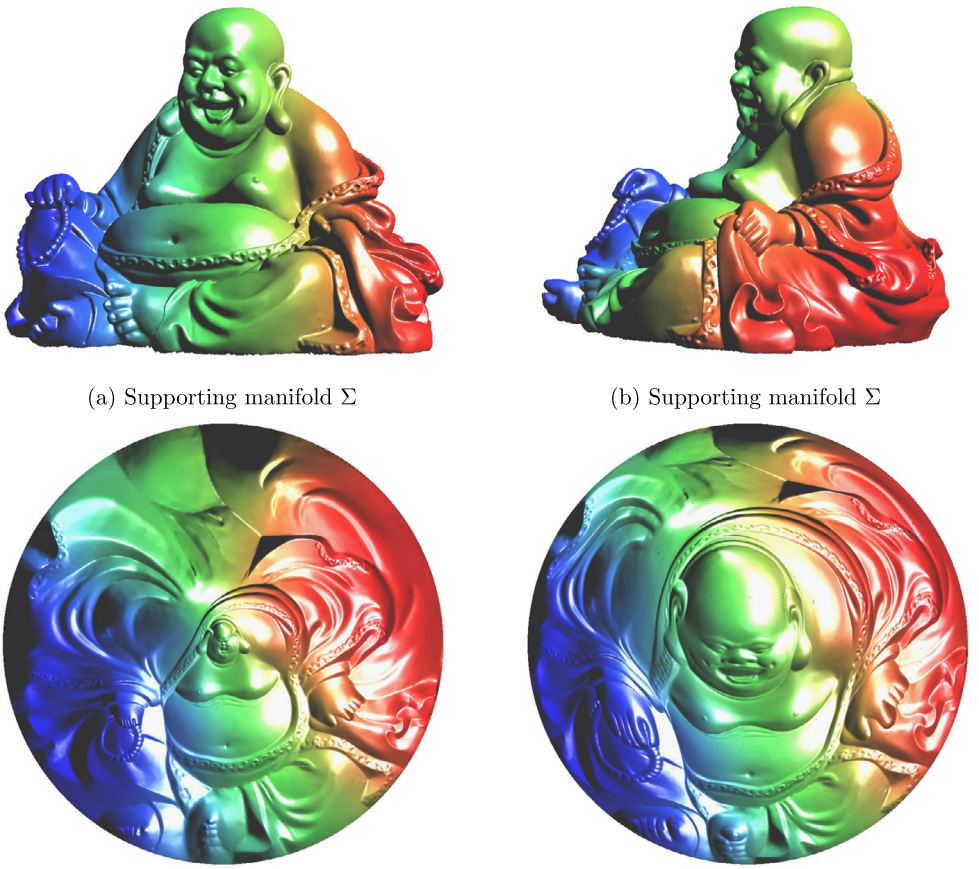
In this experiment, we use the pure geometric method to generate uniform distribution on a surface Σ with complicated geometry. As shown in Fig. 9, the image space $\mathcal{X}$ is the 3 dimensional Euclidean space $\mathbb{R}^3$. The real data samples are



(a) Brenier potential u_h

(b) Optimal transportation map $T : W_i \mapsto y_i$, where W_i is a power diagram cell.

Fig. 8. Geometric model learns the Gaussian mixture distribution.



(a) Supporting manifold Σ

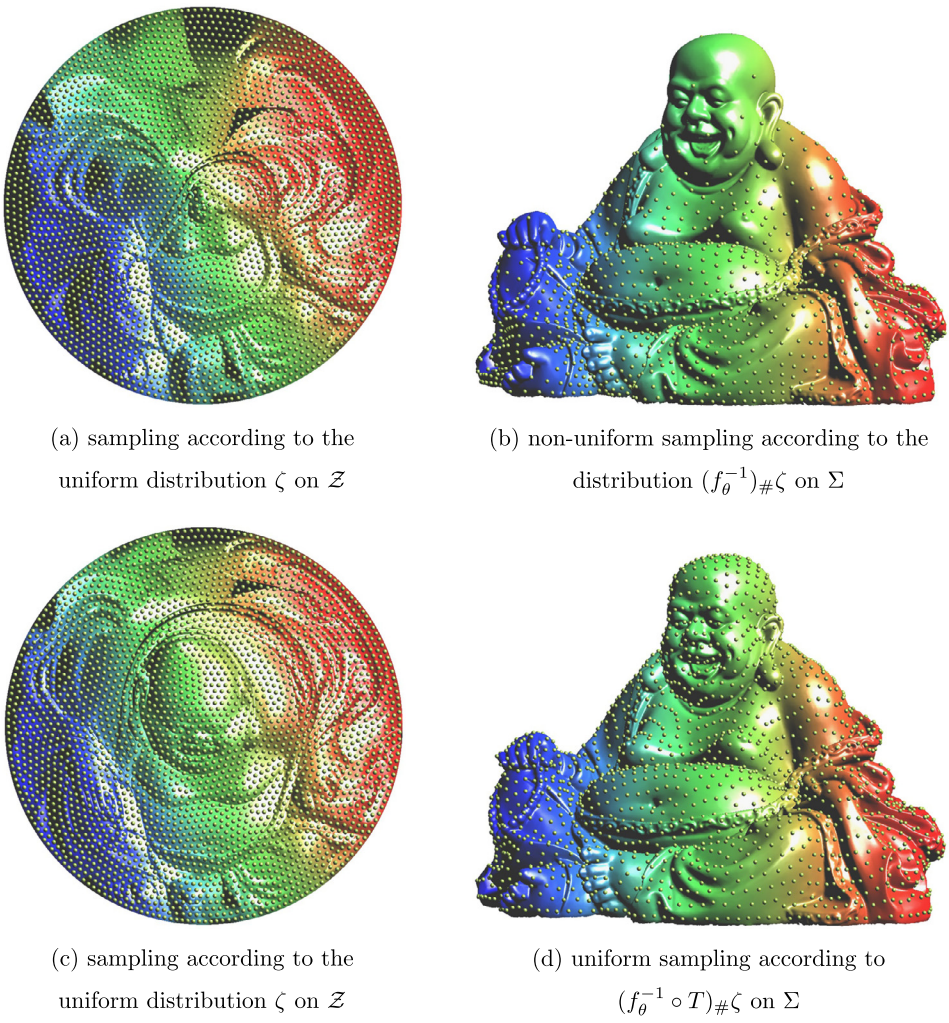
(b) Supporting manifold Σ

(c) Image of encoding map $f_\theta(\Sigma)$

(d) Image of encoding map composed with optimal transportation map $(T^{-1} \circ f_\theta)(\Sigma)$

Fig. 9. Illustration of geometric generative model.

distributed on a submanifold Σ, which is represented as a surface, as illustrated in (a) and (b). The encoding mapping $f_\theta : \Sigma \to \mathcal{Z}$ maps the supporting manifold to the latent space, which is a planar disk. The encoding map f_θ can be computed using discrete surface Ricci flow method (Gu et al., 2018). We color-encode the normals to the surface, and push forward the color function from Σ to the latent space $f_\theta(\Sigma)$, therefore users can directly visualize the correspondence between Σ



(a) sampling according to the uniform distribution ζ on $\mathcal{Z}$

(b) non-uniform sampling according to the distribution $(f_\theta^{-1})_\# \zeta$ on Σ

(c) sampling according to the uniform distribution ζ on $\mathcal{Z}$

(d) uniform sampling according to $(f_\theta^{-1} \circ T)_\# \zeta$ on Σ

Fig. 10. Illustration of geometric generative model.

and its image in $\mathcal{Z}$ as shown in (c). Suppose the Riemannian metric on Σ is $\mathbf{g}$, the conformal map f_θ parameterizes Σ, such that

$$\mathbf{g} = e^{2\lambda(z)} dz d\bar{z},$$

where $dzd\bar{z}$ is the Euclidean metric on the plane, λ is a function defined on the disk. f_θ pushes the area element of Σ to the plane, and obtain a measure $e^{2\lambda(z)} \frac{i}{2} dz \wedge d\bar{z}$, where $\frac{i}{2} dz \wedge d\bar{z}$ is the Euclidean area element on the plane. Then we construct the optimal mass transportation map $T : (\mathcal{Z}, \frac{i}{2} dz \wedge d\bar{z}) \to (\mathcal{Z}, e^{2\lambda(z)} \frac{i}{2} dz \wedge d\bar{z})$, the pre-image of T is shown in (d).

In Fig. 10, we demonstrate the generated distributions. In (a), we generate samples $\{z_1, \ldots, z_k\}$ on the latent space $f_\theta(\Sigma)$ according to the uniform distribution ζ, the samples are pulled back to the surface Σ as $\{f_\theta^{-1}(z_1), \ldots, f_\theta^{-1}(z_k)\}$ as shown in (b), which illustrate the distribution $(f_\theta^{-1})_\# \zeta$. It is obvious that the distribution generated this way is highly non-uniform on the surface. In frame (c), we uniformly generate samples on $(T^{-1} \circ f_\theta)(\Sigma)$, and map them back to the surface Σ as shown in (d). This demonstrates the generated distribution $(f_\theta^{-1} \circ T)_\# \zeta$ on Σ is highly uniform as desired.

7. Discussion and conclusion

In this work, we bridge convex geometry with optimal transportation, then use optimal transportation to analyze generative models. The basic view is that the discriminator computes the Wasserstein distance or equivalently the Kantorovich potential φ_ξ; the generator calculates the transportation map g_θ. By selecting the transportation cost, such as L^2 distance, φ_ξ and g_θ are related by a closed form, hence it is sufficient to train one of them.

L^2 cost function For general transportation cost $c(x, y)$, the explicit relation between φ_ξ and g_θ may not exist, it seems that both training processes are necessary. Hence by using the L^2 distance as the cost function, the efficiency of the system can be improved prominently.

Optimal transportation map Current generative adversarial network model (GAN) computes the Wasserstein distance, which requires the optimality of the transportation map, namely the Brenier map for L^2 cost function. For high dimensional setting, rigorous computational geometric method to compute the optimal transportation map is intractable, due to the maintenance of the complex geometric data structures. Instead, we can use GPU to improve the efficiency of the computational geometric method, or follow the Kantorovich's approach to find the optimal transportation map using the conventional linear programming method, as described in Liu et al. (2018).

Ambient space vs. latent space In the generative models, we need to compute the distance between the data distribution ν and the generated distribution $\mu_\theta = (g_\theta)_\# \zeta$, where $g_\theta : Z \to X$ is the decoding map. Some models compute the distance directly in the ambient space X, such as WGAN (Arjovsky et al., 2017), Least Square GAN (Mao et al., 2017; Radford et al., 2016) so on. Some models push ν forward to the latent space by the encoding map $g_\theta^{-1} : X \to Z$, and compute the distance between $(g_\theta^{-1})_\# \nu$ and ζ in the latent space, such as adversarial autoencoders (Makhzani et al., 2016) and Wasserstein auto-encoders (Tolstikhin et al., 2018). Because the dimension of the latent space is much smaller than that of the ambient space, the computation in the latent space is much more efficient.

Different type of generative models Furthermore, we can design different types of generative models, which solely compute the transportation maps without the optimality requirement. There are many more economical ways to compute (non-optimal) transportation maps, such as stochastic method, sliced optimal transportation method, hierarchical optimal transportation method, Knothe–Rosenblatt maps (Villani, 2003) and so on. Recently the sink-horn method has been introduced by Peyré and Cuturi in Peyré and Cuturi (2018), which greatly improves the computational efficiency.

In the future, we will explore along these directions, and implement the proposed model in a large scale.

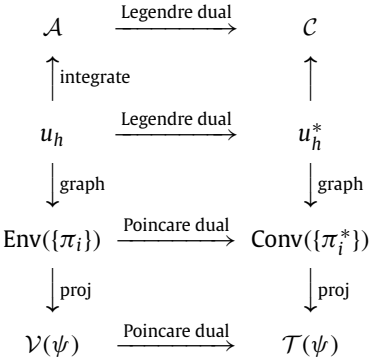
Acknowledgement

We thank the inspiring discussions with Dr. Dimitris Samaras from Stony Brook University, Dr. Shoucheng Zhang from Stanford University, Dr. Limin Chen from Ecole of Central Lyon and so on. Especially, we thank Dr. Cédric Villani for the discussion on the profound insights on Optimal Transportation theory and consistent encouragements. The WGAN experiment is conducted by Mr. Huidong Liu and generalized in the work (Liu et al., 2018). The project is partially supported by NSFC No. 61772105, 61772379, 61720106005, NSF DMS-1418255, NSF CMMI-1762287, AFOSR FA9550-14-1-0193 and the Fundamental Research Funds for the Central Universities No. 2015KJJCB23.

Appendix A

A.1. Commutative diagram

The relations among geometric/functional objects are summarized in the following diagram:



where each two adjacent layers are commutable. These relations can be observed from Fig. 4 as well.

A.2. Symbol list

Table 1 is the symbol list.

Table 1
Symbol list.

$\mathcal{X}$	ambient space, image space	
Σ	support manifold for some distribution	
$\mathcal{Z}$	latent space, feature space	
ζ	a fixed probability measure on $\mathcal{Z}$	
g_θ	generating/decoding map	$g_\theta : \mathcal{Z} \to \mathcal{X}$
φ_ξ	Kantorovich potential	
c	distance between two points	$c(x, y) = \lvert x - y\rvert^p,\ p \geq 1$
W_c	Wasserstein distance	$W_c(\mu, \nu)$
X	source space	
Y	target space	
μ	source probability measure	
ν	target probability measure	
Ω	source domain	$\Omega \subset X$
y_i	the i-th sample in target	$\{y_1, \ldots, y_k\} \in Y$
φ	Kantorovich potential	$\phi^c = \psi,\ \psi^c = \phi$
ψ	power weight	$\psi = (\psi_1, \ldots, \psi_k)$
h	plane heights	$h = (h_1, \ldots, h_k)$
π_i	hyper-plane	$\pi_i^h(x) = \langle y_i, x\rangle + h_i$
π_i^*	dual point of π_i	$\pi_i^* = (y_i, -h)$
pow	power distance	$\mathrm{pow}(x, y_i) = c(x, y_i) - \psi_i$
W_i	power Voronoi cell	$W_i(\psi) = \{x \in X \lvert \mathrm{pow}(x, y_i) \leq \mathrm{pow}(x, y_j)\}$
w_i	the volume of W_i	$w_i(h) = \mu(W_i(h) \cap \Omega)$
u	Brenier potential	$u_h(x) = \max_i\{\langle x, y_i\rangle + h_i\}$
$\mathcal{A}$	Alexandrov potential	$\mathcal{A}(h) = \int^h \sum_i w_i dh_i$
T	transportation map	$T = \nabla u_h$
$\mathcal{C}$	transportation cost	$\mathcal{C}(T) = \int_X c(x, T(x))d\mu$
Env	upper envelope of planes	$\mathrm{Env}(\{\pi_i\})$ graph of u_h
Conv	convex hull of points	$\mathrm{Conv}(\{\pi_i^*\})$ graph of u_h^*
$\mathcal{V}$	power diagram	$\mathcal{V}(\psi) : X = \bigcup_i W_i(\psi)$
$\mathcal{T}$	weighted Delaunay triangulation	

References

Alexandrov, A.D., 2005. Convex Polyhedra. Translated from the 1950 Russian edition by N.S. Dairbekov, S.S. Kutateladze and A.B. Sossinsky. Springer Monographs in Mathematics. Springer-Verlag, Berlin.

Ambrosio, Luigi, Gigli, Nicola, Savaré, Giuseppe, 2008. Gradient Flows in Metric Spaces and in the Space of Probability Measures. Springer Nature.

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

Bonnotte, Nicolas, 2012. From Knothe's rearrangement to Brenier's optimal transport map. arXiv:1205.1099. pp. 1–29.

Brenier, Yann, 1991. Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. 44 (4), 375–417.

Brock, Andrew, Lim, Theodore, Ritchie, James M., Weston, Nick, 2017. Neural photo editing with introspective adversarial networks. In: International Conference on Learning Representations.

Dong, Hao, Neekhara, Paarth, Wu, Chao, Guo, Yike, 2017. Unsupervised image-to-image translation with generative adversarial networks. arXiv preprint arXiv:1701.02676.

Edelsbrunner, Herbert, 1987. Voronoi Diagrams. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 293–333.

Genevay, Aude, Cuturi, Marco, Peyr'e, Gabriel, Bach, Fancis, 2016. Stochastic optimization for large-scale optimal transport. In: Lee, D.D., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Proceedings of NIPS'16, pp. 3432–3440.

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

Gu, Xianfeng, Luo, Feng, Sun, Jian, Yau, Shing-Tung, 2016. Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge–Ampere equations. Asian J. Math. 20 (2), 383–398.

Gu, Xianfeng, Luo, Feng, Sun, Jian, Wu, Tianqi, 2018. A discrete uniformization theorem for polyhedral surfaces. J. Differ. Geom 109, 223–256.

Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, Courville, Aaron, 2017. Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028.

Guo, Xin, Hong, Johnny, Lin, Tianyi, Yang, Nan, 2017. Relaxed Wasserstein with applications to gans. arXiv preprint arXiv:1705.07164.

Gurumurthy, Swaminathan, Sarvadevabhatla, Ravi Kiran, Babu, Venkatesh, Deligan, Radhakrishnan, 2017. Generative adversarial networks for diverse and limited data. In: CVPR.

Hinton, Geoffrey, Srivastava, Nitish, Swersky, Kevin, 2012. Neural Networks for Machine Learning – Lecture 6a – Overview of Mini-Batch Gradient Descent. Lecture notes.

Hoang, Quan, Nguyen, Tu Dinh, Le, Trung, Phung, Dinh, 2017. Multi-generator generative adversarial nets. arXiv preprint arXiv:1708.02556.

Huang, Rui, Zhang, Shu, Li, Tianyu, He, Ran, 2017. Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In: International Conference on Computer Vision.

Iizuka, Satoshi, Simo-Serra, Edgar, Ishikawa, Hiroshi, 2017. Globally and locally consistent image completion. ACM Trans. Graph. 36 (4), 107.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, Efros, Alexei A., 2016. Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004.
Kantorovich, L.V., 1948. On a problem of Monge. Usp. Mat. Nauk 3, 225–226.
Kim, Taeksoo, Cha, Moonsu, Kim, Hyunsoo, Lee, Jungkwon, Kim, Jiwon, 2017. Learning to discover cross-domain relations with generative adversarial networks. arXiv preprint arXiv:1703.05192.
Kumar, Abhishek, Sattigeri, Prasanna, Fletcher, P. Thomas, 2017. Improved semi-supervised learning with gans using manifold invariances. arXiv preprint arXiv:1705.08850.
Larsen, Anders Boesen Lindbo, Sønderby, Søren Kaae, Larochelle, Hugo, Winther, Ole, 2015. Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300.
Ledig, Christian, Theis, Lucas, Huszár, Ferenc, Caballero, Jose, Cunningham, Andrew, Acosta, Alejandro, Aitken, Andrew, Tejani, Alykhan, Totz, Johannes, Wang, Zehan, et al., 2016. Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802.
Li, Jianan, Liang, Xiaodan, Wei, Yunchao, Xu, Tingfa, Feng, Jiashi, Yan, Shuicheng, 2017a. Perceptual generative adversarial networks for small object detection. arXiv preprint arXiv:1706.05274.
Li, Yijun, Liu, Sifei, Yang, Jimei, Yang, Ming-Hsuan, 2017b. Generative face completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Liu, Ming-Yu, Tuzel, Oncel, 2016. Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 469–477.
Liu, Ming-Yu, Breuel, Thomas, Kautz, Jan, 2017. Unsupervised image-to-image translation networks. arXiv preprint arXiv:1703.00848.
Liu, Huidong, Gu, Xianfeng, Samaras, Dimitris, 2018. A two-step computation of the exact GAN Wasserstein distance. In: International Conference on Machine Learning. ICML.
Luc, Pauline, Couprie, Camille, Chintala, Soumith, Verbeek, Jakob, 2016. Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408.
Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, Goodfellow, Ian, Frey, Brendan, 2016. Adversarial autoencoders. arXiv preprint arXiv:1511.05644.
Mao, Xudong, Li, Qing, Xie, Haoran, Lau, Raymond, Wang, Zhen, Paul, Stephen, 2017. Smolley least squares generative adversarial networks. In: ICCV.
Mathieu, Michael, Couprie, Camille, LeCun, Yann, 2015. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440.
Odena, Augustus, 2016. Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583.
Park, Eunbyung, Yang, Jimei, Yumer, Ersin, Ceylan, Duygu, Berg, Alexander C., 2017. Transformation-grounded image generation network for novel 3d view synthesis. arXiv preprint arXiv:1703.02921.
Pathak, Deepak, Krahenbuhl, Philipp, Donahue, Jeff, Darrell, Trevor, Efros, Alexei A., 2016. Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544.
Perarnau, Guim, van de Weijer, Joost, Raducanu, Bogdan, Álvarez, Jose M., 2016. Invertible conditional gans for image editing. arXiv preprint arXiv:1611.06355.
Peyré, Gabriel, Cuturi, Marco, 2018. Computational optimal transport. arXiv:1803.00567.
Radford, Alec, Metz, Luke, Chintala, Soumith, 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
Radford, Alec, Metz, Luke, Chintala, Soumith, 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR2016.
Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, Chen, Xi, 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems, pp. 2234–2242.
Shen, Wei, Liu, Rujie, 2017. Learning residual images for face attribute manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Shu, Zhixin, Yumer, Ersin, Hadap, Sunil, Sunkavalli, Kalyan, Shechtman, Eli, Samaras, Dimitris, 2017. Neural face editing with intrinsic image disentangling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Springenberg, Jost Tobias, 2016. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In: International Conference on Learning Representations.
Taigman, Yaniv, Polyak, Adam, Wolf, Lior, 2016. Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200.
Tolstikhin, Ilya, Gelly, Sylvain, Bousquet, Olivier, Simon-Gabriel, Carl-Johann, Adagan, Bernhard Schölkopf, 2017. Boosting generative models. arXiv preprint arXiv:1701.02386.
Tolstikhin, Ilya, Bousquet, Olivier, Gelly, Sylvain, Schoelkopf, Bernhard, 2018. Wasserstein auto-encoders. In: ICLR.
Villani, Cédric, 2003. Topics in Optimal Transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence, RI.
Villani, Cédric, 2008. Optimal Transport: Old and New, Vol. 338. Springer Science & Business Media.
Vondrick, Carl, Pirsiavash, Hamed, Torralba, Antonio, 2016. Generating videos with scene dynamics. In: Advances in Neural Information Processing Systems, pp. 613–621.
Wang, Xiaolong, Shrivastava, Abhinav, Gupta, Abhinav, 2017. A-Fast-RCNN: hard positive generation via adversary for object detection. arXiv preprint arXiv:1704.03414.
Wu, Jiajun, Zhang, Chengkai, Xue, Tianfan, Freeman, Bill, Tenenbaum, Josh, 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Advances in Neural Information Processing Systems, pp. 82–90.
Yeh, Raymond, Chen, Chen, Lim, Teck Yian, Hasegawa-Johnson, Mark, Do, Minh N., 2017. Semantic image inpainting with perceptual and contextual losses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Zhao, Junbo, Mathieu, Michael, LeCun, Yann, 2016. Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126.
Zhu, Wentao, Xie, Xiaohui, 2016. Adversarial deep structural networks for mammographic mass segmentation. arXiv preprint arXiv:1612.05970.
Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, Efros, Alexei A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593.