SUPPLEMENTARY MATERIAL:
Wasserstein of Wasserstein Loss for Learning Generative Models

Yonatan Dukler [*1] **Wuchen Li** [*1] **Alex Tong Lin** [*1] **Guido Montúfar** [*1 2 3]

A. Wasserstein metrics in continuous sample space

In this section, we briefly review the duality structures of Wasserstein-p in continuous sample space. More details are provided in (Villani, 2009). When $p = 1$, a particular duality structure is shown. When $p = 2$, a metric tensor property will be discussed. These properties will be used intensively throughout the paper.

Given a sample space $\Omega \subset \mathbb{R}^d$, the Wasserstein-$p$ metric introduces a distance between probability density functions $\rho^0, \rho^1 \in \mathcal{P}(\Omega)$ by

$$W_p(\rho^0, \rho^1)^p = \inf_\pi \int_{\Omega\times\Omega} c(x,y)\pi(x,y)dxdy,$$

where the infimum is taken over all joint measures $\pi \geq 0$ with marginals

$$\int_\Omega \pi(x,y)dx = \rho^0(y), \qquad \int_\Omega \pi(x,y)dy = \rho^1(x).$$

Here $c(x,y)$ is a homogenous degree p function. E.g., $c(x,y) = \|x-y\|^p$ with $\|\cdot\|$ the Euclidean norm.

The Dual problem of the linear programming has the form

$$\begin{aligned} &W_p(\rho^0,\rho^1)^p \\ =& \sup_{\Phi^0,\Phi^1\in C(\Omega)} \Big\{ \int_\Omega \Phi^1(x)\rho^1(x) - \Phi^0(x)\rho^0(x)dx \colon \\ &\qquad \Phi^1(y) - \Phi^0(x) \leq c(x,y)\Big\}, \end{aligned}$$

where Φ^0, $\Phi^1 \colon \Omega \to \mathbb{R}$ are the Lagrangian multiplier variables for the constraint of linear programming involving ρ^0, ρ^1. Here Φ^0, Φ^1 are the so-called Kantorovich dual variables.

The Wasserstein metric exhibits special structures for $p = 1$ and $p = 2$. We discuss these in turn.

A.1. Wasserstein-1 metric

If $p = 1$, one can check that $\Phi^1(x) = \Phi^0(x)$. Denote $f(x) = \Phi^1(x)$ the constraint condition for duality problem has the form

$$f(x) - f(y) \leq c(x,y), \quad \text{for any } x, y \in \Omega.$$

This gives the 1-Lipscthiz condition with respect to the norm of metric $c(x,y)$, i.e.

$$\|\operatorname{grad} f(x)\|_c \leq 1.$$

We can apply this condition into the dual problem. We then derive the dual of dual problem as follows:

$$\inf_m \Big\{ \int_\Omega \|m(x)\|dx \colon \operatorname{div}(m) + \rho^1 - \rho^0 = 0\Big\}$$

where m is the flux function, and div is the divergence operator depending on the ground metric c. Here the minimizer of Wasserstein function satisfies

$$\begin{cases} \operatorname{div}(m(x)) = \rho^0(x) - \rho^1(x) \\ \dfrac{m(x)}{\|m(x)\|_c} = \operatorname{grad} f(x), \quad \text{when } \|m(x)\|_c > 0, \end{cases}$$

where div and grad are divergence and gradient operators with respect to the ground metric c. As we can see, the second formula in above system satisfies the Lipschitz-1 condition, i.e. the Eikonal equation

$$\|\operatorname{grad} f(x)\|_c = \|\frac{m(x)}{\|m(x)\|_c}\|_c = 1.$$

Following the direction of flux function $m(x)$ by the direction of grad $f(x)$, one transports ρ^0 to ρ^1. The transport direction follows the characteristic of Eikonal equation, i.e. the geodesic curve in (Ω, d).

[*]Equal contribution [1]Department of Mathematics and [2]Department of Statistics, University of California, Los Angeles, CA 90095. [3]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany. Correspondence to: Guido Montúfar <montufar@math.ucla.edu>, Alex Tong Lin <atlin@math.ucla.edu>, Wuchen Li <wcli@math.ucla.edu>, Yonatan Dukler <ydukler@math.ucla.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

A.2. Wasserstein-2 metric

If $p = 2$, one can relate the duality formula of Φ^1, Φ^0 with the solution of Hamilton-Jacobi equation by the Hopf-Lax formula (Villani, 2009). In other words, $\Phi^0(x)$, $\Phi^1(x)$ are the solution of Hamilton-Jacobi equation at times $t = 0$, $t = 1$:

$$\partial_t \Phi(t,x) + \frac{1}{2}\|\operatorname{grad}\Phi(t,x)\|_c^2 = 0.$$

The minimizer of optimal transport has a form

$$\begin{cases} \partial_t \rho(t,x) + \operatorname{div}\Big(\rho(t,x)\operatorname{grad}\Phi(t,x)\Big) = 0 \\ \partial_t \Phi(t,x) + \frac{1}{2}\|\operatorname{grad}\Phi(t,x)\|_c^2 = 0 \end{cases}$$

with the time zero and one density solution $\rho(0,x) = \rho^0(x)$, $\rho(1,x) = \rho^1(x)$. We notice the fact that the characteristic of continuity equation and Hamilton-Jacobi equation is again the geodesics in pixel space Ω.

Proof of Proposition 1. Combining the properties of Wasserstein-1 and Wasserstein-2 metric, we obtain that the Lipschitz-1 condition w.r.t. Wasserstein-2 metric gives the following fact. The characteristic of characteristic in probability of probability space gives the geodesic in the pixel space.

A.3. Wasserstein-2 gradient

In the last, we formally derive the Wasserstein-2 gradient operator.

Consider Ω is a compact region with the set of smooth and strictly positive densities:

$$\mathcal{P}_+(\Omega) = \Big\{\rho \in C^\infty(\Omega) \colon \rho(x) > 0,\ \int_\Omega \rho(x)dx = 1\Big\}.$$

Denote by $\mathcal{F}(\Omega) := C^\infty(\Omega)$ the set of smooth real valued functions on Ω. The tangent space of $\mathcal{P}_+(\Omega)$ is given by

$$T_\rho\mathcal{P}_+(\Omega) = \Big\{\sigma \in \mathcal{F}(\Omega) \colon \int_\Omega \sigma(x)dx = 0\Big\}.$$

Given $\Phi \in \mathcal{F}(\Omega)$ and $\rho \in \mathcal{P}_+(\Omega)$, define

$$V_\Phi(x) := -\nabla\cdot(\rho(x)\nabla\Phi(x)) \in T_\rho\mathcal{P}_+(\Omega).$$

Here the elliptic operator identifies the function Φ on Ω modulo additive constants with the tangent vector V_Φ in $\mathcal{P}_+(\Omega)$:

$$\mathcal{F}(\Omega)/\mathbb{R} \to T_\rho\mathcal{P}_+(\Omega), \quad \Phi \mapsto V_\Phi.$$

Denote $T_\rho^*\mathcal{P}_+(\Omega) = \mathcal{F}(\Omega)/\mathbb{R}$ as the smooth cotangent space of $\mathcal{P}_+(\Omega)$. Then the L^2-Wasserstein metric tensor on density space is defined as follows:

Definition 8 (Wasserstein-2 metric tensor)**.** *Define the inner product on the tangent space of positive densities* $g_\rho \colon T_\rho\mathcal{P}_+(\Omega) \times T_\rho\mathcal{P}_+(\Omega) \to \mathbb{R}$ *by*

$$g_\rho^W(\sigma_1,\sigma_2) = \int_\Omega \nabla\Phi_1(x)\cdot\nabla\Phi_2(x)\rho(x)dx,$$

where $\sigma_1 = V_{\Phi_1}$, $\sigma_2 = V_{\Phi_2}$ *with* $\Phi_1(x)$, $\Phi_2(x) \in \mathcal{F}(\Omega)/\mathbb{R}$*.*

In (Lafferty, 1988), $(\mathcal{P}_+(\Omega), g_\rho)$ is named density manifold. Following the Riemannian calculus, the gradient operator with respect to the Wassestein-2 metric (Otto, 2001) has the following form.

Proposition 9 (Wasserstein-2 gradient)**.**

$$\operatorname{grad}\mathcal{F}(\rho)(x) = -\nabla\cdot(\rho\nabla\frac{\delta}{\delta\rho(x)}\mathcal{F}(\rho)),$$

and

$$\|\operatorname{grad}\mathcal{F}(\rho)\|_W = \int \|\nabla\frac{\delta}{\delta\rho(x)}\mathcal{F}(\rho)\|^2\rho(x)dx.$$

This proposition is one of the motivation in Theorem 2. We next present the Wasserstein-2 gradient operator defined in a discrete sample space.

B. Wasserstein-2 gradient on discrete sample space

We recall the definition of discrete probability simplex with Wasserstein-2 Riemannian metric. Consider the discrete pixel space $I = \{1,\ldots,n\}$. The probability simplex on I is the set

$$\mathcal{P}(I) = \Big\{(p_1,\cdots,p_n) \in \mathbb{R}^n \colon \sum_{i=i}^n p_i = 1, \quad p_i \geq 0\Big\}.$$

Here $p = (p_1,\ldots,p_n)$ is a probability vector with coordinates p_i corresponding to the probabilities assigned to each node $i \in I$. The probability simplex $\mathcal{P}(I)$ is a manifold with boundary. We denote the interior by $\mathcal{P}_+(I)$. This consists of the strictly positive probability distributions, with $p_i > 0$ for all $i \in I$. To simplify the discussion, we will focus on the interior $\mathcal{P}_+(I)$.

We next define the Wasserstein-2 metric tensor on $\mathcal{P}_+(I)$, which also encodes the metric tensor of discrete states I. We need to give a ground metric notion on sample space. We do this in terms of a undirected graph with weighted edges, $\mathcal{G} = (I,E,\omega)$, where I is the vertex set, $E \subseteq \binom{I}{2}$ is the edge set, and $\omega = (\omega_{ij})_{i,j\in I} \in \mathbb{R}^{n\times n}$ is a matrix of edge weights satisfying

$$\omega_{ij} = \begin{cases} \omega_{ji} > 0, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}.$$

The set of neighbors (adjacent vertices) of i is denoted by $N(i) = \{j \in V \colon (i,j) \in E\}$. The normalized volume form on node $i \in I$ is given by $d_i = \frac{\sum_{j \in N(i)} \omega_{ij}}{\sum_{i=1}^{n} \sum_{i' \in N(i)} \omega_{ii'}}$.

The graph structure $\mathcal{G} = (I, E, \omega)$ induces a graph Laplacian matrix function.

Definition 10 (Weighted Laplacian matrix)**.** *Given an undirected weighted graph $\mathcal{G} = (I, E, \omega)$, with $I = \{1, \ldots, n\}$, the matrix function $L(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ is defined by*

$$L(p) = D^{\mathsf{T}} \Lambda(p) D, \quad p = (p_i)_{i=1}^n \in \mathbb{R}^n,$$

where

- *$D \in \mathbb{R}^{|E| \times n}$ is the discrete gradient operator defined by*

$$D_{(i,j) \in E, k \in V} = \begin{cases} \sqrt{\omega_{ij}}, & \textit{if } i = k,\, i > j \\ -\sqrt{\omega_{ij}}, & \textit{if } j = k,\, i > j \\ 0, & \textit{otherwise} \end{cases},$$

- *$-D^{\mathsf{T}} \in \mathbb{R}^{n \times |E|}$ is the oriented incidence matrix, and*
- *$\Lambda(p) \in \mathbb{R}^{|E| \times |E|}$ is a weight matrix depending on p,*

$$\Lambda(p)_{(i,j) \in E, (k,l) \in E} = \begin{cases} \frac{1}{2}(\frac{1}{d_i} p_i + \frac{1}{d_j} p_j), & \textit{if } (i,j) = (k,l) \in E \\ 0, & \textit{otherwise} \end{cases}.$$

The Laplacian matrix function $L(p)$ is the discrete analog of the weighted Laplacian operator $-\nabla \cdot (\rho \nabla)$ from Definition 8.

We are now ready to present the Wasserstein-2 metric tensor. Consider the tangent space of $\mathcal{P}_+(I)$ at p,

$$T_p \mathcal{P}_+(I) = \Big\{ (\sigma_i)_{i=1}^n \in \mathbb{R}^n \colon \sum_{i=1}^n \sigma_i = 0 \Big\}.$$

Denote the space of *potential functions* on I by $\mathcal{F}(I) = \mathbb{R}^n$, and consider the quotient space

$$\mathcal{F}(I)/\mathbb{R} = \{ [\Phi] \mid (\Phi_i)_{i=1}^n \in \mathbb{R}^n \},$$

where $[\Phi] = \{ (\Phi_1 + c, \cdots, \Phi_n + c) \colon c \in \mathbb{R} \}$ are functions defined up to addition of constants.

We introduce an identification map via the weighted Laplacian matrix $L(p)$ by

$$\mathbf{V} \colon \mathcal{F}(I)/\mathbb{R} \to T_p \mathcal{P}_+(I), \qquad \mathbf{V}_\Phi = L(p)\Phi.$$

We know that $L(p)$ has only one simple zero eigenvalue with eigenvector $c(1, 1, \cdots, 1)$, for any $c \in \mathbb{R}$. This is true since for $(\Phi_i)_{i=1}^n \in \mathbb{R}^n$,

$$\begin{aligned} \Phi^{\mathsf{T}} L(p) \Phi &= (D\Phi)^{\mathsf{T}} \Lambda(p) (D\Phi) \\ &= \sum_{(i,j) \in E} \omega_{ij} (\Phi_i - \Phi_j)^2 (\frac{1}{2}(\frac{1}{d_i} p_i + \frac{1}{d_j} p_j)) = 0 \end{aligned}$$

implies $\Phi_i = \Phi_j$, $(i,j) \in E$. If the graph is connected, as we assume, then $(\Phi_i)_{i=1}^n$ is a constant vector. Thus $V_\Phi \colon \mathcal{F}(I)/\mathbb{R} \to T_p \mathcal{P}_+(I)$ is a well defined map, linear, and one to one. I.e., $\mathcal{F}(I)/\mathbb{R} \cong T_p^* \mathcal{P}_+(I)$, where $T_p^* \mathcal{P}_+(I)$ is the cotangent space of $\mathcal{P}_+(I)$. This identification induces the following inner product on $T_p \mathcal{P}_+(I)$.

Definition 11 (Wasserstein-2 metric tensor)**.** *The inner product $g_p : T_p \mathcal{P}_+(I) \times T_p \mathcal{P}_+(I) \to \mathbb{R}$ takes any two tangent vectors $\sigma_1 = \boldsymbol{V}_{\Phi_1}$ and $\sigma_2 = \boldsymbol{V}_{\Phi_2} \in T_p \mathcal{P}_+(I)$ to*

$$g_p(\sigma_1, \sigma_2) = \sigma_1^{\mathsf{T}} \Phi_2 = \sigma_2^{\mathsf{T}} \Phi_1 = \Phi_1^{\mathsf{T}} L(p) \Phi_2. \tag{1}$$

In other words,

$$g_p(\sigma_1, \sigma_2) := {\sigma_1}^{\mathsf{T}} L(p)^{\dagger} \sigma_2, \quad \textit{for any } \sigma_1, \sigma_2 \in T_p \mathcal{P}_+(I),$$

where $L(p)^{\dagger}$ is the pseudo inverse of $L(p)$.

Following the inner product equation 1, the Wasserstein-2 metric on images $W \colon \mathcal{P}_+(I) \times \mathcal{P}_+(I) \to \mathbb{R}$ is defined by

$$W(p^0, p^1)^2 := \inf_{p(t), \Phi(t)} \Big\{ \int_0^1 \Phi(t)^{\mathsf{T}} L(p(t)) \Phi(t) dt \Big\}. \tag{2}$$

Here the infimum is taken over pairs $(p(t), \Phi(t))$ with $p \in H^1((0,1), \mathbb{R}^n)$ and $\Phi \colon [0,1] \to \mathbb{R}^n$ measurable, satisfying

$$\frac{d}{dt} p(t) - L(p(t)) \Phi(t) = 0, \quad p(0) = p^0, \quad p(1) = p^1.$$

The Wasserstein-2 metric on graph introduces the following gradient operator.

Theorem 12 (Wasserstein gradient on graphs)**.** *Given $\mathcal{F} \in C^1(\mathcal{P}_+(I))$, the gradient operator in Riemannian manifold $(\mathcal{P}_+(I), g)$ satisfies*

$$\operatorname{grad} \mathcal{F}(p) = L(p) d_\rho \mathcal{F}(p),$$

where d is the Euclidean gradient operator.

Proof. As in the definition of Riemannian gradient, we have

$$\operatorname{grad} \mathcal{F}(p) = (L(p)^{\dagger})^{\dagger} d_p \mathcal{F}(p) = L(p) d_p \mathcal{F}(p),$$

which finishes the proof. □

Proof of Proposition 5. *Following the proof of Theorem 2, we prove the proposition 5.*

We last illustrate the Wasserstein metric tensor in unnormalized density space. The new metric tensor induces the gradient operator in unnormalized density space.

In other words, consider

$$\mathcal{M}_+(I) = \Big\{ \mu = (\mu_1, \cdots, \mu_n) \in \mathbb{R}^n \colon \mu_i \geq 0 \Big\}.$$

The tangent space of $\mathcal{M}_+(I)$ at μ forms

$$T_\mu \mathcal{M}_+(I) = \mathbb{R}^n.$$

Definition 13 (Unnormalized Wasserstein-2 metric tensor)**.** *The inner product* $\tilde{g}_\mu : T_\mu\mathcal{M}_+(I) \times T_\mu\mathcal{M}_+(I) \to \mathbb{R}$ *forms*

$$\tilde{g}_\mu(\sigma_1, \sigma_2) := {\sigma_1}^\mathsf{T}\Big(L(p)^\dagger + \frac{1}{\alpha}\mathbf{1}\mathbf{1}^T\Big)\sigma_2,$$

for any $\sigma_1, \sigma_2 \in T_p\mathcal{P}_+(I)$*.*

It is clear that $(\mathcal{M}_+(I), \tilde{g})$ is a well defined metric in positive octant. In this case, the unnormalized Wasserstein-2 gradient is given by the following theorem.

Theorem 14 (Unnormalized Wasserstein-2 gradient on graphs)**.** *Given* $\mathcal{F} \in C^1(\mathcal{M}_+(I))$*, the gradient operator in Riemannian manifold* $(\mathcal{M}_+(I), \tilde{g})$ *satisfies*

$$\operatorname{grad}\mathcal{F}(\mu) = \Big(L(\mu) + \alpha\mathbf{1}\mathbf{1}^T\Big)d_\mu\mathcal{F}(\mu).$$

In other words,

$$\begin{aligned}\operatorname{grad}\mathcal{F}(\mu)_i =& \frac{1}{2}\sum_{j\in N(i)}\omega_{ij}\Big(\frac{\partial}{\partial\mu_i}\mathcal{F} - \frac{\partial}{\partial\mu_j}\mathcal{F}\Big)\big(\frac{\mu_i}{d_i} + \frac{\mu_j}{d_j}\big)\\ &+\alpha\sum_{i=1}^n\frac{\partial}{\partial\mu_i}\mathcal{F}(\mu).\end{aligned}$$

Proof. Notice that

$$L(\mu) = T\begin{pmatrix}0 & & & \\ & \lambda_{sec}(L(\mu)) & & \\ & & \ddots & \\ & & & \lambda_{\max}(L(\mu))\end{pmatrix}T^{-1}\ ,$$

where $0 < \lambda_{sec}(L(\mu)) \leq \cdots \leq \lambda_{\max}(L(\mu))$ are eigenvalues of $L(\rho)$ arranged in ascending order, and T is its corresponding eigenvector matrix. Here the zero eigenvalue correspond to the eigenvector $\mathbf{1}$. Thus

$$\Big(L(\mu)^\dagger + \frac{1}{\alpha}\mathbf{1}\mathbf{1}^T\Big)^{-1} = L(\mu) + \alpha\mathbf{1}\mathbf{1}^T.$$

Then

$$\begin{aligned}\operatorname{grad}\mathcal{F}(\mu) &= \Big(L(\mu)^\dagger + \frac{1}{\alpha}\mathbf{1}\mathbf{1}^\dagger\Big)^{-1}d_\mu\mathcal{F}(\mu)\\ &= L(\mu)d_\mu\mathcal{F}(\mu) + \alpha\mathbf{1}\mathbf{1}^T d_\mu\mathcal{F}(\mu),\end{aligned}$$

which finishes the proof. □

C. Detailed description of the experiments

We run experiments on the CIFAR-10 and CelebA (aligned, cropped, 64 × 64) datasets.

For the experiment measuring discriminator robustness to noise, or hyperparameters for WGAN-GP is,

- DCGAN Architecture, with 3 convolutional layers, and no batch-normalization in the discriminator.
- Adam optimizer, with learning rate 0.0003, and $\beta_1 = 0.5$, and $\beta_2 = 0.9$
- Batch size of 64, and noise vector of dimension 128.

For the WWGAN loss, we use the same hyperparameters as WGAN-GP, and for the WWGAN, we set $\alpha = 1.0$ and $\beta = 50$.

For the noise model, we used RGB salt and pepper noise, which first transforms the $3 \times N \times N$ image to a $3N^2$ vector, and provides a probability of changing any coordinate. Once a change is decided, the coordinate value is set to 0.0 or 1.0 (the max pixel value) with equal probability.

Then the discriminator is evaluated on 64 noisy and clean images. And we see that the discriminator trained with WWGAN is more robust to noise.

We compare the WWGAN loss function with the WGAN-GP loss For both losses, we use a DCGAN architecture, removing the batch-normalization layer in the discriminator. We also train with the Adam optimizer with learning rate $1e-4$ and $\beta_1 = 0.9$, $\beta_2 = 0$.

D. WWGAN generated images

Figures 1 and 2 below show sample images generated from the WWGAN model trained with the settings described in Appendix C.

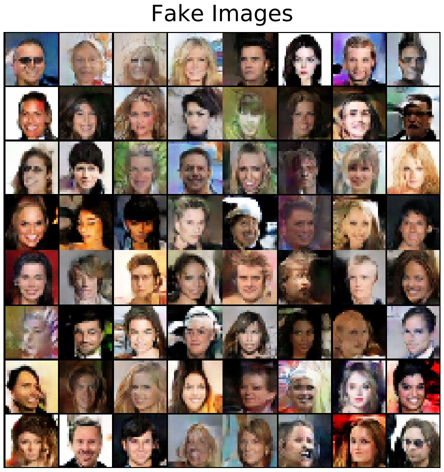

Figure 1. CelebA cropped 64 × 64 WWGAN generated images.

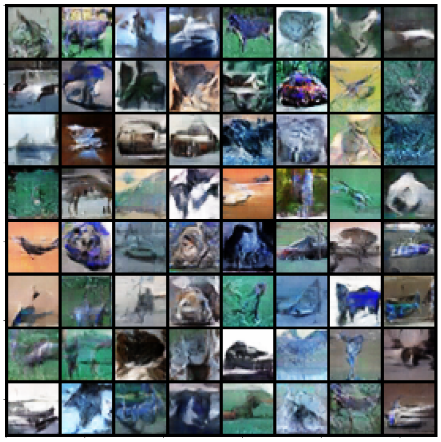

Figure 2. CIFAR-10 WWGAN generated images.

References

Lafferty, J. D. The Density Manifold and Configuration Space Quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.

Otto, F. The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

Villani, C. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.