WASSERSTEIN OF WASSERSTEIN LOSS FOR LEARNING GENERATIVE MODELS

YONATAN DUKLER, WUCHEN LI, ALEX TONG LIN, AND GUIDO MONTÚFAR

Abstract. The Wasserstein distance serves as a loss function for unsupervised learning which depends on the choice of a ground metric on sample space. We propose to use a Wasserstein distance as the ground metric on the sample space of images. This ground metric is known as an effective distance for image retrieval, since it correlates with human perception. We derive the Wasserstein ground metric on image space and define a Riemannian Wasserstein gradient penalty to be used in the Wasserstein Generative Adversarial Network (WGAN) framework. The new gradient penalty is computed efficiently via convolutions on the L^2 (Euclidean) gradients with negligible additional computational cost. The new formulation is more robust to the natural variability of images and provides for a more continuous discriminator in sample space.

1. Introduction

In recent years, optimal transport has become increasingly important in the formulation of training objectives for machine learning applications (Frogner et al., 2015; Montavon et al., 2016; Arjovsky et al., 2017). In contrast to traditional information divergences (arising in maximum likelihood estimation), the Wasserstein distance between probability distributions incorporates the distance between samples, via a ground metric of choice. In this way, it provides a continuous loss function for learning probability models supported on possibly disjoint, lower dimensional subset of the sample space. These properties are especially useful for training implicit generative models, with a prominent example being Generative Adversarial Networks (GAN). The application of the Wasserstein metric to define the objective function of GANs is known as Wasserstein GANs (WGANs) (Frogner et al., 2015; Arjovsky et al., 2017).

When training WGANs, one problem that remains is that of choosing a suitable ground metric for the sample space. The choice of the ground metric plays a crucial role in the training quality of WGANs. Usually, the distance between two sample images is taken to be the mean square difference over the features, i.e., the L^2 (Euclidean) norm. This, however, does not incorporate additional knowledge that we have about the space of natural images. In order to improve training and direct focus to selected features, other Sobolev norms in image space have been studied (Adler and Lunz, 2018). Recent works are also investigating distances based on higher level representations of the samples, which can be obtained by means of techniques such as vector embeddings (Mroueh et al., 2017), auto-encoders, or other unsupervised and semi-supervised feature learning techniques (Nowak et al., 2006). Meanwhile, another distance that has been very successful in comparing images, has

Key words and phrases. Wasserstein metric, GAN, Wasserstein statistical manifold.

remained unnoticed in the context of WGANs, namely the Wasserstein distance (also named Earth Mover's distance or Monge-Kantorvich distance). In particular, this distance has been considered in image retrieval problems (Rubner et al., 2000; Zhang et al., 2007). The Wasserstein distance between images is known to correlate well with human perception for natural images, e.g., being robust to translations and rotations (Engquist and Yang, 2018; Puthawala et al., 2018). See Figure 1. In addition, this distance is very natural and does not require computing higher level representations of the images or any feature selection.

In this paper, we propose to apply the Wasserstein distance over the sample space of images with a ground metric over the discrete space of pixels in the generative model formulation. We call this ground metric the *Wasserstein ground metric*, and call the Wasserstein loss over the Wasserstein ground metric the Wasserstein of Wasserstein loss. At first sight, it may appear overly complicated to define a loss function of this form. Since computing the Wasserstein distance is already quite involved, a Wasserstein loss based on another Wasserstein ground metric may seem infeasible. Nonetheless, we will show that it is possible to derive an equivalent expression in the settings of gradient penalty of WGANs (Petzka et al., 2017). In details, the Wasserstein-2 ground metric exhibits a metric tensor structure (Otto, 2001; Villani, 2009). This introduces a Lipschitz condition based on the Wasserstein norm, rather than the L^2 norm of the standard WGAN setting.

In this work we focus on generative models for images and specifically the WGAN formulation, but the proposed Wasserstein of Wasserstein loss function can be applied to learning with other types of models or other types of data for which a natural distance between features can be introduced.

This paper is organized as follows. In Section 2, we introduce the Wasserstein loss function with Wasserstein ground metric. Based on duality and the metric tensor of the proposed problem, we derive an equivalent practical formulation. In Section 3 we discuss our application to Wasserstein of Wasserstein GANs (WWGANs). Numerical experiments illustrating the benefits of the new gradient norm penalty are provided in Section 4. Related works are reviewed in Section 5.

2. Wasserstein of Wasserstein Loss

In this section, we introduce the Wasserstein ground metric in the Wasserstein loss function. A motivating example is presented to demonstrate the utility of the proposed model.

2.1. **Wasserstein loss.**

Consider a metric sample space $(\mathcal{X}, d_{\mathcal{X}})$. The Wasserstein-$p$ distance is defined as follows. Given a pair $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}_p(\mathcal{X})$ of probability densities with finite p-th moment, let

$$W_{p,d_{\mathcal{X}}}(\mathbb{P}_0, \mathbb{P}_1) = \inf_{\Pi} \left\{ \left(\mathbb{E}_{(X,Y)\sim\Pi} d_{\mathcal{X}}(X,Y)^p \right)^{\frac{1}{p}} \right\}, \tag{1}$$

where Π is a joint distribution of (X, Y) with marginals $X \sim \mathbb{P}_0$, $Y \sim \mathbb{P}_1$. We note that W_p depends on the choice of a distance function $d_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on sample space, which is usually called the ground metric.

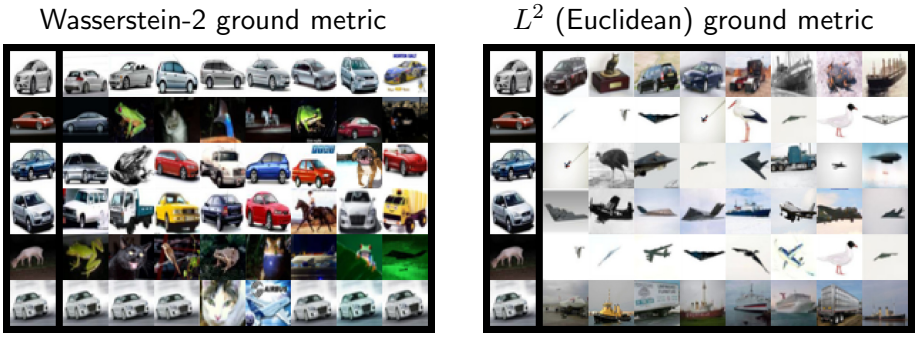

FIGURE 1. Source and 9 nearest neighbors of images from the CIFAR-10 dataset, with respect to the Wasserstein-2 (top) and L^2 (bottom) ground metrics. We note that the Wasserstein-2 distance give rise to neighbors that are similar perceptually and robust to translations and rotations. In contrast, the Euclidean distance is highly sensitive and oftentimes the nearest neighbors are predominantly white images.

In practice, the sample space $\mathcal{X}$ is typically very high dimensional, sometimes even being an (infinite dimensional) Banach space. We focus on the case where $\mathcal{X}$ is the space of images, which can be regarded as a density space over pixels, i.e., $\mathcal{X} = \mathcal{P}(\Omega)$, where $\Omega = [0, M] \times [0, M]$ is a discrete grid of pixels. With this in mind, we will define the distance function between pixels $d\colon \Omega \times \Omega \to \mathbb{R}_+$ as their physical distance.

2.2. **Wasserstein loss function with Wasserstein ground metric.** We further introduce the Wasserstein of Wasserstein loss. Here, the first 'Wasserstein' refers to the Wasserstein loss function over probability distributions on the space of images. The second 'Wasserstein' refers to the ground metric of this loss function. It is chosen as the Wasserstein distance over the space of images defined as histograms over pixels, having a ground metric over pixel locations.

That is, a raster image can be viewed as a 2D histogram with each pixel representing a bin. Defining a ground metric between pixels (e.g., the physical distance between pixels), the Wasserstein distance between images can be introduced. This serves as the new ground metric for defining a Wasserstein distance between probability distributions over images. See Figure 2.

As mentioned in the introduction, the Wasserstein distance is also known as the Earth Mover's distance and is known as an effective metric in distinguishing images (Rubner et al., 2000). Motivated by this fact, we use the Earth Mover's distance (of images) as the ground metric,

$$\begin{aligned} d_{\mathcal{X}}(X, Y) :=& W_{q,d_\Omega}(X, Y) \\ =& \inf_{\pi} \left\{ \left(\mathbb{E}_{(x,y)\sim\pi} d_\Omega(x, y)^q \right)^{\frac{1}{q}} \right\}, \end{aligned} \tag{2}$$

where π is a joint distribution of (x, y) with marginals $x \sim X$, $y \sim Y$ both being images, viewed as histograms over pixels. Here $d_{\mathcal{X}} = W_{q,d_\Omega}(x, y)$ is named Wasserstein-q ground

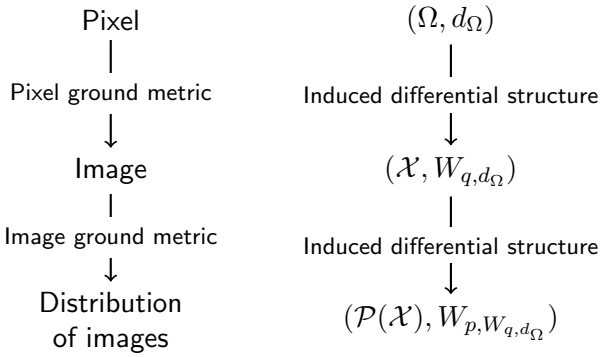

FIGURE 2. Illustration of Wasserstein-p loss function with Wasserstein-q ground metric.

metric. It is defined with the pixel ground metric $d_\Omega \colon \Omega \times \Omega \to \mathbb{R}_+$ assigning distances to pairs of pixels.

In this work, combining the above approaches, we obtain a Wasserstein-p distance with Wasserstein-q ground metric as the loss function for training.

Definition 1. *Given a probability model $\{\mathbb{P}_G \colon G \in \Theta\} \subseteq \mathcal{P}_p(\mathcal{X})$ and a data distribution $\mathbb{P}_r \in \mathcal{P}_p(\mathcal{X})$, we propose the minimization problem*

$$\inf_G W_{p,W_{q,d_\Omega}}(\mathbb{P}_G, \mathbb{P}_r), \tag{3}$$

where $\mathcal{P}_p(\mathcal{X})$ are densities with finite p-th moment, $W_{p,d_\mathcal{X}}$ is defined by (1) and W_{q,d_Ω} is given by (2).

The next example illustrates the difference between the proposed Wasserstein of Wasserstein loss and the Wasserstein loss with L^2 ground metric.

Motivation example: Consider the distribution $\mathbb{P}_r = \delta_X$ which assigns probability one to a single image X. Suppose the generative model tries to mimic this via a distribution of the form $\mathbb{P}_G = \delta_Y$ which assigns probability one to a fake image Y. Now suppose that $X = \delta_x$, $Y = \delta_y$ are images with intensity 1 on pixel locations x, y, respectively, and intensity zero elsewhere. See Figure 3. In this case we have

$$W_{p,d_\mathcal{X}}(\mathbb{P}_r, \mathbb{P}_G) = d_\mathcal{X}(X, Y).$$

We check the following choices of the ground metric $d_\mathcal{X}$ between images X and Y.

(1) Wasserstein-2 ground metric:

$$d_\mathcal{X}(X, Y) = W_{2,d_\Omega}(X, Y) = d_\Omega(x, y);$$

(2) L^2 (Euclidean) ground distance:

$$d_\mathcal{X}(X, Y) = d_{L^2}(X, Y) = \begin{cases} 0 & \text{if } x = y \\ \text{constant} & \text{if } x \neq y \end{cases}.$$

We see that the Wasserstein distance will assign two distant pixels the same cost as two adjacent pixels. This results in a highly discontinuous distance that is sensitive to single

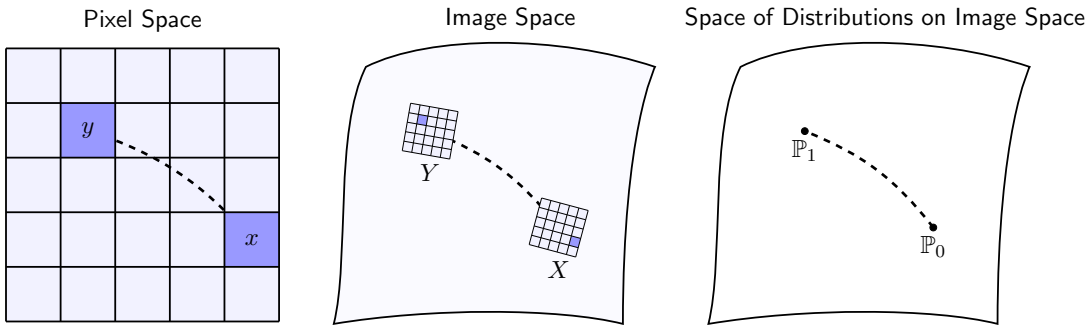

FIGURE 3. Depending on how we measure distances between pixel locations, the distance between images will be determined, and this in turn will determine how distances are measured between probability distributions.

pixel translations! To make matters worse, in the case of continuous domain images, the L^2 distance will be infinite for all non-overlapping pixels. Hence, for learning images with low dimensional support, the Wasserstein of Wasserstein loss function is still well defined, while the Wasserstein loss with L^2 ground metric function is ill-posed. We note, in particular, that the Wasserstein of Wasserstein loss function is continuous with respect to continuous change of pixels in images.

2.3. **Duality formulation and properties.** The computation required for the Wasserstein of Wasserstein loss function as stated in the previous section is unfeasible. To compute (3) one needs to handle a linear programming computation at both the level of probability distributions over images and individual images over pixels.

In this section, we present the Kantorovich duality formulation of Wasserstein of Wasserstein loss function with $p = 1$ and $q = 2$. As is done for Wasserstein GANs (Arjovsky et al., 2017), we consider an equivalent Lipschitz-1 condition, which can be practically applied in the framework of GANs.

Theorem 2 (Duality of Wasserstein of Wasserstein loss function)**.** *The Wasserstein-1 loss function over Wasserstein-2 ground metric has the following equivalent formulation:*

$$W_{1,W_{2,d_\Omega}}(\mathbb{P}_G,\mathbb{P}_r) = \sup_{f\in C(\mathcal{X})} \Big\{\mathbb{E}_{X\sim\mathbb{P}_G} f(X) - \mathbb{E}_{X\sim\mathbb{P}_r} f(X) \colon \int_\Omega \|\nabla_x \delta_X f(X)(x)\|^2_{d_\Omega} X(x)dx \le 1\Big\}, \tag{4}$$

where ∇_x is the gradient operator in pixel space Ω and δ_X is the L^2 gradient in image space $\mathcal{X}$.

Proof. The result is from the duality of Wasserstein-1 metric, together with the Wasserstein-2 metric induced gradient operator. First, the Wasserstein-1 metric has a particular dual formulation, known as the Kantorovich duality:

$$W_{1,d_\mathcal{X}}(\mathbb{P}_0,\mathbb{P}_1) = \sup_f \ \mathbb{E}_{X\sim\mathbb{P}_0} f(X) - \mathbb{E}_{X\sim\mathbb{P}_1} f(X),$$

where the supremum is taken among all $f\colon \mathcal{X} \to \mathbb{R}$ satisfying a 1-Lipschitz condition with respect to the ground metric $d_{\mathcal{X}}$, i.e.,

$$\|\operatorname{grad} f(X)\|_{d_{\mathcal{X}}} \leq 1. \tag{5}$$

Second, consider the ground metric given by the Wasserstein-2 metric $d_{\mathcal{X}} = W_{2,d_\Omega}$ with ground metric d_Ω of pixel space. Then the gradient operator in $(\mathcal{X}, d_{\mathcal{X}})$ is the Wasserstein-2 gradient, i.e.,

$$\operatorname{grad} f(X) = -\nabla_x \cdot (X(x)\nabla_x \delta_X f(X)(x)).$$

The 1-Lipschitz condition for $(\mathcal{X}, d_{\mathcal{X}})$ in (5) gives $\|\operatorname{grad} f(X)\|_{W_{2,d_\Omega}} \leq 1$, i.e.,

$$(\operatorname{grad} f(X), \operatorname{grad} f(X))_{W_{2,d_\Omega}} \leq 1.$$

It is rewritten as the following integral of the Lipschitz-1 condition w.r.t. the Wasserstein ground metric:

$$\int_\Omega \|\nabla_x \delta_X f(X)(x)\|^2_{d_\Omega} X(x) dx \leq 1.$$

Combining the above facts, we derive the formula for Wasserstein of Wasserstein loss function. □

Remark 1. *We note that the Kantorovich duality formula holds for any ground metric. The Wasserstein ground metric introduces differential structures and can be computed from the L^2 gradient. We review the Wasserstein gradient operators in Appendix A.*

The minimizer f in (4) corresponds to an Eikonal equation in image space $(\mathcal{X}, W_{2,d_\Omega})$. In other words, the Lipschitz-1 condition in Wasserstein norm has the form

$$\int_\Omega \|\nabla_x \delta_X f(X)(x)\|^2_{d_\Omega} X(x) dx = 1.$$

We call this equation the Wasserstein Eikonal equation.

Proposition 3 (Wasserstein Eikonal equation)**.** *The characteristic of characteristic for the Wasserstein Eikonal equation is the geodesic in pixel space.*

We defer the proof of the above proposition to Appendix A. Here the characteristic curve of our Eikonal equation is the geodesic curve in Wasserstein space $(\mathcal{X}, W_{2,d_\Omega})$. The characteristic curve of geodesics in Wasserstein space is again a geodesic in pixel space (Ω, d_Ω). We call this fact the **double characteristic property**. This is illustrated in Figure 3. In contrast, the characteristic of geodesics in L^2 space does not depend on pixel space. In the experiments section, we show that with the double characteristic property, the discriminator is continuous with respect to translations in pixel space, and is robust with respect to spatially independent noise added to the samples.

3. Wasserstein of Wasserstein GANs

In this section we apply the Wasserstein of Wasserstein loss function to implicit generative models.

3.1. **Background.** We start by reviewing generative adversarial networks (GAN). GANs are a deep learning approach to generative modelling that has demonstrated significant potential in the realm of image and text synthesis (Yu et al., 2017; Meng et al., 2018). The GAN model is composed of two competing agents: A discriminator and a generator. At each training step the generator produces synthesized images and the discriminator is given a batch of real and synthesized images to be classified as real or fake. The generator is trained to maximize the predictions of the discriminator while the discriminator is trained to classify generated images aside from real images. At the end of training the generator has learned how to trick the discriminator and ideally also the underlying data distribution.

Mathematically if we define a trainable generative model $\mathbb{P}_G$ and discriminator D, the GAN objective formulation is as follows:

$$\min_{\mathbb{P}_G} \max_{D} \Big\{ \mathbb{E}_{x\sim\mathbb{P}_r} \log(D(X)) + \mathbb{E}_{x\sim\mathbb{P}_G} \log(1 - D(X)) \Big\}. \tag{6}$$

Here $\mathbb{P}_r$ is the true, or real, data distribution. The distribution $\mathbb{P}_G$ is defined in terms of a generator parameterized by $\theta \in \mathbb{R}^d$. Let the generator be given by $G_\theta \colon \mathbb{R}^m \to \mathcal{X}; z \mapsto x = G(\theta, z)$. This takes a noise sample $Z \sim p(z) \in \mathcal{P}_2(\mathbb{R}^m)$ to an output sample with density given by $X = G(\theta, Z) \sim \rho(\theta, x) = \mathbb{P}_G$. Here $\mathbb{R}^d$ is the parameter space, $\mathbb{R}^m$ is the latent space, and $\mathcal{X}$ is the sample space.

The approach described above was found to suffer from difficulties at training including lack of convergence and mode collapse, a phenomenon where $\mathbb{P}_G$ restricts to a subset of $\mathbb{P}_r$. The above-mentioned challenges are often the result of the discontinuous nature of the loss in (6). To resolve such problems, Arjovsky et al. (2017) proposed to use the Wasserstein metric with Euclidean ground metric as the objective, formulated as

$$\min_{\mathbb{P}_G} W_{1,L^2}(\mathbb{P}_G, \mathbb{P}_r) = \min_{\mathbb{P}_G} \sup_{f\in C(\mathcal{X})} \Big\{ \mathbb{E}_{X\sim\mathbb{P}_G} f(X) - \mathbb{E}_{X\sim\mathbb{P}_r} f(X) \colon \| \operatorname{grad} f(X)\|_2 \leq 1 \Big\}. \tag{7}$$

The Lipschitz condition in (7) was enforced via weight-clipping, ensuring $\| \operatorname{grad} f(X)\|_2 < C_0$. While now providing GAN with a continuous loss, the WGAN formulation with weight-clipping has suffered cyclic behavior and in-stability which was significantly improved by Gulrajani et al. (2017) by changing the Lipschitz enforcing condition from hard weight-clipping to a soft gradient penalty term,

$$\min_{\mathbb{P}_G} \sup_{f\in C(\mathcal{X})} \Big\{ \mathbb{E}_{X\sim\mathbb{P}_G} f(X) - \mathbb{E}_{X\sim\mathbb{P}_r} f(X) + \lambda \mathbb{E}_{x\sim\mathbb{P}_{\text{interp}}} (\nabla_X f(x) - 1)^2 \Big\}. \tag{8}$$

where $\mathbb{P}_{\text{interp}}$ is an interpolation between $\mathbb{P}_r$ and $\mathbb{P}_G$ and λ is fixed. The gradient penalty term in (8) is not in full compliance with the Kantorovich duality of the problem as it also penalizes a discriminator of Lipschitz constants smaller than 1. To remedy this issue, the works of Petzka et al. (2017) replace the gradient penalty term in (8) by

$$\lambda \mathbb{E}_{x\sim\mathbb{P}_{\text{interp}}} (\max(\nabla_X f(x) - 1, 0))^2$$

We now derive our formulation that improves current methods based on the L_2 ground metric. Following Theorem 2, the Wasserstein of Wasserstein loss function can be rewritten to give the optimization problem

$$\min_{\mathbb{P}_G} W_{1,W_{2,d_\Omega}}(\mathbb{P}_G, \mathbb{P}_r) = \min_{\mathbb{P}_G} \sup_{f\in C(\mathcal{X})} \Big\{ \mathbb{E}_{X\sim\mathbb{P}_G} f(X) - \mathbb{E}_{X\sim\mathbb{P}_r} f(X) \colon \| \operatorname{grad} f(X)\|_{W_{2,d_\Omega}} \leq C \Big\}.$$

The above formulation is suitable for training GANs. Here we call the dual variable, f, the discriminator, while G is the generator. In the setting of GANs, we further apply neural networks to approximate both the discriminator and generator, which results in the optimization problem

$$\min_{\theta}\sup_{w}\Big\{\mathbb{E}_{Z\sim p(z)}f_w(g(\theta,Z))-\mathbb{E}_{X\sim\mathbb{P}_r}f_w(X)\colon \int_{\Omega}\|\nabla_x\delta_X f(X)(x)\|_{d_\Omega}^2 X(x)dx\le 1\Big\}.$$

Here the generator G is expressed as a neural network with parameters $\theta\in\Theta$, and the discriminator is approximated by a neural network with parameters w. Our approach implements the 1-Lipschitz condition in terms of the Wasserstein gradient operator.

3.2. **Discretization.** We next present a discrete version of the Wasserstein-2 gradient. In practice, the image space $\mathcal{X}$ is not infinite dimensional, although it may have a huge dimension. E.g., $\mathcal{X}=\mathbb{R}^{28\times 28}$ or $\mathbb{R}^{32\times 32}$ for MNIST or CIFAR-10 datasets. To discretize, we first review the L^2-Wasserstein metric tensor (matrix) defined on a finite dimensional space. Consider a pixel space graph $\mathcal{G}=(V,E,\omega)$. Here $V=\{1,\cdots,n\}$ is the vertex set (e.g., $n=28\times 28$), E is the edge set, and ω is a matrix of weights associated to the edges, with $\omega_{ij}=\omega_{ji}$, which defines a ground metric of pixels. We denote the neighborhood of node $i\in V$ by $N(i)=\{j\in V\colon (i,j)\in E\}$, and the degree of node i by $d_i=\frac{\sum_{j\in N(i)}\omega_{ij}}{\sum_{i=1}^n\sum_{i'\in N(i)}\omega_{ii'}}$. We can then define a Wasserstein-2 metric W on $\mathcal{X}$ (details in Appendix B), and further introduce the Wasserstein-2 gradient on discrete image space.

Proposition 4 (Wasserstein gradient on pixel space graph)**.** *Given a pixel space graph $\mathcal{G}$, the gradient of $f\in C^1(\mathcal{X})$ w.r.t. $(\mathcal{X},W)$ satisfies*

$$\operatorname{grad} f(X)=L(X)\nabla_X f(X),$$

where ∇_X is the Euclidean gradient operator, and $L(X)\in\mathbb{R}^{n\times n}$ is the weighted Laplacian matrix defined as

$$L(X)_{ij}=\begin{cases}\frac{1}{2}\sum_{k\in N(i)}\omega_{ik}(\frac{X_i}{d_i}+\frac{X_k}{d_k}) & \text{if } i=j;\\ -\frac{1}{2}\omega_{ij}(\frac{X_i}{d_i}+\frac{X_j}{d_j}) & \text{if } j\in N(i);\\ 0 & \text{otherwise.}\end{cases}$$

Moreover, the 1-Lipschitz condition w.r.t. $(\mathcal{X},W)$, $\|\operatorname{grad} f(X)\|_W\le 1$, is equivalent to

$$\nabla_x f(X)^{\mathsf{T}}L(X)\nabla_x f(X)\le 1.$$

Remark 2. *We observe that the 1-Lipschitz condition is exactly the discrete analog of the one in* (4)*,*

$$\nabla_x f(X)^{\mathsf{T}}L(X)\nabla_x f(X)=\sum_{(i,j)\in E}\omega_{ij}(\nabla_{X_j}f(X)-\nabla_{X_i}f(X))^2\frac{X_i/d_i+X_j/d_j}{2}\le 1.$$

In the implementation, we simply times the weighted Laplacian matrix $L(X)$ with Euclidean gradient operator in image space. The product is the Wasserstein gradient operator. We note that the Wasserstein gradient written in this form can be compared with the graph Laplacian on images (Bertozzi and Flenner, 2012; Zheng et al., 2011).

3.3. **Computing the Wasserstein gradient via convolutions.** We utilize the symmetry of the similarity graph of the image space to compute the Wasserstein gradient efficiently via convolutions as illustrated in Algorithm 2. As the optimal transport can be defined for local distances and truncated at a given threshold, this leads to a sparse $\omega_{i,j}$, positive only for nearby pixels. We therefore can calculate all pairs $\nabla_{X_i} f(X) - \nabla_{X_j} f(X)$ with a given neighboring pattern by computing a set of kernels $K_{\mathcal{O}_1} \dots K_{\mathcal{O}_d}$ on the Euclidean gradient $\nabla_X f(X)$. The kernels $K_{\mathcal{O}_1} \dots K_{\mathcal{O}_d}$ are each defined as a convolution with fixed kernel of zeros with 1 and -1 in the corresponding neighbor pattern pixels. By creating a convolution filter for each neighbor pattern (e.g., right or up neighbor) we reach the desired output channels. In practice the different kernels $K_{\mathcal{O}_1} \dots K_{\mathcal{O}_k}$ are grouped to form a single 3D kernel. Likewise we apply the same kernel patterns, now with $\frac{1}{2}, \frac{1}{2}$ in the corresponding neighbor pattern pixels to obtain the terms $\frac{X_i/d_i + X_j/d_j}{2}$ for each i, j. This is done analogously, computing each kernel $M_{\mathcal{O}_k}$ over the images X/d. Applying entry-wise multiplication ($\odot$) and a summation collapsing all pixel locations and channels then yields an efficient and general method of calculating the Wasserstein gradient $\|\operatorname{grad} f\|_{W_{2,d(\Omega)}}$ for general local cost metrics on highly optimized convolution. The specific choice of the graph could serve to enhance different effects, which is a possibility that we leave for future study.

3.4. **Wasserstein gradient regularization in GANs.** We next adopt the gradient penalty into the loss function (cf. Petzka et al., 2017; Gulrajani et al., 2017) as follows:

$$\min_{\theta} \sup_{w} \Big\{ \mathbb{E}_{z\sim p(z)} f_w(g(\theta, z)) - \mathbb{E}_{x\sim \mathbb{P}_r} f_w(x) + \lambda \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}} \Big(\sqrt{\nabla_X f(\hat{X})^{\intercal} L(\hat{X}) \nabla_X f(\hat{X})} - 1 \Big)^2 \Big\},$$

where λ is chosen as a large constant and $\hat{\mathbb{P}}$ is the distribution of $\hat{x}$ taken to be the uniform distributions on "Euclidean" lines connecting points drawn from $\mathbb{P}_G$ and $\mathbb{P}_r$. Our WWGAN training method is summarized in Algorithm 1.

Remark 3. *In practice, the image is often not with the same intensity. We need to consider a gradient operator, which also takes the effect of the change of total intensity. As proposed by Li (2018), we consider*

$$\tilde{L}(X) = \alpha \mathbf{1}\mathbf{1}^T + L(X).$$

Here $\mathbf{1} = (1, \cdots, 1)^T \in \mathbb{R}^n$ *is a constant vector. In Appendix C, we will show that* $\mathbf{1}$ *adds one additional direction into the original normalized Wasserstein metric tensor. Compared to* $L(X)$ *being an inverse metric tensor defined in a simplex (normalized intensity space),* $\tilde{L}(X)$ *is a well defined inverse metric tensor in the positive orthant. In the algorithm, we simply replace matrix function* L *by* $\tilde{L}$ *for general unnormalized intensity space.*

4. Experiments

In this section, we present experiments demonstrating the effects and effectiveness of WWGAN. We preform experiments on the CIFAR-10 and 64×64 cropped-CelebA image datasets. In both experiments the discriminator is a convolutional neural network with 3 hidden layers and leaky ReLU activations. For the generator we utilize a network with 3 hidden de-convolution layers and batch normalization. The dimensionality of the latent variable input of the generator is set at 128. Batch normalization is not applied to the

Algorithm 1 WWGAN Gradient Penalty

Require: The gradient penalty coefficient λ, discriminator iterations per generator iteration $n_{discriminator}$, batch size m, ADAM hyperparameters α, β_1, β_2.
Require: initial discriminator parameters w_0, initial generator parameters θ_0
Require: L matrix-function derived from the graph structure for image space $G = (V, E, \omega)$

while θ has not converged **do**
  for $t = 1, \dots, n_{discriminator}$ **do**
    for $i = 1, \dots, m$ **do**
      Sample real data $\boldsymbol{x} \sim \mathbb{P}_r$, latent variable $\boldsymbol{z} \sim p(\boldsymbol{z})$, a random number $\epsilon \sim U[0, 1]$.
      $\tilde{\boldsymbol{x}} \leftarrow G_\theta(\boldsymbol{z})$
      $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon)\tilde{\boldsymbol{x}}$
      $M^{(i)} \leftarrow D_\omega(\tilde{\boldsymbol{x}}) - D_\omega(\boldsymbol{x}) + \lambda(\sqrt{\nabla_{\hat{\boldsymbol{x}}} D_\omega(\hat{\boldsymbol{x}})^T L(\tilde{\boldsymbol{x}}) \nabla_{\hat{\boldsymbol{x}}} D_\omega(\hat{\boldsymbol{x}})} - 1)^2$
    end for
    $\omega \leftarrow \mathrm{Adam}(\nabla_\omega \frac{1}{m} \sum_{i=1}^m M^{(i)}, \omega, \alpha, \beta_1, \beta_2)$
  end for
  Sample a batch of latent variables $\{\boldsymbol{z}^i\}_{i=1}^m \sim p(\boldsymbol{z})$
  $\theta \leftarrow \mathrm{Adam}(\nabla_\theta \frac{1}{m} \sum_{i=1}^m -D_\omega(G_\theta(\boldsymbol{z}), \theta, \alpha, \beta_1, \beta_2))$
end while

Algorithm 2 Wasserstein gradient norm $\| \operatorname{grad} f(X) \|_W$

Require: The pixel graph: $\mathcal{G} = (V, E, \omega)$; local weights: (w_{ij}); neighbor relations arranged symmetrically: $\mathcal{O}_1 \dots \mathcal{O}_d$
Require: Euclidean gradient $\nabla_X f$

Wasserstein-grad $\leftarrow 0$
for neighbor relations $k = 1, \dots, d$ **do**
  Build kernel $K_{\mathcal{O}_k}$ to compute $\nabla_{X_i} f - \nabla_{X_{\mathcal{O}_k(i)}} f$
  Build corresponding kernel $M_{\mathcal{O}_k}$ to compute $\frac{X_i}{2d_i} + \frac{X_{\mathcal{O}_k}}{2d_{\mathcal{O}_k}}$
  $H \leftarrow K_{\mathcal{O}_k}(\nabla_X f)$
  $V \leftarrow M_{\mathcal{O}_k}(X)$
  $H \leftarrow H \odot H$ (entry-wise multiplication)
  $W \leftarrow H \odot V$
  Wasserstein-grad $\leftarrow$ *Wasserstein-grad* $+ \operatorname{sum}(W)$
end for
Return $\| \operatorname{grad} f(X) \|_W = \sqrt{\textit{Wasserstein-grad}}$

discriminator, in order to avoid dependencies when computing the gradient penalties. The model is then trained with the ADAM optimizer with fixed parameters $(\beta_1, \beta_2) = (0.9, 0)$. More details on the implementation are provided in Appendix C.

Figure 6 shows that in terms of computation time and quality of the generated images as measured by the Frechét Inception Distance (FID), WWGAN is comparable to state of the art WGAN-GP. Next, we take a closer look at the properties of the trained discriminators,

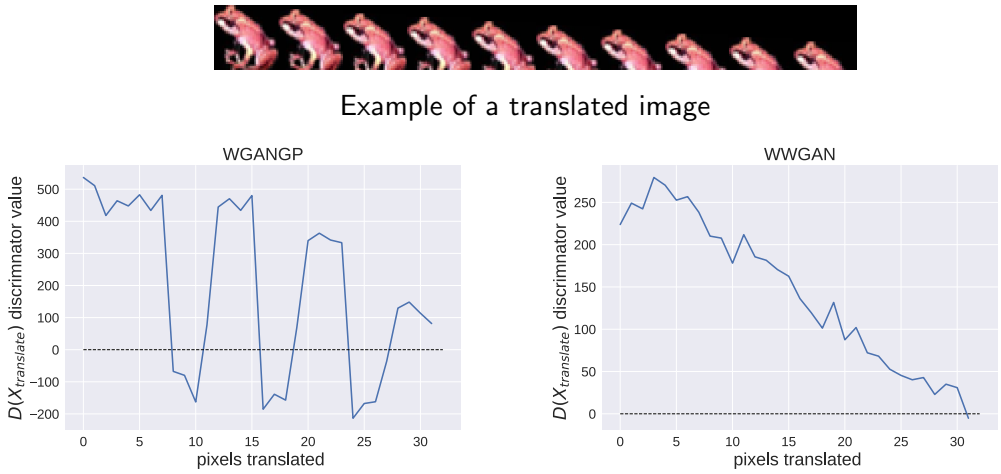

FIGURE 4. Discriminator for CIFAR-10 images translated continuously by a vertical shift from 0 (no shift) to 32 pixels (complete image). The WWGAN discriminator is continuous to natural perturbations, e.g. vertical translation. WGAN-GP discriminator exhibits unpredictable behavior for small vertical perturbations, oscillating between real (positive values) and fake (negative values) labels. Both WWGAN, WGAN-GP discriminators tested were trained identically to reach an FID value of 40.

which also serves to probe the shape of the probability densities over images defined by generators.

4.1. **Perturbation stability.** In this experiment we investigate how the discriminator trained with WWGAN on images benefits from the properties of the Wasserstein ground metric. Specifically, we test whether the discriminator trained with the new gradient penalty is more continuous with respect to natural variations of the images, which is a desirable property. The variability refers continuous transformations of natural images that result in natural looking images, such as translations and rotations. If the transformations are applied gradually, one should expect to observe only gradual changes in the discriminator. The experiment is illustrated in Figure 4, where a randomly selected image from the CIFAR-10 dataset is gradually shifted vertically, shifting all pixels a single pixel downward at each step. In the figure, the sequence of shifted images is passed through the WWGAN and the WGAN-GP discriminators trained with their respective loss to reach an FID value of 40 for the generator. We observe with our WWGAN model, the discriminator values change continuously with the translation of the input image. In contrast, this type of continuity is not observed in models that are trained with the Euclidean Lipschitz condition. We note that WWGAN assigns a positive value to the image and gradually decreases to the end limit when the entire image is shifted away. Unlike WWGAN, WGAN-GP is highly sensitive to perturbations in image space and oscillates wildly, assigning highly positive (real label) and negative (fake labels) to images shifted less than 2 pixels away. We observed the same type of behavior across all images that we tested.

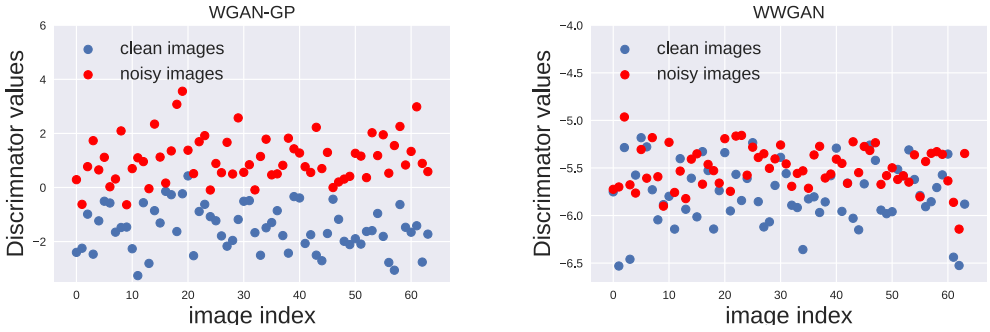

FIGURE 5. Here we test the robustness of the discriminator values to noise on real CIFAR-10 images. The noise is the RGB version of salt and pepper noise, where 15% of the pixels are modified. As we can see, the WGAN-GP discriminator values cluster according to noise, so they give different values to whether a real image is noisy or not. In the bottom figure, the WWGAN discriminator is more robust to noise and changes relatively little.

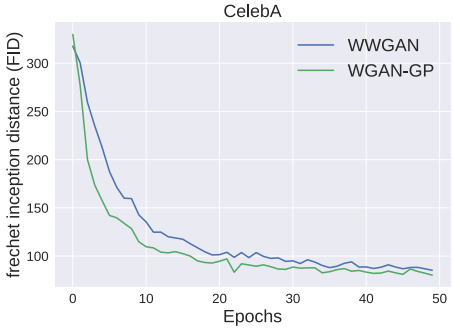

FIGURE 6. WWGAN gives comparable results with state of the art WGAN-GP training in terms of the Frechét Inception Distance (FID) of generated images. In terms of computation time, the overhead of WWGAN is negligible, with average epoch wall-clock times of 218.1 (s) and 236.9 (s), respectively, for the settings of our experiments.

4.2. **Discriminator robustness to noise.** In this experiment, we test the robustness of the discriminator to RGB salt and pepper noise, i.e., every pixel in the image has a probability to be changed to either 0 or 1. We chose our probability to be 15%, so 15% of the pixels are modified. We trained GANs under the WGAN-GP and the WWGAN loss, until each achieved an FID of about 40. We then use the trained discriminators and measure their values on real images with RGB salt and pepper noise. In Figure 5, we see that the WGAN-GP discriminator has separate clusters for noisy and clean images, while the WWGAN discriminator is more robust to the noise and assigns more consistent values to all these images.

5. Related works

In this section, we review the connection between the proposed work and literature.

Ground Metric for space of functions. Banach GAN (Adler and Lunz, 2018) pointed out the importance of ground metric in training Wasserstein loss function in GANs. They apply Sobolev norms and their induced gradient operator for the ground metric. Other than the Sobolev gradient, we apply the optimal transport induced operator (Otto, 2001; Villani, 2009). The gradient operator depends on the new ground metric structure within the sample space. We demonstrate that the optimal transport gradient provides the other practical 1-Lipschitz condition for training GANs.

Connection with Mean field games. Mean field games considers the optimal control problem in Wasserstein space (Cardaliaguet et al., 2015). In potential games, the Hamilton-Jacobi equation in Wasserstein space plays a vital roe (Gangbo et al., 2008). In this paper, we present a new Hamilton-Jacobi equation in Wasserstein space. It is the Ekional equation in Wasserstein space as shown in proposition 3. The new proposed equation has naturally the double characteristics properties as the ones in Mean field games. Here we demonstrate experimentally that the double characteristics property is very suitable for training GANs.

Geometric deep learning and Wasserstein metric on graphs. In geometric deep learning one considers mappings where the input space has a rich geometric structure (Bronstein et al., 2017). In particular one considers the case where the input space consists of functions defined on a graph (Raster images are examples, where the graphs are grids). One reason for doing this is that one can then define convolutions based on the group structures of these graphs.

Here we propose to formulate the graph structure into the weighted Laplacian matrix. This matrix is connected to the *Wasserstein metric tensor on discrete space* (Chow et al., 2012; Maas, 2011; Mielke, 2011). A systemic geometry study is provided by (Li, 2018). The discrete Wasserstein metric tensor gives analytic formulas for proposing the graph structure in sample space into learning loss function. The Wasserstein of Wasserstein loss function is one of examples in this direction.

Wasserstein natural gradients. Recent work has also investigated the notion of natural gradients based on the Riemannian structures derived from optimal transport (Li and Montúfar, 2018). In this case, optimal transport serves to define an optimization method, rather than a loss function as in the present paper. This approach has also been applied to training of GANs, where it leads to an iterative regularizer for the generator (Lin et al., 2018).

6. Discussion

We proposed a Wasserstein loss function with Wasserstein ground metric for learning generative models. The Wasserstein ground metric introduces a graph / manifold structure into the sample space of the model and allows us to introduce meaningful priors to the learning model. Experiments demonstrate that this approach can contribute to making the generator and discriminator in GANs more stable with respect to noise and the natural variability of image data.

We consider the Wasserstein of Wasserstein loss an important advance at a conceptual level. It has a physical intuition. Consider a physical motion or translation in pixel space. It corresponds to a change or motion in image space, and it further changes the distribution over images accordingly. The double characteristic property of the Wasserstein Eikonal equation reflects this intuition analytically. We regard it as surprising that this high level approach can be translated to practical computational methods. Remarkably, our approach has no additional computational cost over the standard Wasserstein loss function with L^2 (Euclidean) ground metric.

In the future, we suggest to explore the consequences of our approach from the statistical and optimization point of view. Also, to continue exploring the role of the graph structure that is chosen to define the Wasserstein ground metric in relation to specific data types.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^{o} 757983).

References

J. Adler and S. Lunz. Banach Wasserstein GAN. *arXiv:1806.06621 [cs, math]*, 2018.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*, 2017.

A. L. Bertozzi and A. Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling & Simulation*, 10(3):1090–1118, 2012.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. ISSN 1053-5888. doi: 10.1109/MSP.2017.2693418.

P. Cardaliaguet, F. Delarue, J.-M. Lasry, and P.-L. Lions. The master equation and the convergence problem in mean field games, 2015.

S.-N. Chow, W. Huang, Y. Li, and H. Zhou. Fokker–Planck Equations for a Free Energy Functional or Markov Process on a Graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.

B. Engquist and Y. Yang. Seismic imaging and optimal transport, 2018.

C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein Loss. *arXiv:1506.05439 [cs, stat]*, 2015.

W. Gangbo, T. Nguyen, and A. Tudorascu. Hamilton-Jacobi equations in the Wasserstein space. *Methods Appl. Anal.*, 15(2):155–184, 06 2008. URL `https://projecteuclid.org:443/euclid.maa/1234536492`.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc., 2017.

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

W. Li. Geometry of probability simplex via optimal transport. *arXiv:1803.06360 [math]*, 2018.

W. Li and G. Montúfar. Natural gradient via optimal transport. *Information Geometry*, 1(2):181–214, Dec 2018. ISSN 2511-249X. doi: 10.1007/s41884-018-0015-3. URL `https://doi.org/10.1007/s41884-018-0015-3`.

A. Lin, W. Li, S. Osher, and G. Montúfar. Wasserstein proximal of GANs. *CAM report 18-53*, 2018.

J. Maas. Gradient Flows of the Entropy for Finite Markov Chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

R. Meng, Q. Cui, and C. Yuan. A survey of image information hiding algorithms based on deep learning. 2018.

A. Mielke. A Gradient Structure for Reaction–diffusion Systems and for Energy-Drift-Diffusion Systems. *Nonlinearity*, 24(4):1329, 2011.

G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein Training of Restricted Boltzmann Machines. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3718–3726. Curran Associates, Inc., 2016.

Y. Mroueh, T. Sercu, and V. Goel. McGan: Mean and covariance feature matching GAN. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2527–2535, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/mroueh17a.html`.

E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European conference on computer vision*, pages 490–503. Springer, 2006.

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

H. Petzka, A. Fischer, and D. Lukovnicov. On the regularization of Wasserstein GANs. *arXiv:1709.08894 [cs, stat]*, 2017.

M. A. Puthawala, C. D. Hauck, and S. J. Osher. Diagnosing forward operator error using optimal transport, 2018.

Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

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

L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017.

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238, 2007.

M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5): 1327–1336, 2011.

Appendix 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 $\Omega \subset \mathbb{R}^d$, Wasserstein-$p$ metric introduces the distance between probability density functions $\rho^0, \rho^1 \in \mathcal{P}(\Omega)$ as follows.

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

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

The dual problem of the linear programming has the form

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

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 $p = 2$.

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 $\operatorname{grad} f(x)$, one transports ρ^0 to ρ^1. The transport direction follows the characteristic of Eikonal equation, i.e. the geodesic curve in (Ω, d).

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 time $t = 0$, $t = 1$:

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

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 5 (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* $\Phi_1(x)$, $\Phi_2(x) \in \mathcal{F}(\Omega)/\mathbb{R}$.

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

Proposition 6 (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.

Appendix 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 $I = \{1, \cdots, n\}$. 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, $G = (I, E, \omega)$, 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 $G = (I, E, \omega)$ induces a graph Laplacian matrix function.

Definition 7 (Weighted Laplacian matrix)**.** *Given an undirected weighted graph $G = (I, E, \omega)$, with $I = \{1, \dots, n\}$, 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*

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

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

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

The Laplacian matrix function $L(p)$ is the discrete analog of the weighted Laplacian operator $-\nabla\cdot(\rho\nabla)$ from Definition 5.

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 $L(p)$:

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

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 $\Phi_i = \Phi_j$, $(i, j) \in E$. It 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 8 (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*

$$g_p(\sigma_1, \sigma_2) = \sigma_1^{\mathsf{T}}\Phi_2 = \sigma_2^{\mathsf{T}}\Phi_1 = \Phi_1^{\mathsf{T}}L(p)\Phi_2. \tag{9}$$

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

$$W(p^0, p^1)^2 := \inf_{p(t),\Phi(t)} \Big\{ \int_0^1 \Phi(t)^{\mathsf{T}} L(p(t))\Phi(t) dt \Big\}. \tag{10}$$

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 9 (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 4. Following the proof of Theorem 2, we prove Proposition 4. □

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 10 (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 11 (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 the fact 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. □

Appendix 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 firs transforms the $3 \times N \times N$ into a $3N^2$ vector, and runs through each vector and provides a probability of

changing any coordinate into 1.0 (the max pixel value) or 0.0 (once a change is decided, the probability of choosing 0.0 or 1.0 is equal.

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 $\beta_1 = 0.9$, $\beta_2 = 0$.

Department of Mathematics, University of California, Los Angeles, USA.

E-mail address: ydukler@math.ucla.edu

Department of Mathematics, University of California, Los Angeles, USA.

E-mail address: wcli@math.ucla.edu

Department of Mathematics, University of California, Los Angeles, USA.

E-mail address: atlin@math.ucla.edu

Department of Mathematics and Department of Statistics, University of California, Los Angeles, USA; Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany.

E-mail address: montufar@math.ucla.edu