Wasserstein GAN with Quadratic Transport Cost

Huidong Liu, Xianfeng Gu, Dimitris Samaras
Stony Brook University
Stony Brook, NY 11794, USA
{huidliu, gu, samaras}@cs.stonybrook.edu

Abstract

Wasserstein GANs are increasingly used in Computer Vision applications as they are easier to train. Previous WGAN variants mainly use the $l_1$ transport cost to compute the Wasserstein distance between the real and synthetic data distributions. The $l_1$ transport cost restricts the discriminator to be $1$-Lipschitz. However, WGANs with $l_1$ transport cost were recently shown to not always converge. In this paper, we propose WGAN-QC, a WGAN with quadratic transport cost. Based on the quadratic transport cost, we propose an Optimal Transport Regularizer (OTR) to stabilize the training process of WGAN-QC. We prove that the objective of the discriminator during each generator update computes the exact quadratic Wasserstein distance between real and synthetic data distributions. We also prove that WGAN-QC converges to a local equilibrium point with finite discriminator updates per generator update. We show experimentally on a Dirac distribution that WGAN-QC converges, when many of the $l_1$ cost WGANs fail to [22]. Qualitative and quantitative results on the CelebA, CelebA-HQ, LSUN and the ImageNet dog datasets show that WGAN-QC is better than state-of-art GAN methods. WGAN-QC has much faster runtime than other WGAN variants.

1. Introduction

Generative Adversarial Networks (GANs) [11] successfully model data distributions, and have been used in many vision applications such as image synthesis [20, 34, 28], image inpainting [39, 40], semantic segmentation [15, 26], etc. While widely used, GANs are known to be hard to train. GANs need to solve a min-max saddle point optimization problem [1]. Due to the competition between the discriminator and the generator, it is difficult to train a GAN to consistently produce meaningful images. Hence, a number of authors have attempted to stabilize GAN training [3, 29, 12, 24, 14]. The Boundary Equilibrium GAN (BEGAN) [3] adopts Proportional Control Theory to balance the training between generator and discriminator. [29] proposes strategies to regularize the gradient of the discriminator, leading to more stable GAN training. The Wasserstein GAN family [2, 12, 19, 24, 13, 10, 9] employs the Wasserstein distance to measure the distance between the distributions of real and synthetic data. The Wasserstein distance guarantees that even if there is no support between the real and generated data distributions, the discriminator still provides gradients to the generator, unlike the Jensen-Shannon (JS) divergence used in the original GAN objective [2].

It is still unclear whether GANs converge. Recent work [25, 23, 22] has shown that analyzing the Jacobian of the gradient field of the GAN parameters near the equilibrium...
of the discriminator computes the exact quadratic Wasserstein distance [36, 8], based on the quadratic transport cost, to stabilize the training process during each generator update.

Previous Wasserstein GAN variants mainly use the $l_1$ transport cost, because the discriminator can be restricted to be 1-Lipschitz [2, 12, 24] so that it can be used to approximate the Wasserstein distance between real and synthetic data distributions. The Sliced WGAN (SWGAN) [8] used the quadratic transport cost and computed the sliced Wasserstein distance [5] between real and synthetic data distributions. The 0-centered gradient penalty methods [22, 35] are proposed to stabilize GAN training, and are shown to be local convergent. However, a large-scale study [7] concluded that the regularization term in [22] actually leads to significant drop of Inception Scores (IS) [31] using the suggested regularization parameter. The regularization term trades off training stability with generated image quality.

In this paper, we improve the stability of GAN training, and, at the same time, optimize the discriminator such that the optimal discriminator can be used to compute the exact quadratic Wasserstein distance [33, 8], by using the quadratic transport cost in WGANs. WGANs with $l_1$ transport cost can be solved by a Two-Step method [19]. However, generalizing from $l_1$ to quadratic transport cost is non-trivial, because the quadratic transport cost does not satisfy the triangle inequality condition [19]. Note that the quadratic transport cost is also used in SWGAN [8], but 1) convergence is not guaranteed in [8], 2) the sliced Wasserstein Distance is a different metric from the Wasserstein distance [5], 3) SWGAN computes the generator from the primal form of Optimal Transport (OT) [36], whereas our proposed method computes the discriminator from the dual form of OT.

In summary, our main contributions are:

- We propose WGAN-QC, a new Wasserstein GAN with quadratic transport cost. In WGAN-QC, we propose a modified two-step computation to optimize the discriminator during each generator update.

- We propose the novel Optimal Transport Regularizer (OTR), based on the quadratic transport cost, to stabilize the training process of WGAN-QC. We prove that the objective of the discriminator computes the exact quadratic Wasserstein distance during each generator update.

- We prove that WGAN-QC can converge to a local equilibrium point given a small enough learning rate.

- We show that, contrary to many $l_1$ cost WGANs, WGAN-QC converges to the real data distribution in the 1-d Dirac distribution example. Qualitative and quantitative results on the CelebA [21], CelebA-HQ [17], LSUN bedroom [38], and the ImageNet dog [30] datasets show that WGAN-QC is better than state-of-the-art GAN methods.

- We show that WGAN-QC is 3.5x and 1.8x faster than WGAN-div which is faster than WGAN-GP on the CelebA and LSUN bedroom datasets, respectively.

We show some randomly generated face images in Fig. 1. These images look very realistic.

2. Optimal Transport

Since our framework is based on the Optimal Transport (OT), we shall briefly review the definition of OT in the Monge-Kantorovich dual formulation [36, 27].

The Monge-Kantorovich dual problem is given below:

**Problem 1.** Given two bounded domains $X$ and $Y$ and their probability measures $\nu \in \mathbb{P}(X)$, $\mu \in \mathbb{P}(Y)$, respectively, find functions $\phi$ and $\psi$ to solve

$$C(\mu, \nu) = \sup_{\phi, \psi \leq c} \left\{ \int \phi(y)d\mu(y) - \int \psi(x)d\nu(x) \right\}$$

(1)

where $c : X \times Y \mapsto [0, +\infty]$ is the transport cost.

In practice, given empirical distributions, we write Problem 1 in the discrete case. Suppose $\tilde{X} = \{x_j\}_{j \in \mathcal{X}}$ sampled from $\nu$ containing $n$ samples and $\tilde{Y} = \{y_i\}_{i \in \mathcal{Y}}$ sampled from $\mu$ containing $m$ samples, where $\mathcal{I}$ and $\mathcal{J}$ are disjoint index sets. Therefore, each element $x_j$ has a Dirac measure of $1/n$, and $y_i$ has a Dirac measure of $1/m$. Hence, the discrete Monge-Kantorovich dual problem is:

$$\max_{\phi, \psi} \frac{1}{m} \sum_{i \in \mathcal{I}} \phi(y_i) - \frac{1}{n} \sum_{j \in \mathcal{J}} \psi(x_j)$$

(2)

s.t. $\phi(y_i) - \psi(x_j) \leq c(x_j, y_i), \forall i \in \mathcal{I}, \forall j \in \mathcal{J}$

Kantorovich showed [16] that if the transport cost $c(\cdot, \cdot)$ satisfies the triangle inequality, then $\phi$ and $\psi$ can be unified into just one function. In WGAN [2], WGAN-GP [12], WGAN-TS [19] etc., the $l_1$ transport cost is used so that $\phi$ and $\psi$ are unified in one function and used as the discriminator. However, $\phi$ and $\psi$ cannot be unified when the quadratic transport cost is applied. The quadratic transport cost is:

$$c(x_j, y_i) = K \frac{1}{2} \|x_j - y_i\|^2_2$$

(3)

where $K$ is any constant positive real number. When Eq. (3) is applied, the optimal objective in Eq. (2) equals to $K \frac{1}{2} W_2^2$, which is the quadratic Wasserstein distance [33, 8].

3. WGAN with Quadratic Transport Cost

It has been recently shown [22] that WGANs with $l_1$ transport cost do not always converge. In this section, we propose WGAN-QC, a Wasserstein GAN with quadratic transport cost. We show in the next section that WGAN-QC always converges to a local equilibrium point.
3.1. Learning the Discriminator From the Kantorovich Potential

Let $D_w$ be the discriminator and $G_\theta$ the generator parameterized by $w$ and $\theta$, respectively. $P_r$ is the real data distribution and $P_z$ is a simple distribution (e.g., Gaussian or Uniform). When the discriminator loss in WGAN-QC is optimized, we want the discriminator to compute the exact quadratic Wasserstein distance. We also use the quadratic transport cost of Eq. (3), since it contributes to local convergence of WGAN-QC. In WGAN-QC, $K$ in Eq. (3) is set to $1/d$, where $d$ is the dimensionality of data $x_j$. We regard $\{y_i\}_{i \in I}$ as real data and $\{x_j\}_{j \in J}$ as synthetic data.

When the quadratic transport cost is used, $\phi$ and $\psi$ in (3) cannot be unified into one function. We need to select either $\phi$ or $\psi$ as the discriminator in WGAN-QC such that the optimal discriminator can be used to compute the quadratic Wasserstein distance. In fact, we care more about the discriminator’s value and gradients on the synthetic samples, because the generator is updated according to their gradients. So, we select $\psi$ to be the discriminator as it is defined on synthetic samples. In fact, Eq. (2) can be solved by linear programming. In this equation, if we substitute $\phi(y_i)$ by $H_i$ and $\psi(x_j)$ by $H_j$, we denote $H^*_i$ and $H^*_j$ to be the optimal solutions for $H_i$ and $H_j$, respectively. So, we can regress each $D_w(x_j)$ in the discriminator, to $H^*_j$.

We regress $\frac{1}{m}\sum_{i \in I} D_w(y_i)$ to $\frac{1}{m}\sum_{i \in I} H^*_i$ such that the optimal discriminator computes the quadratic Wasserstein distance. Thus, the discriminator provides an ascent direction for generator updates. We regress the discriminator as:

$$\min_w \frac{1}{2} \left( \frac{1}{m}\sum_{i \in I} D_w(y_i) - \frac{1}{m}\sum_{i \in I} H^*_i \right)^2$$

$$+ \frac{1}{2} \left( \frac{1}{n}\sum_{j \in J} (D_w(x_j) - H^*_j)^2 \right)$$

The generator loss is:

$$\min_\theta \mathcal{L}(\theta) = -\frac{1}{n}\sum_{j \in J} D_w(G_\theta(z_j))$$

3.2. Optimal Transport Regularization

There could be infinite solutions to Eq. (4). We need to regularize the discriminator. Hence, we introduce the Optimal Transport Regularizer (OTR) to stabilize the training process of WGAN-QC. The empirical optimal transport mapping is computed after the linear programming step:

$$\sigma(j) = \arg\min_{i \in I} \frac{K}{2} ||x_j - y_i||^2_2 + H^*_j - H^*_i$$

Essentially, Eq. (6) tries to find $H^*_{\sigma(j)} - H^*_j = c(x_j, y_{\sigma(j)})$, and Lemma 3.1 in [19] guarantees that for each $x_j$ we can always find a $y_{\sigma(j)}$ such that $H^*_{\sigma(j)} - H^*_j = c(x_j, y_{\sigma(j)})^1$. Therefore, $x_j$ minimizes

$$H^*_{\sigma(j)} = \inf_{j \in J} \{H^*_j + c(x_j, y_{\sigma(j)})\}$$

We use the $D_w(x_j)$ to regress $H^*_j$ and $x_j$ minimizes Eq. (7), thus $x_j$ is a local minimum and the first order derivative of Eq. (7) should be 0 in the continuous case, i.e.,

$$\nabla_x D_w(x_j) + K(x_j - y_{\sigma(j)}) = 0$$

Therefore, we propose the following Optimal Transport Regularizer (OTR) for WGAN-QC:

$$\frac{1}{2n}\sum_{j \in J} (||\nabla_x D_w(x_j)||^2 - K||y_{\sigma(j)} - x_j||^2)$$

where $|| \cdot ||$ is the $l_2$ norm. Eq. (8) holds only when Eq. (3), the quadratic transport cost, is applied in OT. Thus, OTR is specific to WGAN-QC. Eq. (8) has another explanation. According to Brenier’s theorem [6, 18], for an optimal discriminator, if $x_j$ is transformed to $y_{\sigma(j)}$, then Eq. (8) holds.

3.3. The Discriminator Loss of WGAN-QC

The complete discriminator loss of WGAN-QC (Algorithm 1) in each generator update step is:

$$\min_w \mathcal{L}(w)$$

$$= \frac{1}{2} \left( \frac{1}{m}\sum_{i \in I} D_w(y_i) - \frac{1}{m}\sum_{i \in I} H^*_i \right)^2$$

$$+ \frac{1}{2} \left( \frac{1}{n}\sum_{j \in J} (D_w(x_j) - H^*_j)^2 \right)$$

$$+ \frac{\gamma}{\sqrt{Km}} \sum_{j \in J} (||\nabla_x D_w(x_j)||^2 - K||y_{\sigma(j)} - x_j||^2)$$

The coefficient $\gamma$ in OTR balances the regression and regularization terms. The regularization term is obtained by multiplying (9) by $2\gamma/\sqrt{K}$. We found it is easier to set different values of $\gamma$ for different image sizes.

Next we show in Theorem 1 that the optimal discriminator $D^*_w$ during each generator update can be used to compute the exact quadratic Wasserstein distance between the real and synthetic data distributions for any $\gamma > 0$.

**Theorem 1.** If the discriminator in Eq.(10) has sufficient capacity such that the optimal objective of Eq.(10) is 0, then for any $\gamma > 0$, and any optimal solution $D^*_w$ to Eq.(10),

$$\frac{1}{m}\sum_{i \in I} D^*_w(y_i) - \frac{1}{n}\sum_{j \in J} D^*_w(x_j)$$

is the quadratic Wasserstein distance between $\hat{X}$ and $\hat{Y}$.

---

1See the complete proof in the supplementary material
2Please see supplementary material for complete proof
3Please refer to supplementary material for the proof.
Algorithm 1 WGAN-QC

1. Input: Real data $Y$, batch size $m$, $k_D$ and $\gamma$. Adam parameters, $\alpha, \beta_1, \beta_2$
2. Output: $G_{\theta}, D_w$
3. while $\theta$ has not converged do
4. Sample $\{y_t\}_{t \in T} \sim P_r$ from real data.
5. Sample $\{z_t\}_{j \in J} \sim P_z$ random noise.
6. Let $x_t = G_\theta(z_t), \forall j \in J$.
7. Solve the Linear Programming problem in Eq. (2), and obtain $H^*$.
8. $H^* \leftarrow H^* - \left(\sum_{k \in I \cup J} H_k^*\right)/(m + n), \forall t \in T \cup J$.
9. for $t = 0$ to $k_D$ do
10. $g_w \leftarrow$ the gradient of (10).
11. $w \leftarrow$ Adam($g_w, w, \alpha, \beta_1, \beta_2$)
12. end for
13. $g_\theta \leftarrow \nabla_{\theta} - \frac{1}{n} \sum_{j \in J} D_w(G_\theta(z_j))$
14. $\theta \leftarrow$ Adam($g_\theta, \theta, \alpha, \beta_1, \beta_2$)
15. end while

4. Convergence Analysis

WGAN with $l_1$ transport cost cannot always converge [22]. In this section, we analyze the convergence properties of WGAN-QC under finite discriminator iterations per generator iteration. First, we write the loss functions of the discriminator and generator under continuous distributions. Let $H_r(y)$ and $H_s(x)$ be the outputs of the linear programming part. $T : X \mapsto Y$ denotes that $x$ is transported to $y$ using Eq. (6). The loss of the discriminator is then:

$$\min_w L_D(w, \theta) = \frac{1}{2} \left( \mathbb{E}_{P_r(y)}[D_w(y)] - \mathbb{E}_{P_r(y)}[H_r(y)] \right)^2$$

$$+ \frac{1}{2} \mathbb{E}_{P_r(x)}[(D_w(x) - H_s(x))^2]$$

$$+ \frac{\lambda}{2} \mathbb{E}_{P_r(x)}[(\|\nabla_x D_w(x)\| - K\|y_{T(x)} - x\|)^2]$$

(12)

where $\lambda = 2\gamma/\sqrt{K}$, $P_r(x)$ denotes the probability of synthetic data, $P_r(y)$ denotes the probability of real data, and $\mathbb{E}_{[\cdot]}$ denote expectation. The loss of the generator is then:

$$\min_\theta L_G(w, \theta) = -\mathbb{E}_{P_r(x)}[D_w(G_\theta(z))]$$

(13)

In order to analyze the local convergence of WGAN-QC, we analyze the Jacobian of the gradient field of WGAN-QC. For simultaneous gradient descent the gradient field is

$$g(w, \theta) = \begin{pmatrix} \nabla_w L_D(w, \theta) \\ \nabla_\theta L_G(w, \theta) \end{pmatrix}$$

(14)

The gradient update operator is expressed as:

$$U(w, \theta) = \begin{pmatrix} w - \alpha \nabla_w L_D(w, \theta) \\ \theta - \alpha \nabla_\theta L_G(w, \theta) \end{pmatrix}$$

(15)

where $\alpha$ is the learning rate. The Jacobian of the gradient field is expressed as:

$$g'(w, \theta) = \begin{pmatrix} \nabla^2_w L_D(w, \theta) \\ \nabla^2_\theta L_D(w, \theta) \\ \nabla^2_w L_G(w, \theta) \\ \nabla^2_\theta L_G(w, \theta) \end{pmatrix}$$

(16)

We define $M_G$ and $M_D$ as the solution spaces for $G$ and $D$ respectively:

$$M_G := \{\theta | P_r(G_\theta(z)) = P_r(y)\},$$

$$M_D := \{w | L_D(w, \theta^*) = 0, \theta^* \in M_G\}$$

(17)

$(w^*, \theta^*)$ is an equilibrium point if $w^* \in M_D$ and $\theta^* \in M_G$. We define:

$$r(w) = \mathbb{E}_{P_r(y)}[D_w(y)^2 + \|\nabla_y D_w(y)\|^2]$$

(18)

From Eq. (12), we have $D_w(y) = 0$ and $\nabla_y D_w(y) = 0$, and thus $r(w^*) = 0$.

In order to analyze the convergence of our algorithm, we need two assumptions.

**Assumption 1.** We assume that the generator $G$ has sufficient expressive power that $P_r(G_\theta(z)) = P_r(y)$.

**Assumption 2.** If $(w, \theta)$ is not the equilibrium point, then $\partial_w^2 r(w^*) \neq 0$.

Assumption 1 is the feasibility assumption. Assumption 2 means that near the equilibrium point, the discriminator geometry is described by the second order derivative of $r$. The second order derivative of OTR in WGAN-QC is:

**Lemma 1.** The second order derivative of the regularization term

$$\frac{\lambda}{2} \mathbb{E}_{P_r(x)}[(\|\nabla_x D_w(x)\| - K\|y_{T(x)} - x\|)^2]$$

(19)

with respect to $(w, \theta)$ at the equilibrium point is given by:

$$M_R = \lambda \cdot \mathbb{E}_{P_r(x)}[\nabla_{w,x} D_w(x) \nabla_{w,x} D_w(x) \nabla_{w,x} D_w(x) \nabla_{w,x} D_w(x) \nabla_{w,x} D_w(x)]$$

(20)

Next, we give the Jacobian of the gradient field $g(w, \theta)$.

**Lemma 2.** The Jacobian of the gradient field $g(w, \theta)$ at the equilibrium point $(w^*, \theta^*)$ is given by:

$$g'(w^*, \theta^*) = \begin{pmatrix} M_{DD} + M_R & M_{GD} \\ 0 & 0 \end{pmatrix}$$

(21)

where $M_R$ is defined in Lemma 1,

$$M_{DD} = + \mathbb{E}_{P_r(y)}[\nabla_{w} D_w(y)] \cdot \mathbb{E}_{P_r(y)}[\nabla_{w} D_w(y) \nabla_{w} D_w(y)]$$

(22)

$$M_{GD} = -\mathbb{E}_{P_r(x)}[\nabla^2_{w,x} D_w(x) \nabla_{w} G_{\theta}(z)]$$

(23)

and $M_{DD} + M_R$ is positive definite.

4The second assumption is the same as Assumption III (i) in [22]

5Please refer to supplementary for proofs of Lemmas 1-3
then for small enough learning rate $\alpha$.

**Theorem 2.** WGAN-QC converges to a local equilibrium point.

However, in WGAN-QC, we employ an alternating gradient descent algorithm. Therefore, we show in Theorem 2 that if the discriminator converges, then the generator converges.

**Lemma 3.** For simultaneous gradient updates of $(w, \theta)$ in WGAN-QC using Eq. (15), if $w = w^*$, then $\theta = \theta^*$.

Lemma 3 shows that for simultaneous gradient descent, if the discriminator converges, then the generator converges. However, in WGAN-QC, we employ an alternating gradient descent algorithm. Therefore, we show in Theorem 2 that WGAN-QC converges to a local equilibrium point.

**Theorem 2.** Suppose Assumptions 1 and 2 are satisfied, then for small enough learning rate $\alpha$, there exists $\lambda$ such that WGAN-QC converges to a local equilibrium point.\(^6\)

### 5. Experiments

We first study the convergence of WGAN-QC, Critic Regularization GAN (CRGAN) [22] and WGAN-div [37] on a Dirac distribution. Then, we also compare WGAN-QC with state-of-the-art GANs, PGGAN [17], WGAN-GP [12], SWGAN [8], OT-GAN [32], and BigGAN [7] on the CelebA, CelebA-HQ, LSUN bedroom and the ImageNet dog datasets. We use the default published parameters for each method. Architecture details and other experimental settings are in supplementary material.

**Hyperparameter Study** WGAN-QC has a hyperparameter $\gamma$. We investigate the FID scores on the CelebA dataset w.r.t. $\gamma$ in Table 1. When $\gamma = 0.1$ and $\gamma = 1$, WGAN-QC achieves the best performance. Therefore, we suggest tuning $\gamma$ in $[0.01, 1.0]$ for WGAN-QC.

**Results on a Dirac distribution** We test WGAN-QC on a Dirac distribution which is concentrated at 0, with noise $z = -1$ with probability of 1. The generator is $G_{\theta}(z) = \theta \cdot z$. The discriminator is $D_{w}(x) = w \cdot x$. $(w, \theta)$ is initialized as $(0.01, 1.0)$. [22] showed that for this simple problem, the original GAN, WGAN and WGAN-GP do not converge, but CRGAN converges.

Results are in Fig. 2. The x-axis is the number of generator iterations and the y-axis is the output of the generator. Since the real data distribution is a Dirac distribution concentrated at 0, the generator output should converge to 0. WGAN-QC and CRGAN generate the true distribution, i.e., the output of the generator is 0. WGAN-div is oscillating around 0, mainly because the regularization term is very small near 0 according to the suggested parameter $p = 6$.

**Results on the CelebA dataset** Fig. 3 shows randomly generated images by each method. Many faces generated by WGAN-GP and WGAN-div have artifacts and some are incompletely generated. See faces marked with red boxes in Fig. 3 (a) and (b). CRGAN generates much better faces than WGAN-GP and WGAN-div. However, it tends to generate very similar faces (See faces marked with red and yellow boxes in Fig. 3 (c)). This suggests that CRGAN has a mode collapse problem. Fig. 3 (d) shows faces generated by WGAN-QC. All the faces generated by WGAN-QC are complete, smooth and distinct from each other. Almost all appear realistic.

FID scores of different methods on this dataset are in Table 2. WGAN-QC has the best performance of 12.9, which is 15.2% less than the second best method WGAN-div.

**Results on the CelebA-HQ dataset** We resize the face images in CelebA-HQ to $256 \times 256$ and train WGAN-QC on them. We can see that most of the randomly generated images by WGAN-QC in Figs. 1 and 4 look realistic. Even without progressive training, WGAN-QC can still generate

\[ FID \begin{array}{cccccccc} \gamma & 10^{-1} & 10^{-2} & 10^{-3} & 10^{-4} & 10^{-5} & 10^{-6} & 10^{-7} \\ \hline DCGAN & 52.0 & - & 61.1 & \hline PGGAN & 16.3 & 14.1 & 17.8 & \hline SWGAN & 23.2 & - & 52.9 & \hline WGAN-GP & 18.4 & - & 26.8 & \hline WGAN-div & 15.2 & 13.5 & 15.9 & \hline WGAN-QC & 12.9 & 7.7 & 13.9 & \end{array} \]

Table 1. WGAN-QC achieves the lowest FIDs at $\gamma = 0.1$ and 1. So, we suggest tuning $\gamma$ in $[0.01, 1.0]$ on other datasets.

Table 2. FID scores of different methods.

---

\(^6\)Please refer to supplementary material for proof of Theorem 2.
highly realistic 256×256 face images. We measure the performance of WGAN-QC following the strategy of [37] to compute the FID for WGAN-QC. We compare PGGAN, WGAN-div and WGAN-QC in Table 2. WGAN-QC’s FID of 7.7 considerably reduces the FID of WGAN-div by 43%.

In order to verify the smoothness of the face manifold learned by WGAN-QC, we interpolate between two faces randomly generated by WGAN-QC. Fig. 5 shows that the face transitions appear to be smooth. This suggests WGAN-QC captures the face manifold well.

Results on the LSUN dataset WGAN-QC has the smallest FID score of 13.9 in Table 2, 12% less than that of WGAN-div. Fig. 6 shows images generated by these methods. Some images generated by WGAN-GP and CRGAN are hard to recognize as bedrooms. Many images generated by WGAN-div are distorted. Almost all images produced by WGAN-QC are smooth and look like bedrooms.
Figure 4. Faces of size $256 \times 256$ randomly generated by WGAN-QC on CelebA-HQ. (Best seen in color)

Figure 5. Face interpolation by WGAN-QC. Transitions between faces appear good. (Best seen in color)

Table 3. Inception Scores on the ImageNet dog dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>CelebA (i / o)</th>
<th>LSUN (i / o)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWGAN</td>
<td>5.39</td>
<td></td>
</tr>
<tr>
<td>OT-GAN</td>
<td>8.97</td>
<td></td>
</tr>
<tr>
<td>BigGAN</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>WGAN-QC</td>
<td><strong>10.48</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Running time comparison. i / o means running time per generator iteration / overall training time.

<table>
<thead>
<tr>
<th>Method</th>
<th>CelebA (i / o)</th>
<th>LSUN (i / o)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGAN-GP</td>
<td>36.2ms / 5.0 days</td>
<td>47.8ms / 6.6 days</td>
</tr>
<tr>
<td>WGAN-div</td>
<td>30.6ms / 2.1 days</td>
<td>41.0ms / 2.4 days</td>
</tr>
<tr>
<td>WGAN-QC</td>
<td><strong>14.0ms / 0.6 days</strong></td>
<td><strong>18.6ms / 1.3 days</strong></td>
</tr>
</tbody>
</table>

Results on the ImageNet dog subset The Inception Scores (IS) achieved by state-of-the-art GAN methods are shown in Table 3. WGAN-QC is much better than SWGAN and OT-GAN on this dataset. WGAN-QC gives slightly higher IS than BigGAN, even though current version of WGAN-QC is unsupervised learning while BigGAN is supervised learning.

Run Time Comparison We run all comparisons on the same NVIDIA TITAN Xp under the same batch size of 64 on the CelebA and LSUN datasets. In Table 4 we show runtimes for WGAN-GP, WGAN-div and WGAN-QC. On both datasets WGAN-QC is the fastest one per iteration. Also, WGAN-QC requires the least overall training time on both datasets. WGAN-QC is 3.5x and 1.8x faster than WGAN-div on the CelebA and LSUN bedroom datasets, respectively.

6. Conclusions and Future Work

In this paper, we proposed WGAN-QC, a WGAN with quadratic transport cost whose discriminator is regularized
by optimal transport. We showed that the objective of the discriminator during each generator update computes the exact quadratic Wasserstein distance. We also proved that for small enough learning rates, WGAN-QC converges to a local equilibrium point. Consequently, we improved the state-of-the-art on four datasets while executing much faster than other WGAN variants.

In future work, we will extend WGAN-QC to the conditional version accepting image labels and investigate the performance of WGAN-QC on other large-scale datasets and higher-resolution images.

Acknowledgements

Chao Chen and Zhixin Shu provided valuable comments. This work was supported by a gift from Adobe, NSF grants CNS-1718014, IIS-1763981, 1762287, 1418255, 1737812 and DMS 1737876, the Partner University Fund, the SUNY2020 Infrastructure Transportation Security Center, and NSFC Grants 61772105, 61720106005 and 61432003.
References


