Training Wasserstein generative adversarial networks without gradient penalties

Dohyun Kwon

Department of Mathematics,

University of Wisconsin-Madison

October 24, 2021

This is Joint Work with Guido Montúfar (UCLA / Max Planck Institute), Yeoneung Kim and Insoon Yang (Seoul National University)

Overview

Overview

Wasserstein Generative Adversarial Networks

- Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have seen remarkable success in generating synthetic images. The generator G and the discriminator D compete with each other:

$$\min_G \max_D V(G, D)$$

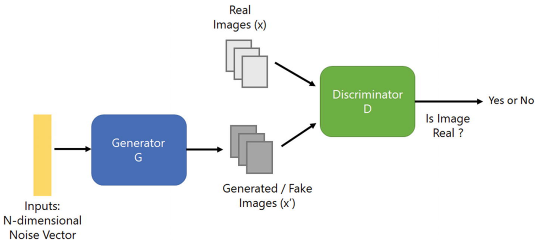


Figure: *The architecture of GANs [Salvaris-Dean-Tok, 2018]*

Here, $V(G, D) = E_{x\sim\text{data}}[log(D(x))] + E_{z\sim\text{noise}}[log(1 - D(G(z))]$.

- In the Wasserstein GAN framework proposed by Arjovsky, Chintala, and Bottou (2017), the training objective for the generator network is the Wasserstein distance to the target distribution.

Wasserstein Generative Adversarial Networks

- Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have seen remarkable success in generating synthetic images. The generator G and the discriminator D compete with each other:

$$\min_G \max_D V(G, D)$$

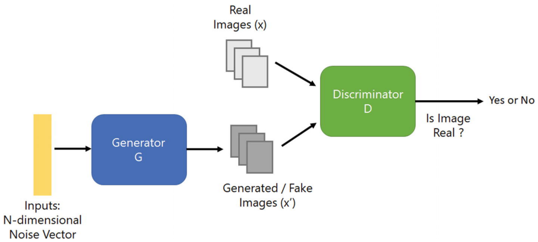


Figure: *The architecture of GANs [Salvaris-Dean-Tok, 2018]*

Here, $V(G, D) = E_{x\sim\text{data}}[log(D(x))] + E_{z\sim\text{noise}}[log(1 - D(G(z))]$.

- In the Wasserstein GAN framework proposed by Arjovsky, Chintala, and Bottou (2017), the training objective for the generator network is the Wasserstein distance to the target distribution.

Wasserstein Generative Adversarial Networks

The main objective of WGANs

For $0 < m << n$, let $\mu \in \mathscr{P}(\mathbb{R}^n)$ be a target distribution and $\rho \in \mathscr{P}(\mathbb{R}^m)$ be a source distribution. Find a parametrized generator $G_\theta : \mathbb{R}^m \to \mathbb{R}^n$ so that

$$W_p(\mu, G_\theta \# \rho) \approx 0.$$

- For $\mu, \nu \in \mathscr{P}_p(\Omega)$, the p-Wasserstein distance between two probability measures μ and ν in $\mathcal{P}(\Omega)$ is defined as

$$W_p(\mu, \nu) := \min \left\{ \int_{\Omega \times \Omega} |x - y|^p \mathrm{d}\gamma : \gamma \in \Pi(\mu, \nu) \right\}.$$

- Computing the Wasserstein distance has been a difficult task.

  A non-exhaustive list:
  [Benamou-Brenier, Numer. Math. 2000] The Benamou-Brenier formula
  [Cuturi, NIPS 2013] Sinkhorn distances
  [Benamou-Froese-Oberman, JCP 2014] The Monge-Ampére equation
  [Jacobs-Leger, Numer. Math. 2020] The back-and-forth method

  and much more...

Wasserstein Generative Adversarial Networks

The main objective of WGANs

For $0 < m << n$, let $\mu \in \mathscr{P}(\mathbb{R}^n)$ be a target distribution and $\rho \in \mathscr{P}(\mathbb{R}^m)$ be a source distribution. Find a parametrized generator $G_\theta : \mathbb{R}^m \to \mathbb{R}^n$ so that

$$W_p(\mu, G_\theta \# \rho) \approx 0.$$

- For $\mu, \nu \in \mathscr{P}_p(\Omega)$, the p-Wasserstein distance between two probability measures μ and ν in $\mathcal{P}(\Omega)$ is defined as

$$W_p(\mu, \nu) := \min \left\{ \int_{\Omega \times \Omega} |x - y|^p \mathrm{d}\gamma : \gamma \in \Pi(\mu, \nu) \right\}.$$

- Computing the Wasserstein distance has been a difficult task.

  A non-exhaustive list:
  [Benamou-Brenier, Numer. Math. 2000] The Benamou-Brenier formula
  [Cuturi, NIPS 2013] Sinkhorn distances
  [Benamou-Froese-Oberman, JCP 2014] The Monge-Ampére equation
  [Jacobs-Leger, Numer. Math. 2020] The back-and-forth method

  and much more...

Training WGANs if $p = 1$

- If $p = 1$, then $\phi^c = -\phi$ for all $\phi \in Lip_1$ and thus

$$W_1(\mu, \nu) = \sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}.$$

WGAN-WC [Arjovsky-Chintala-Bottou, 2017]

- clamp all the weights in the network of ϕ to a fixed box,
- but this can overly restrict the class of functions

WGAN-GP [Gulrajani-Ahmed-Arjovsky, 2017]

$$\inf_\theta \sup_\eta \left\{ \int_\Omega \phi_\eta(\mathrm{d}\mu - \mathrm{d}G_\theta \# \rho) + \lambda \int_\Omega (|D\phi_\eta| - 1)^2 \, \mathrm{d}\omega \right\}$$

- $\|D\phi\| = 1$ is not necessarily satisfied globally,
- applying the gradient penalty only at sample points is insufficient [Wei et al., 2018],
- WGAN-GP computes the minimum of a different optimal transport problem related to the congested transport [Milne-Nachman, 2021]

Training WGANs if $p = 1$

- If $p = 1$, then $\phi^c = -\phi$ for all $\phi \in Lip_1$ and thus

$$W_1(\mu, \nu) = \sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}.$$

WGAN-WC [Arjovsky-Chintala-Bottou, 2017]

- clamp all the weights in the network of ϕ to a fixed box,
- but this can overly restrict the class of functions

WGAN-GP [Gulrajani-Ahmed-Arjovsky, 2017]

$$\inf_\theta \sup_\eta \left\{ \int_\Omega \phi_\eta(\mathrm{d}\mu - \mathrm{d}G_\theta \# \rho) + \lambda \int_\Omega (|D\phi_\eta| - 1)^2 \,\mathrm{d}\omega \right\}$$

- $\|D\phi\| = 1$ is not necessarily satisfied globally,
- applying the gradient penalty only at sample points is insufficient [Wei et al., 2018],
- WGAN-GP computes the minimum of a different optimal transport problem related to the congested transport [Milne-Nachman, 2021]

Training WGANs if $p = 1$

- If $p = 1$, then $\phi^c = -\phi$ for all $\phi \in Lip_1$ and thus

$$W_1(\mu, \nu) = \sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}.$$

WGAN-WC [Arjovsky-Chintala-Bottou, 2017]

- clamp all the weights in the network of ϕ to a fixed box,
- but this can overly restrict the class of functions

WGAN-GP [Gulrajani-Ahmed-Arjovsky, 2017]

$$\inf_\theta \sup_\eta \left\{ \int_\Omega \phi_\eta(\mathrm{d}\mu - \mathrm{d}G_\theta \# \rho) + \lambda \int_\Omega (|D\phi_\eta| - 1)^2 \,\mathrm{d}\omega \right\}$$

- $\|D\phi\| = 1$ is not necessarily satisfied globally,
- applying the gradient penalty only at sample points is insufficient [Wei et al., 2018],
- WGAN-GP computes the minimum of a different optimal transport problem related to the congested transport [Milne-Nachman, 2021]

Our main questions

How to

- estimate the Wasserstein distance
- make an algorithm perform well in the generative setting
- enforce the Lipschitz constraint efficiently

A partial list of WGANs

WGAN-LP (Lipschitz Penalty) [Petzka-Fischer-Lukovnikov, 2018]

$$\inf_{\theta} \sup_{\eta} \left\{ \int_{\Omega} \phi_\eta (\mathrm{d}\mu - G_\theta \# \rho) + \int_{\Omega} \left(\max \left\{ 0, |D\phi_\eta|^2 - 1 \right\} \right)^2 d\omega \right\}$$

CT-GAN [Wei et al, 2018]

WGANs based c-transform:

$$\int_{\Omega} \phi \mathrm{d}\mu + \int_{\Omega} \phi^c \mathrm{d}\nu$$

- This method allows for a more accurate estimation of the true Wasserstein metric, but it does not perform well in the generative setting [Mallasto-Montúfar-Gerolin, 2019].

Overview

Revisit of the admissible condition (1/2)

Recall

$$W_1(\mu,\nu) = \sup\left\{\int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega)\right\}.$$

- The maximizer ϕ can take any values at $x \in (\mathrm{supp}\ (\mu) \cup \mathrm{supp}\ (\nu))^c$ as long as $\phi \in Lip_1(\Omega)$.
- Instead of the Lipschitz condition, we consider the following admissible condition:

$$\phi(x) - \phi(y) \le |x-y| \text{ for all } (x,y) \in \mathrm{supp}(\mu) \times \mathrm{supp}(\nu), \qquad \text{(A)}$$

- If both supp(μ) and supp(ν) are equal to Ω, then (A) is equivalent to the 1-Lipschitzness on Ω, which rarely happens in real-world data.
- Using (A) is more efficient if $\mathrm{supp}\ (\mu), \mathrm{supp}\ (\nu) \subset M$ for some manifold M such that $\dim(M) << \dim(\mathbb{R}^n) = n$.

Revisit of the admissible condition (1/2)

Recall

$$W_1(\mu,\nu) = \sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}.$$

- The maximizer ϕ can take any values at $x \in (\mathrm{supp}\ (\mu) \cup \mathrm{supp}\ (\nu))^c$ as long as $\phi \in Lip_1(\Omega)$.
- Instead of the Lipschitz condition, we consider the following admissible condition:

$$\phi(x) - \phi(y) \le |x-y| \text{ for all } (x,y) \in \mathrm{supp}(\mu) \times \mathrm{supp}(\nu), \qquad \text{(A)}$$

- If both supp(μ) and supp(ν) are equal to Ω, then (A) is equivalent to the 1-Lipschitzness on Ω, which rarely happens in real-world data.
- Using (A) is more efficient if $\mathrm{supp}\ (\mu), \mathrm{supp}\ (\nu) \subset M$ for some manifold M such that $\dim(M) << \dim(\mathbb{R}^n) = n$.

Revisit of the admissible condition (1/2)

Recall

$$W_1(\mu,\nu) = \sup\left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}.$$

- The maximizer ϕ can take any values at $x \in (\mathrm{supp}\ (\mu) \cup \mathrm{supp}\ (\nu))^c$ as long as $\phi \in Lip_1(\Omega)$.
- Instead of the Lipschitz condition, we consider the following admissible condition:

$$\phi(x) - \phi(y) \leq |x - y| \text{ for all } (x,y) \in \mathrm{supp}(\mu) \times \mathrm{supp}(\nu), \qquad \text{(A)}$$

- If both $\mathrm{supp}(\mu)$ and $\mathrm{supp}(\nu)$ are equal to Ω, then (A) is equivalent to the 1-Lipschitzness on Ω, which rarely happens in real-world data.
- Using (A) is more efficient if $\mathrm{supp}\ (\mu), \mathrm{supp}\ (\nu) \subset M$ for some manifold M such that $\dim(M) << \dim(\mathbb{R}^n) = n$.

Revisit of the admissible condition (2/2)

- For ϕ satisfies (A) and a transport plan γ satisfying $\gamma(A \times \Omega) = \mu(A)$ and $\gamma(\Omega \times A) = \nu(A)$ for all measurable subsets $A \subset \Omega$,

$$\int_{\Omega} \phi(\mathrm{d}\mu - \mathrm{d}\nu) = \int_{\Omega\times\Omega} \phi(x) - \phi(y)\mathrm{d}\gamma \leq \int_{\Omega\times\Omega} |x - y|\mathrm{d}\gamma$$

- As a consequence,

$$\sup\left\{\int_{\Omega} \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \text{ satisfies (A)}\right\} \leq \inf_{\gamma\in\Pi(\mu,\nu)}\left\{\int_{\Omega\times\Omega} |x - y|\mathrm{d}\gamma\right\} = W_1(\mu, \nu)$$

Revisit of the admissible condition (2/2)

- For ϕ satisfies (A) and a transport plan γ satisfying $\gamma(A \times \Omega) = \mu(A)$ and $\gamma(\Omega \times A) = \nu(A)$ for all measurable subsets $A \subset \Omega$,

$$\int_{\Omega} \phi(\mathrm{d}\mu - \mathrm{d}\nu) = \int_{\Omega \times \Omega} \phi(x) - \phi(y) \mathrm{d}\gamma \leq \int_{\Omega \times \Omega} |x - y| \mathrm{d}\gamma$$

- As a consequence,

$$\sup \left\{ \int_{\Omega} \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \text{ satisfies (A)} \right\} \leq \inf_{\gamma \in \Pi(\mu,\nu)} \left\{ \int_{\Omega \times \Omega} |x - y| \mathrm{d}\gamma \right\} = W_1(\mu, \nu)$$

c-transform on mini-batch

- In practice, one does not have access to the true distribution, but rather to mini-batches that are sampled from the available training data set.

$$\phi^c(y;\mu_n) := \inf_{x\in \operatorname{supp}(\mu_n)} \{|x-y| - \phi(x)\} \text{ for } y \in \Omega.$$

Here, μ_n is an empirical measures based on n i.i.d. observations X_1, X_2, . . . , X_n distributed according to μ.

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

- We use the c-transform on the support of η: for $\eta \in \mathcal{P}(\Omega)$, a function $\phi^c(\cdot;\eta) : \Omega \to \mathbb{R}$ is given by

$$\phi^c(y;\eta) := \inf_{x\in \operatorname{supp}(\eta)} \{|x-y| - \phi(x)\} \text{ for } y \in \Omega.$$

- Note that the original c-transform is defined as

$$\phi^c(y) := \inf_{x\in\Omega} \{|x-y| - \phi(x)\}.$$

c-transform on mini-batch

- In practice, one does not have access to the true distribution, but rather to mini-batches that are sampled from the available training data set.

$$\phi^c(y;\mu_n) := \inf_{x\in \mathrm{supp}(\mu_n)} \{|x-y| - \phi(x)\} \text{ for } y \in \Omega.$$

Here, μ_n is an empirical measures based on n i.i.d. observations X_1, X_2, ..., X_n distributed according to μ.

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

- We use the c-transform on the support of η: for $\eta \in \mathcal{P}(\Omega)$, a function $\phi^c(\cdot;\eta) : \Omega \to \mathbb{R}$ is given by

$$\phi^c(y;\eta) := \inf_{x\in \mathrm{supp}(\eta)} \{|x-y| - \phi(x)\} \text{ for } y \in \Omega.$$

- Note that the original c-transform is defined as

$$\phi^c(y) := \inf_{x\in\Omega} \{|x-y| - \phi(x)\}.$$

c-transform on mini-batch

- In practice, one does not have access to the true distribution, but rather to mini-batches that are sampled from the available training data set.

$$\phi^c(y;\mu_n) := \inf_{x \in \text{supp}(\mu_n)} \{|x-y| - \phi(x)\} \text{ for } y \in \Omega.$$

Here, μ_n is an empirical measures based on n i.i.d. observations X_1, X_2, ..., X_n distributed according to μ.

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

- We use the c-transform on the support of η: for $\eta \in \mathcal{P}(\Omega)$, a function $\phi^c(\cdot;\eta) : \Omega \to \mathbb{R}$ is given by

$$\phi^c(y;\eta) := \inf_{x \in \text{supp}(\eta)} \{|x-y| - \phi(x)\} \text{ for } y \in \Omega.$$

- Note that the original c-transform is defined as

$$\phi^c(y) := \inf_{x \in \Omega} \{|x-y| - \phi(x)\}.$$

Comparison between objective functions (1/2)

For two empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ and $\nu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i}$,

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega (-\phi)\mathrm{d}\nu_n,$$

$$\mathcal{J}_2(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot;\mu_n)\mathrm{d}\nu_n,$$

$$\mathcal{J}_3(\phi) := \int_\Omega (-\phi)^c(\cdot;\nu_n)\mathrm{d}\mu_n + \int_\Omega (-\phi)\mathrm{d}\nu_n,$$

$$\mathcal{J}_4(\phi) := \int_\Omega (-\phi)^c(\cdot;\nu_n)\mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot;\mu_n)\mathrm{d}\nu_n.$$

If ϕ satisfies the admissibility condition (A), then

$$-\phi(y) \le \phi^c(\cdot;\mu_n)$$

for all $y \in \mathrm{supp}\ (\nu)$.

Lemma

If ϕ satisfies the admissibility condition (A), then we have

$$\mathcal{J}_1(\phi) \le \mathcal{J}_2(\phi) \le \mathcal{J}_4(\phi) \text{ and } \mathcal{J}_1(\phi) \le \mathcal{J}_3(\phi) \le \mathcal{J}_4(\phi).$$

Comparison between objective functions (1/2)

For two empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ and $\nu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i}$,

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega (-\phi)\mathrm{d}\nu_n,$$

$$\mathcal{J}_2(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot;\mu_n)\mathrm{d}\nu_n,$$

$$\mathcal{J}_3(\phi) := \int_\Omega (-\phi)^c(\cdot;\nu_n)\mathrm{d}\mu_n + \int_\Omega (-\phi)\mathrm{d}\nu_n,$$

$$\mathcal{J}_4(\phi) := \int_\Omega (-\phi)^c(\cdot;\nu_n)\mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot;\mu_n)\mathrm{d}\nu_n.$$

If ϕ satisfies the admissibility condition (A), then

$$-\phi(y) \leq \phi^c(\cdot;\mu_n)$$

for all $y \in \mathrm{supp}\ (\nu)$.

Lemma

If ϕ satisfies the admissibility condition (A), then we have

$$\mathcal{J}_1(\phi) \leq \mathcal{J}_2(\phi) \leq \mathcal{J}_4(\phi) \text{ and } \mathcal{J}_1(\phi) \leq \mathcal{J}_3(\phi) \leq \mathcal{J}_4(\phi).$$

Comparison between objective functions (2/2)

Lemma

If ϕ satisfies the admissibility property (A)*, then we have*

$$\mathcal{J}_1(\phi) \leq \mathcal{J}_2(\phi) \leq \mathcal{J}_4(\phi) \text{ and } \mathcal{J}_1(\phi) \leq \mathcal{J}_3(\phi) \leq \mathcal{J}_4(\phi).$$

- Equivalently, if $\mathcal{J}_1 > \mathcal{J}_2$ or $\mathcal{J}_1 > \mathcal{J}_3$, then ϕ does not satisfy (A).

Lemma

If $\mathcal{J}_1(\phi) \leq \mathcal{J}_2(\phi)$ for all μ_n and ν_n, then ϕ satisfies the admissibility property (A)*. Here, μ_n and ν_n are empirical measures from μ and ν.*

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega (-\phi) \mathrm{d}\nu_n,$$
$$\mathcal{J}_2(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \mathrm{d}\nu_n,$$
$$\mathcal{J}_3(\phi) := \int_\Omega (-\phi)^c(\cdot; \nu_n) \mathrm{d}\mu_n + \int_\Omega (-\phi) \mathrm{d}\nu_n,$$
$$\mathcal{J}_4(\phi) := \int_\Omega (-\phi)^c(\cdot; \nu_n) \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \mathrm{d}\nu_n.$$

Comparison between objective functions (2/2)

Lemma

If ϕ satisfies the admissibility property (A)*, then we have*

$$\mathcal{J}_1(\phi) \leq \mathcal{J}_2(\phi) \leq \mathcal{J}_4(\phi) \text{ and } \mathcal{J}_1(\phi) \leq \mathcal{J}_3(\phi) \leq \mathcal{J}_4(\phi).$$

- Equivalently, if $\mathcal{J}_1 > \mathcal{J}_2$ or $\mathcal{J}_1 > \mathcal{J}_3$, then ϕ does not satisfy (A).

Lemma

If $\mathcal{J}_1(\phi) \leq \mathcal{J}_2(\phi)$ for all μ_n and ν_n, then ϕ satisfies the admissibility property (A)*. Here, μ_n and ν_n are empirical measures from μ and ν.*

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega (-\phi) \mathrm{d}\nu_n,$$
$$\mathcal{J}_2(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \mathrm{d}\nu_n,$$
$$\mathcal{J}_3(\phi) := \int_\Omega (-\phi)^c(\cdot; \nu_n) \mathrm{d}\mu_n + \int_\Omega (-\phi) \mathrm{d}\nu_n,$$
$$\mathcal{J}_4(\phi) := \int_\Omega (-\phi)^c(\cdot; \nu_n) \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \mathrm{d}\nu_n.$$

The orignal c-transform vs c-transform on mini-batch

- In fact, if ϕ is Lipschitz continuous, then $\phi^c = -\phi$. Therefore,

$$W_1(\mu, \rho) = \sup_{\phi \in Lip_1} \mathcal{I}_1 = \sup_{\phi} \mathcal{I}_2 = \sup_{\phi} \mathcal{I}_3 = \sup_{\phi} \mathcal{I}_4.$$

where

$$\mathcal{I}_1(\phi) = \textstyle\int_\Omega \phi \mathrm{d}\mu + \int_\Omega (-\phi) \mathrm{d}\nu, \qquad \mathcal{I}_2(\phi) = \displaystyle\int_\Omega \phi \mathrm{d}\mu + \int_\Omega \phi^c \mathrm{d}\nu,$$

$$\mathcal{I}_3(\phi) = \textstyle\int_\Omega (-\phi)^c \mathrm{d}\mu + \int_\Omega (-\phi) \mathrm{d}\nu, \qquad \mathcal{I}_4(\phi) = \displaystyle\int_\Omega (-\phi)^c \mathrm{d}\mu + \int_\Omega \phi^c \mathrm{d}\nu.$$

- However, the relation $\phi^c \leq -\phi$ does not hold for $\phi^c(\cdot;\eta)$ in general.
- As a consequence, $\phi^c(\cdot;\eta)$ is not necessarily equal to $-\phi$ even if ϕ is a 1-Lipschitz function.
- Similarly, $\mathcal{J}_1$ is not necessarily equal to $\mathcal{J}_2$ or $\mathcal{J}_3$ even though our discriminator is optimal.

Overview

How should we find the minimizer?

$\inf_{\nu \in P(\Omega)} W_1(\mu, \nu)$

1. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}$$

2. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \text{ satisfies (A)} \right\}$$

3. $$\sup_\phi \left\{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1] : \phi \text{ satisfies (A)} \right\}$$

4. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1] :$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

How should we find the minimizer?

$\inf_{\nu \in P(\Omega)} W_1(\mu, \nu)$

1. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}$$

2. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \text{ satisfies (A)} \right\}$$

3. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[\mathcal{J}_1] : \phi \text{ satisfies (A)} \}$$

4. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[\mathcal{J}_1] :$$

   $\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n)$ and

   $\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n)$ for all empirical measures $f_n \sim \mu, g_n \sim \nu \}$

How should we find the minimizer?

$\inf_{\nu \in P(\Omega)} W_1(\mu, \nu)$

1. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}$$

2. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \text{ satisfies (A)} \right\}$$

3. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1] : \phi \text{ satisfies (A)} \}$$

4. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1] :$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

How should we find the minimizer?

$\inf_{\nu \in P(\Omega)} W_1(\mu, \nu)$

1. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}$$

2. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \text{ satisfies (A)} \right\}$$

3. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[\mathcal{J}_1] : \phi \text{ satisfies (A)} \}$$

4. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[\mathcal{J}_1] :$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

How should we find the minimizer?

$\inf_{\nu \in P(\Omega)} W_1(\mu, \nu)$

1. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \in Lip_1(\Omega) \right\}$$

2. $$\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu - \mathrm{d}\nu) : \phi \text{ satisfies (A)} \right\}$$

3. $$\sup_\phi \left\{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[\mathcal{J}_1] : \phi \text{ satisfies (A)} \right\}$$

4. $$\sup_\phi \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[\mathcal{J}_1] :$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

Comparison based WGAN training

$$\inf_{\nu \in P(\Omega)} \sup_{\phi} \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1] :$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

Algorithm 1: CoWGAN

for *iter of training iterations* **do**
  for $t = 1, 2, \ldots, N_{critic}$ **do**
    if $\mathcal{J}_2 < \mathcal{J}_1$ **then**
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_2(\phi)$; increase $\mathcal{J}_2$
    else if $\mathcal{J}_3 < \mathcal{J}_1$ **then**
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_3(\phi)$; increase $\mathcal{J}_3$
    else
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_1(\phi)$; increase $\mathcal{J}_1$
  $\nu \leftarrow \nu - \tau \nabla_\nu \mathcal{J}_1$; decrease $\mathcal{J}_1$

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \, d\mu_n + \int_\Omega (-\phi) \, d\nu_n,$$
$$\mathcal{J}_2(\phi) := \int_\Omega \phi \, d\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \, d\nu_n,$$
$$\mathcal{J}_3(\phi) := \int (-\phi)^c(\cdot; \nu_n) \, d\mu_n + \int (-\phi) \, d\nu_n,$$

Comparison based WGAN training

$$\inf_{\nu \in P(\Omega)} \sup_{\phi} \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[\mathcal{J}_1] :$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

Algorithm 1: CoWGAN

for *iter of training iterations* **do**
  for $t = 1, 2, \ldots, N_{critic}$ **do**
    if $\mathcal{J}_2 < \mathcal{J}_1$ **then**
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_2(\phi)$; increase $\mathcal{J}_2 \leftarrow 1$
    else if $\mathcal{J}_3 < \mathcal{J}_1$ **then**
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_3(\phi)$; increase $\mathcal{J}_3 \leftarrow 1$
    else
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_1(\phi)$; increase $\mathcal{J}_1$
  $\nu \leftarrow \nu - \tau \nabla_\nu \mathcal{J}_1$; decrease $\mathcal{J}_1$

Step 1: Enforcing the admissible condition

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \, \mathrm{d}\mu_n + \int_\Omega (-\phi) \, \mathrm{d}\nu_n,$$
$$\mathcal{J}_2(\phi) := \int_\Omega \phi \, \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \, \mathrm{d}\nu_n,$$

Comparison based WGAN training

$$\inf_{\nu \in P(\Omega)} \sup_{\phi} \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1] :$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$
$$\mathcal{J}_1(\phi; f_n, g_n) \leq \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

Algorithm 1: CoWGAN

for *iter of training iterations* **do**
- **for** $t = 1, 2, \ldots, N_{critic}$ **do**
  - **if** $\mathcal{J}_2 < \mathcal{J}_1$ **then**
    - $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_2(\phi)$; increase $\mathcal{J}_2$
  - **else if** $\mathcal{J}_3 < \mathcal{J}_1$ **then**
    - $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_3(\phi)$; increase $\mathcal{J}_3$
  - **else**
    - $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_1(\phi)$; increase $\mathcal{J}_1 \leftarrow 2$
- $\nu \leftarrow \nu - \tau \nabla_\nu \mathcal{J}_1$; decrease $\mathcal{J}_1$

Step 2: Solving the maximization problem $\sup_\phi \mathcal{J}_1$

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega (-\phi) \mathrm{d}\nu_n,$$
$$\mathcal{J}_2(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \mathrm{d}\nu_n,$$

Comparison based WGAN training

$$\inf_{\nu \in P(\Omega)} \sup_{\phi} \{ \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1] :$$

$$\mathcal{J}_1(\phi; f_n, g_n) \le \mathcal{J}_2(\phi; f_n, g_n) \text{ and}$$

$$\mathcal{J}_1(\phi; f_n, g_n) \le \mathcal{J}_3(\phi; f_n, g_n) \text{ for all empirical measures } f_n \sim \mu, g_n \sim \nu \}$$

Algorithm 1: CoWGAN

for *iter of training iterations* **do**
  for $t = 1, 2, \dots, N_{critic}$ **do**
    if $\mathcal{J}_2 < \mathcal{J}_1$ **then**
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_2(\phi)$; increase $\mathcal{J}_2$
    else if $\mathcal{J}_3 < \mathcal{J}_1$ **then**
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_3(\phi)$; increase $\mathcal{J}_3$
    else
      $\phi \leftarrow \phi + \tau \nabla_\phi \mathcal{J}_1(\phi)$; increase $\mathcal{J}_1$
  $\nu \leftarrow \nu - \tau \nabla_\nu \mathcal{J}_1$; decrease $\mathcal{J}_1$ ← 3

Step 3: Solving the minimization problem w.r.t. ν

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \, d\mu_n + \int_\Omega (-\phi) \, d\nu_n,$$

$$\mathcal{J}_2(\phi) := \int_\Omega \phi \, d\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \, d\nu_n,$$

Comparison based WGAN training

$$\inf_\theta W_1(\mu, G_\theta\#\rho) = \inf_\theta \sup_\eta \left\{ \int_\Omega \phi_\eta \mathrm{d}(\mu - G_\theta\#\rho) : \phi_\eta \text{ satisfies (A)} \right\}.$$

Algorithm 2: CoWGAN

for *iter of training iterations* **do**
  for $t = 1, 2, \ldots, N_{critic}$ **do**
    if $\mathcal{J}_2 < \mathcal{J}_1$ **then**
      $\eta \leftarrow \text{Adam}(-\mathcal{J}_2, \eta)$
    else if $\mathcal{J}_3 < \mathcal{J}_1$ **then**
      $\eta \leftarrow \text{Adam}(-\mathcal{J}_3, \eta)$
    else
      $\eta \leftarrow \text{Adam}(-\mathcal{J}_1, \eta)$
  $\theta \leftarrow \text{Adam}(\mathcal{J}_1, \theta)$

$$\mathcal{J}_1(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega (-\phi)\mathrm{d}\nu_n,$$

$$\mathcal{J}_2(\phi) := \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot;\mu_n)\mathrm{d}\nu_n,$$

$$\mathcal{J}_3(\phi) := \int_\Omega (-\phi)^c(\cdot;\nu_n)\mathrm{d}\mu_n + \int_\Omega (-\phi)\mathrm{d}\nu_n,$$

$$\mathcal{J}_4(\phi) := \int_\Omega (-\phi)^c(\cdot;\nu_n)\mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot;\mu_n)\mathrm{d}\nu_n.$$

Overview

Task 1: Estimate the Wasserstein metric (Mini-batch size 256)

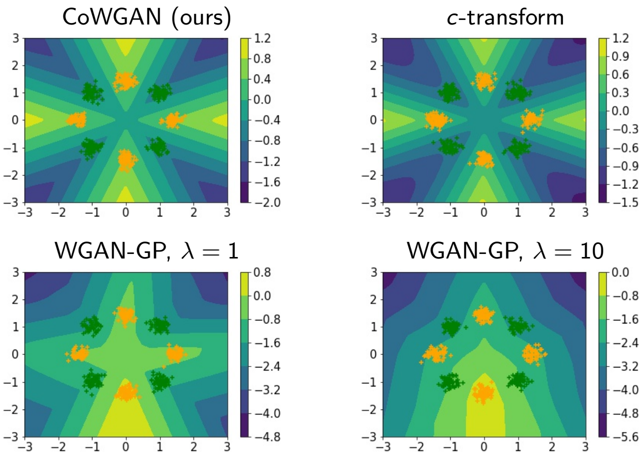


Figure: *The Kantorovich potential ϕ for two mixtures of 4 Gaussians (samples shown as green and yellow dots) after 2000 iterations with different methods and mini-batch size 256.*

Task 1: Estimate the Wasserstein metric (Mini-batch size 256)

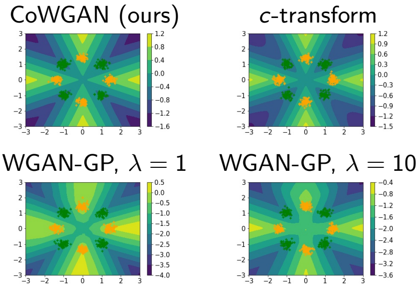


Figure: *The discriminator ϕ after 10,000 iterations with mini-batches of size 256.*

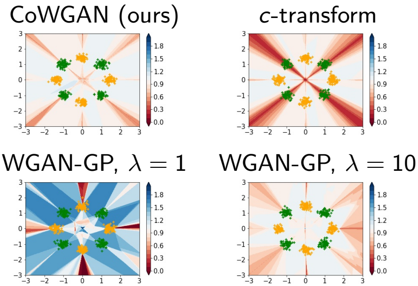


Figure: *Shown is $\|D\phi\|$ after 10,000 iterations with mini-batches of size 256.*

Task 1: Estimate the Wasserstein metric (Mini-batch size 256)

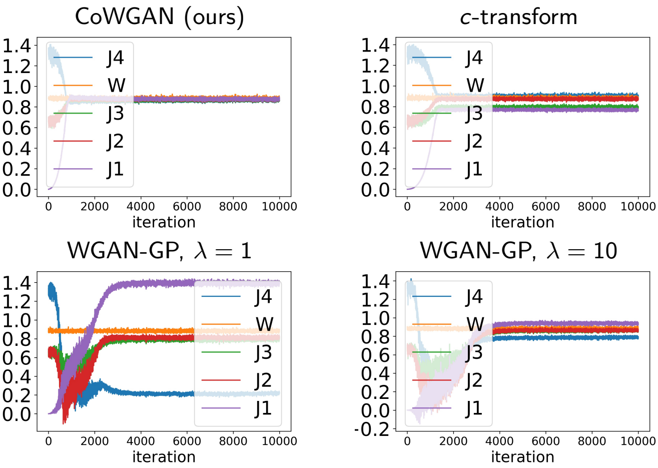


Figure: *The $\mathcal{J}_i$'s and the true Wasserstein distance (W).*

Task 1: Estimate the Wasserstein metric (Mini-batch size 8)

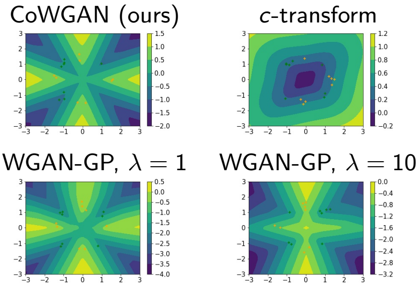


Figure: *The discriminator ϕ after 10,000 iterations with mini-batches of size 8.*

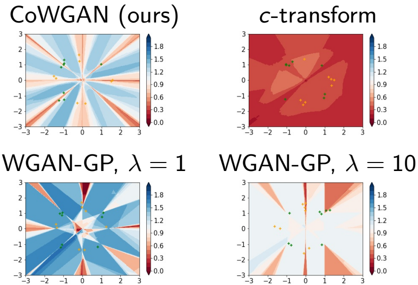


Figure: *Shown is $\|D\phi\|$ after 10,000 iterations with mini-batches of size 8.*

Task 1: Estimate the Wasserstein metric (Mini-batch size 8)

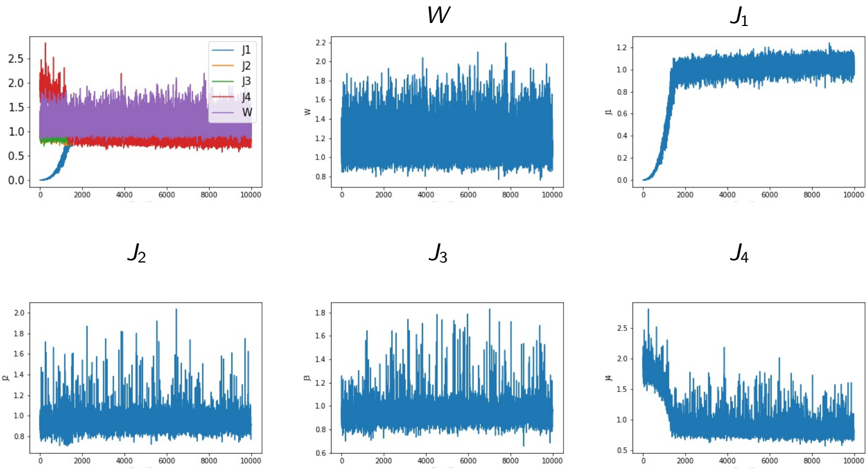


Figure: *The $\mathcal{J}_i$'s and the true Wasserstein distance (W) after 10,000 iterations with mini-batches of size 8*

Task 1: Estimate the Wasserstein metric (MNIST)

We sampled 5,000 images of digit 1 and 5,000 images of digit 2 from the MNIST dataset.

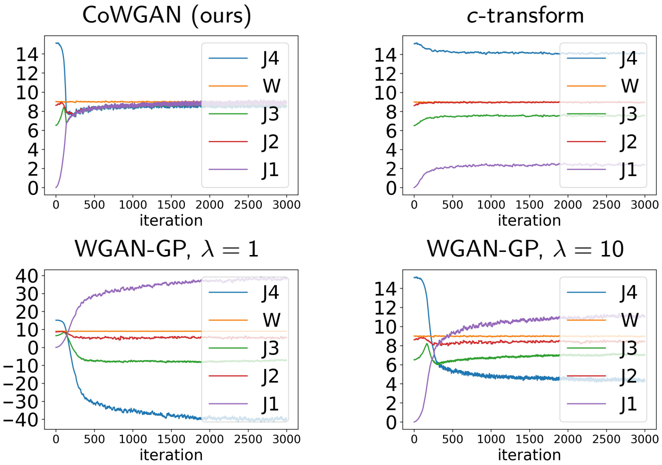


Figure: *The $\mathcal{J}_i$'s and the true Wasserstein distance (W) for the MNIST dataset.*

Task 2: Perform well in the generative setting

CoWGAN (ours)

WGAN-GP

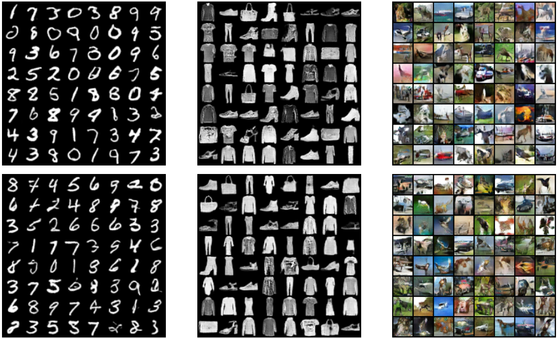


Figure: *From left to right the training data was MNIST, F-MNIST, and CIFAR-10. Visually, the generated images are of similar quality, but our algorithm runs six times faster in wall-clock time.*

Task 2: Perform well in the generative setting

CoWGAN (ours)

WGAN-GP

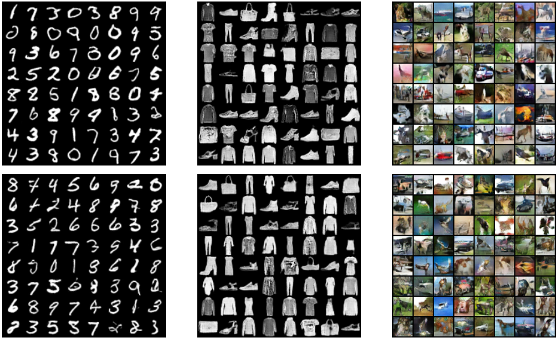


Figure: *From left to right the training data was MNIST, F-MNIST, and CIFAR-10. Visually, the generated images are of similar quality, but our algorithm runs six times faster in wall-clock time.*

Task 2: Perform well in the generative setting

The Fréchet inception distance (FID): the squared Wasserstein metric between two multidimensional Gaussian distributions

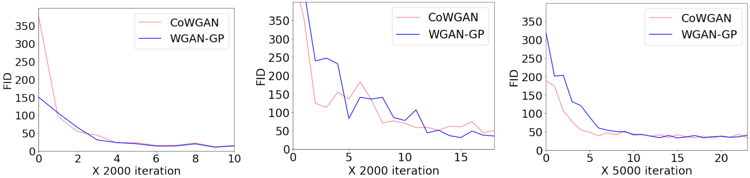


Figure: *FID; MNIST (left), F-MNIST(middle) and CIFAR10 (right).*

Task 3: Enforce the Lipschitz constraint

Compute

$$\sup_{x\sim\mu, y\sim G_\theta \# \rho} \frac{|\phi(x)-\phi(y)|}{|x-y|}.$$

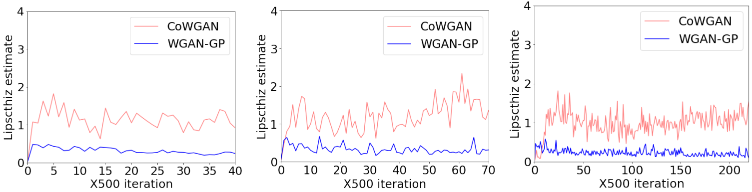


Figure: *Lipschitz constant; MNIST (left), F-MNIST(middle) and CIFAR10 (right)*

Overview

Which J_i's should be minimize?

$\inf_{\nu \in P(\Omega)} \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[W_1(\mu_n, \nu_n)]$

1.
$$\mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} \left[\sup \left\{ \int_\Omega \phi(\mathrm{d}\mu_n - \mathrm{d}\nu_n) : \phi(x_i) - \phi(y_j) \le |x_i - y_j| \text{ for all } 1 \le i, j \le n \right\} \right.$$

2.
$$\mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} \left[\sup_\phi \left\{ \int_\Omega \phi \mathrm{d}\mu_n + \int_\Omega \phi^c(\cdot; \mu_n) \mathrm{d}\nu_n \right\} \right]$$

3.
$$\mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu} \sup_\phi [J_2(\phi; \mu_n, \nu_n)]$$

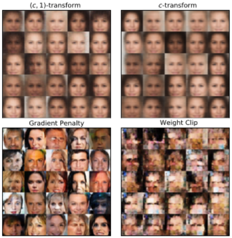
Which J_i's should be minimize?

$$\inf_{\nu \in P(\Omega)} \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[W_1(\mu_n, \nu_n)].$$

- The question is if an optimal ν is similar with the given probability measure μ.
- The answer is no as illustrated in the following lemma.

Lemma

Assume that $d = n = 1$ and $\mu \in \mathcal{P}_m(\Omega)$ for $m > 1$. Then, for any median y of μ, $\nu = \delta_y$ is a global minimizer of the above problem.



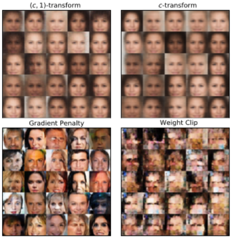
Which J_i's should be minimize?

$$\inf_{\nu \in P(\Omega)} \mathbb{E}_{\mu_n \sim \mu, \nu_n \sim \nu}[W_1(\mu_n, \nu_n)].$$

- The question is if an optimal ν is similar with the given probability measure μ.
- The answer is no as illustrated in the following lemma.

Lemma

Assume that $d = n = 1$ and $\mu \in \mathcal{P}_m(\Omega)$ for $m > 1$. Then, for any median y of μ, $\nu = \delta_y$ is a global minimizer of the above problem.



Controlling the centrality

For $\epsilon \in (0,1)$, consider

$$\inf_{\nu \in P(\Omega)} \sup_{\phi \in \mathcal{A}} E_{\mu_n \sim \mu, \nu_n \sim \nu} \left[(1-\epsilon)\mathcal{J}_1 + \epsilon \mathcal{J}_2 \right]$$

Here, ϵ is a parameter controlling the centrality of points according to a new probability measure ν.

$$\inf_{\nu \in P(\Omega)} \sup_{\phi \in \mathcal{A}} E_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_1]$$

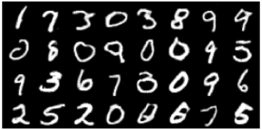


Figure: *CoWGAN;* $\epsilon = 0$

$$\inf_{\nu \in P(\Omega)} \sup_{\phi \in \mathcal{A}} E_{\mu_n \sim \mu, \nu_n \sim \nu} [\mathcal{J}_2]$$

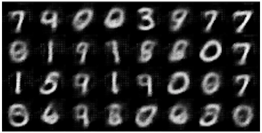


Figure: *Using* $\mathcal{J}_2$ *and* $\mathcal{J}_3$ *only;* $\epsilon = 1$

WGANs with the 2-Wasserstein distance

- Using the 2-Wasserstein distance has many advantages in theoretical perspectives as well as applications.
- For instance, the optimal map can be recovered from ϕ. This also can be useful when computing the Wasserstein gradient flow.
- However, in the generative setting it does not perform as good as the one with the 1-Wasserstein distance.

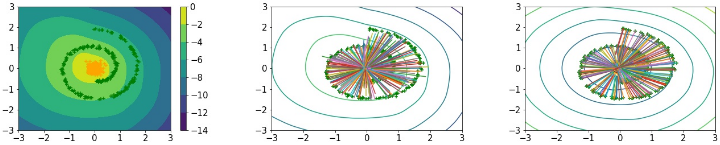


Figure: *The optimal map from yellow points to green points (middle), the optimal map from green points to yellow points (right)*

Summary

- Our comparison based WGAN training algorithm enforces a 1-Lipschitz bound without the need of introducing a gradient penalty.
- Consequently, no hyperparameter tuning for such a penalty is needed.
- Our new algorithm generates realistic synthetic images and works well with various types of data. Concretely, 8-Gaussians, MNIST, Fashion MNIST and CIFAR-10.

Thank you for your attention!

Kantorovich duality, $p = 2$

- Recall

$$
\begin{aligned}
W_2(\mu,\rho) &= \inf_T \sup_\phi \left\{ \int_\Omega |x - T(x)|^2 \mathrm{d}\rho(x) + \int_\Omega \phi \mathrm{d}\mu - \int_\Omega \phi \circ T \mathrm{d}\rho \right\}, \\
&= \sup_\phi \inf_T \left\{ \int_\Omega |x - T(x)|^2 \mathrm{d}\rho(x) + \int_\Omega \phi \mathrm{d}\mu - \int_\Omega \phi \circ T \mathrm{d}\rho \right\}, \\
&= \sup_\phi \left\{ \int_\Omega \phi \mathrm{d}\mu + \int_\Omega \inf_T \left\{ |x - T(x)|^2 - \phi \circ T \right\} \mathrm{d}\rho(x) \right\}.
\end{aligned}
$$

- Consequently,

$$W_2(\mu,\rho) = \sup_\phi \left\{ \int_\Omega \phi \mathrm{d}\mu + \int_\Omega \phi^c \mathrm{d}\nu \right\}$$

where ϕ^c is the c-transform of ϕ defined as

$$\phi^c(y) := \inf_{x\in\Omega} \left\{ |x-y|^2 - \phi(x) \right\}.$$

- ϕ^c is also not easy to compute.

Kantorovich duality, $p = 2$

- Recall

$$W_2(\mu,\rho) = \inf_T \sup_\phi \left\{ \int_\Omega |x - T(x)|^2 \mathrm{d}\rho(x) + \int_\Omega \phi \mathrm{d}\mu - \int_\Omega \phi \circ T \mathrm{d}\rho \right\},$$
$$= \sup_\phi \inf_T \left\{ \int_\Omega |x - T(x)|^2 \mathrm{d}\rho(x) + \int_\Omega \phi \mathrm{d}\mu - \int_\Omega \phi \circ T \mathrm{d}\rho \right\},$$
$$= \sup_\phi \left\{ \int_\Omega \phi \mathrm{d}\mu + \int_\Omega \inf_T \left\{ |x - T(x)|^2 - \phi \circ T \right\} \mathrm{d}\rho(x) \right\}.$$

- Consequently,

$$W_2(\mu,\rho) = \sup_\phi \left\{ \int_\Omega \phi \mathrm{d}\mu + \int_\Omega \phi^c \mathrm{d}\nu \right\}$$

where ϕ^c is the c-transform of ϕ defined as

$$\phi^c(y) := \inf_{x\in\Omega} \left\{ |x - y|^2 - \phi(x) \right\}.$$

- ϕ^c is also not easy to compute.