Bregman and Wasserstein, with Applications to Generative Adversarial Networks (GANs) and beyond

Xin Guo

Joint work with Johnny Hong, Tianyi Lin and Nan Yang

University of California, Berkeley

IMS-FIPS 2018, September 10, King's College, London

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
    - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.
  - $\mathbb{P}_r$ could be complicated as d increases.

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.
  - $\mathbb{P}_r$ could be complicated as d increases.
- **Goal:** How to learn $\mathbb{P}_r$ from data?

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.
  - $\mathbb{P}_r$ could be complicated as d increases.
- **Goal:** How to learn $\mathbb{P}_r$ from data?
- **Idea:** Construct a sequence of parametric probability distributions $\mathbb{P}_\theta$ to approximate $\mathbb{P}_r$.

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.
  - $\mathbb{P}_r$ could be complicated as d increases.
- **Goal:** How to learn $\mathbb{P}_r$ from data?
- **Idea:** Construct a sequence of parametric probability distributions $\mathbb{P}_\theta$ to approximate $\mathbb{P}_r$.
  - $\mathbb{P}_\theta$ is a parametric distribution over $\mathcal{X}$.

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
    - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
    - $\mathbb{P}_r$ is unknown.
    - $\mathbb{P}_r$ could be complicated as d increases.
- **Goal:** How to learn $\mathbb{P}_r$ from data?
- **Idea:** Construct a sequence of parametric probability distributions $\mathbb{P}_\theta$ to approximate $\mathbb{P}_r$.
    - $\mathbb{P}_\theta$ is a parametric distribution over $\mathcal{X}$.
    - $\mathbb{P}_\theta$ is **structured**!

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.
  - $\mathbb{P}_r$ could be complicated as d increases.
- **Goal:** How to learn $\mathbb{P}_r$ from data?
- **Idea:** Construct a sequence of parametric probability distributions $\mathbb{P}_\theta$ to approximate $\mathbb{P}_r$.
  - $\mathbb{P}_\theta$ is a parametric distribution over $\mathcal{X}$.
  - $\mathbb{P}_\theta$ is **structured**!
  - $\mathbb{P}_\theta$ is **known**.

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.
  - $\mathbb{P}_r$ could be complicated as d increases.
- **Goal:** How to learn $\mathbb{P}_r$ from data?
- **Idea:** Construct a sequence of parametric probability distributions $\mathbb{P}_\theta$ to approximate $\mathbb{P}_r$.
  - $\mathbb{P}_\theta$ is a parametric distribution over $\mathcal{X}$.
  - $\mathbb{P}_\theta$ is **structured**!
  - $\mathbb{P}_\theta$ is **known**.
- **Question 1:** How to generate $\mathbb{P}_\theta$?

Problem Set-Up

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$ where $X_i \in \mathbb{R}^d$.
  - $\mathbb{P}_r$ is the true distribution over sample $\mathcal{X}$.
  - $\mathbb{P}_r$ is unknown.
  - $\mathbb{P}_r$ could be complicated as d increases.
- **Goal:** How to learn $\mathbb{P}_r$ from data?
- **Idea:** Construct a sequence of parametric probability distributions $\mathbb{P}_\theta$ to approximate $\mathbb{P}_r$.
  - $\mathbb{P}_\theta$ is a parametric distribution over $\mathcal{X}$.
  - $\mathbb{P}_\theta$ is **structured**!
  - $\mathbb{P}_\theta$ is **known**.
- **Question 1:** How to generate $\mathbb{P}_\theta$?
- **Question 2*:** How to evaluate the quality of $\mathbb{P}_\theta$?

Roadmap

1. Bregman Divergence Function
2. Generative Adversarial Networks (GANs)
3. Wasserstein Divergence and GANs
4. Relaxed Wasserstein
   - Moment Estimate, Concentration Inequality, and Duality
   - Continuity, Differentiability
   - Gradient Descent Scheme
5. Empirical Results
   - Experiment Setup
   - MNIST and Fashion-MNIST datasets
   - CIFAR-10 and ImageNet datasets
6. Conclusions

A curious and simple math puzzle

- Given a random variable X, and a flitration $\mathcal{G}$, find (all?) loss/divergence functions $F(x, y)$ such that

$$\arg\min_{Y \in \mathcal{G}} E[F(X, Y)] = E[X|\mathcal{G}].$$

A curious and simple math puzzle

- Given a random variable X, and a flitration $\mathcal{G}$, find (all?) loss/divergence functions $F(x, y)$ such that

$$\arg\min_{Y \in \mathcal{G}} E[F(X, Y)] = E[X|\mathcal{G}].$$

- *Example:* L^2 function: $\arg\min_{Y \in \mathcal{G}} E[(X - Y)^2] = E[X|\mathcal{G}]$

A curious and simple math puzzle

- Given a random variable X, and a flitration $\mathcal{G}$, find (all?) loss/divergence functions $F(x, y)$ such that

$$\arg\min_{Y \in \mathcal{G}} E[F(X, Y)] = E[X|\mathcal{G}].$$

- *Example:* L^2 function: $\arg\min_{Y \in \mathcal{G}} E[(X - Y)^2] = E[X|\mathcal{G}]$
- *Counter-example:* L^1 function

A curious and simple math puzzle

- Given a random variable X, and a flitration $\mathcal{G}$, find (all?) loss/divergence functions $F(x, y)$ such that

$$\arg\min_{Y \in \mathcal{G}} E[F(X, Y)] = E[X|\mathcal{G}].$$

- *Example:* L^2 function: $\arg\min_{Y \in \mathcal{G}} E[(X - Y)^2] = E[X|\mathcal{G}]$
- *Counter-example:* L^1 function
- Is L^2 the unique choice?

- **Answer:** Bregman Loss Fucntions $D_\phi(x, y)$ (Banerjee, G. and Wang (2005))

- **Answer:** Bregman Loss Fucntions $D_\phi(x, y)$ (Banerjee, G. and Wang (2005))
    - Sufficient
$$\arg\min_{Y \in \mathcal{G}} E[D_\phi(X, Y)] = E[X|\mathcal{G}].$$

- **Answer:** Bregman Loss Fucntions $D_\phi(x, y)$ (Banerjee, G. and Wang (2005))
  - Sufficient
  $$\arg\min_{Y \in \mathcal{G}} E[D_\phi(X, Y)] = E[X|\mathcal{G}].$$
  - Necessary: If for all X
  $$\arg\min_{y \in R^d} E[F(X, y)] = E[X].$$
  then with proper regularity conditions and up to an additive constant,
  $$F(x, y) = D_\phi(x, y)$$

What is Bregman Divergence Function?

- BDF $D_\phi(x, y)$

What is Bregman Divergence Function?

- BDF $D_{\phi}(x, y)$
  - Let $\phi : R^d \mapsto R$ be a strictly convex, differentiable function

What is Bregman Divergence Function?

- BDF $D_\phi(x, y)$
  - Let $\phi : R^d \mapsto R$ be a strictly convex, differentiable function
  - Then, $D_\phi : R^d \times R^d \mapsto R$ is defined as

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle.$$

What is Bregman Divergence Function?

- BDF $D_\phi(x, y)$
  - Let $\phi : R^d \mapsto R$ be a strictly convex, differentiable function
  - Then, $D_\phi : R^d \times R^d \mapsto R$ is defined as

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle.$$

- For any $x, y \in R^d$, $D_\phi(x, y) \geq 0$, the equality holds iff $x = y$.

Some examples of BDFs

- $\phi(x) = x^2$, then $D_\phi(x, y) = (x - y)^2$

Some examples of BDFs

- $\phi(x) = x^2$, then $D_\phi(x, y) = (x - y)^2$
- Let $p \doteq (p_1, \ldots, p_d)$ be a probability distribution

Some examples of BDFs

- $\phi(x) = x^2$, then $D_\phi(x, y) = (x - y)^2$
- Let $p \doteq (p_1, \ldots, p_d)$ be a probability distribution
  - $\sum_{j=1}^{d} p_j = 1$, with $\phi(p) \doteq \sum_{j=1}^{d} p_j \log p_j$ (negative Shannon entropy) is strictly convex on the d-simplex.

Some examples of BDFs

- $\phi(x) = x^2$, then $D_\phi(x, y) = (x - y)^2$
- Let $p \doteq (p_1, \ldots, p_d)$ be a probability distribution
  - $\sum_{j=1}^{d} p_j = 1$, with $\phi(p) \doteq \sum_{j=1}^{d} p_j \log p_j$ (negative Shannon entropy) is strictly convex on the d-simplex.
  - Let $q = (q_1, \ldots, q_d)$ be another probability distribution

Some examples of BDFs

- $\phi(x) = x^2$, then $D_\phi(x, y) = (x - y)^2$
- Let $p \doteq (p_1, \ldots, p_d)$ be a probability distribution
  - $\sum_{j=1}^d p_j = 1$, with $\phi(p) \doteq \sum_{j=1}^d p_j \log p_j$ (negative Shannon entropy) is strictly convex on the d-simplex.
  - Let $q = (q_1, \ldots, q_d)$ be another probability distribution
  -

$$
\begin{aligned}
D_\phi(p, q) &= \sum_{j=1}^d p_j \log p_j - \sum_{j=1}^d q_j \log q_j \\
&\quad -\langle p - q, \nabla\phi(q)\rangle \\
&= \sum_{j=1}^d p_j \log\left(p_j / q_j\right),
\end{aligned}
$$

is the KL-divergence between p and q

Proof of sufficiency

Let Y be any $\mathcal{G}$-measurable random variable, and $Y^* \doteq E[X|\mathcal{G}]$.

- Then

$$\begin{aligned} & E[D_\phi(X, Y)] - E[D_\phi(X, Y^*)] \\ = \; & E[\phi(Y^*) - \phi(Y) - \langle X - Y, \nabla\phi(Y)\rangle \\ & + \langle X - Y^*, \nabla\phi(Y^*)\rangle]. \end{aligned}$$

Proof of sufficiency

Let Y be any $\mathcal{G}$-measurable random variable, and $Y^* \doteq E[X|\mathcal{G}]$.

- Then

$$\begin{aligned} & E[D_\phi(X, Y)] - E[D_\phi(X, Y^*)] \\ = \; & E[\phi(Y^*) - \phi(Y) - \langle X - Y, \nabla\phi(Y)\rangle \\ & + \langle X - Y^*, \nabla\phi(Y^*)\rangle]. \end{aligned}$$

- Notice

$$\begin{aligned} E[\langle X - Y, \nabla\phi(Y)\rangle] \quad & = E[E[\langle X - Y, \nabla\phi(Y)\rangle|\mathcal{G}]] \\ & = E[\langle Y^* - Y, \nabla\phi(Y)\rangle] \end{aligned}$$

Proof of sufficiency

Let Y be any $\mathcal{G}$-measurable random variable, and $Y^* \doteq E[X|\mathcal{G}]$.

- Then

$$\begin{aligned} & E[D_\phi(X,Y)] - E[D_\phi(X,Y^*)] \\ = \; & E[\phi(Y^*) - \phi(Y) - \langle X - Y, \nabla\phi(Y)\rangle \\ & + \langle X - Y^*, \nabla\phi(Y^*)\rangle]. \end{aligned}$$

- Notice

$$\begin{aligned} E[\langle X - Y, \nabla\phi(Y)\rangle] \quad & = E[E[\langle X - Y, \nabla\phi(Y)\rangle|\mathcal{G}]] \\ & = E[\langle Y^* - Y, \nabla\phi(Y)\rangle] \end{aligned}$$

- Thus $E[\langle X - Y^*, \nabla\phi(Y^*)\rangle] = 0$,

Proof of sufficiency

Let Y be any $\mathcal{G}$-measurable random variable, and $Y^* \doteq E[X|\mathcal{G}]$.

- Then

$$\begin{aligned} & E[D_\phi(X, Y)] - E[D_\phi(X, Y^*)] \\ = \; & E[\phi(Y^*) - \phi(Y) - \langle X - Y, \nabla\phi(Y)\rangle \\ & + \langle X - Y^*, \nabla\phi(Y^*)\rangle]. \end{aligned}$$

- Notice

$$\begin{aligned} E[\langle X - Y, \nabla\phi(Y)\rangle] & = E[E[\langle X - Y, \nabla\phi(Y)\rangle|\mathcal{G}]] \\ & = E[\langle Y^* - Y, \nabla\phi(Y)\rangle] \end{aligned}$$

- Thus $E[\langle X - Y^*, \nabla\phi(Y^*)\rangle] = 0$,
- And $E[D_\phi(X, Y)] - E[D_\phi(X, Y^*)] = E[D_\phi(Y^*, Y)] \geq 0$.

More facts about BDFs

- Introduced and studied in the context of projection (Csiszar (1975))

More facts about BDFs

- Introduced and studied in the context of projection (Csiszar (1975))
- Pythagoras theorem holds for BDF (Censor and Lent (1981))

More facts about BDFs

- Introduced and studied in the context of projection (Csiszar (1975))
- Pythagoras theorem holds for BDF (Censor and Lent (1981))
- Bijection between family of exponential distributions and BDFs, via Legendre duality (Merugu, Banerjee, Dhillon, Ghosh (2003))

More facts about BDFs

- Introduced and studied in the context of projection (Csiszar (1975))
- Pythagoras theorem holds for BDF (Censor and Lent (1981))
- Bijection between family of exponential distributions and BDFs, via Legendre duality (Merugu, Banerjee, Dhillon, Ghosh (2003))
- Widely applied to data analysis and machine learning, such as K-means clustering

More facts about BDFs

- Introduced and studied in the context of projection (Csiszar (1975))
- Pythagoras theorem holds for BDF (Censor and Lent (1981))
- Bijection between family of exponential distributions and BDFs, via Legendre duality (Merugu, Banerjee, Dhillon, Ghosh (2003))
- Widely applied to data analysis and machine learning, such as K-means clustering
- Well adopted in convex optimization

Generator Network [Goodfellow et al., 2014]

Generate the samples according to $\mathbb{P}_\theta$.

Generator Network [Goodfellow et al., 2014]

Generate the samples according to $\mathbb{P}_\theta$.

- The real samples $\mathcal{X}$ is inaccessible.

Generator Network [Goodfellow et al., 2014]

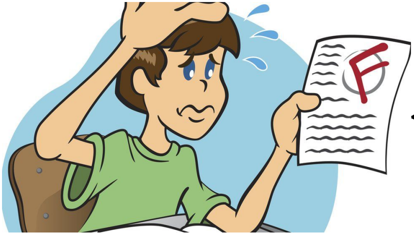
Generate the samples according to $\mathbb{P}_\theta$.

- The real samples $\mathcal{X}$ is inaccessible.
- Generate more compelling copies of $\mathcal{X}$.

Generator Network [Goodfellow et al., 2014]

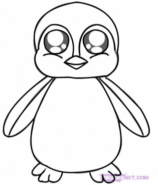
Generate the samples according to $\mathbb{P}_\theta$.

- The real samples $\mathcal{X}$ is inaccessible.
- Generate more compelling copies of $\mathcal{X}$.



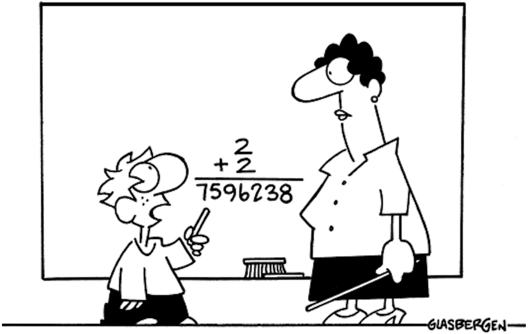
inaccessible

generate



How to Make Generator Network Better?

A knowledgeable mentor (discriminator)—



Discriminator Network [Goodfellow et al., 2014]

Determines whether the samples are generated or not.

Discriminator Network [Goodfellow et al., 2014]

Determines whether the samples are generated or not.

- has access to the real samples $\mathcal{X}$.

Discriminator Network [Goodfellow et al., 2014]

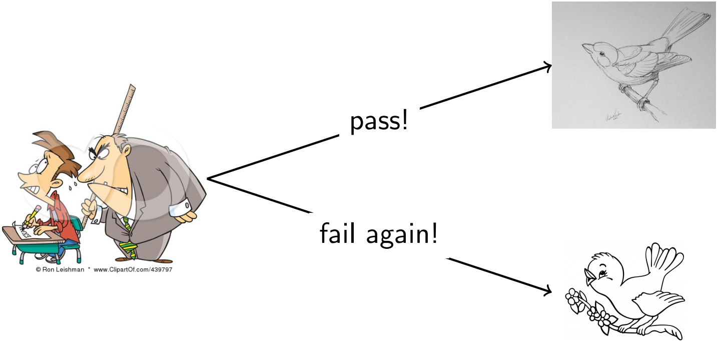
Determines whether the samples are generated or not.

- has access to the real samples $\mathcal{X}$.
- optimizes the generator network by identifying **faked** samples.

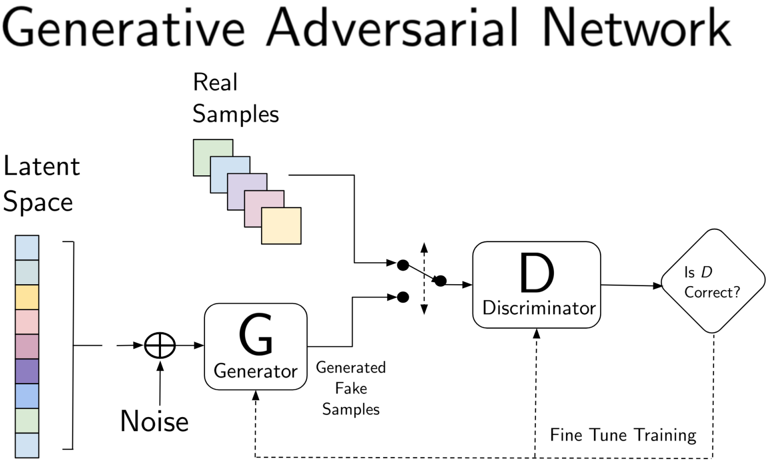
Discriminator Network [Goodfellow et al., 2014]

Determines whether the samples are generated or not.

- has access to the real samples $\mathcal{X}$.
- optimizes the generator network by identifying **faked** samples.



Graphical Model



Generative modeling

The procedure of generative modeling is to construct a class of suitable parametric probability distributions $\mathbb{P}_\theta$.

Generative modeling

The procedure of generative modeling is to construct a class of suitable parametric probability distributions $\mathbb{P}_\theta$.

- Generates latent variable $Z \in \mathcal{Z}$ with a fixed probability distribution $\mathbb{P}_Z$.

Generative modeling

The procedure of generative modeling is to construct a class of suitable parametric probability distributions $\mathbb{P}_\theta$.

- Generates latent variable $Z \in \mathcal{Z}$ with a fixed probability distribution $\mathbb{P}_Z$.
    - $\mathbb{P}_Z$ is **known** and **simple**, e.g., uniform distribution.

Generative modeling

The procedure of generative modeling is to construct a class of suitable parametric probability distributions $\mathbb{P}_\theta$.

- Generates latent variable $Z \in \mathcal{Z}$ with a fixed probability distribution $\mathbb{P}_Z$.
  - $\mathbb{P}_Z$ is **known** and **simple**, e.g., uniform distribution.
- Generates a sequence of parametric functions $g_\theta : \mathcal{Z} \to \mathcal{X}$.

Generative modeling

The procedure of generative modeling is to construct a class of suitable parametric probability distributions $\mathbb{P}_\theta$.

- Generates latent variable $Z \in \mathcal{Z}$ with a fixed probability distribution $\mathbb{P}_Z$.
    - $\mathbb{P}_Z$ is **known** and **simple**, e.g., uniform distribution.
- Generates a sequence of parametric functions $g_\theta : \mathcal{Z} \to \mathcal{X}$.
    - g_θ is **complicated** but **structured**.

Generative modeling

The procedure of generative modeling is to construct a class of suitable parametric probability distributions $\mathbb{P}_\theta$.

- Generates latent variable $Z \in \mathcal{Z}$ with a fixed probability distribution $\mathbb{P}_Z$.
  - $\mathbb{P}_Z$ is **known** and **simple**, e.g., uniform distribution.
- Generates a sequence of parametric functions $g_\theta : \mathcal{Z} \to \mathcal{X}$.
  - g_θ is **complicated** but **structured**.
  - g_θ is the reason why the generative modeling is **powerful**.

Generative modeling

The procedure of generative modeling is to construct a class of suitable parametric probability distributions $\mathbb{P}_\theta$.

- Generates latent variable $Z \in \mathcal{Z}$ with a fixed probability distribution $\mathbb{P}_Z$.
  - $\mathbb{P}_Z$ is **known** and **simple**, e.g., uniform distribution.
- Generates a sequence of parametric functions $g_\theta : \mathcal{Z} \to \mathcal{X}$.
  - g_θ is **complicated** but **structured**.
  - g_θ is the reason why the generative modeling is **powerful**.
- Construct $\mathbb{P}_\theta$ as the probability distribution of $g_\theta(Z)$. More specifically,

$$\mathbb{P}_\theta(dx) = \int_{\mathcal{Z}} 1_{\{g_\theta(z)=dx\}} \mathbb{P}_Z(dz) = \mathbb{E}_Z \left[1_{\{g_\theta(Z)=dx\}} \right].$$

GANs: different divergence functions

- **GANs:**

GANs: different divergence functions

- **GANs:**
  - LSGANs [Mao et al., 2016]: Least square loss.
  - DRAGANs [Kodali et al., 2017]: Regret minimization.
  - CGANs [Mirza and Osindero, 2014]: Conditional extension.
  - InfoGANs [Chen et al., 2016]: Information-theoretic extension.
  - ACGANs [Odena et al., 2017] Structured latent space.
  - EBGANs [Zhao et al., 2016]: New perspective of the energy.
  - BEGANs [Berthelot et al., 2017]: Auto-encoder extension.

GANs: different divergence functions

- **GANs:**
  - LSGANs [Mao et al., 2016]: Least square loss.
  - DRAGANs [Kodali et al., 2017]: Regret minimization.
  - CGANs [Mirza and Osindero, 2014]: Conditional extension.
  - InfoGANs [Chen et al., 2016]: Information-theoretic extension.
  - ACGANs [Odena et al., 2017] Structured latent space.
  - EBGANs [Zhao et al., 2016]: New perspective of the energy.
  - BEGANs [Berthelot et al., 2017]: Auto-encoder extension.
- **GANs training:** [Arjovsky and Bottou, 2017]

GANs: different divergence functions

- **GANs:**
  - LSGANs [Mao et al., 2016]: Least square loss.
  - DRAGANs [Kodali et al., 2017]: Regret minimization.
  - CGANs [Mirza and Osindero, 2014]: Conditional extension.
  - InfoGANs [Chen et al., 2016]: Information-theoretic extension.
  - ACGANs [Odena et al., 2017] Structured latent space.
  - EBGANs [Zhao et al., 2016]: New perspective of the energy.
  - BEGANs [Berthelot et al., 2017]: Auto-encoder extension.
- **GANs training:** [Arjovsky and Bottou, 2017]
- **Wasserstein GANs:**

GANs: different divergence functions

- **GANs:**
  - LSGANs [Mao et al., 2016]: Least square loss.
  - DRAGANs [Kodali et al., 2017]: Regret minimization.
  - CGANs [Mirza and Osindero, 2014]: Conditional extension.
  - InfoGANs [Chen et al., 2016]: Information-theoretic extension.
  - ACGANs [Odena et al., 2017] Structured latent space.
  - EBGANs [Zhao et al., 2016]: New perspective of the energy.
  - BEGANs [Berthelot et al., 2017]: Auto-encoder extension.
- **GANs training:** [Arjovsky and Bottou, 2017]
- **Wasserstein GANs:**
  - WGANs [Arjovsky et al., 2017]: Wasserstein L^1 divergence.
  - Improved WGANs [Gulrajani et al., 2017]: Gradient Penalty.

Several Choices of Divergence

The divergences to measure the difference between $\mathbb{P}$ and $\mathbb{Q}$ inlcude

Several Choices of Divergence

The divergences to measure the difference between $\mathbb{P}$ and $\mathbb{Q}$ inlcude

- Kullback-Leibler divergence:

$$KL(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \mathbb{P}(dx) \cdot \log\left(\frac{\mathbb{P}(dx)}{\mathbb{Q}(dx)}\right).$$

Several Choices of Divergence

The divergences to measure the difference between $\mathbb{P}$ and $\mathbb{Q}$ inlcude

- Kullback-Leibler divergence:

$$KL(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \mathbb{P}(dx) \cdot \log\left(\frac{\mathbb{P}(dx)}{\mathbb{Q}(dx)}\right).$$

- Jensen-Shannon (JS) divergence:

$$JS(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\left[KL(\mathbb{P}, \frac{\mathbb{P}+\mathbb{Q}}{2}) + KL(\mathbb{Q}, \frac{\mathbb{P}+\mathbb{Q}}{2})\right].$$

Several Choices of Divergence

The divergences to measure the difference between $\mathbb{P}$ and $\mathbb{Q}$ inlcude

- Kullback-Leibler divergence:

$$KL(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \mathbb{P}(dx) \cdot \log\left(\frac{\mathbb{P}(dx)}{\mathbb{Q}(dx)}\right).$$

- Jensen-Shannon (JS) divergence:

$$JS(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\left[KL(\mathbb{P}, \frac{\mathbb{P}+\mathbb{Q}}{2}) + KL(\mathbb{Q}, \frac{\mathbb{P}+\mathbb{Q}}{2})\right].$$

- Wasserstein divergence/distance of order p

$$W_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}\times\mathcal{X}} m(x,y)^p \; \pi(dx, dy)\right)^{\frac{1}{p}},$$

  with m a metric such as $m(x, y) = ||x - y||_q$ for $q \geq 1$.

Discussions on these divergences

- **Example:** Given $\theta \in [0, 1]$, assume that $\mathbb{P}$ and $\mathbb{Q}$ satisfy

$$\forall (x, y) \in \mathbb{P}, \;\; x = 0, \;\; y \sim \text{Uniform}(0, 1),$$
$$\forall (x, y) \in \mathbb{Q}, \;\; x = \theta, \;\; y \sim \text{Uniform}(0, 1),$$

Discussions on these divergences

- **Example:** Given $\theta \in [0, 1]$, assume that $\mathbb{P}$ and $\mathbb{Q}$ satisfy

$$\forall (x, y) \in \mathbb{P}, \ x = 0, \ y \sim \text{Uniform}(0, 1),$$
$$\forall (x, y) \in \mathbb{Q}, \ x = \theta, \ y \sim \text{Uniform}(0, 1),$$

- As $\theta \neq 0$,

$$KL(\mathbb{P}, \mathbb{Q}) = KL(\mathbb{Q}, \mathbb{P}) = +\infty, \ JS(\mathbb{P}, \mathbb{Q}) = \log(2), \ W_1(\mathbb{P}, \mathbb{Q}) = |\theta| \, .$$

Discussions on these divergences

- **Example:** Given $\theta \in [0,1]$, assume that $\mathbb{P}$ and $\mathbb{Q}$ satisfy

$$\forall (x,y) \in \mathbb{P}, \ x = 0, \ y \sim \text{Uniform}(0,1),$$
$$\forall (x,y) \in \mathbb{Q}, \ x = \theta, \ y \sim \text{Uniform}(0,1),$$

- As $\theta \neq 0$,

$$KL(\mathbb{P},\mathbb{Q}) = KL(\mathbb{Q},\mathbb{P}) = +\infty, \ JS(\mathbb{P},\mathbb{Q}) = \log(2), \ W_1(\mathbb{P},\mathbb{Q}) = |\theta|.$$

- As $\theta = 0$,

$$KL(\mathbb{P},\mathbb{Q}) = KL(\mathbb{Q},\mathbb{P}) = JS(\mathbb{P},\mathbb{Q}) = W_1(\mathbb{P},\mathbb{Q}) = 0.$$

Remark

- KL is **infinite** when two distributions are disjoint;

Remark

- *KL* is **infinite** when two distributions are disjoint;
- *JS* has sudden jump, **discontinuous** at $\theta = 0$;

Remark

- *KL* is **infinite** when two distributions are disjoint;
- *JS* has sudden jump, **discontinuous** at $\theta = 0$;
- W_1 is continuous and relatively smooth;

Remark

- KL is **infinite** when two distributions are disjoint;
- JS has sudden jump, **discontinuous** at $\theta = 0$;
- W_1 is continuous and relatively smooth;
- Wasserstein L^1 divergence outperforms KL and JS divergences but lacks the flexibility.

Remedy: Relaxed Wasserstein

Definition (G., Hong, Lin, and Yang 2018)

The Relaxed Wasserstein divergence between the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$W_{D_\phi}(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} D_\phi(x, y) \, \pi(dx, dy),$$

where D_ϕ is the Bregman divergence with a strictly convex and differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle$$

1. $W_{D_\phi}(\mathbb{P}, \mathbb{Q}) \geq 0$ and $= 0$ iff $\mathbb{P} = \mathbb{Q}$ almost everywhere.

Remedy: Relaxed Wasserstein

Definition (G., Hong, Lin, and Yang 2018)

The Relaxed Wasserstein divergence between the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$W_{D_\phi}(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} D_\phi(x, y) \, \pi(dx, dy),$$

where D_ϕ is the Bregman divergence with a strictly convex and differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle$$

1. $W_{D_\phi}(\mathbb{P}, \mathbb{Q}) \geq 0$ and $= 0$ iff $\mathbb{P} = \mathbb{Q}$ almost everywhere.
2. $W_{D_\phi}(\mathbb{P}, \mathbb{Q})$ is a metric, as it is asymmetric.

Remedy: Relaxed Wasserstein

Definition (G., Hong, Lin, and Yang 2018)

The Relaxed Wasserstein divergence between the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$W_{D_\phi}(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} D_\phi(x, y) \ \pi(dx, dy),$$

where D_ϕ is the Bregman divergence with a strictly convex and differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle$$

1. $W_{D_\phi}(\mathbb{P}, \mathbb{Q}) \geq 0$ and $= 0$ iff $\mathbb{P} = \mathbb{Q}$ almost everywhere.
2. $W_{D_\phi}(\mathbb{P}, \mathbb{Q})$ is a metric, as it is asymmetric.
3. $W_{D_\phi}(\mathbb{P}, \mathbb{Q})$ includes W_{KL} with $\phi(x) = -x^\top \log(x)$.

Relaxed Wasserstein as Divergence

Question: Is W_ϕ a good divergence?

Relaxed Wasserstein as Divergence

Question: Is W_ϕ a good divergence?

- **Point 1:** $W_\phi(\mathbb{P}, \mathbb{Q})$ should be small when $\mathbb{P}$ and $\mathbb{Q}$ are close.

Relaxed Wasserstein as Divergence

Question: Is W_ϕ a good divergence?

- **Point 1:** $W_\phi(\mathbb{P}, \mathbb{Q})$ should be small when $\mathbb{P}$ and $\mathbb{Q}$ are close.
- **Requirement:** $W_\phi(\mathbb{P}, \mathbb{Q})$ should be dominated by standard divergence,

$$TV(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{B}} \left| \mathbb{P}(A) - \mathbb{Q}(A) \right| .$$

Relaxed Wasserstein as Divergence

Question: Is W_ϕ a good divergence?

- **Point 1:** $W_\phi(\mathbb{P}, \mathbb{Q})$ should be small when $\mathbb{P}$ and $\mathbb{Q}$ are close.
- **Requirement:** $W_\phi(\mathbb{P}, \mathbb{Q})$ should be dominated by standard divergence,

$$TV(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{B}} |\mathbb{P}(A) - \mathbb{Q}(A)| .$$

- **Point 2:** $W_\phi(\mathbb{P}_n, \mathbb{P}_r) \to 0$ as $n \to \infty$ where $\mathbb{P}_r$ is a true distribution $\mathbb{P}_r$ and $\mathbb{P}_n$ is the empirical distribution based on $\mathcal{X} = (X_1, X_2, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$.

Relaxed Wasserstein as Divergence

Question: Is W_ϕ a good divergence?

- **Point 1:** $W_\phi(\mathbb{P}, \mathbb{Q})$ should be small when $\mathbb{P}$ and $\mathbb{Q}$ are close.
- **Requirement:** $W_\phi(\mathbb{P}, \mathbb{Q})$ should be dominated by standard divergence,

$$TV(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{B}} \left| \mathbb{P}(A) - \mathbb{Q}(A) \right| .$$

- **Point 2:** $W_\phi(\mathbb{P}_n, \mathbb{P}_r) \to 0$ as $n \to \infty$ where $\mathbb{P}_r$ is a true distribution $\mathbb{P}_r$ and $\mathbb{P}_n$ is the empirical distribution based on $\mathcal{X} = (X_1, X_2, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_r$.
- **Requirement:** $W_\phi(\mathbb{P}_n, \mathbb{P}_r)$ should have the moment estimate and concentration inequality, i.e., there exist $\alpha, \beta > 0$ such that

$$\begin{aligned} \mathbb{E}\left[W_{D_\phi}\left(\mathbb{P}_n, \mathbb{P}_r\right)\right] &= O(n^{-\alpha}) && \text{(Moment Estimate)}, \\ \text{Prob}\left(W_{D_\phi}\left(\mathbb{P}_n, \mathbb{P}_r\right) \geq \epsilon \right) &= O(n^{-\beta}) && \text{(Concentration Inequality)}. \end{aligned}$$

Dominated by TV and Standard Wasserstein

Theorem (G., Hong, Lin, and Yang 2018)

Assume that $\phi : \mathcal{X} \to \mathbb{R}$ is a strictly convex and smooth function with an L-Lipschitz continuous factor,

$$
\begin{aligned}
W_{D_\phi}(\mathbb{P}, \mathbb{Q}) &\leq L\,[\mathrm{diam}(\mathcal{X})]^2 \cdot TV(\mathbb{P}, \mathbb{Q}) \\
W_{D_\phi}(\mathbb{P}, \mathbb{Q}) &\leq \frac{L}{2} W_{L^2}(\mathbb{P}, \mathbb{Q})^2
\end{aligned}
$$

where $\mathbb{P}$ and $\mathbb{Q}$ are two probability distributions supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$.

Table of Contents

Moment Estimate for RW

Theorem (G, Hong, Lin, and Yang 2018)

Assume that

$$M_q(\mathbb{P}_r) = \int_{\mathcal{X}} \|x\|_2^q \, \mathbb{P}_r(dx) < +\infty$$

for some $q > 2$, then there exists a constant $C(q,d) > 0$ such that, for $n \geq 1$,

$$\mathbb{E}\left[W_{D_\phi}\left(\mathbb{P}_n, \mathbb{P}_r\right)\right] \leq \frac{C(q,d) L M_q^{\frac{2}{q}}(\mathbb{P}_r)}{2} \cdot \begin{cases} n^{-\frac{1}{2}} + n^{-\frac{q-2}{q}}, & 1 \leq d \leq 3, \ q \neq 4, \\ n^{-\frac{1}{2}} \log(1+n) + n^{-\frac{q-2}{q}}, & d = 4, \ q \neq 4, \\ n^{-\frac{2}{d}} + n^{-\frac{q-2}{q}}, & d \geq 5, \ q \neq d/(d-2). \end{cases}$$

Concentration Inequality for RW

Theorem (G., Hong, Lin, and Yang 2018)

Assume that

$$\mathcal{E}_{\alpha,\gamma}(\mathbb{P}_r) = \int_{\mathcal{X}} \exp\left(\gamma \|x\|_2^{\alpha}\right) \ \mathbb{P}_r(dx).$$

and one of the three following conditions holds,

$$\begin{aligned} & \exists\ \alpha > 2,\ \exists\ \gamma > 0,\ \mathcal{E}_{\alpha,\gamma}(\mathbb{P}_r) < \infty, \\ \textit{or} \quad & \exists\ \alpha \in (0,2)\,,\ \exists\ \gamma > 0,\ \mathcal{E}_{\alpha,\gamma}(\mathbb{P}_r) < \infty, \\ \textit{or} \quad & \exists\ q > 4,\ M_q(\mathbb{P}_r) < \infty. \end{aligned}$$

Then for $n \geq 1$ *and* $\epsilon > 0$*, there exist the scalar* $a(n,\epsilon)$ *and* $b(n,\epsilon)$ *such that*

$$\text{Prob}\left(W_{D_\phi}\left(\mathbb{P}_n, \mathbb{P}_r\right) \geq \epsilon\right) \leq a(n,\epsilon) 1_{\{\epsilon \leq \frac{L}{2}\}} + b(n,\epsilon).$$

Duality Representation for RW

Theorem (G., Hong, Lin, and Yang 2018)

Assume that two probability distributions $\mathbb{P}$ *and* $\mathbb{Q}$ *satisfy*

$$\int_{\mathcal{X}} \|x\|_2^2 \ (\mathbb{P}+\mathbb{Q})(dx) < +\infty.$$

Then there exists a Lipschitz continuous function $f : \mathcal{X} \to \mathbb{R}$ *such that the RW divergence has a duality representation as*

$$\begin{aligned} W_{D_\phi}(\mathbb{P},\mathbb{Q}) &= \int_{\mathcal{X}} \phi(x) \ (\mathbb{P}-\mathbb{Q})(dx) + \int_{\mathcal{X}} \langle \nabla\phi(x), x\rangle \ \mathbb{Q}(dx) \\ &\quad - \left(\int_{\mathcal{X}} f(x) \ \mathbb{P}(dx) + \int_{\mathcal{X}} f^*(\nabla\phi(x)) \ \mathbb{Q}(dx) \right), \end{aligned}$$

where f^* *is the conjugate of* f, *i.e.,*

$$f^*(y) = \sup_{x\in\mathbb{R}^d} \ \langle x, y\rangle - f(x).$$

Key element for proof of duality

- The classical duality representation for the standard Wasserstein distance

Key element for proof of duality

- The classical duality representation for the standard Wasserstein distance
- The RW can be decomposed in terms of a distorted squared Wasserstein-L^2 distance of order 2, plus some residual terms that are independent of the choice of the coupling π.

Table of Contents

Relaxed Wasserstein for GANs

Question: Is W_ϕ tractable for GANs?

Relaxed Wasserstein for GANs

Question: Is W_ϕ tractable for GANs?

- **Requirement 1:** $W_\phi(\mathbb{P}_r, \mathbb{P}_\theta)$ should be continuous and differentiable w.r.t. θ.

Relaxed Wasserstein for GANs

Question: Is W_ϕ tractable for GANs?

- **Requirement 1:** $W_\phi(\mathbb{P}_r, \mathbb{P}_\theta)$ should be continuous and differentiable w.r.t. θ.
- **Requirement 2:** $W_\phi(\mathbb{P}_r, \mathbb{P}_\theta)$ should have the easily computed or approximated gradient evaluation, i.e.,

$$\nabla_\theta \left[W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta) \right] = F\left(g_\theta, \phi, Z, \ldots\right).$$

where F is an abstract mapping.

Continuity and Differentiablity

Theorem (G., Hong, Lin, and Yang 2018)

Continuity and Differentiablity

Theorem (G., Hong, Lin, and Yang 2018)

1. $W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)$ *is continuous in* θ *if* g_θ *is continuous in* θ.

Continuity and Differentiablity

Theorem (G., Hong, Lin, and Yang 2018)

1. *$W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous in θ if g_θ is continuous in θ.*
2. *$W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)$ is differentiable almost everywhere if g_θ is locally Lipschitz with a constant $\bar{L}(\theta, z)$ such that $\mathbb{E}\left[\bar{L}(\theta, Z)^2\right] < \infty$, i.e., for each given (θ_0, z_0), there exists a neighborhood $\mathcal{N}$ such that*

$$\|g_\theta(z) - g_{\theta_0}(z_0)\|_2 \leq L(\theta_0, z_0)\left(\|\theta - \theta_0\|_2 + \|z - z_0\|_2\right).$$

for any $(\theta, z) \in \mathcal{N}$.

Table of Contents

Gradient Descent Scheme

Corollary (G., Hong, Lin, and Yang 2018)

Assume that g_θ is locally Lipschitz with a constant $L(\theta, z)$ such that $\mathbb{E}\left[L(\theta, Z)^2\right] < \infty$, and $\int_{\mathcal{X}} \|x\|_2^2 \,(\mathbb{P}_r + \mathbb{P}_\theta)(dx) < +\infty$. Then there exists a Lipschitz continuous solution $f : \mathcal{X} \to \mathbb{R}$ such that the gradient of the RW divergence has an explicit form, i.e.,

$$\nabla_\theta\left[W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)\right] = \mathbb{E}_Z\left[\left[\nabla_\theta g_\theta(Z)\right]^\top \nabla^2\phi(g_\theta(Z))g_\theta(Z)\right] + \mathbb{E}_Z\left[\nabla_\theta f\left(\nabla\phi(g_\theta(Z))\right)\right].$$

Table of Contents

Experiment Setup

- **RW:** KL divergence where $\phi(x) = -x^\top \log(x)$.

Experiment Setup

- **RW:** KL divergence where $\phi(x) = -x^\top \log(x)$.
- **Approach:** RMSProp [Tieleman and Hinton, 2012].

Experiment Setup

- **RW:** KL divergence where $\phi(x) = -x^\top \log(x)$.
- **Approach:** RMSProp [Tieleman and Hinton, 2012].
- **Experiment I:**
  - **Baselines:** WGANs, CGANs, InfoGANs, GANs, LSGANs, DRAGANs, BEGANs, EBGANs and ACGANs.
  - **Datasets:**
    - MNIST: 60000 (train) and 10000 (test).
    - Fashion-MNIST: 60000 (train) and 10000 (test).

Experiment Setup

- **RW:** KL divergence where $\phi(x) = -x^\top \log(x)$.
- **Approach:** RMSProp [Tieleman and Hinton, 2012].
- **Experiment I:**
  - **Baselines:** WGANs, CGANs, InfoGANs, GANs, LSGANs, DRAGANs, BEGANs, EBGANs and ACGANs.
  - **Datasets:**
    - MNIST: 60000 (train) and 10000 (test).
    - Fashion-MNIST: 60000 (train) and 10000 (test).
- **Experiment II:**
  - **Baselines:** WGANs and WGANs-GP.
  - **Datasets:**
    - CIFAR-10 (color): 50000 (train) and 10000 (test).
    - ImageNet (color): 14197122.

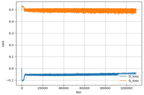
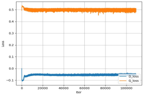
Metric for performance

The inception score is defined as follows:

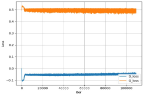
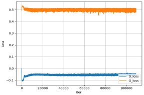
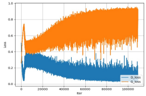
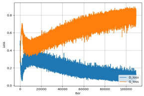
$$\text{Inception\_Score} = \exp\left\{\mathbb{E}_x\left[D_{\mathrm{KL}}(p(y|x), p(y)\right]\right\},$$

Table of Contents

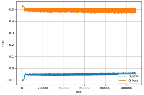
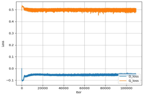
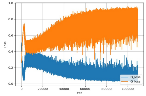
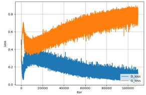
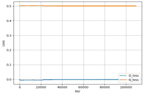
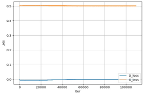
Empirical Results on MNIST and Fashion-MNIST datasets

Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs					

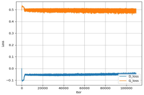
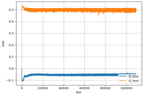
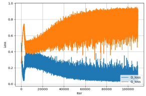
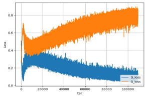
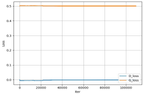
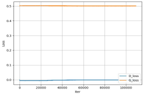
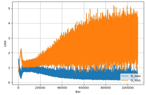
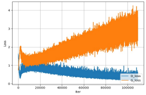
Empirical Results on MNIST and Fashion-MNIST datasets

Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		

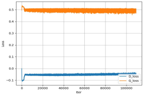
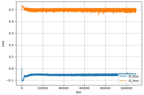
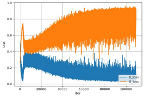
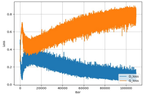
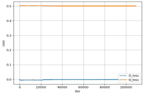
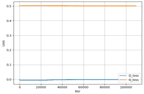
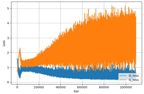
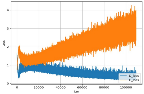
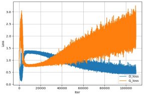
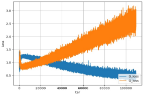
Empirical Results on MNIST and Fashion-MNIST datasets

Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs					

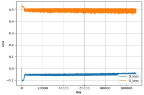
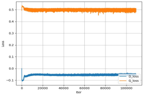
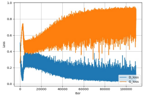
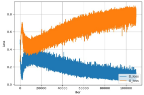
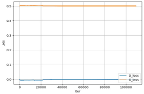
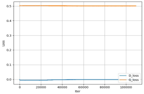
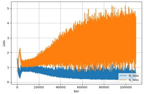
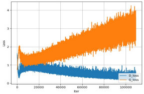
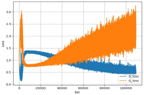
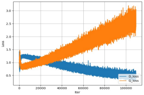
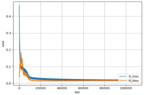
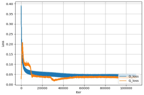
Empirical Results on MNIST and Fashion-MNIST datasets

Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs			DRAGANs		

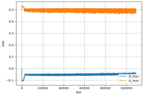
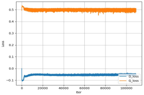
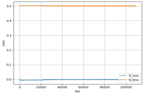
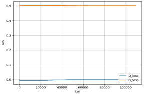
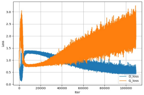
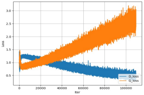
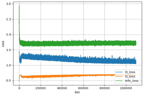
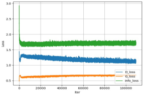
Empirical Results on MNIST and Fashion-MNIST datasets

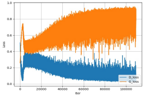
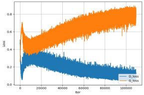
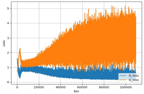
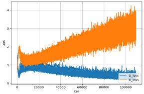
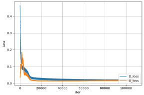
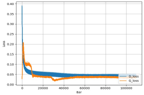
Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs			DRAGANs		
CGANs					

Empirical Results on MNIST and Fashion-MNIST datasets

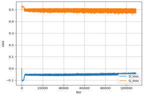
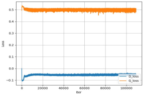
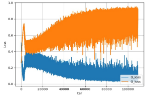
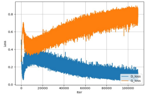
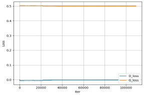
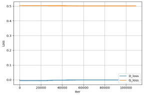
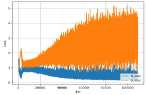
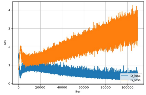
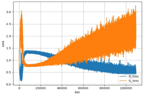
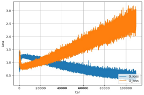
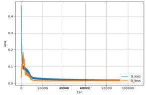
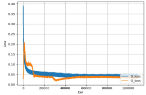
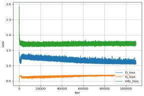
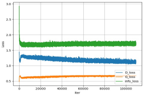
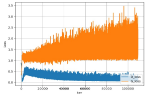
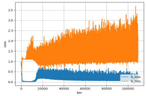
Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs			DRAGANs		
CGANs			BEGANs		

Empirical Results on MNIST and Fashion-MNIST datasets

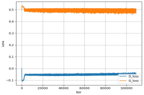
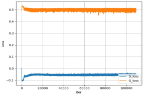
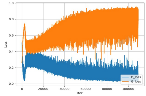
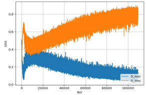
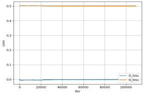
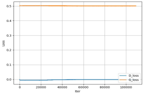
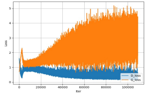
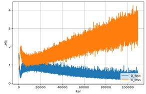
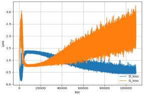
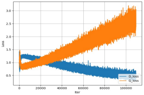
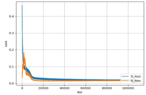
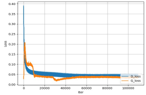
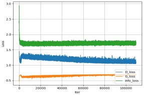
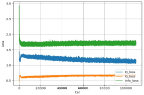
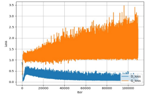
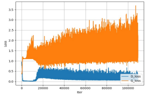
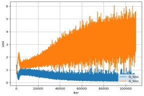
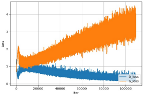
Method	MNIST	Fashion-MNIST
RWGANs		
WGANs		
CGANs		
InfoGANs		

Method	MNIST	Fashion-MNIST
LSGANs		
DRAGANs		
BEGANs		

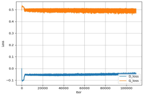
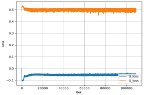
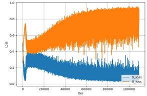
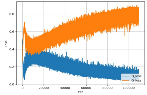
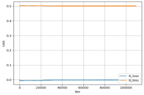
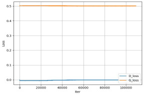
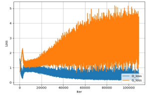
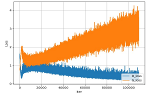
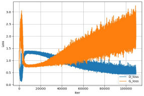
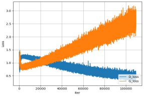
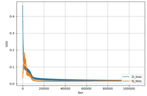
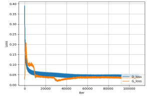
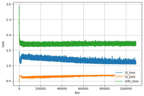
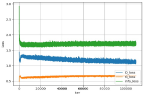
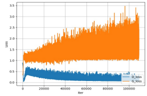
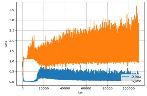
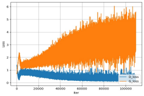
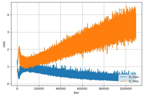
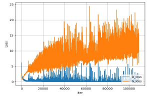
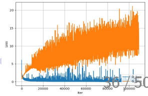
Empirical Results on MNIST and Fashion-MNIST datasets

Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs			DRAGANs		
CGANs			BEGANs		
InfoGANs			EBGANs		

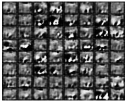
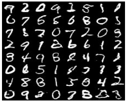
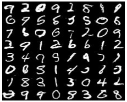
Empirical Results on MNIST and Fashion-MNIST datasets

Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs			DRAGANs		
CGANs			BEGANs		
InfoGANs			EBGANs		
GANs					

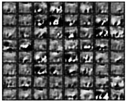
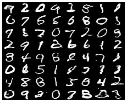
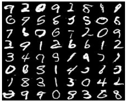
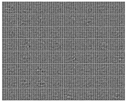
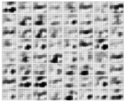
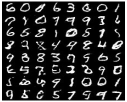
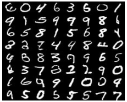
Empirical Results on MNIST and Fashion-MNIST datasets

Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs			DRAGANs		
CGANs			BEGANs		
InfoGANs			EBGANs		
GANs			ACGANs		

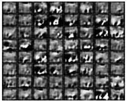
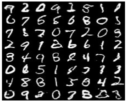
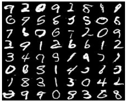
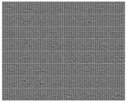
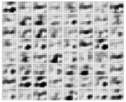
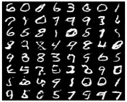
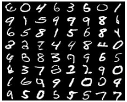
Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
RWGANs				

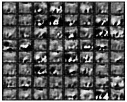
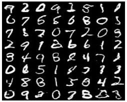
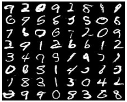
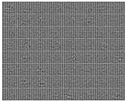
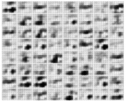
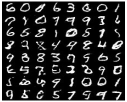
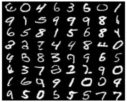
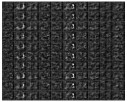
Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
RWGANs				
WGANs				

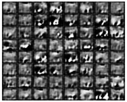
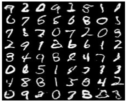
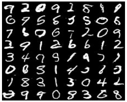
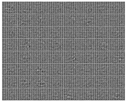
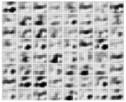
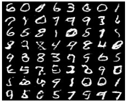
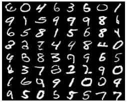
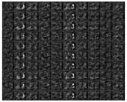
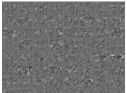
Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
RWGANs				
WGANs				
CGANs				

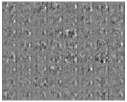
Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
RWGANs				
WGANs				
CGANs				
InfoGANs				

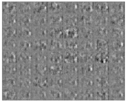
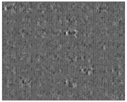
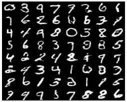
Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
RWGANs				
WGANs				
CGANs				
InfoGANs				
GANs				

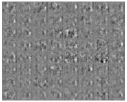
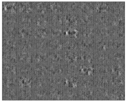
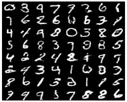
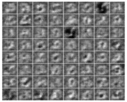
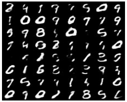
Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
LSGANs				

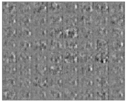
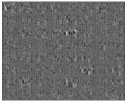
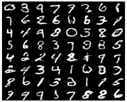
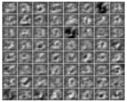
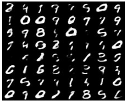
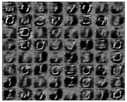
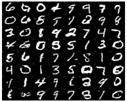
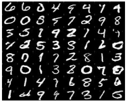
Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
LSGANs				
DRAGANs				

Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
LSGANs				
DRAGANs				
BEGANs				

Empirical Results on MNIST dataset

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
LSGANs				
DRAGANs				
BEGANs				
EBGANs				

Empirical Results on MNIST dataset

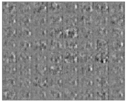
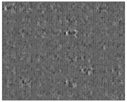
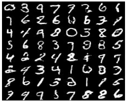
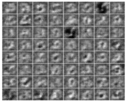
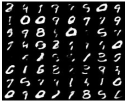
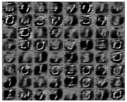
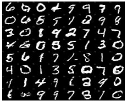
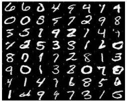
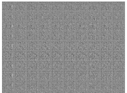
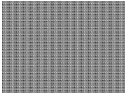
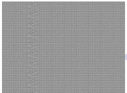
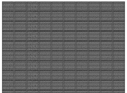
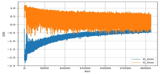
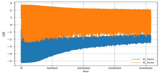
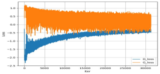
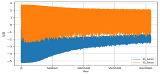
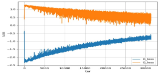
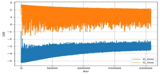
Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
LSGANs				
DRAGANs				
BEGANs				
EBGANs				
ACGANs				

Table of Contents

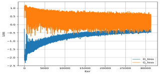
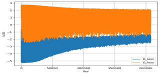
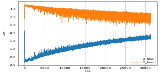
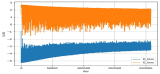
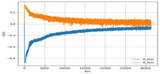
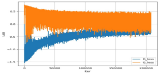
Empirical Results on CIFAR-10 and ImageNet datasets

Method	Architecture	CIFAR-10	ImageNet
RWGANs	DCGAN		

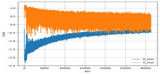
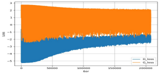
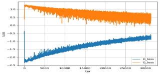
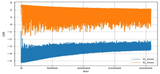
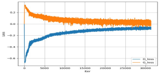
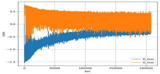
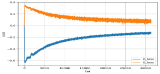
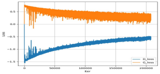
Empirical Results on CIFAR-10 and ImageNet datasets

Method	Architecture	CIFAR-10	ImageNet
RWGANs	DCGAN		
	MLP		

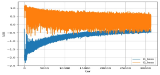
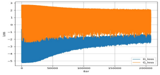
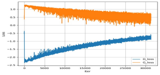
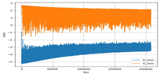
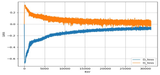
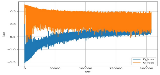
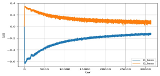
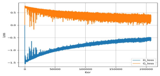
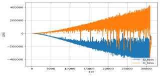
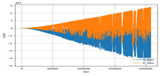
Empirical Results on CIFAR-10 and ImageNet datasets

Method	Architecture	CIFAR-10	ImageNet
RWGANs	DCGAN		
	MLP		
WGANs	DCGAN		

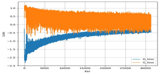
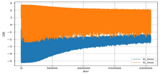
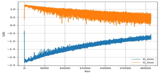
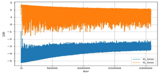
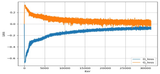
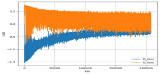
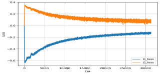
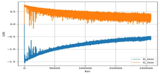
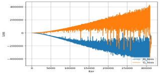
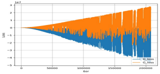
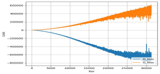
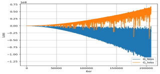
Empirical Results on CIFAR-10 and ImageNet datasets

Method	Architecture	CIFAR-10	ImageNet
RWGANs	DCGAN		
	MLP		
WGANs	DCGAN		
	MLP		

Empirical Results on CIFAR-10 and ImageNet datasets

Method	Architecture	CIFAR-10	ImageNet
RWGANs	DCGAN		
	MLP		
WGANs	DCGAN		
	MLP		
WGANs-GP	DCGAN		

Empirical Results on CIFAR-10 and ImageNet datasets

Method	Architecture	CIFAR-10	ImageNet
RWGANs	DCGAN		
	MLP		
WGANs	DCGAN		
	MLP		
WGANs-GP	DCGAN		
	MLP		

Empirical Results on Inception Score

Architecture	Method	CIFAR-10		ImageNet	
		First 5 epochs	Last 10 epochs	First 3 epochs	Last 5 epochs
DCGAN	RWGANs	**1.8606**	2.3962	2.0430	2.7008
	WGANs	1.6329	**2.4246**	2.2070	**2.7972**
	WGANs-GP	1.7259	2.3731	**2.2749**	2.7331
MLP	RWGANs	**1.3126**	2.1710	**2.0025**	**2.4805**
	WGANs	1.2798	1.9007	1.7401	2.2304
	WGANs-GP	1.2711	**2.2192**	1.8845	2.3448

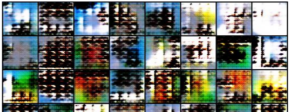
Empirical Results on ImageNet dataset

Method	$N = 1$	
	DCGAN	MLP
RWGANs		

Empirical Results on ImageNet dataset

Method	$N = 1$	
	DCGAN	MLP
RWGANs		
WGANs		

Empirical Results on ImageNet dataset

Method	$N = 1$	
	DCGAN	MLP
RWGANs		
WGANs		
WGANs-GP		

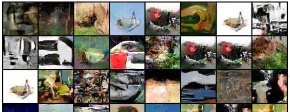
Empirical Results on ImageNet dataset

Method	$N = 25$	
	DCGAN	MLP
RWGANs		

Empirical Results on ImageNet dataset

Method	$N = 25$	
	DCGAN	MLP
RWGANs		
WGANs		

Empirical Results on ImageNet dataset

Method	$N = 25$	
	DCGAN	MLP
RWGANs		
WGANs		
WGANs-GP		

Conclusions

In Summary,

- We propose a novel class of statistical divergence called *Relaxed Wasserstein* (RW) divergence. This RW shares the same critical probabilistic properties as Wasserstein distance, without possible asymmetry.

Conclusions

In Summary,

- We propose a novel class of statistical divergence called *Relaxed Wasserstein* (RW) divergence. This RW shares the same critical probabilistic properties as Wasserstein distance, without possible asymmetry.
- RW divergence provides a lot of flexibility and possibilities in generative modeling by using a class of strictly convex and differentiable functions which contain **different curvature information**.

Conclusions

In Summary,

- We propose a novel class of statistical divergence called *Relaxed Wasserstein* (RW) divergence. This RW shares the same critical probabilistic properties as Wasserstein distance, without possible asymmetry.
- RW divergence provides a lot of flexibility and possibilities in generative modeling by using a class of strictly convex and differentiable functions which contain **different curvature information**.
- We present a gradient-based optimization framework to learn RWGAN and attain an encouraging results on image generation.

Future directions:

- Does some **optimal** choice of ϕ exist in real problems?

Future directions:

- Does some **optimal** choice of ϕ exist in real problems?
- Does ϕ depend on the data samples or the problem structure?

Future directions:

- Does some **optimal** choice of ϕ exist in real problems?
- Does ϕ depend on the data samples or the problem structure?
- Applications to Finance: JP Morgan on-going project using GANs.

Future directions:

- Does some **optimal** choice of ϕ exist in real problems?
- Does ϕ depend on the data samples or the problem structure?
- Applications to Finance: JP Morgan on-going project using GANs.
- In the theory of optimal transport and stochastic games, relaxed Wasserstein is more natural than Wasserstein distance: the same nice mathematical properties, without the symmetry constraint.

References

Arjovsky, M. and Bottou, L. (2017).
Towards principled methods for training generative adversarial networks.
ArXiv Preprint: 1701.04862.

Arjovsky, M., Chintala, S., and Bottou, L. (2017).
Wasserstein generative adversarial networks.
ICML, pages 214–223.

Aude, G., Cuturi, M., Peyré, G., and Bach, F. (2016).
Stochastic optimization for large-scale optimal transport.
ArXiv Preprint:1605.08527.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2017).
Inference in generative models using the wasserstein distance.
ArXiv Preprint: 1701.05146.

Berthelot, D., Schumm, T., and Metz, L. (2017).
BEGAN: Boundary equilibrium generative adversarial networks.
ArXiv Preprint: 1703.10717.

Blanchet, J., Kang, Y., and Murthy, K. (2016).
Robust wasserstein profile inference and applications to machine learning.

References

Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. (2017).
Convergence of entropic schemes for optimal transport and gradient flows.
SIAM Journal on Mathematical Analysis, 49(2):1385–1418.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016).
InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets.
NIPS, pages 2172–2180.

Esfahani, P. M. and Kuhn, D. (2015).
Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.
Mathematical Programming, pages 1–52.

Gao, R., Chen, X., and Kleywegt, A. J. (2017).
Wasserstein distributional robustness and regularization in statistical learning.
ArXiv Preprint: 1712.06050.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative adversarial nets.

References

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017).
Improved training of Wasserstein GANs.
ArXiv Preprint: 1704.00028.

Guo, X., Hong, J., Lin, T., and Yang, N. (2017).
Relaxed Wasserstein with applications to GANs.
ArXiv Preprint: 1705.07164.

Karazeev, A. (Aug 17, 2017).
Generative Adversarial Networks (GANs): Engine and Applications.
https://blog.statsbot.co/generative-adversarial-networks-gans-engine-and-applications-f96291965b47.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. (2017).
How to train your DRAGAN.
ArXiv Preprint: 1705.07215.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. (2016).
Least squares generative adversarial networks.
ArXiv Preprint: 1611.04076.

Mirza, M. and Osindero, S. (2014).

References

Odena, A., Olah, C., and Shlens, J. (2017).
Conditional image synthesis with auxiliary classifier GANs.
ICML, pages 2642–2651.

Peyré, G. (2015).
Entropic approximation of Wasserstein gradient flows.
SIAM Journal on Imaging Sciences, 8(4):2323–2351.

Ramdas, A., Trillos, N. G., and Cuturi, M. (2017).
On Wasserstein two-sample testing and related families of nonparametric tests.
Entropy, 19(2):47.

Tieleman, T. and Hinton, G. (2012).
Lecture 6.5-RMSProp: divide the gradient by a running average of its recent magnitude.
COURSERA: Neural Networks for Machine Learning, 4(2).

Zhao, J., Mathieu, M., and LeCun, Y. (2016).
Energy-based generative adversarial network.
ArXiv Preprint: 1609.03126.

Thank you for your attention !