From Generative Model to ...

Yimeng Chen

University of Chinese Academy of Science

November 15th, 2017

Main idea

- Introducing the theory framework of GAN
- Introducing the Wasserstein GAN
- Talk about some new work on this topic
- Raise some questions

Outline

Section 1

Generative Model vs Discriminative Model

In order to determine the label y of x:

- Discriminative model learns $p(y|x)$ directly.
- Generative model
  Learns $p(x|y)$ (and $p(y)$), then use Bayes rule

$$arg \max_y p(y|x) = arg \max_y p(x|y)p(y).$$

Generative Model vs Discriminative Model

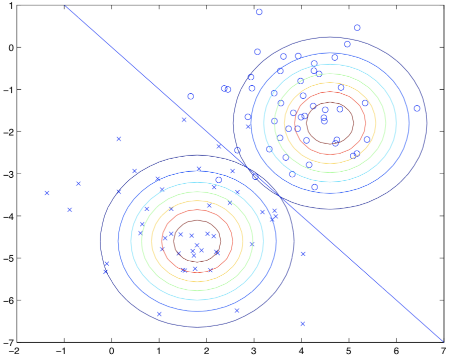

Figure 1: Two Gaussian distribution.

CS 229 lecture notes, Andrew Ng.

Classical Generative Model

Classical way to learn a probability density:

Classical Generative Model

Classical way to learn a probability density:

- Defining a parametric family of densities $(p_\theta)_{\theta \in \mathbb{R}_d}$

Classical Generative Model

Classical way to learn a probability density:

- Defining a parametric family of densities $(p_\theta)_{\theta \in \mathbb{R}_d}$
- Do maximal likelihood estimation on real data samples $\{x^{(i)}\}_{i=1}^m$:

$$\max_{\theta \in \mathbb{R}_d} \frac{1}{m} \Sigma_{i=1}^m \log p_\theta(x^{(i)})$$

Classical Generative Model

Classical way to learn a probability density:

- Defining a parametric family of densities $(p_\theta)_{\theta \in \mathbb{R}_d}$
- Do maximal likelihood estimation on real data samples $\{x^{(i)}\}_{i=1}^m$:

$$\max_{\theta \in \mathbb{R}_d} \frac{1}{m} \Sigma_{i=1}^m \log p_\theta(x^{(i)})$$

- That's equivalent to minimize the KL divergence $KL(\mathbb{P}_{r-emp} \| \mathbb{P}_\theta)$

KL divergence

Definition 1.1 (KL divergence)

$$KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \int \log(\frac{p_r(x)}{p_\theta(x)}) p_r(x) d\mu(x)$$

KL divergence

Definition 1.1 (KL divergence)

$$KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \int \log(\frac{p_r(x)}{p_\theta(x)}) p_r(x) d\mu(x)$$

- Asymmetric

KL divergence

Definition 1.1 (KL divergence)

$$KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \int \log(\frac{p_r(x)}{p_\theta(x)}) p_r(x) d\mu(x)$$

- Asymmetric
- When $P_\theta(x) = 0$ and $P_r(x) > 0$, it is infinite.

KL divergence

Definition 1.1 (KL divergence)

$$KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \int \log(\frac{p_r(x)}{p_\theta(x)}) p_r(x) d\mu(x)$$

- Asymmetric
- When $P_\theta(x) = 0$ and $P_r(x) > 0$, it is infinite.
- Typical remedy is to add a noise component, but it will degrade the quality of the samples.

Section 2

Generative Adversarial Networks

- Generator

  Noise variable $\mathbf{z} \sim p_z(\mathbf{z})$

Generative Adversarial Networks

- Generator

  Noise variable $\mathbf{z} \sim p_z(\mathbf{z})$
  Parametric function(NN) $G(\mathbf{z}; \theta_g) : \mathcal{Z} \to \mathcal{X}$

Generative Adversarial Networks

- Generator

  Noise variable $\mathbf{z} \sim p_z(\mathbf{z})$
  Parametric function(NN) $G(\mathbf{z}; \theta_g) : \mathcal{Z} \rightarrow \mathcal{X}$

- Discriminator

  Parametric function(NN) $D(\mathbf{x}; \theta_d) : \mathcal{X} \rightarrow [0, 1]$

Generative Adversarial Networks

- Generator

  Noise variable $\mathbf{z} \sim p_z(\mathbf{z})$
  Parametric function(NN) $G(\mathbf{z}; \theta_g) : \mathcal{Z} \to \mathcal{X}$

- Discriminator

  Parametric function(NN) $D(\mathbf{x}; \theta_d) : \mathcal{X} \to [0, 1]$

Analogy

Currency Counterfeiters and the Police
Key idea: *Policy update*

Objective of GAN

GAN Objective

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

Objective of GAN

GAN Objective

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

- Also a Maximum Likelihood Estimation.

Objective of GAN

GAN Objective

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

- Also a Maximum Likelihood Estimation.
- There exists an unique optimal D^*, $D_G^*(\mathbf{x}) = \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}$.

Optimal D

Under optimal D, the objective function

$$V(D^*, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}] + \mathbb{E}_{\mathbf{x} \sim p_g}[\log \frac{p_g(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}]$$

Optimal D

Under optimal D, the objective function

$$V(D^*, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}] + \mathbb{E}_{\mathbf{x} \sim p_g}[\log \frac{p_g(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}]$$

In view of JS divergence, we have

$$V(D^*, G) = -\log 4 + 2 \cdot JSD(p_{data} \| p_g)$$

JS divergence

Definition 2.1 (JS divergence)

$$JS(\mathbb{P}_r \| \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_m \| \mathbb{P}_g),$$

$$\mathbb{P}_m = (\mathbb{P}_r + \mathbb{P}_g)/2$$

- Symmetric
- $0 \leq JSD(P \| Q) \leq \log 2$

Training of GAN

- $D(\mathbf{x}; \theta_d)$
  In every step, use a mini-batch of samples of $p_{data}(\mathbf{x})$ and $p_g(\mathbf{z})$.

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(\mathbf{x}^{(i)}) + \log(1 - D(G(\mathbf{z}^{(i)})))]$$

Training of GAN

- $D(\mathbf{x}; \theta_d)$
  In every step, use a mini-batch of samples of $p_{data}(\mathbf{x})$ and $p_g(\mathbf{z})$.

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(\mathbf{x}^{(i)}) + \log(1 - D(G(\mathbf{z}^{(i)})))]$$

- $G(\mathbf{z}; \theta_g)$
  In every step, use a mini-batch of samples of $p_g(\mathbf{z})$.

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(\mathbf{z}^{(i)})))$$

Questions

Questions

Question 1

What's the difference between the empirical distribution

$$\frac{1}{m}\sum_{i}^{m}\delta_{x^i}$$

and the distribution we compute in the training of GAN? How about the real distribution?

Questions

Question 1

What's the difference between the empirical distribution

$$\frac{1}{m}\sum_{i}^{m}\delta_{x^i}$$

and the distribution we compute in the training of GAN? How about the real distribution?

Question 2

What's the relationship between G and D?

Unstability

Unstability

Lemma 1 (low dimensionality)

Let $g : \mathcal{Z} \to \mathcal{X}$ be a function composed by affine transformations and pointwise nonlinearities, then $g(\mathcal{Z})$ is contained in a countable union of manifolds of dimension at most $\dim \mathcal{Z}$.

Lemma 2 (perfectly align)

Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that don't have full dimension. Let η, η' be arbitrary independent continuous random variables. We therefore define the perturbed manifolds as $\widetilde{\mathcal{M}} = \mathcal{M} + \eta$ and $\widetilde{\mathcal{P}} = \mathcal{P} + \eta'$. Then

$$\mathbb{P}_{\eta,\eta'}(\widetilde{\mathcal{M}} \textit{ does not perfectly align with } \widetilde{\mathcal{P}}) = 1$$

Unstability

Theorem 3

Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions whose support lies in two manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't have full dimension and don't perfectly align. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds. Then

$$JSD(\mathbb{P}_r \| \mathbb{P}_g) = \log 2,$$

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = +\infty,$$

$$KL(\mathbb{P}_g \| \mathbb{P}_r) = +\infty.$$

Unstability

Theorem 4 (Vanishing gradients on the generator)

Let G induces $\mathbb{P}_g$. $\mathbb{P}_r$ is the real data distribution. Under the same condition in theorem 3, and when $\|D - D^*\| < \epsilon$, $\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\|\nabla_\theta g_\theta(\mathbf{z})\|_2^2] \leq M^2$, we have

$$\|\nabla_\theta \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D(g_\theta(z)))]\|_2 < M\frac{\epsilon}{1-\epsilon}$$

Section 3

Optimal transport distance

Definition 3.0 (Kantorovich problem)

Given $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$, and $c : X \times Y \to [0, +\infty]$, we consider the problem

$$\min\{\int_{X \times Y} c(x, y) d\gamma : \gamma \in \Pi(\mu, \nu)\}$$

Here $\Pi(\mu, \nu)$ is the set of transport plans

$$\Pi(\mu, \nu) = \{\gamma \in \mathscr{P}(X \times Y) : (\pi_x)_\sharp \gamma = \mu, (\pi_y)_\sharp \gamma = \nu\}$$

Wasserstein distances

Definition 3.1 (Wasserstein Distances on Ω)

For $\Omega \in \mathbb{R}^d$, $\mathscr{P}_p(\Omega) := \{\mu \in \mathscr{P}(\Omega) : \int |x|^p d\mu < +\infty\}$
For $\forall \mu, \nu \in \mathscr{P}_p(\Omega)$,

$$W_p(\mu, \nu) := \min\{\int_{\Omega \times \Omega} |x - y|^p d\gamma : \gamma \in \Pi(\mu, \nu)\}^{\frac{1}{p}}$$

Wasserstein distances

Definition 3.1 (Wasserstein Distances on Ω)

For $\Omega \in \mathbb{R}^d$, $\mathscr{P}_p(\Omega) := \{\mu \in \mathscr{P}(\Omega) : \int |x|^p d\mu < +\infty\}$
For $\forall \mu, \nu \in \mathscr{P}_p(\Omega)$,

$$W_p(\mu, \nu) := \min\{\int_{\Omega\times\Omega} |x - y|^p d\gamma : \gamma \in \Pi(\mu, \nu)\}^{\frac{1}{p}}$$

- Equivalence between the convergence for $W_p(p < \infty)$ and for W_1:

$$W_1(\mu, \nu) \leq W_p(\mu, \nu) \leq CW_1(\mu, \nu)^{\frac{1}{p}}$$

Wasserstein distances

Definition 3.1 (Wasserstein Distances on Ω)

For $\Omega \in \mathbb{R}^d$, $\mathscr{P}_p(\Omega) := \{\mu \in \mathscr{P}(\Omega) : \int |x|^p d\mu < +\infty\}$
For $\forall \mu, \nu \in \mathscr{P}_p(\Omega)$,

$$W_p(\mu, \nu) := \min\{\int_{\Omega\times\Omega} |x-y|^p d\gamma : \gamma \in \Pi(\mu, \nu)\}^{\frac{1}{p}}$$

- Equivalence between the convergence for $W_p (p < \infty)$ and for W_1:
$$W_1(\mu, \nu) \leq W_p(\mu, \nu) \leq C W_1(\mu, \nu)^{\frac{1}{p}}$$
- When $p < q$, $\mathscr{P}_p(\Omega) \subset \mathscr{P}_q(\Omega)$

Weak convergence

Definition 3.2 (Total variation)

Denote $\mathcal{P}(X)$ the space of all the probability measures on X.

- Total variation norm: For $\forall \mu \in \mathcal{P}(X)$, $\|\mu\|_{TV} = \sup_{A \subseteq \mathcal{X}} |\mu(A)|$, A is any Borel set in $\mathcal{X}$.
- Total variation distance: For $\forall \mu, \nu \in \mathcal{P}(X)$, $\delta(\mu, \nu) = \|\mu - \nu\|_{TV}$

Weak convergence

Definition 3.2 (Total variation)

Denote $\mathcal{P}(X)$ the space of all the probability measures on X.

- Total variation norm: For $\forall \mu \in \mathcal{P}(X)$, $\|\mu\|_{TV} = \sup_{A \subseteq \mathcal{X}} |\mu(A)|$, A is any Borel set in $\mathcal{X}$.
- Total variation distance: For $\forall \mu, \nu \in \mathcal{P}(X)$, $\delta(\mu, \nu) = \|\mu - \nu\|_{TV}$

Definition 3.3 (The weak-* convergence of probability measures)

For compact spaces X, $\mathcal{M}(X)$ is isomorphic to the dual space of $C(X)$. The convergence of $\mathcal{M}(X)$ in duality with $C(X)$ is weak-* convergence.

Relationship

By Pinsker's inequality [2],

$$\delta(P, Q) \leq \frac{1}{2}\sqrt{D_{KL}(P, Q)}$$

Relationship

By Pinsker's inequality [2],

$$\delta(P,Q) \leq \frac{1}{2}\sqrt{D_{KL}(P,Q)}$$

On compact subset of $\mathbb{R}^d$[1],

$$\mathbb{P}_n \xrightarrow{TV} \mathbb{P} \Leftrightarrow \mathbb{P}_n \xrightarrow{JSD} \mathbb{P},$$

$$\mathbb{P}_n \overset{*}{\rightharpoonup} \mathbb{P} \Leftrightarrow \mathbb{P}_n \xrightarrow{W_p} \mathbb{P}$$

Relationship

By Pinsker's inequality [2],

$$\delta(P, Q) \leq \frac{1}{2}\sqrt{D_{KL}(P, Q)}$$

On compact subset of $\mathbb{R}^d$[1],

$$\mathbb{P}_n \xrightarrow{TV} \mathbb{P} \Leftrightarrow \mathbb{P}_n \xrightarrow{JSD} \mathbb{P},$$

$$\mathbb{P}_n \overset{*}{\rightharpoonup} \mathbb{P} \Leftrightarrow \mathbb{P}_n \xrightarrow{W_p} \mathbb{P}$$

On separable spaces [3],

$$\mathbb{P}_n \overset{*}{\rightharpoonup} \mathbb{P} \Leftrightarrow \mathbb{P}_n \xrightarrow{D} \mathbb{P}$$

Relationship

On compact subset Ω of $\mathbb{R}^d$,

$$\mathbb{P}_n \xrightarrow{D_{KL}} \mathbb{P} \Rightarrow \mathbb{P}_n \xrightarrow{JSD} \mathbb{P} \Leftrightarrow \mathbb{P}_n \xrightarrow{TV} \mathbb{P} \Rightarrow \mathbb{P}_n \overset{*}{\rightharpoonup} \mathbb{P} \Leftrightarrow \mathbb{P}_n \xrightarrow{W_p} \mathbb{P}$$

Continuity

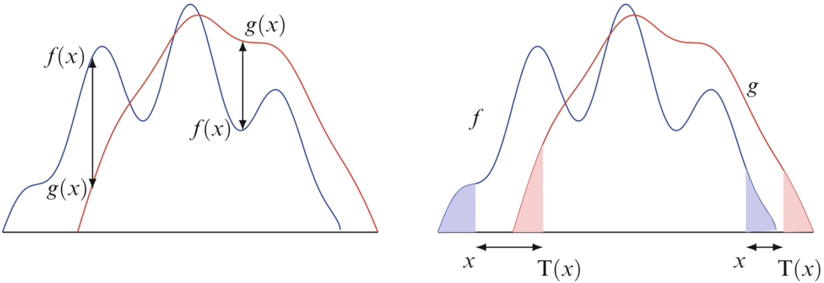

Figure 2: vertical vs horizontal

Continuity

Theorem 5

Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let P_θ denote the distribution of $g_\theta(Z)$. Then,
1. If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$.
2. If g is locally Lipschitz and satisfies regularity assumption 1, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable almost everywhere.
3. Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.

Continuity

Difinition 3.4 (Lipschitz continuity)

Given two metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called K-Lipschitz continuous if there exists a real constant $K \geq 0$ such that, for all x_1 and x_2 in X,

$$d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2).$$

Continuity

Difinition 3.4 (Lipschitz continuity)

Given two metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called K-Lipschitz continuous if there exists a real constant $K \geq 0$ such that, for all x_1 and x_2 in X,

$$d_Y(f(x_1), f(x_2)) \leq Kd_X(x_1, x_2).$$

Definition 3.5 (locally Lipschitz)

A function is called locally Lipschitz continuous if for every x in X there exists a neighborhood U of x such that f restricted to U is Lipschitz continuous.
If X is a locally compact metric space, then f is **locally Lipschitz** if and only if it is Lipschitz continuous on every compact subset of X.

Continuity

Lemma 6 (Regularity assumption 1)

Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be locally Lipschitz between finite dimensional vector spaces, i.e. for a given pair (θ, z) there is a constant $L(\theta, z)$ and an neighborhood U s.t. $\forall(\theta', z') \in U$ we have

$$\|g_\theta(z) - g_{\theta'}(z')\| \leq L(\theta, z)(\|\theta - \theta\| + \|z - z'\|)$$

We say that g satisfies assumption 1 for a certain probability distribution p over Z if $\mathbb{E}_{z \sim p(z)}[L(\theta, z)] < +\infty$

Continuity

Theorem 7 (Continuity of NNs)

Let g_θ be any feed-forward neural network(a function composed by affine transformations and pointwise nonlinearities which are smooth Lipschitz functions) parameterized by θ, and $p(z)$ a prior over z such that $\mathbb{E}_{z\sim p(z)}[\|z\|] < \infty$ (e.g. Gaussian, uniform, etc.). Then assumption 1 is satisfied and therefore $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere.

Duality form of W_1

$$W_1(\mu, \nu) = max\{\int_\Omega \varphi d\mu - \int_\Omega \varphi d\nu : \varphi \in Lip_1(\Omega)\}$$

Duality form of W_1

$$W_1(\mu,\nu) = max\{\int_\Omega \varphi d\mu - \int_\Omega \varphi d\nu : \varphi \in Lip_1(\Omega)\}$$

- We can let our Discriminator act the role of φ.

Duality form of W_1

$$W_1(\mu,\nu) = max\{\int_\Omega \varphi d\mu - \int_\Omega \varphi d\nu : \varphi \in Lip_1(\Omega)\}$$

- We can let our Discriminator act the role of φ.
- By the dual form of W_1,

$$W_1(\mathbb{P}_r,\mathbb{P}_\theta) = \sup_{\|D\|_L \leq 1} \mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[D(x)] - \mathbb{E}_{\mathbf{x}\sim p_\theta(\mathbf{x})}[D(x)]$$

Objective of WGAN

Definition 3.7 (Objective of WGAN)

$$\min_{G}\max_{\omega\in\mathcal{W}} \mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[D_\omega(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{x}}\sim p_g(\tilde{\mathbf{x}})}[D_\omega(\tilde{\mathbf{x}})]],$$

Here $\mathcal{W}$ is bounded to a fixed box like $[-0.01, 0.01]^l$. In this way we restrict D in a compact subset of $\mathbb{R}^d$, to make it Lipschitz.

Improved Objective of WGAN

$$\min_D \max_G \mathbb{E}_{\tilde{\mathbf{x}} \sim p_g(\tilde{\mathbf{x}})}[D(\tilde{\mathbf{x}})] - \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[D(\mathbf{x})] + \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim p_g(\hat{\mathbf{x}})}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2]$$

- A differentiable function is 1-Lipschitz if and only if it has gradients with norm at most 1 everywhere.
- The last item is the gradient penalty.

Questions

Question 3

What is the difference between the Discriminator family and all the Lipschitz-1 functions? Can the Discriminator represent the optimal function?

Section 4

GAN with the wind?

Theorem 8 (Stricly convex costs)

Given μ and ν probability measures on a compact domain $\Omega \in \mathbb{R}^d$, there exists an optimal transport plan γ for the cost $c(x, y) = h(x - y)$ with h strictly convex. It is unique and of the form $(id, T)_{\#}\mu$, provided μ is absolutely continuous and $\delta\Omega$ is negligible. Moreover, there exists a Kantorovich potential φ, and T and the potentials φ are linked by

$$T(x) = x - \nabla(h)^{-1}(\nabla(\varphi(x)))$$

GAN with the wind?

Theorem 8 (Stricly convex costs)

Given μ and ν probability measures on a compact domain $\Omega \in \mathbb{R}^d$, there exists an optimal transport plan γ for the cost $c(x, y) = h(x - y)$ with h strictly convex. It is unique and of the form $(id, T)_{\#}\mu$, provided μ is absolutely continuous and $\delta\Omega$ is negligible. Moreover, there exists a Kantorovich potential φ, and T and the potentials φ are linked by

$$T(x) = x - \nabla(h)^{-1}(\nabla(\varphi(x)))$$

- All the costs of the form $c(x, y) = |x - y|^p$ with $p > 1$ can be dealt with via Theorem 5.

GAN with the wind?

- W_2: Good Geometrical Significance
  By theorem 8, when $c(x, y) = \frac{1}{2}|x - y|^2$

$$T(x) = x - \nabla\varphi(x) = \nabla(\frac{x^2}{2} - \varphi(x)) = \nabla u(x).$$

$u(x)$ is called Brenier's potential. φ is called Kantorovich's potential.

GAN with the wind?

- W_2: Good Geometrical Significance
  By theorem 8, when $c(x, y) = \frac{1}{2}|x - y|^2$

  $$T(x) = x - \nabla\varphi(x) = \nabla(\frac{x^2}{2} - \varphi(x)) = \nabla u(x).$$

  $u(x)$ is called Brenier's potential. φ is called Kantorovich's potential.
- Compute via convex geometry method or numerical method

Geometric generative model

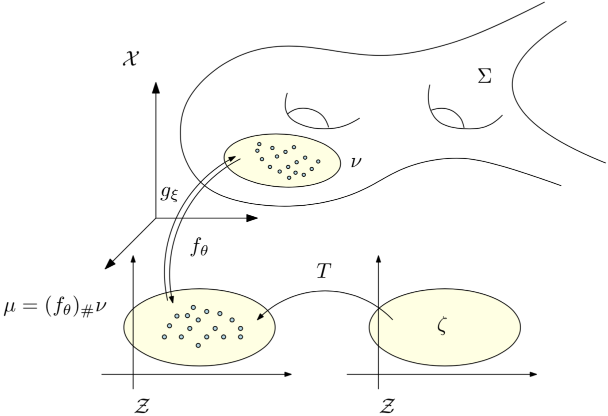

Figure 3: Geometric generative model.

Wind?

Question 1*

An empirical distribution...

Is that what we want?

Wind?

Question 1*

An empirical distribution...

Is that what we want?

Question 2

What's the relationship between G and D?

WGAN is fake

The Answer for the Question 3

The function family in the objective of WGAN is not the same as the family of Lipschitz-1 function.
The objective of WGAN is not the Wasserstein distance.

WGAN is fake

The Answer for the Question 3

The function family in the objective of WGAN is not the same as the family of Lipschitz-1 function.
The objective of WGAN is not the Wasserstein distance.

A part of answer for the Question 1

Definition of "Generalization of GAN".

WGAN is fake

The Answer for the Question 3

The function family in the objective of WGAN is not the same as the family of Lipschitz-1 function.
The objective of WGAN is not the Wasserstein distance.

A part of answer for the Question 1

Definition of "Generalization of GAN".

A kind of answer for the Question 2

Game theory and the equilibrium between G and D

F distance

Definition 4.1 (F distance)

Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $[0,1]$ and φ be a concave measuring function. Then the $\mathcal{F}$-divergence with respect to ϕ between two distribution μ and ν supported on $\mathbb{R}^d$ is defined as

$$d_{\mathcal{F},\phi}(\mu,\nu) = \sup_{D\in\mathcal{F}} |\mathbb{E}_{x\sim\mu}[\phi(D(x))] + \mathbb{E}_{x\sim\nu}[\phi(1 - D(x))]| - 2\phi(1/2)$$

F distance

Definition 4.1 (F distance)

Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $[0,1]$ and φ be a concave measuring function. Then the $\mathcal{F}$-divergence with respect to ϕ between two distribution μ and ν supported on $\mathbb{R}^d$ is defined as

$$d_{\mathcal{F},\phi}(\mu,\nu) = \sup_{D\in\mathcal{F}} |\mathbb{E}_{x\sim\mu}[\phi(D(x))] + \mathbb{E}_{x\sim\nu}[\phi(1-D(x))]| - 2\phi(1/2)$$

- When $\phi(t)=t$, $\mathcal{F}$-distance is a pseudo-metric(Integral Probability Metric, IPM)

F distance

Definition 4.1 (F distance)

Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $[0,1]$ and φ be a concave measuring function. Then the $\mathcal{F}$-divergence with respect to ϕ between two distribution μ and ν supported on $\mathbb{R}^d$ is defined as

$$d_{\mathcal{F},\phi}(\mu,\nu) = \sup_{D\in\mathcal{F}} |\mathbb{E}_{x\sim\mu}[\phi(D(x))] + \mathbb{E}_{x\sim\nu}[\phi(1-D(x))]| - 2\phi(1/2)$$

- When $\phi(t) = t$, $\mathcal{F}$-distance is a pseudo-metric(Integral Probability Metric, IPM)
- When $\phi(t) = \log(t)$ and $\mathcal{F} = \{$all functions from $\mathbb{R}^d$ to $[0,1]\}$, then $d_{\mathcal{F},\phi} = JSD$.

F distance

Definition 4.1 (F distance)

Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $[0,1]$ and φ be a concave measuring function. Then the $\mathcal{F}$-divergence with respect to ϕ between two distribution μ and ν supported on $\mathbb{R}^d$ is defined as

$$d_{\mathcal{F},\phi}(\mu,\nu) = \sup_{D\in\mathcal{F}} |\mathbb{E}_{x\sim\mu}[\phi(D(x))] + \mathbb{E}_{x\sim\nu}[\phi(1-D(x))]| - 2\phi(1/2)$$

- When $\phi(t) = t$, $\mathcal{F}$-distance is a pseudo-metric(Integral Probability Metric, IPM)
- When $\phi(t) = \log(t)$ and $\mathcal{F} = \{$all functions from $\mathbb{R}^d$ to $[0,1]\}$, then $d_{\mathcal{F},\phi} = JSD$.
- When $\phi(t) = t$ and $F = \{$all 1-Lipschitz functions from $\mathbb{R}^d$ to $[0,1]\}$, then $d_{\mathcal{F},\phi} = W_1$.

Neural net distance

Suppose $\mathcal{F}$ is the set of neural networks, and $\phi(t) = t$, then the objective function used empirically in Arjovsky et al. [2017] is equivalent to

$$\min_G d_{\mathcal{F}}(\hat{P}_{real}, \hat{P}_G)$$

Neural net distance

Suppose $\mathcal{F}$ is the set of neural networks, and $\phi(t) = t$, then the objective function used empirically in Arjovsky et al. [2017] is equivalent to

$$\min_G d_{\mathcal{F}}(\hat{P}_{real}, \hat{P}_G)$$

Definition 4.2 (NN distance)

When $\mathcal{F}$ is a neural net, we refer $d_{\mathcal{F},\phi}$ as the **neural net distance**.

Generalization of GAN

Definition 4.3 (Generalization)

We say a divergence or distance $d(\cdot, \cdot)$ between distribution generalizes with m training examples and error ϵ if for the learned distribution $\mathbb{P}_G$, the following holds with high probability

$$d(P_{real}, P_G) - d(\hat{P}_{real}, \hat{P}_G) \leq \epsilon$$

Generalization of GAN

Definition 4.3 (Generalization)

We say a divergence or distance $d(\cdot, \cdot)$ between distribution generalizes with m training examples and error ϵ if for the learned distribution $\mathbb{P}_G$, the following holds with high probability

$$d(P_{real}, P_G) - d(\hat{P}_{real}, \hat{P}_G) \leq \epsilon$$

- JS divergence and Wasserstein distance don't generalize.

Generalization of GAN

Definition 4.3 (Generalization)

We say a divergence or distance $d(\cdot, \cdot)$ between distribution generalizes with m training examples and error ϵ if for the learned distribution $\mathbb{P}_G$, the following holds with high probability

$$d(P_{real}, P_G) - d(\hat{P}_{real}, \hat{P}_G) \leq \epsilon$$

- JS divergence and Wasserstein distance don't generalize.
- Neural Net distance generalizes.

Lack of diversity

The neural net distance $d_{NN}(\mu, \nu)$ can be small even if μ, ν are not very close.

Lack of diversity

The neural net distance $d_{NN}(\mu, \nu)$ can be small even if μ, ν are not very close.

Theorem 9 (Low-capacity discriminators cannot detect lack of diversity)

Let $\hat{\mu}$ be the empirical version of distribution μ with m samples. There is a some constant c such that when $m \leq c$, we have that with high probability

$$d_{\mathcal{F},\phi}(\mu, \hat{\mu}) \leq \epsilon.$$

Game theory and equilibrium

For a class of generators $\{G_u, u \in \mathcal{U}\}$ and a class of discriminators $\{D_v, v \in \mathcal{V}\}$, we can define the payoff $F(u, v)$ of the game between generator and discriminator

$$F(u, v) = \mathbb{E}_{x \sim P_{real}}[\phi(D_v(x))] + \mathbb{E}_{x \sim P_{G_u}}[\phi(1 - D_v(x))].$$

Game theory and equilibrium

For a class of generators $\{G_u, u \in \mathcal{U}\}$ and a class of discriminators $\{D_v, v \in \mathcal{V}\}$, we can define the payoff $F(u, v)$ of the game between generator and discriminator

$$F(u, v) = \mathbb{E}_{x \sim P_{real}}[\phi(D_v(x))] + \mathbb{E}_{x \sim P_{G_u}}[\phi(1 - D_v(x)))].$$

A mixed strategy for the generator is just a distribution $\mathcal{S}_u$ supported on U, and one for discriminator is a distribution $\mathcal{S}_v$ supported on V.

Game theory and equilibrium

For a class of generators $\{G_u, u \in \mathcal{U}\}$ and a class of discriminators $\{D_v, v \in \mathcal{V}\}$, we can define the payoff $F(u, v)$ of the game between generator and discriminator

$$F(u, v) = \mathbb{E}_{x \sim P_{real}}[\phi(D_v(x))] + \mathbb{E}_{x \sim P_{G_u}}[\phi(1 - D_v(x)))].$$

A mixed strategy for the generator is just a distribution $\mathcal{S}_u$ supported on U, and one for discriminator is a distribution $\mathcal{S}_v$ supported on V.

Theorem 10 (Mixed Equilibrium)

Then there exists a value V , and a pair of mixed strategies $(\mathcal{S}_u, \mathcal{S}_v)$ such that

$$\forall v, \mathbb{E}_{u \sim \mathcal{S}_u}[F(u, v)] \leq V \textit{ and } \forall v, \mathbb{E}_{v \sim \mathcal{S}_v}[F(u, v)] \geq V$$

Approximate equilibrium

A pair of mixed strategies $(\mathcal{S}_u, \mathcal{S}_v)$ is an ϵ-approximate equilibrium, if for some value V

$$\forall v \in \mathcal{V}, \mathbb{E}_{u \sim \mathcal{S}_u}[F(u, v)] \leq V + \epsilon;$$

$$\forall u \in \mathcal{U}, \mathbb{E}_{v \sim \mathcal{S}_v}[F(u, v)] \geq V - \epsilon$$

Theorem 11

Suppose ϕ is L_ϕ-Lipschitz and bounded, the generator and discriminators are L-Lipschitz with respect to the parameters and L'-Lipschitz with respect to inputs, then for any ϵ, there exists $T(\epsilon)$ generators $G_{u_1}, \ldots G_{u_T}$ and T discriminators $D_{u_1}, \ldots D_{u_T}$, let $\mathcal{S}_u$ be a uniform distribution on u_i and $\mathcal{S}_v$ be a uniform distribution on v_i, then $(\mathcal{S}_u, \mathcal{S}_v)$ is an ϵ-approximate equilibrium. Furthermore, in this equilibrium the generator "wins", meaning discriminators cannot do better than random guessing.

...

MIX+GAN?

...

MIX+GAN?

Bayes GAN?

...

MIX+GAN?
Bayes GAN?
TO BE CONTINUE...

Thanks for your attention!

For Further Reading

[1] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015

[2] Pinsker's inquality

[3] Lévy–Prokhorov metric