Wasserstein generative adversarial networks

Arjovsky, M., Chintala, S., & Bottou, L. (2017). In International conference on machine learning, 214-223. PMLR.

小川　大智

2023/6/5

理論談話会#14

交通・都市・国土学研究室 M2

研究のサマリー・良い点・悪い点

- 要旨
  低次元多様体上の確率分布を効率的に学習するための理論的な検討と GAN による実装
- 良い点
  - 深層学習の論文だが，モデルの性質についての理論的な裏付けが詳しく説明されている．
  - 他の GAN との理論的な比較が明快．
  - 抽象度が高いので色々な応用が効きそう．
- 悪い点
  - 判別関数のリプシッツ制約について，モデル実装上は解決できていない．
  - 数学用語が多い．

新規性・有用性・信頼度

- 新規性
  - GAN の安定性について分布間の距離の観点から分析した．
  - Wasserstein 距離の性質の理論的な検討と，その性質を利用した機械学習モデルの提案を行った．
- 有用性
  - 生成系モデル一般の学習の安定性について分析する枠組みを提供している．
- 信頼度
  - 数学的な裏付けがあるので信頼度は高い．

論文の構成

1. Introduction
2. Different Distances
3. Wasserstein GAN
4. Emperical Results
5. Related Works
6. Conclusion

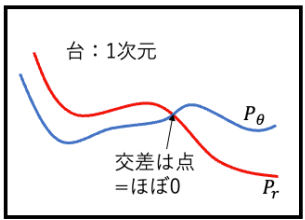
1. Introduction

1. 教師なし学習

教師なし学習：データの確率分布を学習する

- 学習=モデルのパラメータを決めること
- データが真に従う分布に近くなるように確率密度関数のパラメータを最適化する
- E.g.) 最尤推定法
  - $\{x^{(i)}\}_{i=1}^{m}$：実データのサンプル
  - $(P_\theta)_{\theta \in \mathbb{R}^d}$：パラメタライズされた確率密度族

$$\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$



データ：2次元

- これは，真の分布 P_r と P_θ の"距離"を最小化することに相当
  E.g.) Kullback-Leibler divergence $KL(\mathbb{P}_r \| \mathbb{P}_\theta)$
- しかし，P_r の台（定義域の $P_r \neq 0$ の部分）が低次元多様体のとき，P_r と P_θ はほとんど交差しない．$\rightarrow KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \infty$

1. 生成系モデル

通常は P_r を直接近似するのではなく，潜在変数 z を導入して周辺分布を近づける．

- z：確率分布 $p(z)$ に従う潜在変数．$p(z)$ は外生的に与える．
- $g_\theta : \mathcal{Z} \to \mathcal{X}$：生成系モデル
- $P_\theta(x) = \int_{\mathcal{Z}} p(z) g_\theta(z) dz$：生成されたデータの周辺分布（$z$ で周辺化されている．）

この方法は 2 点で優れている．

- P_θ の台が低次元多様体となることを表現できる．
- 確率密度がわかることよりもサンプルを簡単に生成できることの方が多くの場面で有用．（一般に任意の高次元分布に従う乱数を発生させるのは困難）

有名な生成系モデルとしては，

- Variational Auto-Encoder (VAE)[1]：周辺尤度最大化に基づく．
- Generative Adversarial Networks (GAN)[2]：目的関数（P_r と P_θ の距離）を柔軟に変えられる．e.g.) Jensen-Shannon divergence, f-divergence

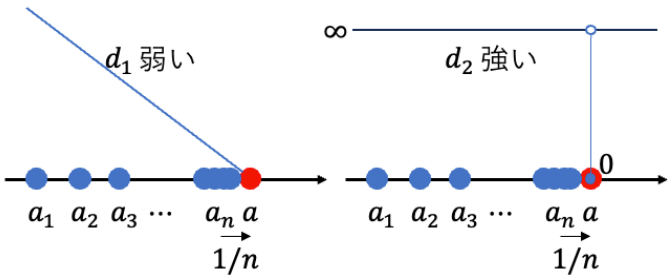
1. 本研究の位置付け

- Jensen-Shannon divergence や f-divergence では根本的に学習が失敗しやすい
  → 教師なし学習のための，Earth Mover (EM) 距離の性質の理論的検討
- GAN に EM 距離を導入
  →Wasserstein GAN の定式化と学習の安定性についての理論的検討
- GAN における種々の問題点の解決 e.g.) ハイパーパラメータに対する敏感性，モード崩壊
  →Wasserstein GAN の実験的評価

2. Different Distances

2. 用語の説明

- $\mathcal{X}$：データの定義域，コンパクト距離空間（e.g. $[0,1]^d$ のような閉集合）
- Σ：$\mathcal{X}$ のボレル集合（$\mathcal{X}$ の部分集合族で大きさが測れるような性質を持つもの）
- μ：$\mathcal{X}$ 上の確率測度（集合の大きさ）
- "距離（位相）が弱い"：その距離のもとで点列が収束しやすい．



2. 距離の定義 1

生成分布 P_θ と実分布 P_r の距離　※ divergence は距離ではないがここでは並列して述べる

- Total Variation (TV) distance（全変動距離）

$$\delta(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{A \in \Sigma} |P_r(A) - P_\theta(A)|$$

- Kullback-Leibler (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \int_{\mathcal{X}} P_r(x) \log \frac{P_r(x)}{P_\theta(x)} d\mu(x)$$

  - KL divergence は $\mathbb{P}_r$ と $\mathbb{P}_\theta$ を入れ替えると値が変わる．
  - $P_\theta(x) = 0$ かつ $P_r(x) > 0$ のような x があると $KL(\mathbb{P}_r \| \mathbb{P}_\theta) = \infty$ になり得る．

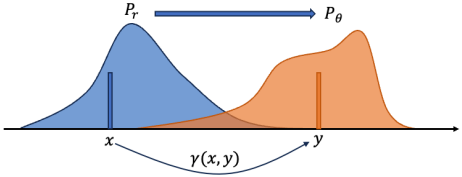
2. 距離の定義 2

- Jensen-Shannon (JS) divergence

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

$$\mathbb{P}_m = \frac{1}{2}(\mathbb{P}_r + \mathbb{P}_\theta)$$

  - **対称性は満たされる．**
  - **普通の GAN の目的関数に使われる．**



- Earth Mover (EM) distance（Wasserstein-1 **距離**）

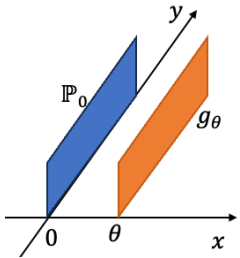
$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

  - $\Pi(\mathbb{P}_r, \mathbb{P}_\theta)$ **は，** $\mathbb{P}_r$ **と** $\mathbb{P}_\theta$ **を周辺分布とするような** x **と** y **の同時分布の集合．**
  - **総量が** $\gamma(x, \cdot) \sim \mathbb{P}_r, \gamma(\cdot, y) \sim \mathbb{P}_\theta$ **となるように，砂山を移動させるときの最小コスト．**

2. 各距離の性質

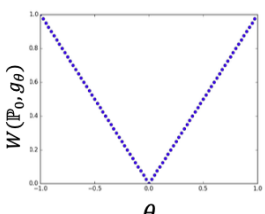
例）2 次元平面上の直線の学習

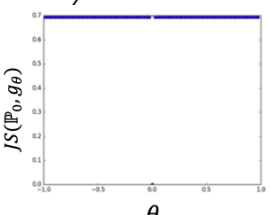
- $Z \sim U[0,1]$：$[0,1]$ 上の一様分布
- $\mathbb{P}_0 = (0, Z)$：$x = 0$ 上にのみ一様分布の確率が存在する分布
- $\mathbb{P}_\theta = (\theta, Z)$：$x = \theta$ 上にのみ一様分布の確率が存在する分布



このとき，

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$
- $JS\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$
- $KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \infty & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$
- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$





$\theta_t \to 0$ のもとで $(\mathbb{P}_{\theta_t})_{t \in \mathbb{N}} \to \mathbb{P}_0$ となるのは W のみ．

2. 定理 1

Wasserstein 距離は JS 距離よりも弱く，θ に対して性質の良い関数となっている．

定理 1

$\mathbb{P}_0$ を $\mathcal{X}$ 上の固定された分布，Z を $\mathcal{Z}$ 上の確率変数とする．関数 $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ を $g_\theta(z)$ と書くこととし，$\mathbb{P}_\theta$ を周辺分布 $g_\theta(Z)$ とする．このとき，

1. g が連続であるならば，$W(\mathbb{P}_0, \mathbb{P}_\theta)$ は θ の連続関数である．
2. g が局所的に Lipschitz 連続で，かつ仮定 1 を満たすならば，$W(\mathbb{P}_0, \mathbb{P}_\theta)$ は全ての点で連続で，ほとんど全ての点で微分可能である．
3. 1，2 は JS 距離や KL ダイバージェンスには成り立たない．

仮定 1

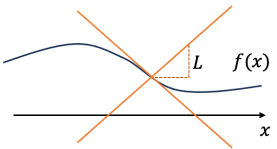
$\mathcal{Z}$ 上のある分布 p について，局所リプシッツ定数 $L(\theta, z)$ が以下を満たす．

$$\mathbb{E}_{z \sim p}\left[L(\theta, z)\right] < \infty$$

2. （補足）リプシッツ連続

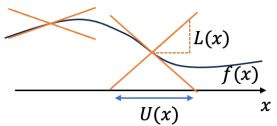
- **リプシッツ連続**
  任意の $\theta, \theta' \in \mathbb{R}^d$ と $z, z' \in \mathcal{Z}$ に対して，ある定数 L が存在して，

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



- **局所リプシッツ連続**
  任意の $\theta \in \mathbb{R}^d$ と $z \in \mathcal{Z}$ に対して，(θ, z) を含むようなある開集合 U が存在して，任意の $\theta', z' \in U$ に対して，ある定数 $L(\theta, z)$ が存在して，

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



2. 系 1

g_θ としてニューラルネットワークを用いても良いことがわかる．

系 1

g_θ を θ をパラメータとするニューラルネットワークとし，$p(z)$ が $\mathbb{E}_{z\sim p(z)}[\|z\|] < \infty$ を満たすとする．このとき，仮定 1 が満たされるので，$W(\mathbb{P}_0, \mathbb{P}_\theta)$ は全ての点で連続で，ほとんど全ての点で微分可能である．

2. 定理 2

$KL > JS, TV > EM$ の順に弱くなる．

定理 2

$\mathbb{P}$ をコンパクト空間 $\mathcal{X}$ 上の確率分布，$(\mathbb{P}_n)_{n\in\mathbb{N}}$ を $\mathcal{X}$ 上の確率分布の列とする．$n \to \infty$ の場合を考える．

1. 以下は同値
   - $\delta(\mathbb{P}_n, \mathbb{P}) \to 0$
   - $JS(\mathbb{P}_n, \mathbb{P}) \to 0$
2. 以下は同値
   - $W(\mathbb{P}_n, \mathbb{P}) \to 0$
   - $\mathbb{P}$ は $\mathbb{P}_n$ に収束する．
3. $KL(\mathbb{P}_n \| \mathbb{P}) \to 0$ または $KL(\mathbb{P} \| \mathbb{P}_n) \to 0$ ならば，1 が成り立つ．
4. 1 が成り立つなら 2 が成り立つ．

3. Wasserstein GAN

最適輸送問題

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

l 個の点からなる空間の場合には，同時分布 $\Gamma \in \mathbb{R}^{l\times l}$，距離行列 $D \in \mathbb{R}^{l\times l}$ を用いて，

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \inf_{\Gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_\theta)} \langle \Gamma, D \rangle_F$$

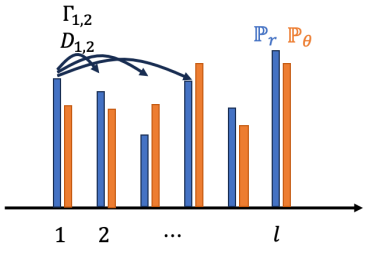
線形計画問題に変換 ($c = \mathrm{vec}(D), x = \mathrm{vec}(\Gamma), b = (P_r, P_\theta)^T$)

$$\min_x c^T x \quad \text{s.t. } Ax = b, x \geq 0$$

制約条件が周辺分布のため，扱いづらい．$\rightarrow$ 相対問題

$$\max_\lambda b^T \lambda \quad \text{s.t. } A^T \lambda \leq c$$

目的関数が周辺分布 b の関数になっているので扱いやすい．



連続量の場合の双対問題

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{y\sim\mathbb{P}_\theta}[f(y)]$$

【証明】

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \inf_\gamma \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] + \underbrace{\sup_f \mathbb{E}_{s\sim\mathbb{P}_r}[f(s)] - \mathbb{E}_{t\sim\mathbb{P}_\theta}[f(t)] - (f(x) - f(y))}_{= \begin{cases} 0 & \text{if } \gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_\theta) \\ \infty & \text{otherwise} \end{cases}}$$

$$= \inf_\gamma \sup_f \mathbb{E}_{s\sim\mathbb{P}_r}[f(s)] - \mathbb{E}_{t\sim\mathbb{P}_\theta}[f(t)] + \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] - (f(x) - f(y))$$

$$= \sup_f \inf_\gamma \mathbb{E}_{s\sim\mathbb{P}_r}[f(s)] - \mathbb{E}_{t\sim\mathbb{P}_\theta}[f(t)] + \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] - (f(x) - f(y))$$

$$= \sup_f \mathbb{E}_{s\sim\mathbb{P}_r}[f(s)] - \mathbb{E}_{t\sim\mathbb{P}_\theta}[f(t)] + \underbrace{\inf_\gamma \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] - (f(x) - f(y))}_{= \begin{cases} 0 & \text{if } \|f\|_L \leq 1 \\ -\infty & \text{otherwise} \end{cases}}$$

$$= \sup_{\|f\|_L \leq 1} \mathbb{E}_{s\sim\mathbb{P}_r}[f(s)] - \mathbb{E}_{t\sim\mathbb{P}_\theta}[f(t)]$$

3. Kantorovich-Rubinstein 双対性

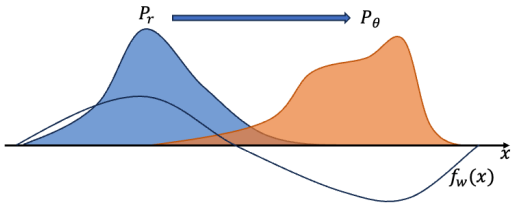
最適輸送問題

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{y\sim\mathbb{P}_\theta}[f(y)]$$

$\mathcal{X}$ 上の 1-Lipschitz な関数 f についての最適化問題．
$\rightarrow$1-Lipschitz な関数 f_w をニューラルネットで作り，最大化問題を解けば W がわかる．

$$\max_w \mathbb{E}_{x\sim\mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z\sim p(z)}[f_w(g_\theta(z))]$$

さらに，第二項を θ で微分すれば W の最小化ができる．



3. 定理 3

W の最小化問題は機械学習で解ける．

定理 3

$\mathbb{P}_r$ を任意の分布，$\mathbb{P}_\theta$ を $g_\theta(Z)$ の確率変数 $z \sim p(z)$ に対する周辺分布とする．分布 p は仮定 1 を満たすとする．このとき，

- 最適解 $f : \mathcal{X} \to \mathbb{R}$ が存在する．

$$\max_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

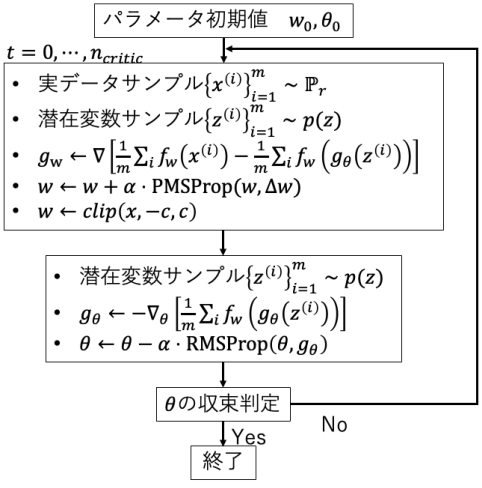
- $\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))]$

WGAN

- 判別関数：f_w，リプシッツ制約を満たすため，w をクリッピングする．
- 生成器：g_θ
- ハイパーパラメータ：α，c，m，n_{critic}．

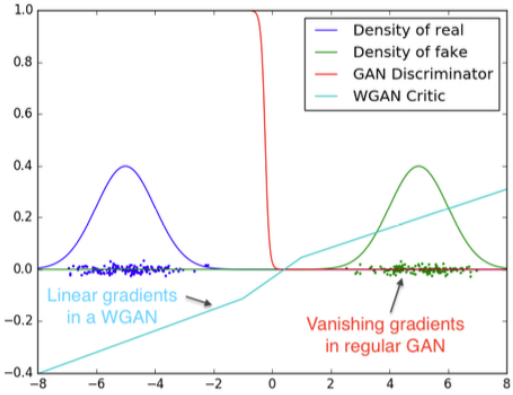
ただ，ウェイトクリッピングは難しい

- クリップしすぎる
  $\rightarrow f_w$ の表現力が足りない．
- クリップが甘い
  $\rightarrow$ リプシッツ制約を満たさない．



3. WGAN の性質のよさ

- 最適な判別関数になるまで学習しても生成器の勾配が消失しない．（定理 1）
- モード崩壊しない．
  ← モード崩壊は判別器の学習が止まってしまうことに起因するため．

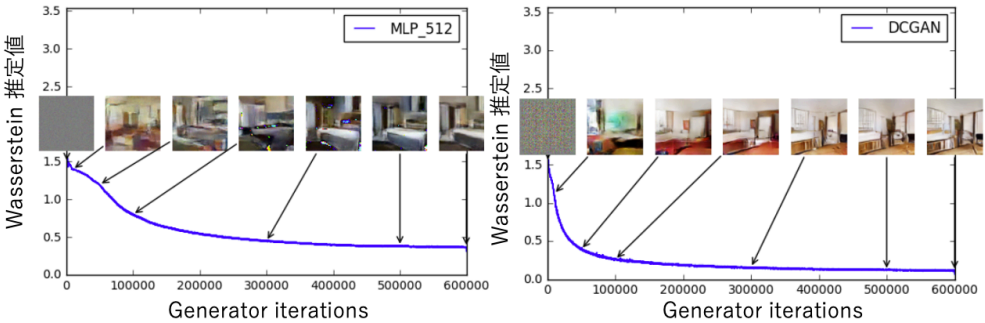


4. Empirical Results

4. 実験設定

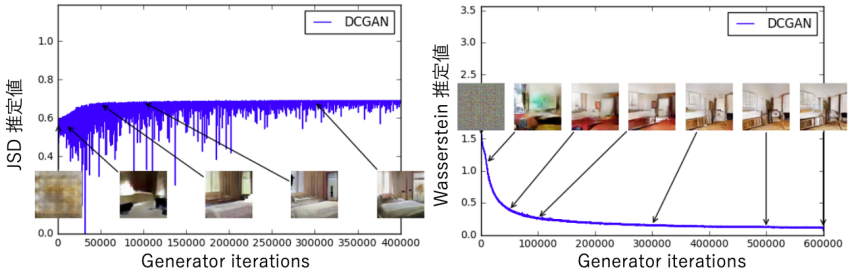
- **データセット**：LSUN-Bedrooms dataset
- **比較対象**：DCGAN
- $\mathcal{X}$：$3 \times 64 \times 64$ **の画像**
- **ハイパーパラメータ**
  - α：0.00005
  - c：0.01
  - m：64
  - n_{critic}：5

4. 損失関数としての評価



- 異なるモデルの生成器を用いても（判別関数は同一モデル）EM 距離（推定値）は生成画像の品質を反映する．
- 普通の GAN では，生成器と判別器が同時に学習するため，判別関数の値は生成器の質を反映しない．

4. 損失関数としての評価



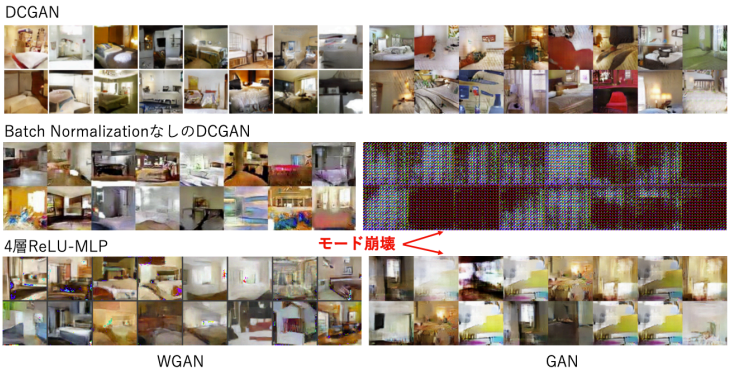
- 普通の GAN は JS 距離の最大化を目指す．
  $L(D, g_\theta) = \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x)] + \mathbb{E}_{x\sim\mathbb{P}_\theta}[\log(1 - D(x))]$
- 判別器の学習が進むにつれ，$\log 2$ に近づいていく．この場合意味のある画像が生成されるが，そうならないときはモード崩壊している．

※プロットは $\frac{1}{2}L(D, g_\theta) + \log 2$（JS 距離の下限）

4. 安定性の評価

WGAN の安定性

- GAN：生成器と判別器のバランスをとりながら学習
- WGAN：判別器の性能が良いほど生成器の学習も進む．
  → モード崩壊は発生せず．



5. Related Works

5. Integral probability metrics

IPMs

$\mathcal{X}$ から $\mathbb{R}$ への関数の集合 $\mathcal{F}$ が与えられたとき，

$$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

- $\mathcal{F}$ が 1-Lipschitz な関数の集合 $\to d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = W(\mathbb{P}_r, \mathbb{P}_\theta)$
- $\mathcal{F}$ が $[-1, 1]$ 内の値を持つ連続関数 $\to d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \delta(\mathbb{P}_r, \mathbb{P}_\theta)$
- Energy-based GAN：判別関数が $\delta(\mathbb{P}_r, \mathbb{P}_\theta)$ を学習．TV 距離は JS と同じ強さ．
- $\mathcal{F}$ が再生核ヒルベルト空間のとき，$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$（MMD）は TV と同等かそれより強い．
  ※ RKHS：再生性（$\forall f \in H, \forall x \in \mathcal{X}, \exists \phi_x \in H; \langle f, \phi_x \rangle = f(x)$）を満たすヒルベルト空間，再生核 $k(y, x) := \phi_x(y)$ を用いて内積 $\langle \phi_x, \phi_y \rangle = k(x, y)$
- Generative Moment Matching Network：MMD を用いた生成モデル．MMD は推定のためのモデルを必要としないが，計算量がサンプル数の自乗に比例するため適用が限られる．

5. Wasserstein Training of Restricted Bolzmann Machines[4]

Restricted Boltzmann Machine (RBM) の学習に Wasserstein 距離を利用

- RBM：$\mathcal{X} = 0, 1^d$ 上の分布 $p_\theta(x)$ を観測データの分布 $\hat{p}$ に近づける．
- 通常は KL divergence で学習するが，これまでの議論と同様に，Wasserstein 距離の方が滑らかであり，θ が収束しやすい．
- Wasserstein 距離の計算には Sinkhorn アルゴリズムを用いる．
  - 線形計画問題の目的関数にエントロピー正則化項を導入し，収束計算により解く．
  - エントロピー項は狭義凸関数であるため，最適化問題の解が一意に定まる．

5. Stochastic optimization for large-scale optimal transport[5]

最適輸送問題の Shinkhorn アルゴリズムに変わる高速な解法の提案

- Sinkhorn アルゴリズムはサンプル数の自乗のオーダーの計算量 $\rightarrow$ 確率的勾配降下法の利用
- 連続分布を扱う場合は RKHS 上での確率的勾配降下法を考える.

$\rightarrow$ より高速な生成モデルの評価方法となる可能性.

6. Conclusion

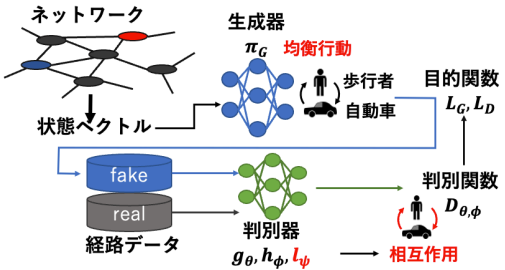
6. 結論

- WGAN を提案した．
  - 高い安定性（生成器と判別器の学習のスケジューリングが不要，モード崩壊しない）
  - 学習曲線が意味のある指標になる．$\rightarrow$ デバッグやハイパーパラメータのチューニングに役立つ．
- WGAN の安定性について理論的な裏付けを行い，他の距離との関係を示した．

自分の研究への適用

歩行者と自動車の相互作用

- 現状
  - Multi-agent AIRL で歩車の動きを学習
  - 歩車の間のナッシュ均衡戦略が学習される
  - 相互作用項の定式化 $r_{ext,i}(a_i, a_{i-1}, s)$
- 課題
  - 学習の安定性 $\rightarrow$ 今回の方法を用いれば解決



Wasserstein IRL はすでにある．（$\mathcal{X}$ を状態 $\mathcal{S}$ と行動 $\mathcal{A}$ に置き換え，f を報酬期待値としたもの．）

- 相互作用項は以前と同様に定式化できそう．
- 複数主体の場合に WIRL が何の均衡になっているかは不明．

参考文献

- Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advanc in Neural.
- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
- Arjovsky, M., & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862.
- Montavon, G., Müller, K. R., & Cuturi, M. (2016). Wasserstein training of restricted Boltzmann machines. Advances in Neural Information Processing Systems, 29.
- Genevay, A., Cuturi, M., Peyré, G., & Bach, F. (2016). Stochastic optimization for large-scale optimal transport. Advances in neural information processing systems, 29.