深層生成モデルの理論と応用

福水 健次

統計数理研究所 ／ Preferred Networks

第5回 統計・機械学習若手シンポジウム

2020年12月3日

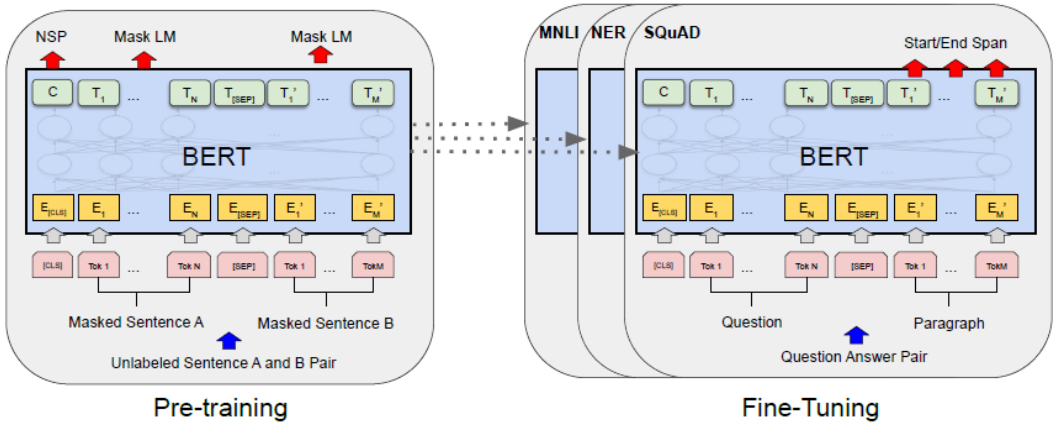
深層学習の隆盛

AlphaGo/AlphaZero

画像生成

Glow [Kingma and Dhariwal, NeurIPS2018]

自然言語処理



Bert [Devlin et al 2019]

画像変換

Style GAN [Zhu, Park, Isola, and Efros ICCV2017.]

深層生成モデルの発展：顔画像

2014
（Goodfellow et al)

2015

2016

2017
(Karas et al ProgressiveGAN)

2018
(Karas et al StyleGAN)

(Goodfellow, ICLR2019より)

今日の内容

1. 統計的推論から見た深層生成モデル

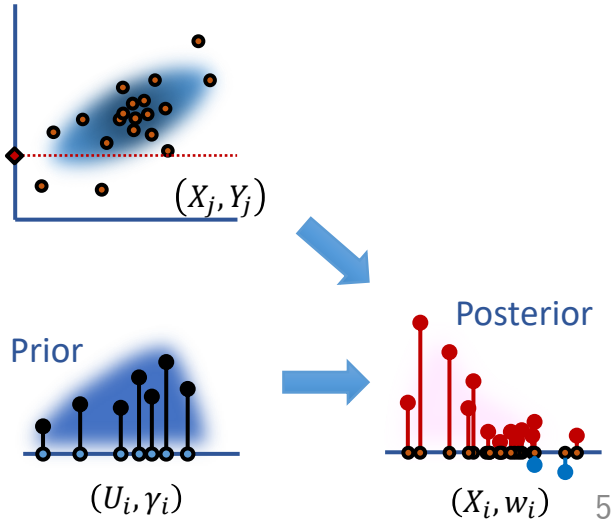
2. 深層生成モデルの理論的考察
   -- GANの安定性に関する理論解析 --

統計的推論

統計的推論：データの背後にある確率分布に関する推論

「確率分布」をどう表現するか？

- 確率密度関数による表現
  - パラメータ族 $f(x;\theta)$
  - カーネル密度推定 $\hat{p}_h(x) = \frac{1}{N}\sum_i k_\sigma(x - X_i)$
- サンプル表現
  - 重み付き粒子表現： e.g. Particle filter
  - カーネル平均： グラム行列による推論
- 生成モデルによる表現



生成モデル

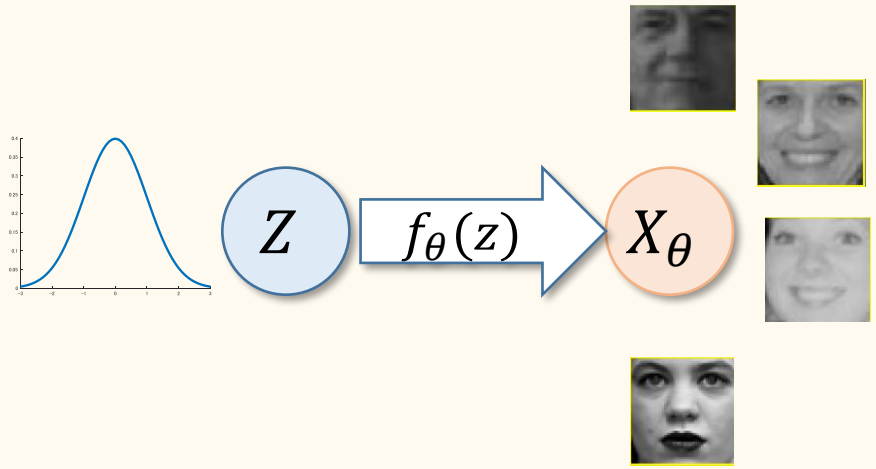
生成モデル：　ある確率分布に従うサンプル X の生成過程のモデル

- Explicit generative model: 密度関数で記述する． e.g. グラフィカルモデル

$$X \sim p(x;\theta)dx$$

- Implicit generative model:
  潜在変数 $Z \sim Q$ (e.g. 標準正規分布)
  からの写像として表現.

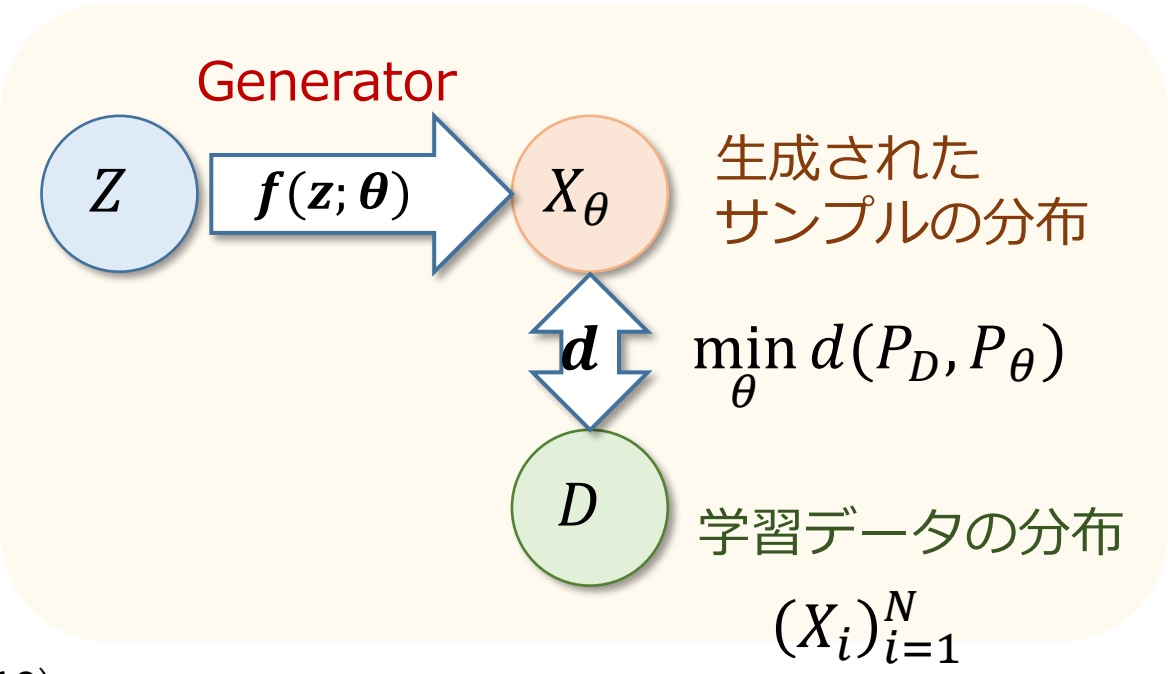
$$X = f(Z;\theta), \qquad Z \sim Q$$



- 深層生成モデル：$f(z;\theta)$ として深層ニューラルネットを使う.
  一般に密度関数は書けない

生成モデルの学習

- 分布を近くすればよい
  - d: 分布の相違をはかる尺度
    - f-divergence
      → f-GAN (Nowozin et al NIPS2016)
    - Wasserstein距離
      → Wasserstein-GAN
      (Arjovsky et al ICLR2017)
    - MMD（カーネル法）
      → MMD-GAN
      (Li et al NIPS2017; Binkowski et al ICLR2018)



GAN：Generative Adversarial Networks
(Goodfellow et al 2014)

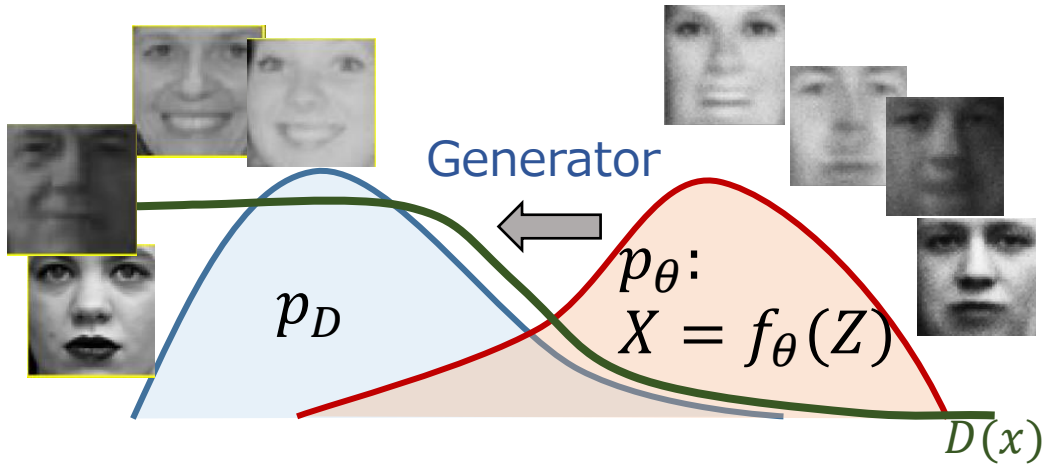
- 分布の距離：Jensen-Shannon divergence

$$JS(p_D, p_w) = KL\left(p_D \middle\| (p_D + p_\theta)/2\right) + KL\left(p_\theta \middle\| (p_D + p_\theta)/2\right)$$

- Generator：分布間距離の最小化

$$\min_{f_\theta} JS(p_D, p_\theta)$$

p_D: データ（真）の密度関数
p_θ: 生成モデルの密度関数　$X = f_\theta(Z)$



- Discriminator：JS-div.は計算困難 → 識別問題に還元

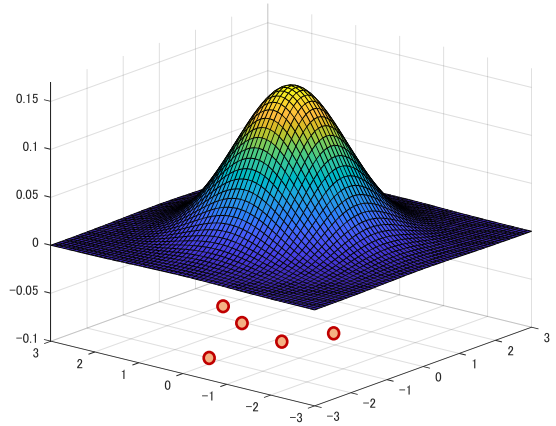
$$JS(p_D, p_\theta) = \sup_{D(x)\text{: DNN}} E_{Y\sim p_D}[\log D(Y)] - E_{X=f_\theta(Z)}[\log(1 - D(X))] + const.$$

− Logistic loss

$D(x) := \frac{p_D(x)}{p_D(x)+p_\theta(x)}$が最適

サンプラーとみなす

- 高性能サンプラー



DNN

Generated by http://www.whichfaceisreal.com/（StyleGAN)
by Jevin West and Carl Bergstrom, U. Washington

- サンプリングによる統計的推論に使おう
- 特に高次元が得意か？

Figure 5: Images generated with U-Net GAN trained on COCO-Animals with resolution 128 × 128.

U-Net GAN [Schönfeld et al CVPR2020]

GANによる統計的推論

例1) 生成モデルのモデル選択：
生成モデルを学習し，データへの適合度をみる．
→ 因果推論の例

例2) ベイズ推論：
事後確率を生成モデルで実現 → 尤度なし推論の例

深層生成モデルによる因果方向推定

- Causal Generative Neural Networks (CGNN, Goudet et al 2017)

  Data： $\mathbf{X} = \{(X_1^i, X_2^i)\}_{i=1}^m$

  生成モデル(A), (B)をMMDにより学習（GMMN）.

  モデルのデータへの適合度によって向きを判定

  - Fitting:

    (A) の場合：$(X_1, \hat{X}_2^A) = (X_1, f(X_1, Z)),\ Z \sim N(0,1)$

    $\hat{\mathbf{X}}^A = \{(X_1^i, \hat{X}_2^{i,A})\}_{i=1}^m$

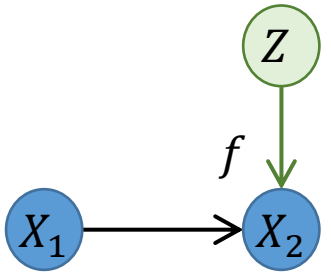
    $MMD_{emp}(\mathbf{X}, \hat{\mathbf{X}}^A)$　を目的関数に用いる

  - 向きの判定

    If $MMD_{emp}(\mathbf{X}, \hat{\mathbf{X}}^A) \leq MMD_{emp}(\mathbf{X}, \hat{\mathbf{X}}^B)$　→　choose (A)

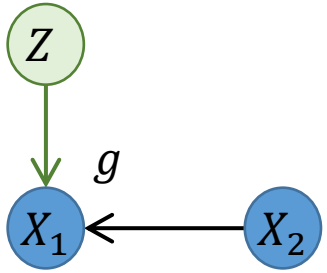
    Else　choose (B).

(A)



$X_2 = f(X_1, Z)$

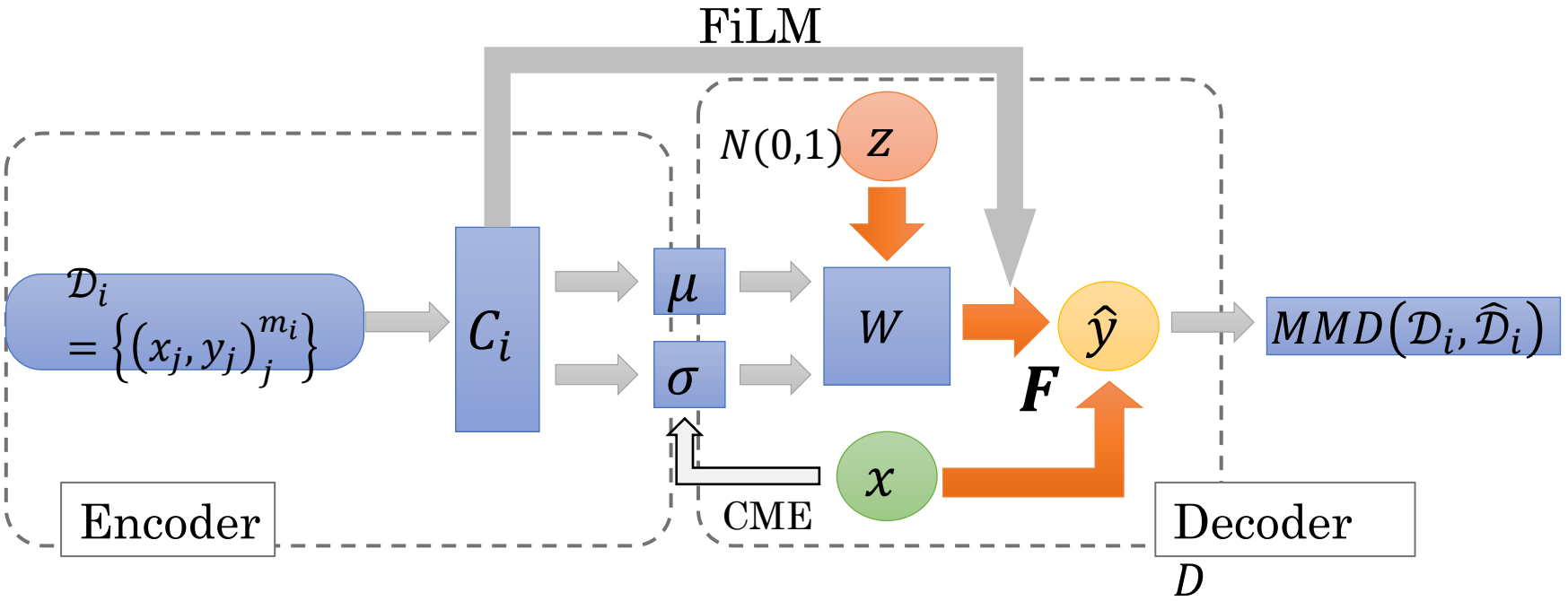
(B)



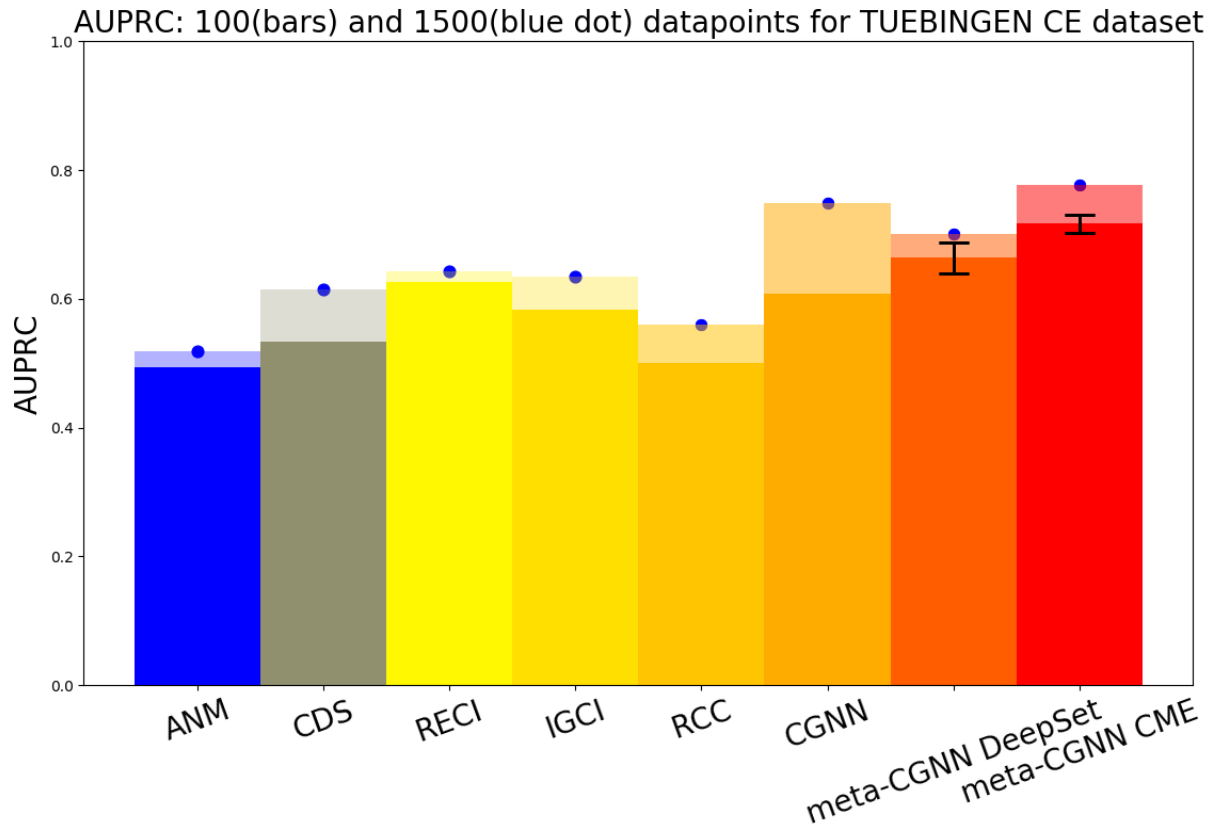
$X_1 = g(X_2, Z)$


J.-F. Ton (Oxford)

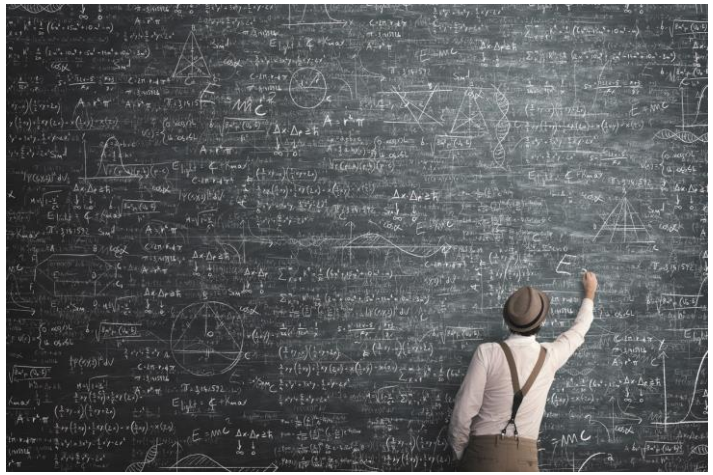
- データ駆動の因果推論：Meta-CGNN (Ton, Sejdinovic, Fukumizu. AAAI 2021)
  - Meta-leaning への拡張：N個ののcause-effect データセット
    $\mathcal{D}_i = \{(X_j^i, Y_j^i)\}_{j=1}^{m_i} \quad i = 1, \ldots, N$　因果 $X^i \longrightarrow Y^i$.
  - $\hat{Y}_j^i = F(X_j^i, Z_j^i, C_i)$　Datasetに共通の生成モデル F を作る.
    $C_i = \Phi(\mathcal{D}_i)$：Dataset feature. $\mathcal{D}_i$の同時分布を表す特徴
  - 向きの判定：$(\mathbf{X}_{test}, \hat{\mathbf{Y}}_{test})$ と $(\mathbf{Y}_{test}, \hat{\mathbf{X}}_{test})$ の適合度を比較.



- Tuebingen Cause Effect Pair データベース（99データセット）



深層生成モデルによるBayes推論



Beyes推論

$$q(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{\int p(y|\theta)\pi(\theta)d\theta}$$

大問題「どう計算するか？」

- サンプリング： Markov Chain Monte Carlo (MCMC), Sequential MC /粒子法, ...
- 近似計算： Laplace近似，変分ベイズ，etc

→ 深層生成モデルを使う.
特に，尤度が陽にかけないケースに焦点をあてる（近似ベイズ計算，ABC）
Tran et al NIPS 2017; Yang et al. NeurIPS2018.

D. Tran, R. Ranganath, D. Blei. Hierarchical Implicit Models and Likelihood-Free Variational Inference (NIPS 2017)
Y. Yang, B. Dai, N. Kiyavash N. He. Predictive Approximate Bayesian Computation via Saddle Points (NeurIPS 2018)

生成モデルによるベイズ推論

仮定：Likelihood-free, $p(y|\theta)$ が陽にかけない

- 同時分布の比較

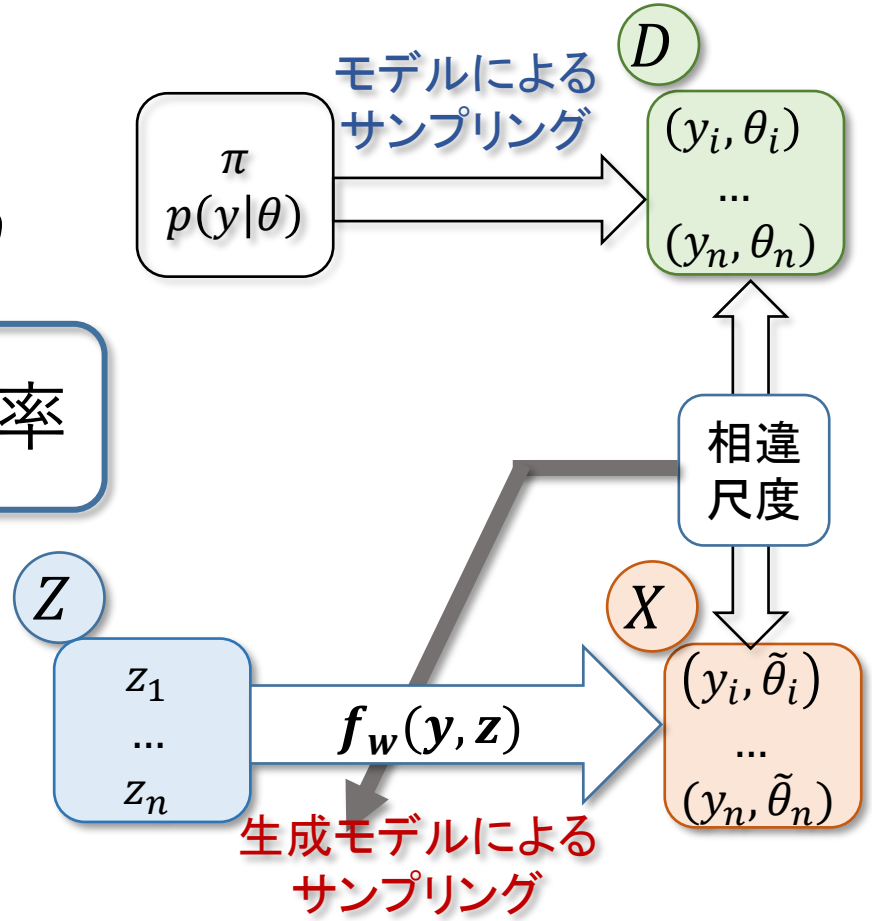
$$p(\theta, y) = p(y|\theta)\pi(\theta)$$
$$q(\theta, y) \coloneqq q(\theta|y)p(y),\ p(y) = \int p(y|\theta)\pi(\theta)d\theta$$

$p(\theta, y) = q(\theta, y)$ ならば $q(\theta|y)$ は事後確率

サンプリング ← Likelihood-free

$p(\theta, y)$: $\theta_i \sim \pi(\theta)d\theta$,
$y_i \sim p(y|\theta_i)dy$

$q(\theta, y)$: $\tilde{\theta}_i = f_w(y_i, z_i)$, 深層生成モデル
$z_i \sim p_0(z)dz$ (ガウスなど)



変分ベイズ ≒ GAN

Likelihood-Free Variational Inference（Tran, Ranganath, Blei, NIPS2017）

- 変分Bayes：事後確率の有力な近似手法

$$\log p(y) = \log \int p(y|\theta)\pi(\theta)d\theta$$

$q(y,\theta) = q(\theta|y)p(y)$により定義

$$\geq \int q(\theta|y) \log \frac{p(y,\theta)}{q(y,\theta)} d\theta + \log q(y)$$

(ELBO)

$q(\theta|y)$ は任意の分布．等号成立 ⇔ $q(\theta|y) = p(\theta|y)$

ELBOを最大にする $q(\theta|y)$ によって事後確率を近似する

- $p(y|\theta)$ の関数形は陽に書けない → サンプリング
- $q(\theta|y)$：生成モデルで表現 → $\theta = f_w(y), y \sim p(y)$ によりサンプリング

- $\log \frac{p(\theta,y)}{q(\theta,y)}$ の推定 → GANのDiscriminatorにより可能

命題　識別問題

$$\max_r E_{p(y,\theta)}\left[\log \sigma\left(r(Y,\theta)\right)\right] - E_{q(y,\theta)}\left[\log\left(1-\sigma\left(r(Y,\theta)\right)\right)\right] \qquad \sigma(t) = \frac{1}{1+e^{-t}}$$

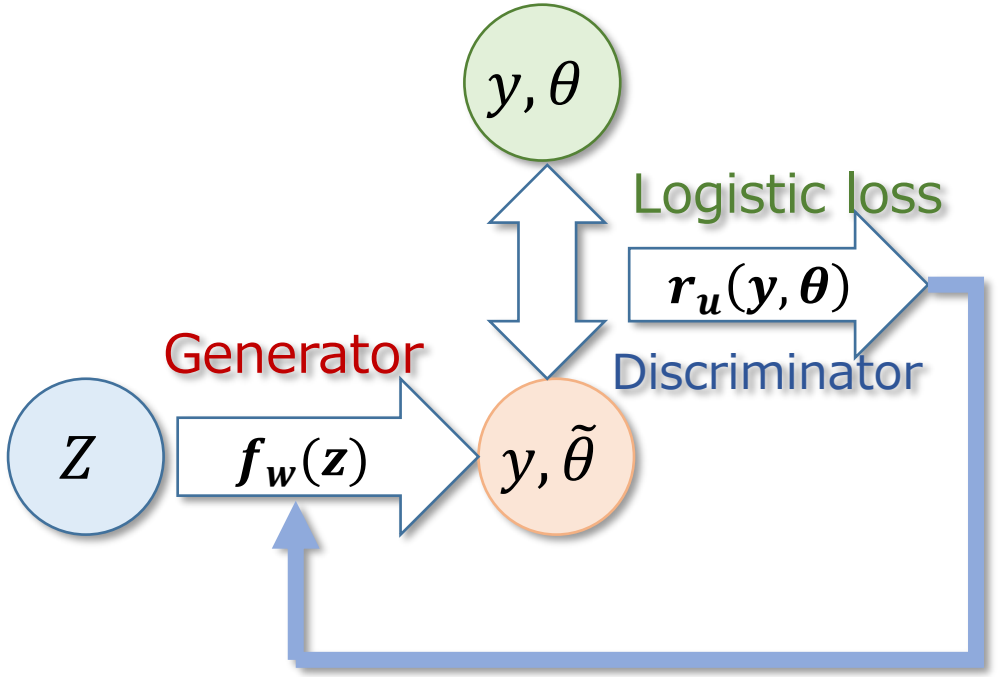
の解は　$r_{opt}(y,\theta) = \log \frac{p(y,\theta)}{q(y,\theta)}$

$r(y,\theta)$ をDNNで構成

- Generator: 変分Bayes（max ELBO）

$$\max_{q:\theta=f_w(y^*,z)} \int q(\theta|y^*) r(y^*,\theta) d\theta$$

y^*: 観測値

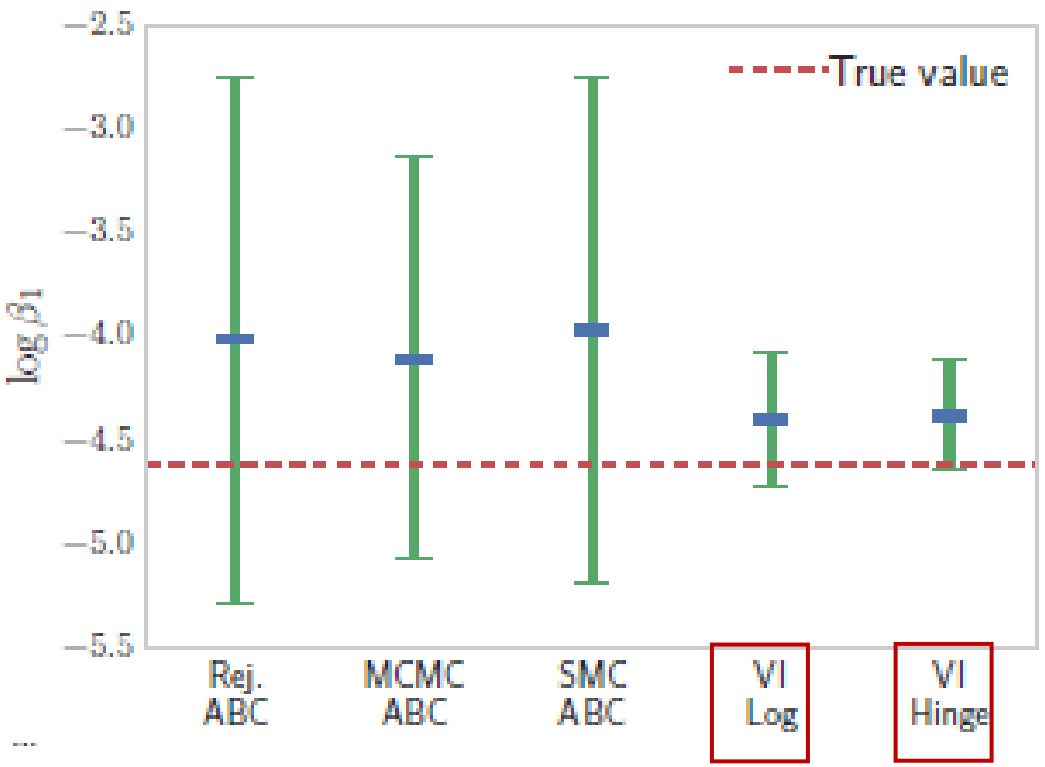


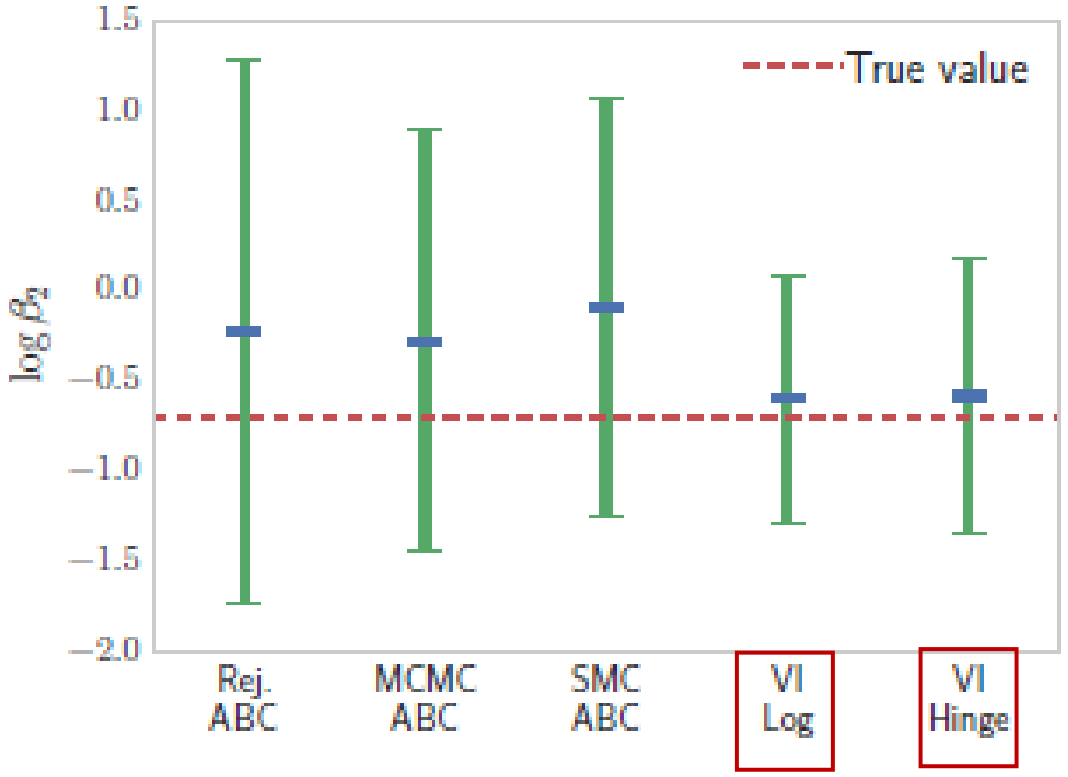
近似Bayes推論（ABC）における比較

Lotka-Volterra Predator-Prey Simulator

$$\frac{dx_1}{dt} = \beta_1 x_1 - \beta_2 x_1 x_2 + \epsilon_1$$
$$\frac{dx_2}{dt} = -\beta_1 x_2 + \beta_3 x_1 x_2 + \epsilon_2$$

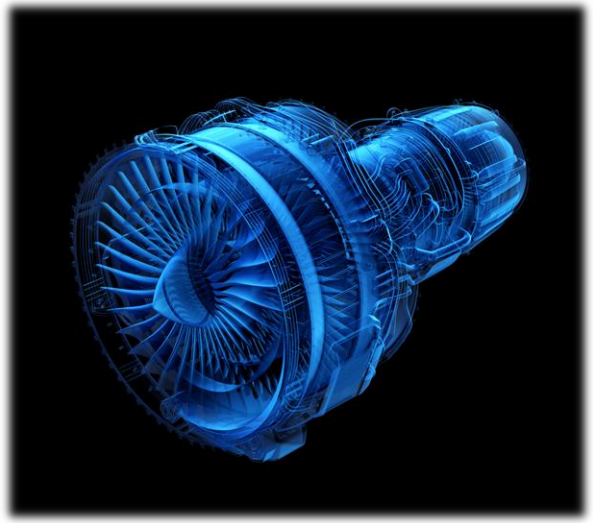
$$\epsilon_1, \epsilon_2 \sim N(0,10)$$





[Tran et al NIPS2017]

Simulation-based inference

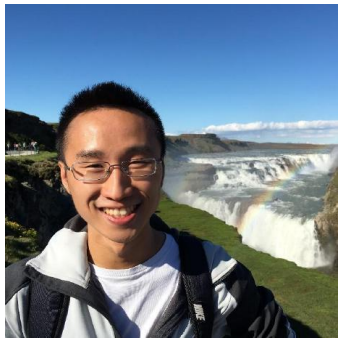


- 高次元データの場合は？
  GANの特長を活かせるか？

- 時系列構造を使えるか？
  シミュレータは（確率）微分方程式のことが多い.

GANの安定性に関する理論解析

C. Chu, K. Minami, K. Fukumizu (2020)
Smoothness and Stability in GANs
ICLR 2020

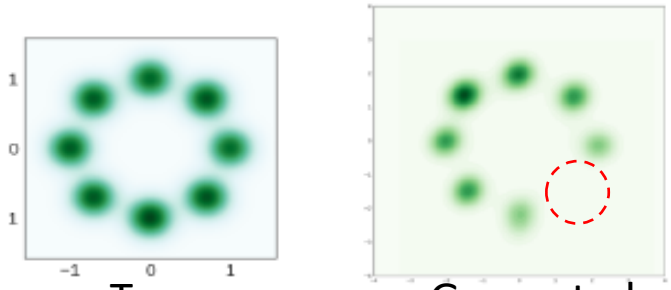


Casey Chu
Stanford U.
Preferred Networks Inern.

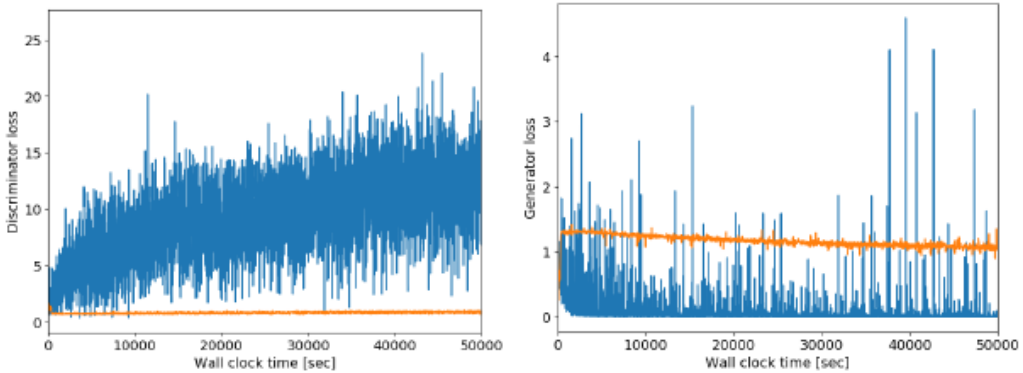
Kentaro Minami
Preferred Networks, Inc.

GANの学習は難しい

- Mode collapse
  いくつかの山が再現されない



- Non-convergence / divergence
  最小化ではなく均衡点を見つけたい.



(a) Discriminator loss　(b) Generator loss

- 勾配消失 (Arjovsky & Bottou ICLR 2017)
  オリジナルのGANではdiscriminator が強いと,
  generatorの勾配が消えやすい.　(後述)

目的：安定化の技法はなぜ有効か?

- GAN安定化の技法
  - 距離関数の選択
    - Wasserstein距離 (Arjovsky et al. ICML2017)
  - 正則化
    - Gradient penalty (Gulrajani et al. NIPS2017)
    - Spectral normalization (Miyato et al. ICLR2018)

Discriminator
の正則化

研究の目的:　Generatorの学習の観点から，これら安定化技法の意味を理論的に明らかにする.

分布間の距離関数とGAN

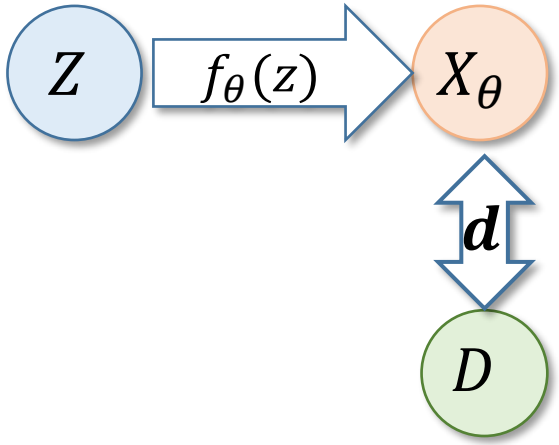
- いろいろなGAN: 距離（ダイバージェンス）の選択により定まる
  - オリジナルGAN (Goodfellow et al 2014) : Jensen-Shannon divergence

    $\mathrm{JSD}(P_D, P_\theta) := \frac{1}{2} KL(P_D || P_m) + \frac{1}{2} KL(P_\theta || P_m), \quad P_m := \frac{P_D + P_\theta}{2}.$
  - Non-saturating GAN (NS-GAN) (Goodfellow et al 2014)

    $\frac{1}{2} KL(P_m || P_D)$
  - f-GAN (Nowozin et al 2016): f-divergence

    $D_f(P_D || P_\theta) := \int dP_\theta f\left(\frac{p_D}{p_\theta}\right) \qquad f: \text{convex}, f(1) = 0$

    オリジナルGANなどが含まれる広い定式化.



- Wasserstein GAN: 1-Wasserstein距離 (Arjovsky et al. ICML2017)

$$W_1(P_D, P_\theta) = \sup_{\varphi:\, ||\varphi||_{Lip} \le 1} \int \varphi(x) dP_D(x) - \int \varphi(x) dP_\theta(x)$$

$$\|\varphi\|_{Lip} := \inf\{L > 0 \mid |\varphi(x) - \varphi(y)| \le L\|x - y\|\}$$

φ：critic (φ_{opt} :Kantorovich potential).

- Generative Moment Matching Network (Li et al. 2015)/ MMD-GAN (Li et al NIPS2017)　MMD: Maximum Mean Discrepancy (Gretton et al 2005).

$$MMD_k^2(P_D, P_\theta) = \left\| E_{P_D}[k(\cdot, X)] - E_{P_\theta}[k(\cdot, X')] \right\|_{H_k}^2$$

k: 正定値カーネル

$$= \sup_{h:\, ||h||_{H_k} \le 1} \int h(x) dP_D(x) - \int h(x) dP_\theta(x)$$

陽に計算できる! (Discriminator不要)

多くの場合,

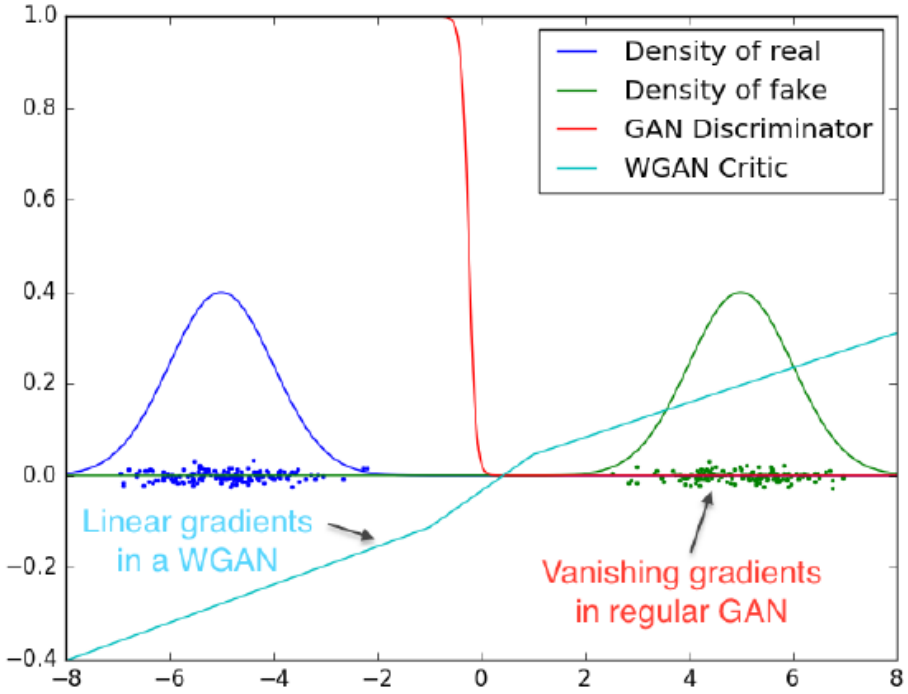
$$\sup_{k \in \mathcal{K}} MMD_k^2(P_D, P_\theta)$$

$\mathcal{K}$:カーネルの族, e.g., $k(g_\xi(x), g_\xi(y))$,　$g_\xi(x)$: NN

安定化手法

(1) Wasserstein GAN (Arjovsky et al ICML 2017)

- 勾配消失:
  JS-ダイバージェンス (ロジスティック回帰)の性質として generatorの微分は消失しやすい.
- Wasserstein-GAN は $||\varphi||_{Lip} \leq 1$ の制約によりそれを防ぐ.
- Weight clipping:
  実際は, $\|\eta\|_\infty \leq c$ の制約を持つニューラルネット$h_\eta(y)$ で代用.



Example 1 in Arjovsky et al ICML 2017

Original GANの目的関数（discriminator固定 $h_{\eta_*}(y)$）

$$\sum_i \log\sigma\left(h_{\eta_*}(Y_i)\right) + \sum_j \log\left(1-\sigma\left(h_{\eta_*}\left(f_\theta(Z_j)\right)\right)\right) =: F(\theta)$$

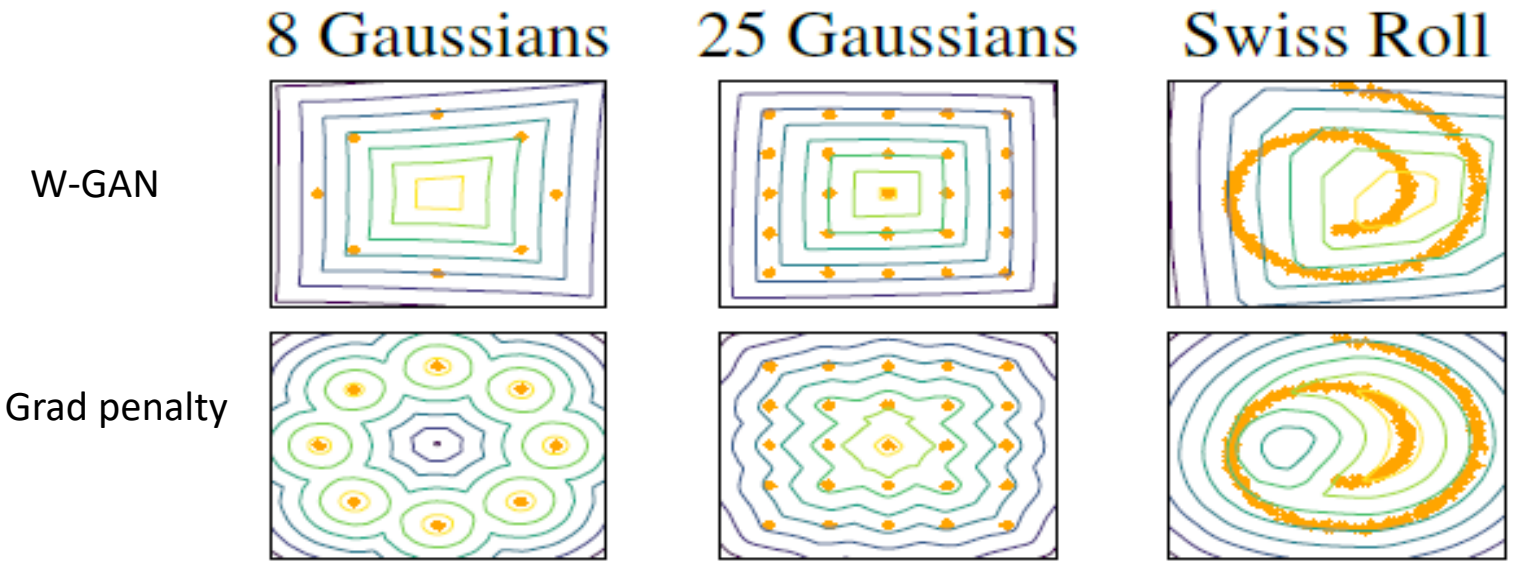
Generatorの勾配: $\nabla_\theta F(\theta)$

$$= -\sum_j \sigma\left(h_{\eta_*}\left(f_\theta(Z_j)\right)\right) \nabla_x h_{\eta_*}\left(f_\theta(Z_j)\right) \nabla_\theta f_\theta(Z_j).$$

Discriminatorの入力に関する微分

(2) Gradient penalty (Gulrajani et al. NIPS2017)

- 最適なcritic φ^*の勾配は，区分的線形に近い単純な形をとることが多い．



Gulrajani et al. NIPS2017

- Gradient penalty

$$\min_{\varphi} \int \varphi(x) dP_D(x) - \int \varphi(x) dP_\theta(x) + \lambda \int (\|\nabla \varphi(\hat{x})\|_2 - 1)^2 dQ(\hat{x})$$

$\|\nabla \varphi(\hat{x})\|_2 = 1$ となる正則化をおいて勾配消失を防ぐ

(3) Spectral normalization (Miyato et al ICLR2018)

第k層のdiscriminator: $\boldsymbol{y}_k = \psi_k(W_k\,\boldsymbol{y}_{k-1})$ $(k = 1, \ldots, L)$

活性化関数 ψ_k は1-Lipschitzと仮定（e.g. ReLU）.

重みを正規化

$$\widetilde{W} = \frac{W}{\sigma(W)}$$

$\sigma(W)$: W の最大特異値

→ Discriminator $h_\eta(x)$ は1-Lipschitzになる.

確率分布の学習としてのGAN

確率分布の学習:

符号付測度 $\mathcal{M}(\mathcal{X})$: ベクトル空間を考えたい

J: 符号付測度 $\mathcal{M}(\mathcal{X})$ 上定義された凸な損失関数

$$\min_{\theta} J(P_\theta)$$

例）

- オリジナルGAN: $J(P_\theta) = \mathrm{JSD}(P_D, P_\theta)$ （Jensen-Shannon divergence）
- Wasserstein GAN: $J(P_\theta) = W_1(P_D, P_\theta)$ （Wasserstein distance）
- GMMN: $J(P_\theta) = \left\| E_{P_\theta}[k(\cdot, X)] - E_{P_D}[k(\cdot, Y)] \right\|_{H_k}^2$

- 双対表現 (f-GAN, Nowozin et al 2016)

  定義域 $\mathcal{M}(\mathcal{X})$ (符号付測度). 双対空間 = $C(\mathcal{X})$ (連続関数の空間)

$$J(\mu) = \sup_{\varphi \in C(\mathcal{X})} \int \varphi(x) d\mu(x) - J^*(\varphi)$$

J^*: 凸共役関数

$$\inf_{P \in \mathcal{P}(X)} J(P) = \inf_{P \in \mathcal{P}(\mathcal{X})} \boxed{\sup_{\varphi \in C(\mathcal{X})} \int \varphi(x) dP(x) - J^*(\varphi)}$$

敵対的学習

Discriminator φ の学習は双対表現から来る

- 仮定 (最適識別性):

  任意のgenerator μ に対し, 最適なdiscriminator $\Phi_\mu(x)$ が得られると仮定する.

$$\boxed{J(\mu) = \int \Phi_\mu(x) d\mu(x) - J^*(\Phi_\mu)}$$

この仮定の下, μ の勾配学習の性質を調べる.

- 例

	Loss function	Optimal discriminator $\Phi_\mu(x)$
Original GAN	$D_{JS}(\mu\|\|\mu_0)$	$\frac{1}{2}\log\frac{\mu(x)}{\mu(x)+\mu_0(x)}$
Non-saturating GAN	$D_{KL}(\frac{1}{2}\mu+\frac{1}{2}\mu_0\|\|\mu_0)$	$\frac{1}{2}\log\frac{\mu(x)}{\mu(x)+\mu_0(x)}$
Wasserstein GAN	$W_1(\mu,\mu_0)$	$\arg\max_{f\in\mathrm{Lip}_1} E_\mu[f(Y)]-E_{\mu_0}[f(X)]$
GMMN, MMD-GAN	$\frac{1}{2}\mathrm{MMD}^2(\mu,\mu_0)$	$E_\mu[K(x,Y)]-E_{\mu_0}[K(x,X)]$

勾配学習の安定性

Def.（L-平滑性）

$L > 0$. 微分可能な関数 $F\colon \mathbf{R}^p \to \mathbf{R}$ が L-平滑 であるとは，任意のx, y に対して

$$\|\nabla F(x) - \nabla F(y)\|_2 \leq L\|x - y\|$$

が成り立つことをいう．∇F が L-Lipschitz であることと同値.

以下の有名な事実が理論のコア.

命題（安定性の十分条件）

$F\colon \mathbf{R}^p \to \mathbf{R}$ が L-smooth かつ下に有界とする． 点列 $(x_n)_{n=1}^{\infty}$ を

$$x_{n+1} := x_n - \frac{1}{L}\nabla F(x_n) \qquad \text{[勾配法．} 1/L\text{：学習係数]}$$

で定めるとき, $\|\nabla F(x_n)\| \to 0 \quad (n \to \infty)$.

- L-平滑性は，安定性の十分条件.
- 極限 x_∞ が存在すれば，それは安定点.

GAN学習の安定性

ω: $\mathcal{Z}$上の確率.　$f_\theta: \mathcal{Z} \to \mathcal{X}$ generatorの族.　P_θ: f_θ により ω:から誘導される確率分布 $P_\theta = f_{\theta\#}\omega$

定理1 (GANの安定性. Chu, Minami, Fukumizu. ICLR 2020)

$J: \mathcal{M}(\mathcal{X}) \to \overline{\mathbf{R}} := \mathbf{R} \cup \{+\infty\}$ 凸な損失関数.　$\Phi_\mu: \mathcal{X} \to R$ 最適な識別関数

仮定

(D1) $x \mapsto \Phi_\mu(x)$: α-Lipschitz $\forall\mu$

(D2) $x \mapsto \nabla_x \Phi_\mu(x)$: β_1-Lipschitz $\forall\mu$

(D3) $\mu \mapsto \nabla_x \Phi_\mu(x)$: 1-Wasserstein距離に関して β_2-Lipschitz $\forall x$

(D1-D3) がgeneratorの学習にどのように影響するかを見る.

最適なdiscriminatorに関する条件

(G1) $\theta \mapsto f_\theta(z)$: 期待値に関して A-Lipschitz.　i.e. $E_{z\sim\omega}[\|f_{\theta_1}(z) - f_{\theta_2}(z)\|] \le A\|\theta_1 - \theta_2\|$

(G2) $\theta \mapsto D_\theta f_\theta(z)$: 期待値に関して B-Lipschitz,

i.e. $E_{z\sim\omega}[\|D_\theta f_{\theta_1}(z) - D_\theta f_{\theta_2}(z)\|] \le B\|\theta_1 - \theta_2\|$

generator familyに関する条件

このとき, $\theta \mapsto J(P_\theta)$ は L-平滑.　$(L = \alpha B + A^2(\beta_1 + \beta_2))$ → 収束（勾配→0）

問題1: どのGANが(D1)-(D3)を満たすか?

問題2: 一般の損失関数に対し，(D1)-(D3)をどのように保証するか?

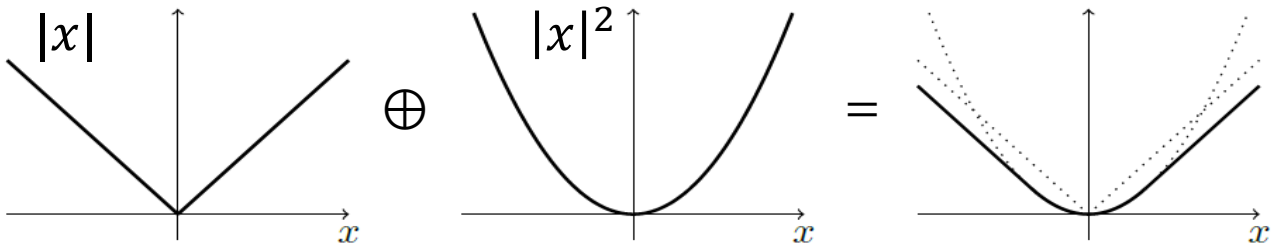
Q1に対する回答:

	(D1)	(D2)	(D3)
Original GAN	X	X	X
NS-GAN	X	X	X
W-GAN	✓	X	?
MMD-GAN	✓*	✓	✓

(D1) $x \mapsto \Phi_\mu(x)$: α-Lipschitz
(D2) $x \mapsto \nabla_x \Phi_\mu(x)$: β_1-Lipschitz
(D3) $\mu \mapsto \nabla_x \Phi_\mu(x)$: 1-Wasserstein距離に関して β_2-Lipschitz

* Lipschitz定数αが次元に依存する.

Inf-Convolution



Def. Inf-convolution

V: ベクトル空間. $F, G: V \to \overline{\mathbf{R}}$.

F と G の infimum convolution：

$$(F \oplus G)(x) := \inf_{y \in V} F(y) + G(x-y)$$

命題.

- [可換律] $F \oplus G = G \oplus F$
- [結合律] $(F \oplus G) \oplus H = F \oplus (G \oplus H)$
- Inf-conv は最小点を変えない.

例: Moreau envelop $F^\beta(x) = F \oplus \frac{1}{2\beta}\|\cdot\|_2^2$ 微分可能で β-平滑

命題. $J, R: \mathbf{R}^d \to \overline{\mathbf{R}}$: proper, lower semi continuous, $\min J$ が存在し $\min J > -\infty$. $R(0) = 0$, 狭義増加関数 $\psi: [0,\infty) \to \mathbf{R}$が あって $R(x) \geq \psi(\|x\|_2)$ と仮定. このとき,

$\min J \oplus R = \min J$ かつ

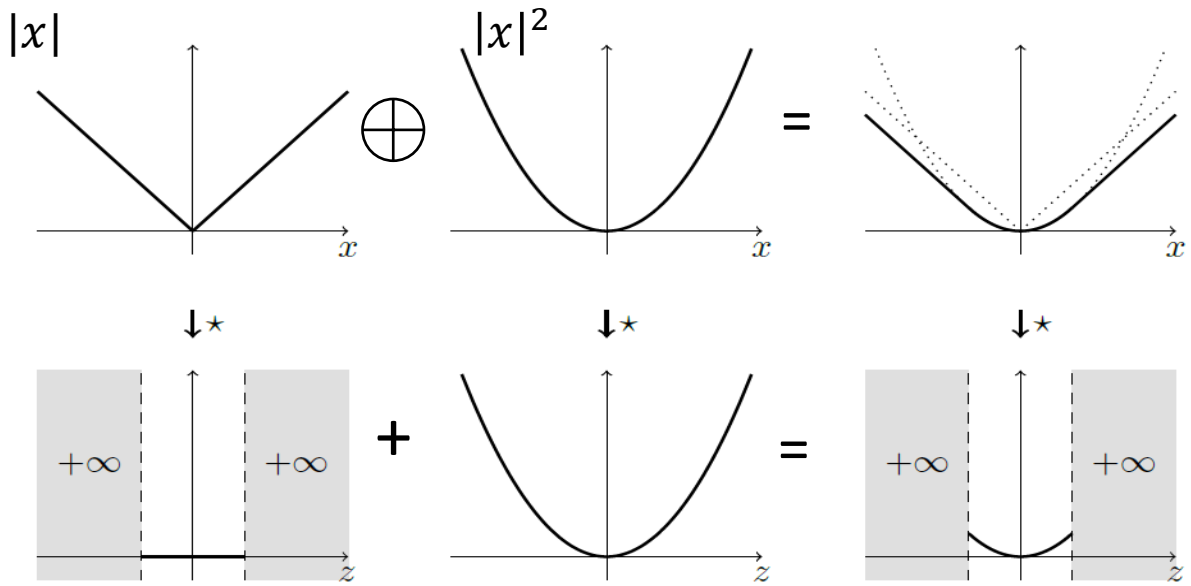
$\arg\min J \oplus R = \arg\min J$.

Inf-convolutionの共役

復習：凸共役　凸関数 J に対し　$J^*(z) := \sup_x \langle x, z\rangle - J(x)$.

命題.　凸関数 J と R に対し inf-convolution $J \oplus R$ は凸であり，共役に関して

$$(J \oplus R)^*(z) = J^*(z) + R^*(z)$$



$$\begin{aligned}
(J \oplus R)^*(z) &= \sup_x \langle x, z\rangle - J \oplus R(x) \\
&= \sup_x \langle x, z\rangle - \inf_y \{J(x-y) + R(y)\} \\
&= \sup_x \sup_y \{\langle x, z\rangle - J(x-y) - R(y)\} \\
&= \sup_y \sup_x \{\langle x-y, z\rangle - J(x-y) \\
&\qquad\qquad + \langle y, z\rangle - R(y)\} \\
&= \sup_y J^*(z) + \langle y, z\rangle - R(y) \\
&= J^*(z) + R^*(z)
\end{aligned}$$

Generatorの目的関数の改変

目的：Discriminatorに対する条件(D1)-(D3)を満たすようにしたい

提案法：Generatorの目的関数を変更

$$\tilde{J}(\mu) \coloneqq (J \oplus R_1 \oplus R_2 \oplus R_3)(\mu)$$

R_1, R_2, R_3 は後で定める

Discriminatorの目的関数：

$$\inf_{\mu} \tilde{J}(\mu) = \inf_{\mu} \sup_{h \in C(X)} \{ \int h(x) d\mu(x) - \underbrace{\{J^*(h) + R_1^*(h) + R_2^*(h) + (h)\}}_{\tilde{J}^*(\mu)}$$

$$= \inf_{\mu} \sup_{h \in C(X)} \mathcal{L}_J(\mu, h) - R_1^*(h) - R_2^*(h) - R_3^*(h)$$

元の損失関数 J に関するGANのminmax

Discriminatorに対する正則化項

- 正則化の具体形

$R(\mu)$	$R^*(h)$	Purpose	GAN Techniques
$\alpha\|\mu\|_{KR}$	$I\{h \mid \alpha$–Lipschitz$\}$	(D1)	→ Spectral norm / W-GAN
$\beta_1\|\mu\|_{\mathcal{S}_1}$	$I\{h \mid \beta_1$–smooth$\}$	(D2)	→ Spectral norm
$\frac{\beta_2}{2}\left\|\mathcal{E}_\mu\right\|_{H_k}^2$	$\frac{1}{2\beta_2}\|h\|_{H_k}^2$ (a)	(D3)	→ Gradient penalty

- $\mathcal{S}_1 \coloneqq \{f: \mathcal{X} \to \mathbf{R} \mid f$:1-smooth$\}$
- Gaussカーネルの場合 (a) から $\mu \mapsto \nabla_x \Phi_\mu(x)$ の β_2-Lipschitzが導かれる
- $\mathcal{E}_\mu \coloneqq \int k(\cdot, x) d\mu(x). \quad k = \exp(-\pi\|x-y\|^2)$:正定値カーネル

- 改変された目的関数 $\tilde{J}(\mu)$ によって安定性の十分条件(D1-D3) が満たされる.
- よく使われる安定化技法は, (D1-D3)を保証する正則化項と関係する.
- 安定化技法は合わせて用いるとよいことが示唆される.

最適点の不変性

定理 (Chu, Minami, F. ICLR2020)

凸な損失関数 $J: \mathcal{M}(\mathcal{X}) \to \overline{\mathbf{R}}$ は唯一の最小点 μ_0 を持ち，$J(\mu_0) = 0$ かつ $J(\mu) \geq c\|\mathcal{E}_\mu - \mathcal{E}_{\mu_0}\|_{H_k}^2$ $(c > 0)$ と仮定する．このとき，

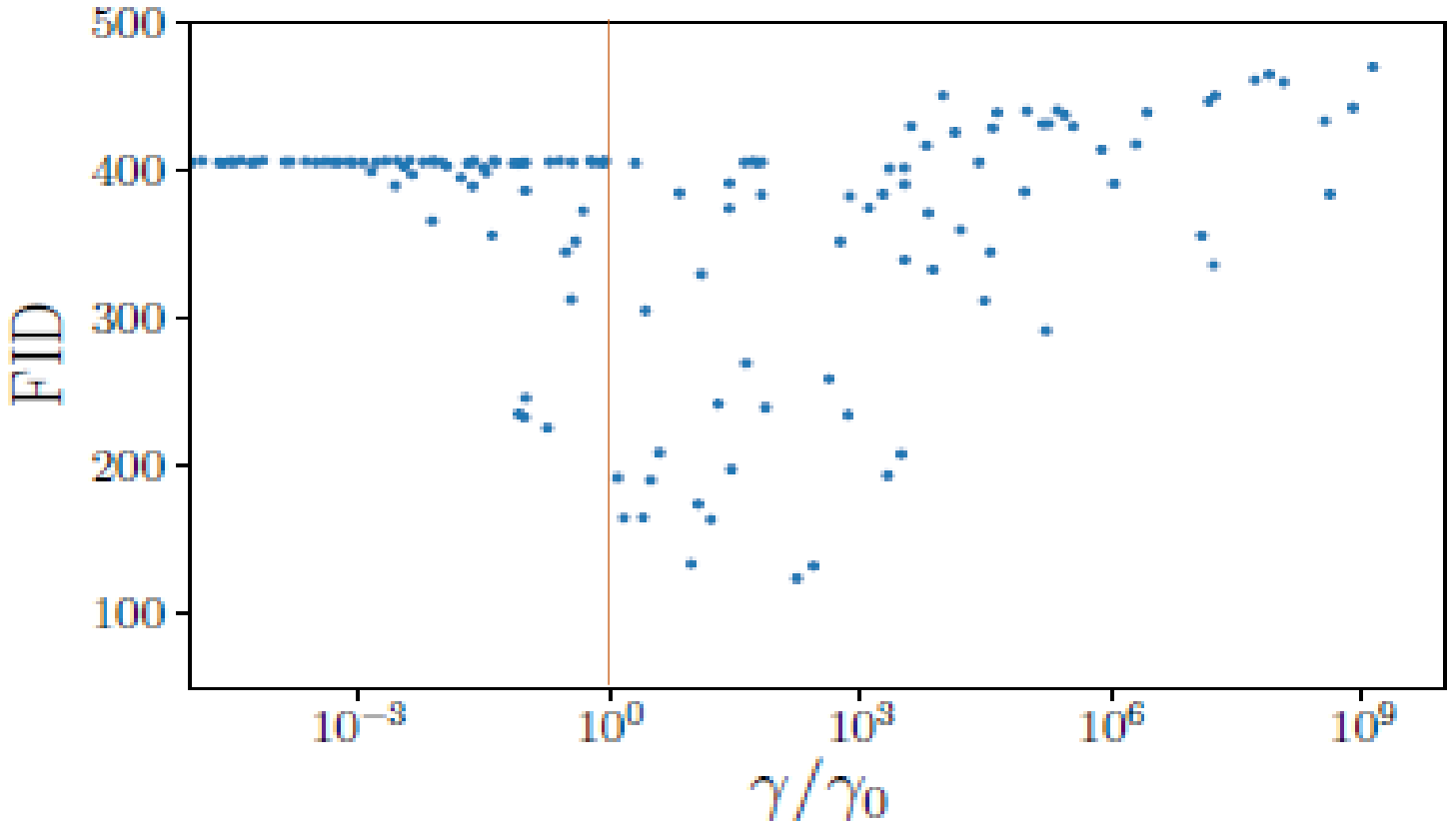
$\tilde{J} = J \oplus R_1 \oplus R_2 \oplus R_3$ は唯一の最小点 μ_0 において最小値 0 をとる．

- $J(\mu) \geq c\|\mathcal{E}_\mu - \mathcal{E}_{\mu_0}\|_{H_k}^2$ の条件は，JSD (オリジナルGAN), KR norm (W-GAN), MMDの2乗 (GMMN)で成立する．
- ３つの正則化によって，GANの最小点を変更せずに，安定化させることができる．既存の安定化手法の（一定の）理論的正当化．

数値実験

- Lipschitz 定数 L は，収束を保証する学習係数 γ を与える.
- 実験：学習係数 γ とFréchet Inception Distance (FID) の関係を調べる
  - CIFAR-10.
  - 単純なgenerator (N個の点からランダムに選択), (G1) (G2)を満たす.
  - Discriminator： 7層 CNN.
  - 損失関数: $\ell(\mu) = I(\mu = P_D)$
  - 目的関数: $\min_{\mu} \max_{\varphi} E_{\mu}[\varphi(X)] - E_{P_D}[\varphi(X')] - \frac{1}{2\beta_2}\|\varphi\|_{H_k}^2$

    $\|\varphi\|_{H_k}^2$ is approximated by an expansion of *k*
  - Spectral normalization： $\|\varphi\|_{Lip} \leq \alpha$.
  - 勾配法で100000 ステップ学習
  - 収束保証を与える学習係数 $\gamma_0 = \frac{1}{7\alpha+\beta_2}$.



Better

True learning rate / theoretical value

FIDs は $\gamma/\gamma_0 \approx 1\sim10^3$ 付近で最小

まとめ

- 深層生成モデルを用いた統計的推論
  - 高次元データのよいサンプルが得られる
  - 統計的推論の要素として組むことが可能 → Simulation-based inference.
    - 高次元サンプラーとして機能するか？
    - 時系列の場合？

- 深層生成モデルの理論
  - GANの学習に関する理論解析
    - 安定性の解析，安定化手法の意義

JST CREST

- JST CREST「数理知能表現による深層構造学習モデルの革新」
  代表者・福水健次.
  主たる共同研究者：鈴木大慈，原田達也
  2020年11月 – 2026年3月
  （数理的情報活用基盤，上田修功総括）

特任研究員（特任助教）募集中！