深層生成モデルによる統計的推論

福水 健次
統計数理研究所　統計的機械学習研究センター長

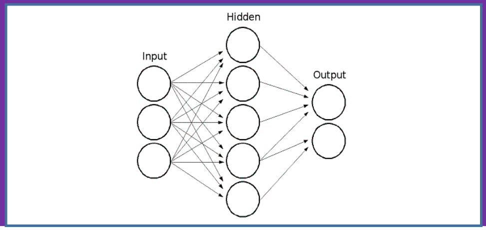
統計数理研究所 創立75周年記念チュートリアル講演
2019年6月5日

（極私的）統計的機械学習の歴史

1980　1990　2000　2010　2020

1980-90 第２次ニューロブーム

1986 PDP本



高い性能
学習コスト高い（計算時間，チューニング）
理論保証困難

1990 - **統計的機械学習**
SVM／カーネル法
グラフィカルモデリング
ベイズ推論（近似手法）
etc

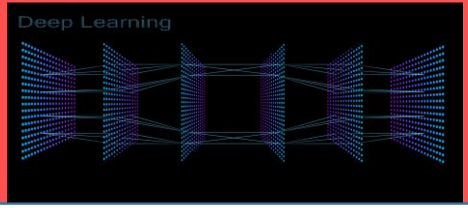
理論保証
凸最適化／近似
大規模・高次元に困難

2012 ImageNet 物体認識コンペティション
誤認識率 15.3%，２位 26.2%

ビッグデータの時代

2007(?) NIPS
非公式Mini-Symposium.
Deep Belief Network

2006 - **深層学習**

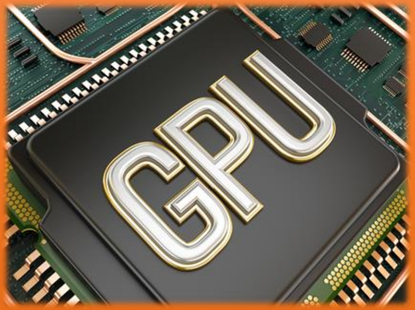


大量データが利用できる
高い性能
学習コスト高い（計算時間，チューニング）
理論保証困難

深層学習の隆盛

大量データ

計算能力



高い性能

プラットフォーム

TensorFlow

Chainer

PYTORCH

theano

講演の概要

1. はじめに

2. 深層生成モデル

3. 深層生成モデルによるBayes推論

4. おわりに

GAN: 敵対的生成ネットワーク

Generative Adversarial Nets. (Goodfellow et al. NIPS2014)

- 生成モデルの深層学習
- 「敵対的学習」という枠組み
- 良好な画像の生成に成功

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)

State-of-the-Art

Cycle-GAN (Zhu et al ICCV2017)

https://www.youtube.com/watch?v=9reHvktowLY

- Everybody Dance Now (Chan et al 2018)

YouTube: Everybody Dance Now.
https://youtu.be/PCBTZh41Ris

生成モデル

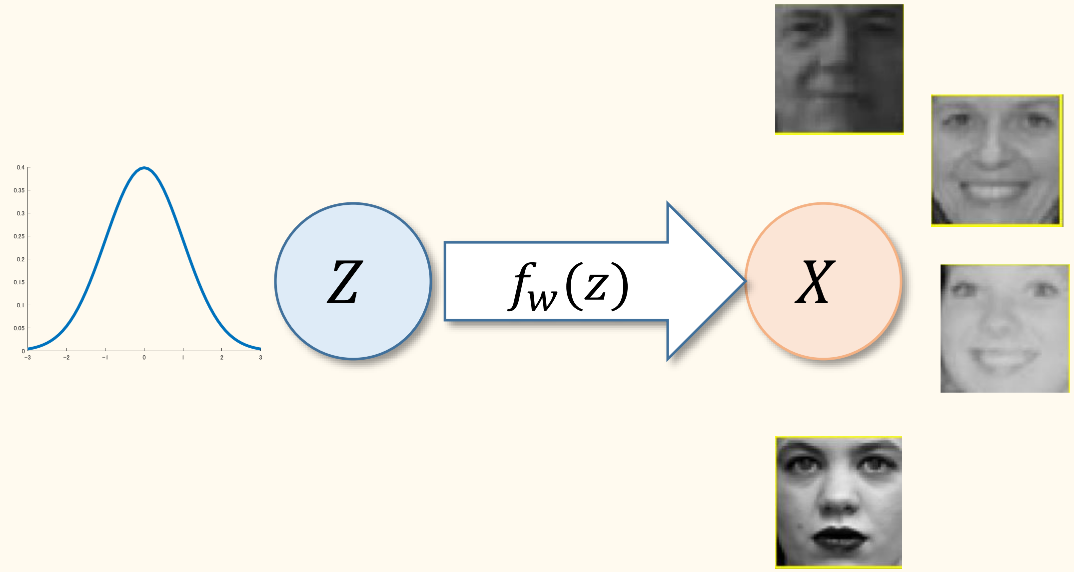
- 生成モデル：　データを生成する確率分布のモデル

  Graphical model
  Mixture model
  カーネル密度推定　etc….

- 深層生成モデル

  $X = f_w(Z)$

  Z: 簡単な確率分布 e.g. $N(0, I_d)$
  $f_w(z)$： 深層ニューラルネット

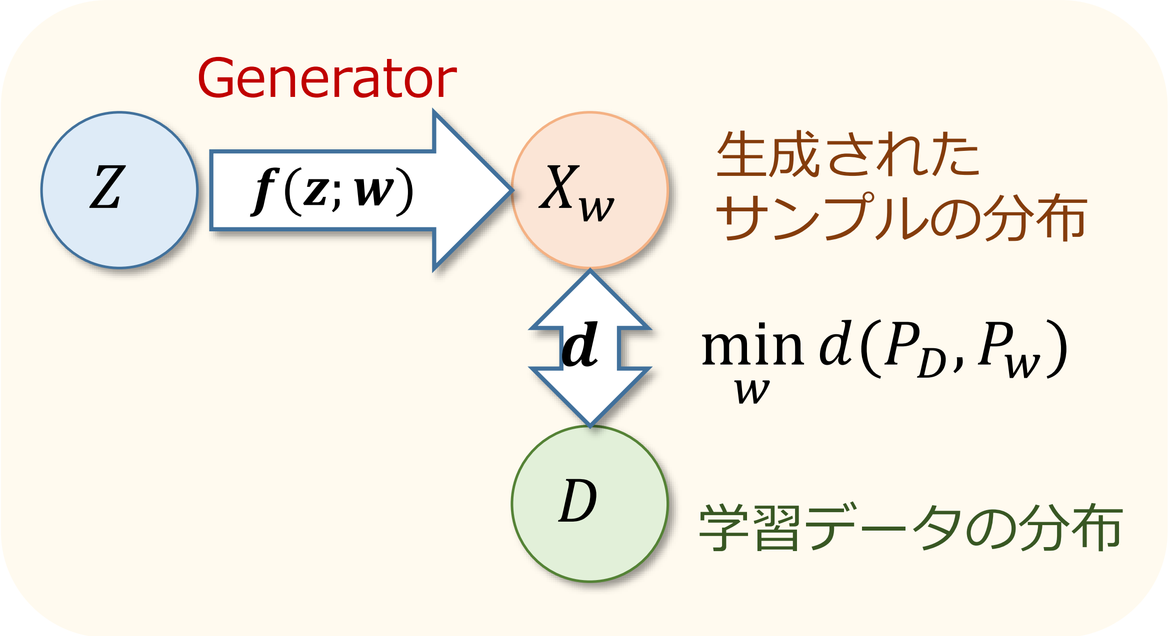


生成モデルの学習

- 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：敵対的学習

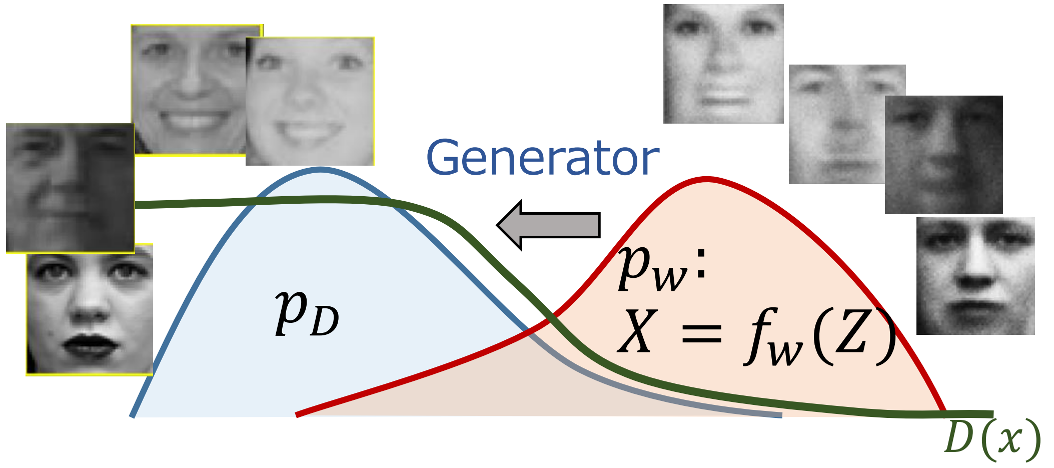
p_D：データ（真）の密度関数
p_w：生成モデルの密度関数　$X = f_w(Z)$

- 分布の距離：Jensen-Shannon divergence

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

- 生成ネット：分布間距離の最小化

$$\min_{f_w} JS(p_D, p_w)$$



- 識別ネット：JS-距離は計算困難 → 識別問題に還元

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

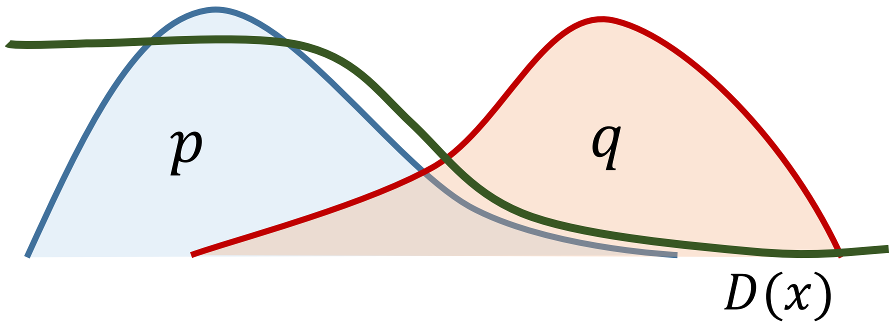
− Logistic loss

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

- GAN：minmaxによる学習

  - Discriminator: ロジスティック判別

$$\max_D E_{X\sim p}[\log D(X)] - E_{X\sim q}[\log(1 - D(X))]$$

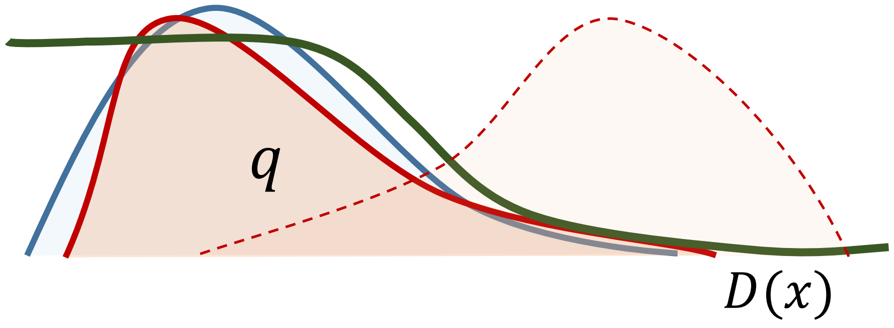


$D(x)$ の値: p で１, q で０が有利

  - Generator: 判別できなくしたい

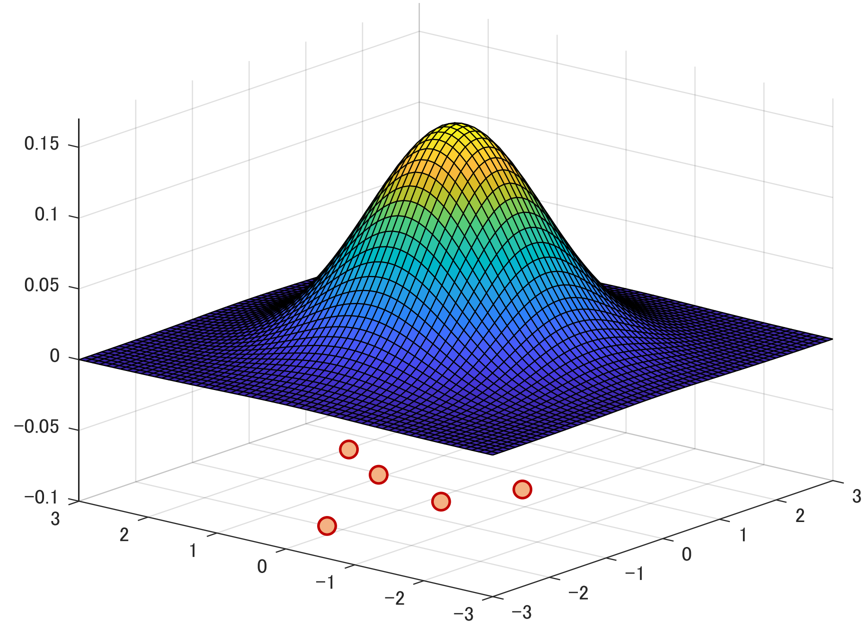
$$\min_{q:X=g(Z)} \max_D E_{X\sim p}[\log D(X)] - E_{X\sim q}[\log(1 - D(X))]$$

$D(x) = 1$ の領域で確率大が有利



サンプリング

- 高次元で高性能なサンプラーができた！



DNN

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

- 画像、映像の生成

  アニメーション、実写、アート、etc

Edmond de Belamy (Oblivious)

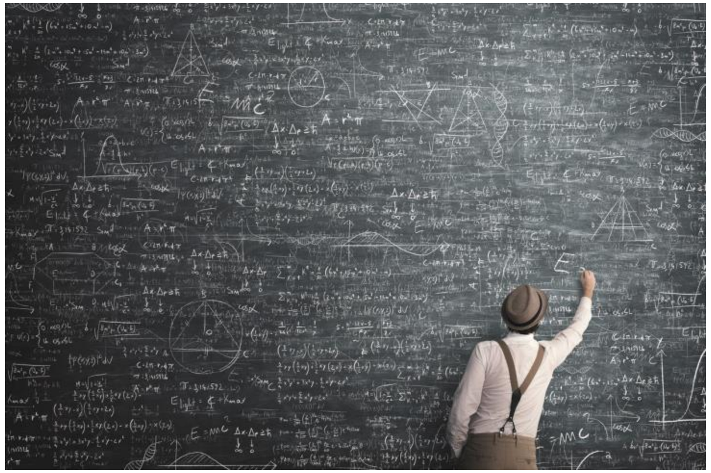
Cycle GAN (Zhu et al 2017)

Yanghua Jin et al (Comiket 92, 2017)

Crypko
by Preferred Networks

- サンプリングによる統計的推論

深層生成モデルによるBayes推論



Beyes推論

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

$\pi(\theta)$: 事前確率（prior）
$p(y|\theta)$: 尤度 （likelihood）
y: 観測データ

Thomas Bayes (1701-1761)

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

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

→ GANによるサンプリングを使おう！

(Tran et al NIPS 2017; Yang et al NeurIPS2018)

生成モデルによるベイズ推論

- 同時分布の比較

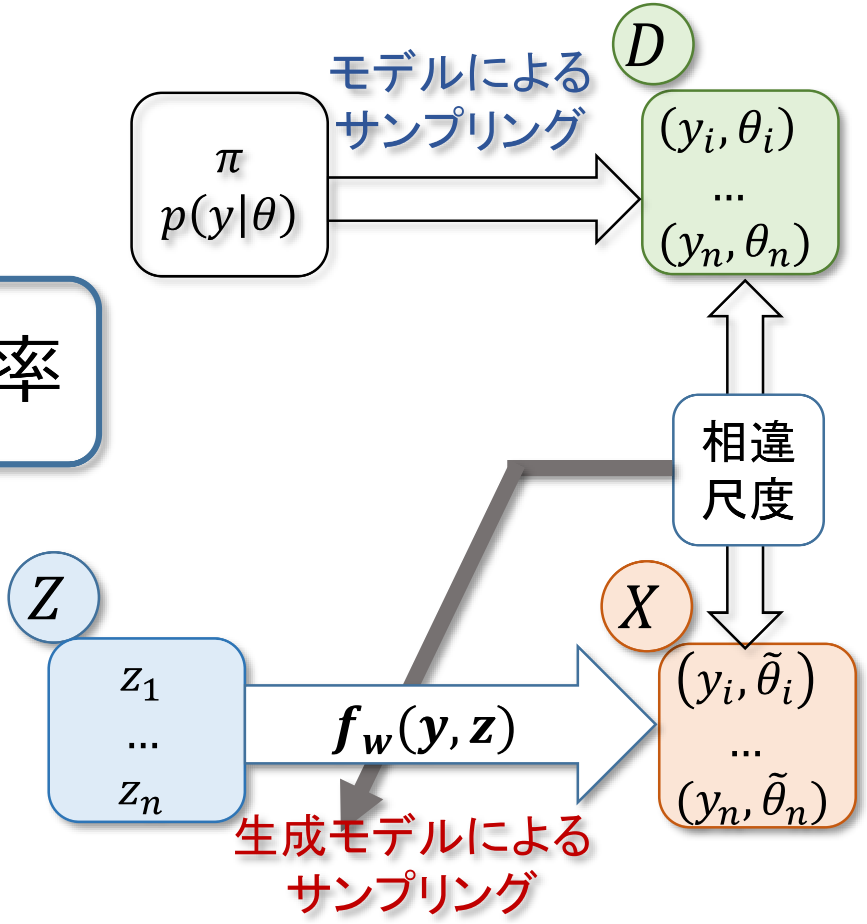
$$p(\theta, y) = p(y|\theta)\pi(\theta)$$
$$q(\theta, y) := q(\theta|y)p(y),\ p(y) = \int p(y|\theta)\pi(\theta)d\theta$$

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

サンプリング → GAN的学習

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

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



- 分布の違いをどう測るか？

1. KL-divergence (Yang et al NeurIPS2018)

$$\int q(y,\theta) \log \frac{\boxed{p(y,\theta)}}{\boxed{q(y,\theta)}} d\theta$$

学習データ

生成データ

2. 変分Bayesからの導出（Tran et al NIPS 2017）

変分ベイズとしての見方 Hierarchical Implicit Models and Likelihood-Free Variational Inference (Tran, Ranganath, Blei, NIPS2017)

- 変分Bayes：事後確率の有力な近似手法

$$\log p(y) = \int p(y|\theta)\pi(\theta)d\theta$$
$$\geq -KL\big(q(\theta|y)||\pi(\theta)\big) + E_{q(\theta|y)}[\log p(y|\theta)]$$

Evidence lower bound (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 によって事後確率を近似する

- ELBO = $\int q(\theta|y) \log \frac{p(y,\theta)}{q(y,\theta)} d\theta + \log p(y)$ → 生成モデルの目的関数！

＊深層生成モデルの尤度関数は計算不能　（*c.f.* Autoregressive flow, Glow）

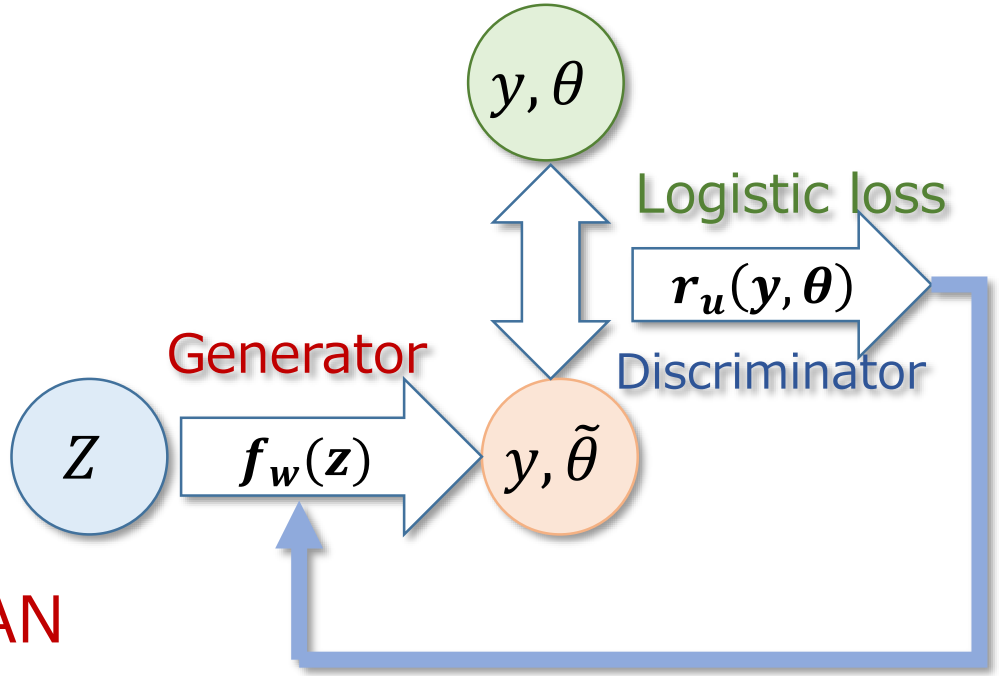
- $\log \dfrac{p(\theta,y)}{q(\theta,y)}$ の推定 → 　Logistic判別により可能　（≒ GANの識別機）

<u>命題</u>　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　$r_{opt}(y,\theta) = \log \dfrac{p(y,\theta)}{q(y,\theta)}$

$r(y,\theta)$ をDNNで構成する



変分Bayes ≒ GAN

変分ベイズとGAN

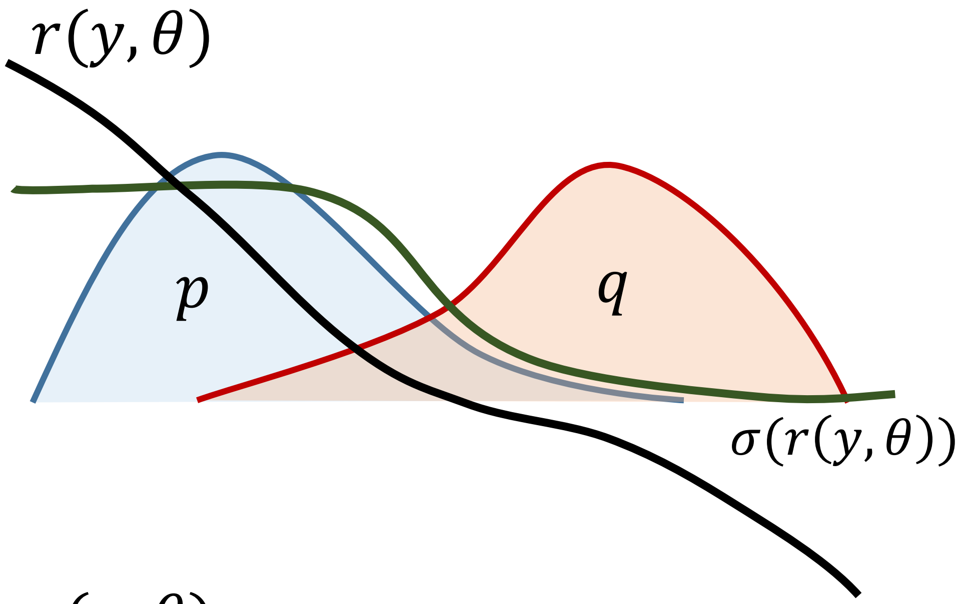
- 変分Bayes + 深層生成モデル
  - Discriminator: ロジスティック判別

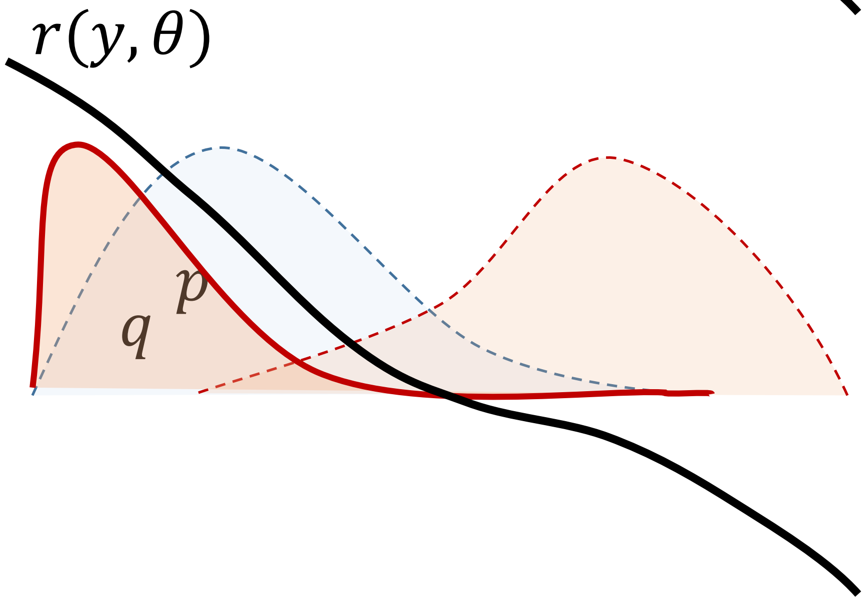
$$\max_{r(y,\theta)} E_{(y,\theta)\sim p}[\log \sigma(r(y,\theta))] - E_{(y,\tilde{\theta})\sim q}\left[\log\left(1-\sigma\left(r(y,\tilde{\theta})\right)\right)\right]$$

r の値：p で ＋，q で － が有利

  - Generator: 変分Bayes（max ELBO）

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

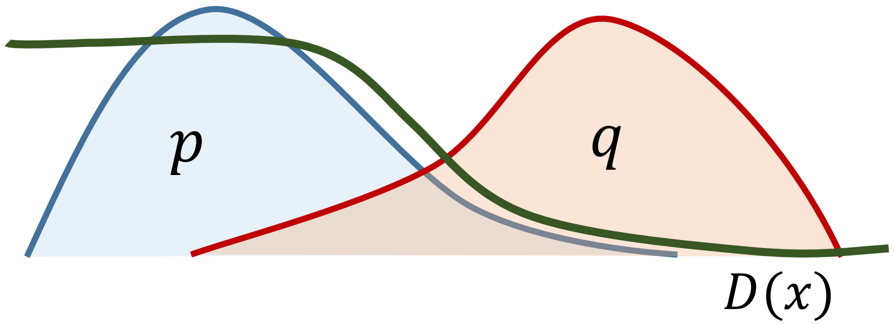




- GAN（振り返り）
  - Discriminator: ロジスティック判別

    $$\max_{D} E_{X\sim p}[\log D(X)] - E_{X\sim q}[\log(1 - D(X))]$$

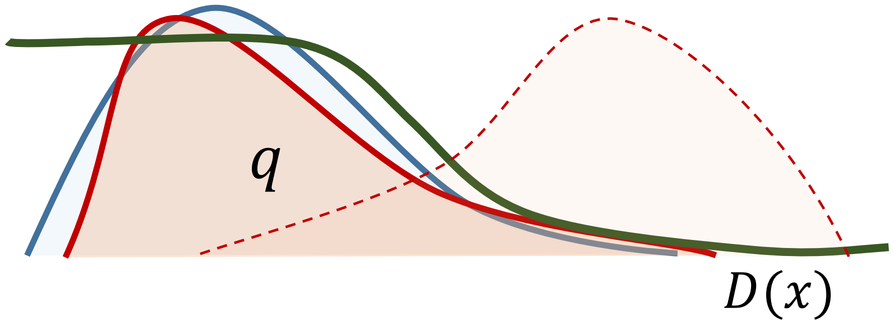
    $D(x)$ の値： p で 1，q で 0 が有利



  - Generator: Discriminatorの与える値が目的関数

    $$\min_{q:X=f(Z)} \max_{D} E_{X\sim p}[\log D(X)] - E_{X\sim q}[\log(1 - D(X))]$$

    $D(x) = 1$ の領域で確率大が有利



近似Bayes推論（ABC）における比較

Lotka-Volterra Predator-Prey Simulator (確率微分方程式)

See Tran et al NIPS2017

おわりに： 興味ある方向性

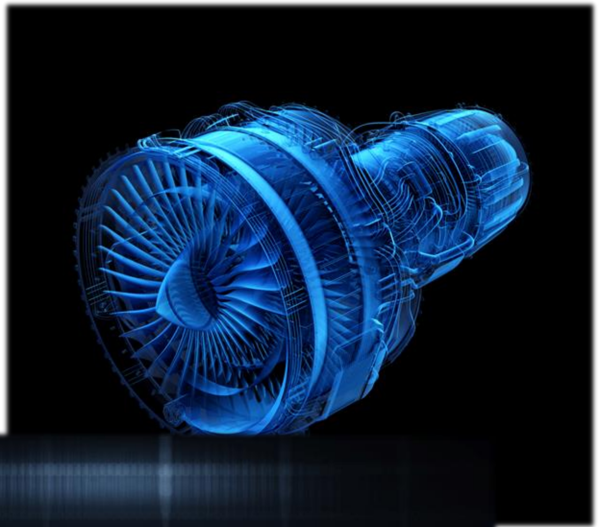
- GANによるサンプリング＋Bayes推論

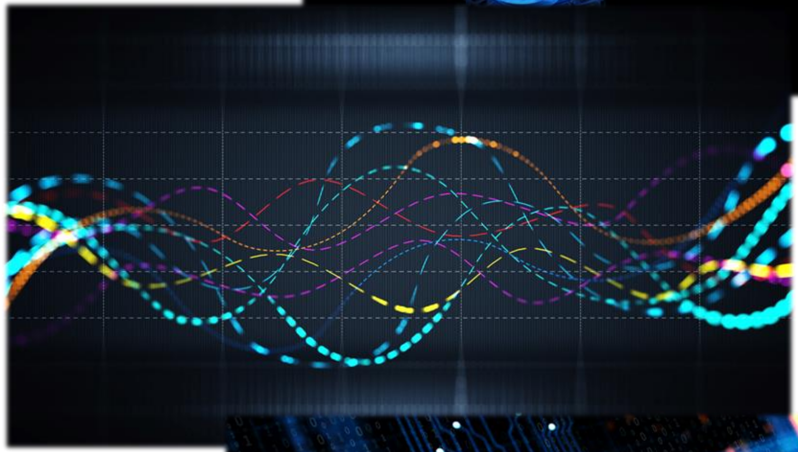
  シミュレーター

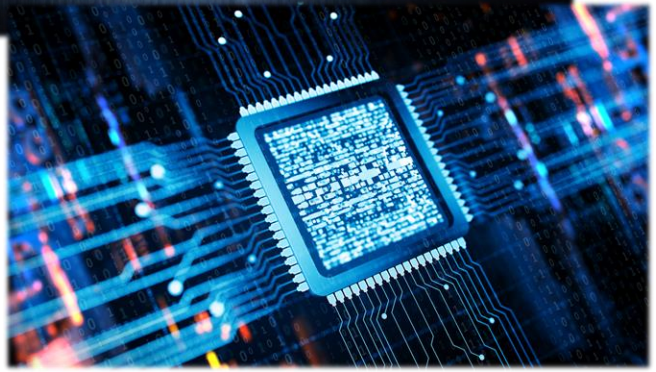
  時系列解析（粒子フィルター）

  etc

- 非線形時系列モデル
  - センサーデータなど、多くのデータは時系列
  - HMM/時空間モデルを超えた時系列解析
  - NEDOプロジェクト「確率深層コンピューティング」
    （理研・泰地代表，アルゴリズム：樋口，福水，今泉）







- より計算量の少ないモデルへ
  - 社会の隅々までのAI化には必要
  - よりよい関数族、関数空間？
  - 可変基底 vs 固定基底を超えられるか？

（参考）学習時間　CycleGANの顔画像：
Tesla V100 GPU 8個で1週間

- 学習ダイナミクスと汎化誤差
  - 確率降下法の理解
  - 汎化誤差とFlat minimaの関係?
    (Fukumizu et al 2019, in preparation)