FY2024/2024年度

High-Dimensional Causal Analysis Team

Masaaki Imaizumi

高次元因果解析チーム　今泉 允聡

RIKEN Center for Advanced Intelligence Project

チーム概要：

大自由度・高次元な
データ構造を数学的に理解

データ構造（因果構造など）
を解析する理論の構築

基礎的研究
- 高次元統計学
- 過剰パラメータ理論
- 深層学習の数理

応用的研究
- 因果推論
- 意志決定論
- 大規模データ解析

基礎的研究：高次元統計・深層モデルの解析

問題：複雑なランダム高次元行列の期待値推定

→従属性や確率的な不安定性の元での平均収束

$M_1, \dots, M_n \in \mathbb{R}^{p \times p}$: ランダム行列の列

列の(弱)従属性

$$\sup_{g \in \mathrm{Lip}} |\mathbb{E}[g(M_{\ell+1}, \dots, M_n) | \mathcal{F}_\ell] - \mathbb{E}[g(M_{\ell+1}, \dots, M_n)]| \leq \Gamma$$

$\mathcal{F}_\ell = \sigma(X_1, \dots, X_\ell)$

ランダム行列の裾確率

$$G(t) = \max_\ell \int_t^\infty \mathbb{P}(\|M_\ell\| > u) \mathrm{d}u$$

アプローチ：変分不等式を用いた摂動評価

→高次元性・従属性を制御

$$\mathbb{E}[\exp(\lambda h(M) - \lambda \mathbb{E}[h(M)])] \leq C \exp\left(\frac{\lambda^2 L^2 \sum_{\ell=1}^n (\kappa + \Gamma)^2}{n^2}\right)$$

結果：高次元性に依存しない経験平均の収束

定理（経験平均の次元フリー収束）

確率 $1 - p(t, \tau)$ 以上で以下が成立：

$$\underbrace{\left\| \frac{1}{n} \sum_{\ell=1}^n M_\ell - \mathbb{E}[M_\ell] \right\|}_{\text{経験平均の誤差}} = O\underbrace{\left(\|\Sigma\|(\tau + \Gamma) \sqrt{\frac{\mathbf{r}(\Sigma) + t}{n}} + G(\tau) \right)}_{\text{次元}p\text{に依存しない誤差}}$$

n: 観測行列の数
$\tau, t > 0$: ハイパラ
$\mathbf{r}(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{\|M\|}$: 実効ランク
$p(t, \tau)$: ある関数

問題：可逆関数を可逆関数で推定

可逆関数を可逆なまま推定する複雑性の定量化

$X \in [0,1]^2$: 入力, $Y \in [0,1]^2$: 出力, ε: ノイズ, f_*: 可逆関数

$$Y = f_*(X) + \varepsilon.$$

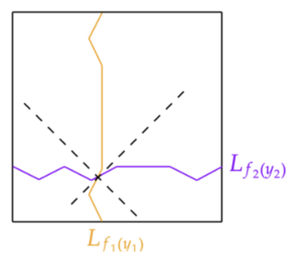
自身が可逆な推定量$\hat{f}(\cdot)$を構成

応用例：
Normalizing Flowなど
可逆性による生成モデル

アプローチ：ミニマックス可逆リスクレートの解析

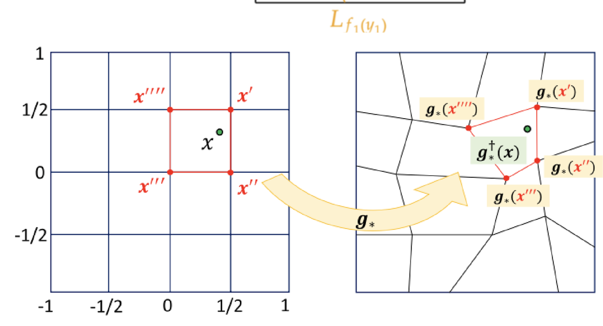
可逆リスク $\mathrm{R}(\hat{f}, f_*) = E\left[\|\hat{f} - f_*\|^2 + \|\hat{f}^{-1} - f_*^{-1}\|^2\right]$

下界：レベルセットの共通部分で
最悪単調関数を構成→誤差導出



上界：可逆推定量の具体的構成
- 入出力空間の格子間写像
- ”ねじれ“の漸近的解消

格子の消失を極限で解消



結果：レート導出と複雑性評価

定理1（可逆リスクレートの上下界）

$$\min_{\hat{f}} \max_{f_*} \mathrm{R}(\hat{f}, f_*) = \tilde{\Theta}(n^{-2/(2+d)})$$

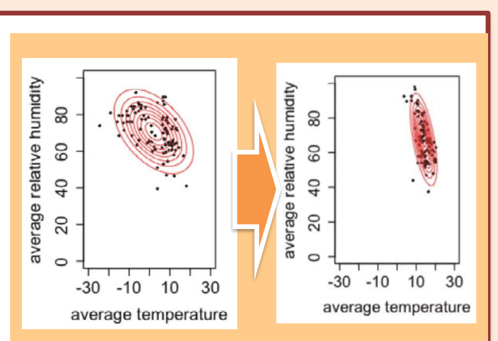
可逆性のないレートと同一
→可逆関数は推定の複雑性を
増加させない

応用的研究：機械学習の諸問題への貢献

問題：ガウス分布間回帰

$(\mu_i, \nu_i), i = 1, \dots, n$: 入出力ガウス分布の組
（μ_i, ν_i のそれぞれが$\mathbb{R}^d$上の確率分布）

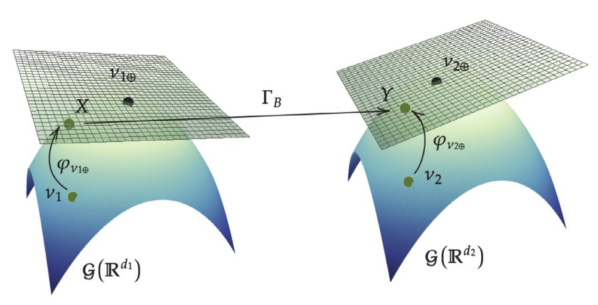
目的：μ_iからν_iを予測する回帰モデル構築



例：気候分布回帰

アプローチ：接空間上での線形回帰モデル

1. 分布空間(Wasserstein空間)の
   接空間に分布を転移
   ガウス分布により等長性が保障
2. 接空間(線型空間)で
   共分散行列間の線形回帰
   テンソルパラメータを用いた回帰モデリング



分布空間(曲面)から接空間(平面)へ

結果：高速・理論付き回帰モデルの提案

定理（推定誤差評価）

$$\underbrace{R(T)}_{\text{推定誤差}} = \tilde{O}\underbrace{\left(\sqrt{K(2d)}/\sqrt{n}\right)}_{\text{次元}p\text{に依存しない誤差}}$$

n: データ数
$d \geq 2$: 分布の定義域の次元
K: テンソルパラメータのランク

多次元分布間回帰の成功例

トピック：敵対的生成ネットワーク(GAN)

f: 識別器, θ: 生成器, μ_θ: 生成した分布
$\mathcal{V}(f, \mu_\theta)$: 識別目的関数, $\mathcal{J}(\theta; f)$: 生成目的関数

2プレイヤー最適化

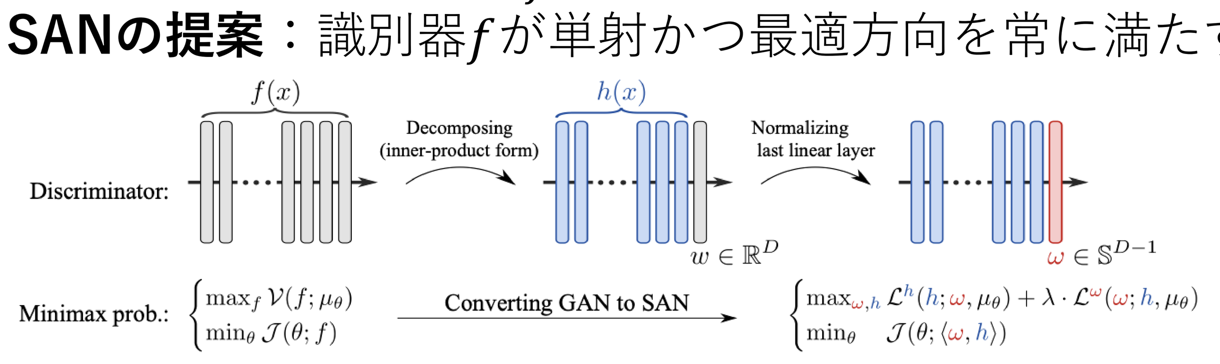
$$\max_f \mathcal{V}(f, \mu_\theta), \min_\theta \mathcal{J}(\theta; f)$$

最適化の不安定性を解決する
新しい目的関数の提案

アプローチ：識別器を一様に距離化可能

背景：最適値以外の識別器fのもとで識別能力が低下
→ 学習途中の識別器fでも十分に識別可（距離化）な設計へ

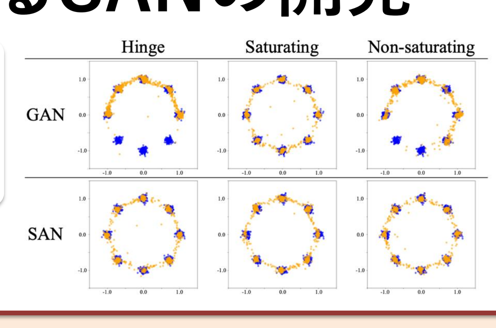
SANの提案：識別器fが単射かつ最適方向を常に満たす目的関数



結果：高精度GANであるSANの開発

定理（距離化可能定理）

SANの識別器は距離化可能



モード崩壊の防止

Rank	Model	FID ↓
1	MDT-XL/2	1.79
2	DiT-XL/2-G++	1.83
3	ViT-XL	2.06
4	StyleSAN-XL	2.14
5	DiT-XL/2	2.27
6	StyleGAN-XL	2.30

ImageNet生成で
GANの中でSOTA

Selected Papers

S.Nakakita, P.Alquier, M. Imaizumi (2024), "Dimension-free Bounds for Sum of Dependent Matrices and Operators with Heavy-Tailed Distribution", Electronic Journal of Statistics, 18(1).
A. Okuno, M.Imaizumi (2024), "Minimax Analysis for Inverse Risk in Nonparametric Planer Invertible Regression", Electronic Journal of Statistics, 18(1).
R.Okano, M.Imaizumi (2024), "Distribution-on-Distribution Regression with Wasserstein Metric: Multivariate Gaussian Case", Journal of Multivariate Analysis, 203.
Y.Takida, M.Imaizumi, et al. (2024), "SAN: Inducing Metrizability of GAN with Discriminative Normalized Linear Layer", International Conference on Learning Representations.