# Outlier-Robust Neural Network Training: Efficient Optimization of Transformed Trimmed Loss with Variation Regularization (Submitted.)

奥野彰文[1,2,3], 柳下翔太郎[1]

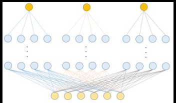
[1]統計数理研究所, [2]理研AIP, [3]理研CBS

# 自己紹介

- 研究遍歴：数理統計 ⇒ 統計的機械学習 ⇒ 統計学一般
- 学部–博士の指導教員：下平英寿教授.
- 博士 (情報学，京都大学，2020年9月)

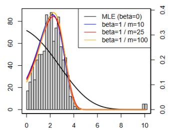
統計数理研究所（立川市，東京）

## Computation (Machine Learning)

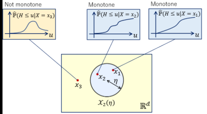


WAIC + NN
Okuno and Yano (JCGS2023)

$$\mathbb{E}_T\left(\left\|\frac{\partial}{\partial\theta} d_\beta(\hat{Q}, P_{\theta(T)})\right\|_2^2\right) \xrightarrow{\text{in prob.}} 0, \quad (T \to \infty).$$

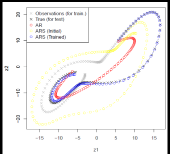



Okuno (2307.05251)

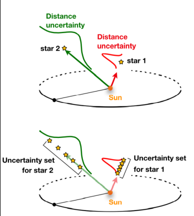



Okuno and Harada (2303.17823)

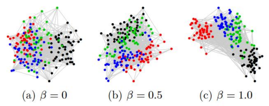
## Application (Scientific Collaborations)



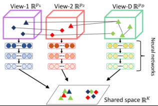

Okuno et al. (2306.16593)



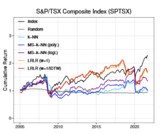

Okuno and Hattori (2204.08205)
Hattori, Okuno, Roederer (ApJ)




Okuno and Shimodaira (AISTATS2019a)




Okuno et al. (ICML2018)

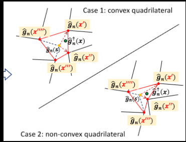



$$\mathbb{V}(\hat{f}_{n,h}(x, x')) = O\left(n^{-\min\{2s, 2\beta+s\}/\{s+d\}}\right)$$

Okuno and Yano (SPL2023)

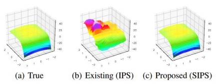
$$\mathcal{E}(\hat{g}_n) = O(n^{-2\beta/(2\beta+d)})$$

Okuno and Shimodaira (NeurIPS2020)
Cao, Okuno et al. (2112.13951)

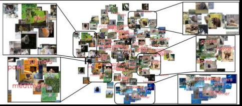



$$\inf_{\hat{f}_n} \sup_{f_* \in \mathcal{F}^{\text{Lip.}}_{\text{INV}}} \mathrm{R}_{\text{INV}}(\hat{f}_n, f_*) \asymp n^{-2/(2+d)}$$

Okuno and Imaizumi (2112.00213)




Okuno et al. (AISTATS2019b)

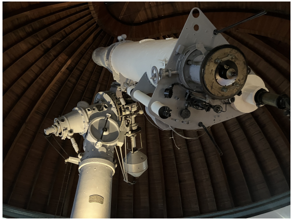


Mizutani, Okuno et al. (2005.00670)

## Visualization/Integration (Representation Learning)

## Theory (Mathematical Statistics)

# 最近は科学系の人たちとワイワイやっています



THE ASTROPHYSICAL JOURNAL, 946:48 (32pp), 2023 March 20 https://doi.org/10.3847/1538-4357/acb93b


**Finding r-II Sibling Stars in the Milky Way with the Greedy Optimistic Clustering Algorithm**

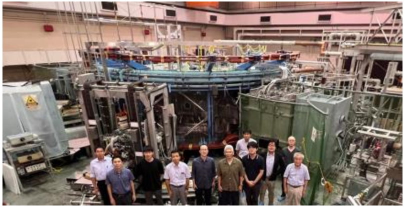
Kohei Hattori[1,2,3], Akifumi Okuno[2,4], and Ian U. Roederer[3,5]
[1] National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
[2] The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan
[3] Department of Astronomy, University of Michigan, 1085 S. University Avenue, Ann Arbor, MI 48109, USA
[4] RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan
[5] Joint Institute for Nuclear Astrophysics–Center for the Evolution of the Elements (JINA-CEE), 640 S. Shaw Lane, East Lansing, MI 48824 USA
*Received 2022 July 8; revised 2023 February 3; accepted 2023 February 3; published 2023 March 28*

**Abstract**

*R*-process enhanced stars with [Eu/Fe] $\geqslant +0.7$ (so-called *r*-II stars) are believed to have formed in an extremely neutron-rich environment in which a rare astrophysical event (e.g., a neutron-star merger) occurred. This scenario is supported by the existence of an ultra-faint dwarf galaxy, Reticulum II, where most of the stars are highly enhanced in *r*-process elements. In this scenario, some small fraction of dwarf galaxies around the Milky Way were *r* enhanced. When each r-enhanced dwarf galaxy accreted to the Milky Way, it deposited many *r*-II stars in the Galactic halo with similar orbital actions. To search for the remnants of the r-enhanced systems, we analyzed the distribution of the orbital actions

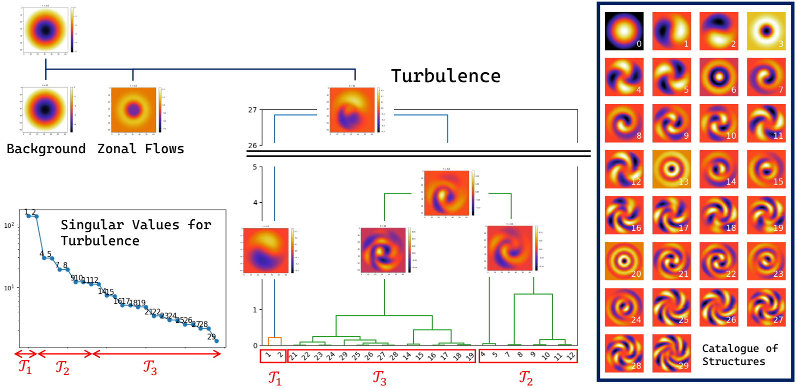
Hattori, Okuno, and Roederer (2023, Astrophysical Journal)

# 最近は科学系の人たちとワイワイやっています


https://statplasma.github.io

# 最近は科学系の人たちとワイワイやっています




Okuno, Kodahara, and Sasaki (2024, Plasma and Fusion Research: Rapid Communications)

- かっこいい図がいろいろ作れるので楽しい.

# 宣伝

- 日本統計学会春季集会（3月8日）@筑波大東京キャンパス

**AM-B「諸科学における統計学的アプローチ」**

10:10〜11:50 119教室 Zoom B

オーガナイザー:奥野彰文(統計数理研究所),三分一史和(統計数理研究所),矢野恵佑(統計数理研究所),横山雅之(核融合科学研究所)

講演1:服部公平(国立天文台/統計数理研究所)
「天文学における統計科学」
講演2:近藤洋平(名古屋大学)
「生物画像に対する力学系モデリングと逆問題」
講演3:加納将行(東北大学)
「情報科学を用いた測地データ解析と断層すべりモニタリングの高度化」
講演4:森下侑哉(京都大学)
「統計科学を活用した核融合プラズマの理解と予測」

https://jss2025spring.ywstat.jp/

- ぜひお越しください.

# 計算系の人たちともワイワイやっています

## 計算技術による学際的統計解析ワークショップ

**開催概要**

高度な計算技術を活用した統計手法研究を目指し，計算技術・統計手法に関する研究発表を通じた研究者の交流を目的とします．口頭発表はオンライン配信予定（質疑応答は対面を最優先します）．

- 日時：2025年2月17日・18日.
- 場所：統計数理研究所セミナー室1（D305）
- オーガナイザ：奥野彰文（統数研/理研AIP）

参加登録はこちら

**スケジュール**

(*口頭発表は全て招待講演)

**2月17日**

**13:00-13:20**
**奥野彰文（統計数理研究所）**
開催趣旨: 計算技術と統計手法研究の融和を目指して
本ワークショップでは主に統計手法・数値解析・最適化・計算代数など様々な分野で活躍されており，また異分野との交流
- expand -

**13:20-14:00**
**寺田吉壱（大阪大学）**
大規模データに対するクラスタリング法について
大規模データに対しては, k-means法などの単純なクラスタリング法を用いることが多い. しかし, 単純な方法では, データの背
- expand -

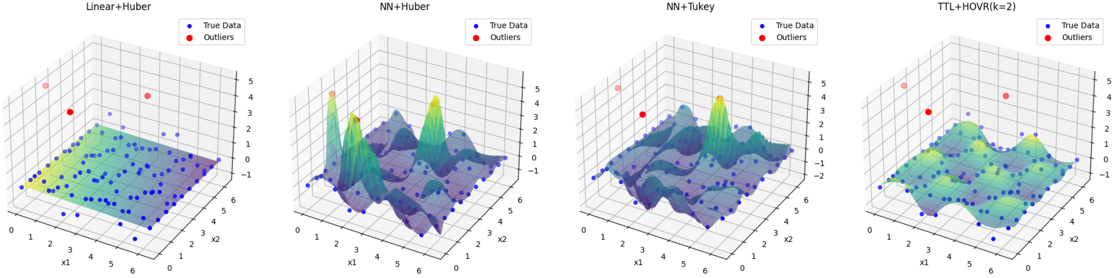
14:00-14:10 休憩

https://okuno.net/events/ISACT2025

# Outlier-Robust Neural Network Training

# なにをしたのか

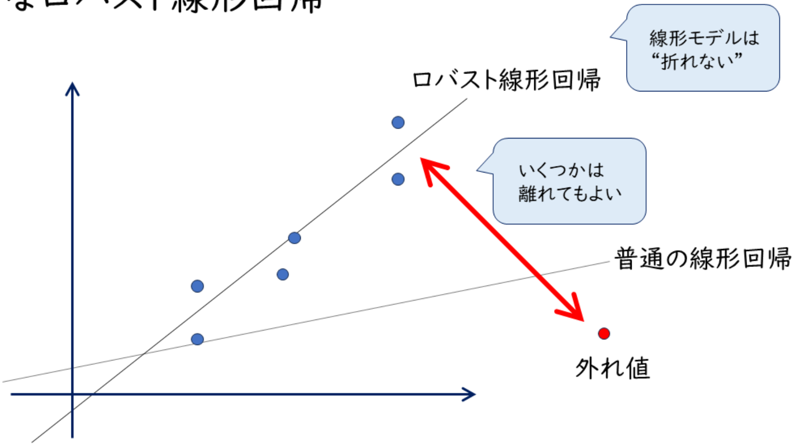
- 外れ値にロバストかつフレキシブルな予測法の提案




(a) 線形モデル+Huber (b) NN+Huber (c) NN+Tukey (d) 提案法

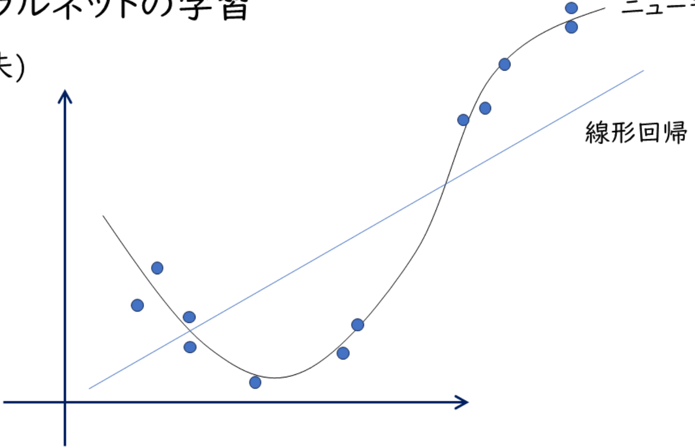
Figure: $f_*(x) = \sin(2x_1)\cos(2x_2)$.

# 典型的なロバスト線形回帰



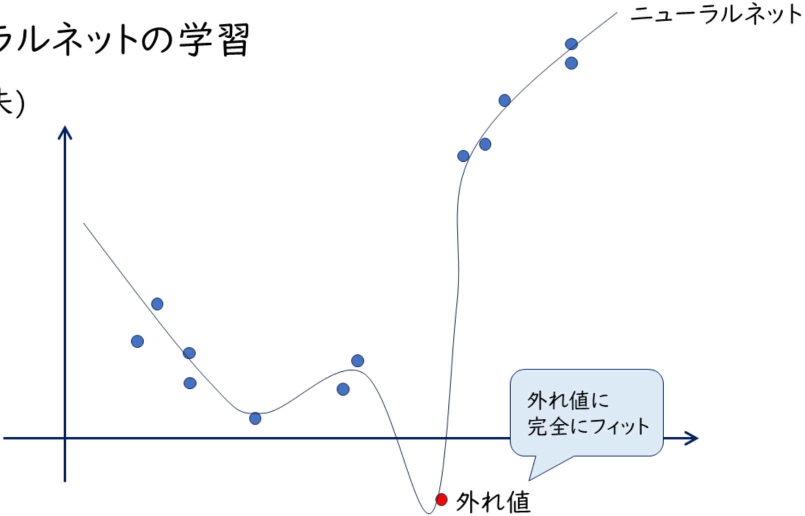
# ニューラルネットの学習

(二乗損失)

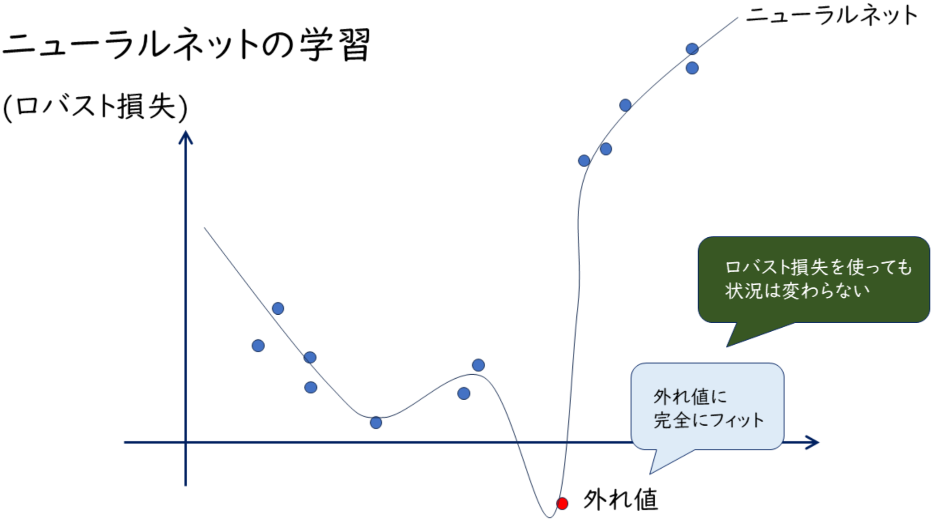


# ニューラルネットの学習

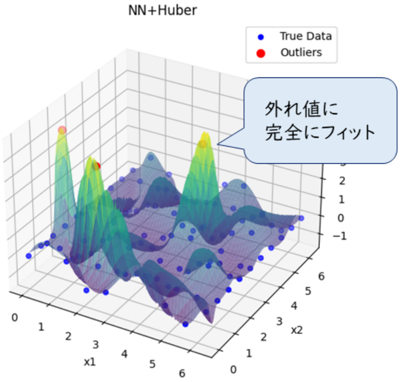
(二乗損失)

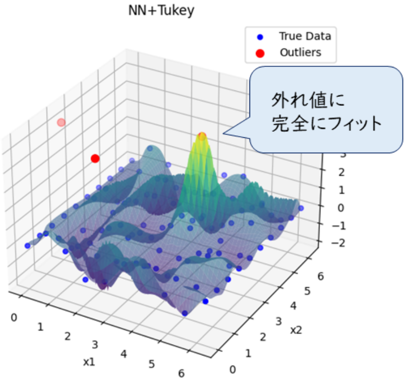


# ニューラルネットの学習

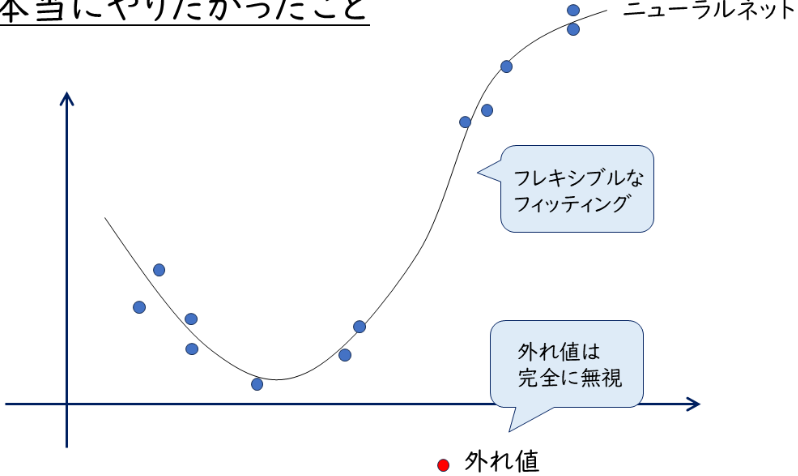


# 実際のロバスト損失+NN (+勾配法)



# 我々が本当にやりたかったこと



# つまり何が言いたいのかというと

- 予測関数 $f_\theta(x)$ はいい感じにグニャグニャしてほしいが,
- 過剰にグニャグニャしないでほしい.

予測関数に "硬さ" のようなものを入れたい.

# 関数の変動を抑える

Higher-Order Variation Regularization (HOVR; Okuno, arXiv:2308.02293v1):

$$C_{k,q}(f_\theta) := \int_\Omega \left| \frac{\partial^k f_\theta(x)}{\partial^k x} \right|^q \mathrm{d}x.$$

例えば正規直交基底を用いた回帰関数 $f_\theta(x) = \sum_j \theta_j \phi_j(x)$ の場合には，

$$C_{k,2}(f_\theta) \lesssim \|\theta\|_2^2$$

なので，パラメータ正則化の一般化と思える.

- 導関数は autograd で計算可能.
- 積分は(明示的に計算せず)確率的に回避できる. <u>特殊な確率的勾配法</u>.

# つまりどうこと?

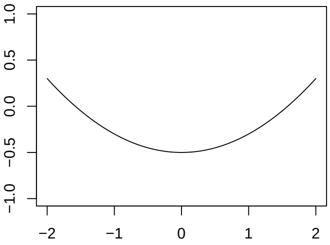



Figure: $C_{2,2}(f) \approx 0.64$

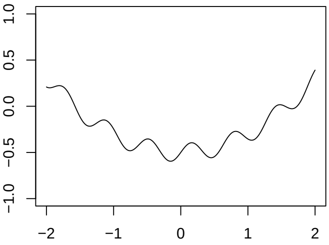



Figure: $C_{2,2}(f) \approx 197$

- 激しいグニャグニャを抑制できる.
- 線形/カーネル回帰モデルなどではパラメータ正則化に対応.
- ニューラルネットでは???

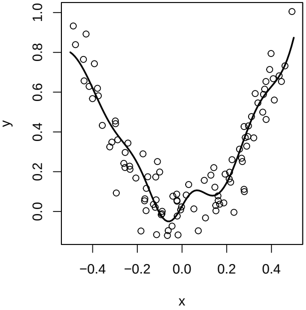



Figure: Weight decay (要するにリッジ)

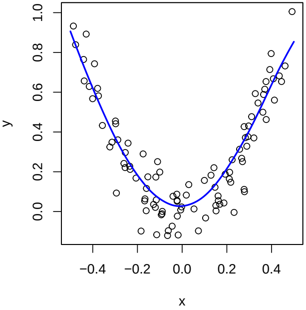



Figure: HOVR

- Akifumi Okuno, “A stochastic optimization approach to train non-linear neural networks with regularization of higher-order total variation”, arXiv:2308.02293v1.
- 確率的勾配法を用いると，積分型の正則化を厳密に最小化できる.
- デスクリジェクトされて萎えた．放置 ⇒ 気付いたら1年が経過していました···

残差を $r_i(\theta) := y_i - f_\theta(x_i)$ とし，添え字を並びかえて

$$|r_{(1;\theta)}(\theta)| \leq |r_{(2;\theta)}(\theta)| \leq \cdots \leq |r_{(n;\theta)}(\theta)|$$

とすると，

$$T_h(r(\theta)) = \frac{1}{n} \sum_{i=1}^{h} r_{(i;\theta)}(\theta)^2$$

を Trimmed Loss と呼ぶ．ロバスト損失として知られている．

- 柳下さんはこのあたりの最適化にとても詳しい．

# アイデア

- 合体させると良いのでは・・・？
- 正則化によりグニャグニャを防ぎ，ロバスト損失で外れ値を無視する.

$$\text{損失関数：}\quad \underbrace{\frac{1}{2}T_h(r(\theta))}_{\text{ロバスト損失}} + \underbrace{\lambda \cdot C_{k,q}(f_\theta)}_{\text{関数を硬くする正則化}}$$

- (伝統的なロバスト指標である)破局点が大きい.
- 効率的な最適化アルゴリズムが作れる.

# 効率的な最適化：柳下さんによる技術

- 余分なパラメータ$\xi$を入れると，Trimmed Loss $T_h$ とパラメータ $\theta$ を分離できる．

Yagishita (arXiv:2410.04554)

$$\frac{1}{2}T_h(r(\theta)) = \underbrace{\min_{\xi \in \mathbb{R}^n}\left\{\frac{1}{n}\|r(\theta) - \xi\|_2^2 + T_h(\xi)\right\}}_{\text{Transformed Trimmed Loss (TTL)}}$$

- これは非常に最適化しやすい形になっている．

# 合体

$$\underbrace{\frac{1}{2}T_h(r(\theta))}_{\text{ロバスト損失}} + \underbrace{\lambda \cdot C_{k,q}(f_\theta)}_{\text{関数を硬くする正則化}} = \min_{\xi}\{U_\lambda(\theta,\xi) - V_h(\xi)\}$$

ただし

$$U_\lambda(\theta,\xi) = \frac{1}{n}\{\|r(\theta)-\xi\|_2^2 + \|\xi\|_2^2\} + \lambda C_{k,q}(f_\theta) \text{ は Nonconvex and smooth,}$$

$$V_h(\xi) = \frac{1}{n}\|\xi\|_2^2 - T_h(\xi) \text{ は Convex and nonsmooth.}$$

つまりTrimmed Loss + HOVR の最小化は，拡張パラメータ$(\theta,\xi)$についての

**Augmented and Regularized Trimmed Loss (ARTL)**

$$F_{h,\lambda}(\theta,\xi) = U_\lambda(\theta,\xi) - V_h(\xi)$$

の最小化と等しい．最適化手法が作れて，DCアルゴリズム風の収束証明ができる．

# 最適化アルゴリズムの提案

- $u_\lambda^{(t)}(\theta^{(t)}, \xi^{(t)})$は$U_\lambda(\theta)$の不偏確率勾配，
- $v_h(\xi)$は$V_h(\xi)$の劣勾配とする．
- $g_{h,\lambda}^{(t)}(\theta, \xi) = u_\lambda^{(t)}(\theta, \xi) - (0, v_h(\xi))$.

**Stochastic Gradient-Supergradient Descent (SGSD)**

$$(\theta^{(t+1)}, \xi^{(t+1)}) \leftarrow (\theta^{(t)}, \xi^{(t)}) - \omega_t g_\lambda^{(t)}(\theta^{(t)}, \xi^{(t)}).$$

- 積分項の問題や微分不可能点の問題を突破して，厳密な収束が示せる．

# 詳細は省略しますが…

- $\theta$についての微分可能性は仮定 (e.g., Sigmoid activation)
- 勾配の不偏性などいくつか仮定.

## 収束定理のシンプルな形

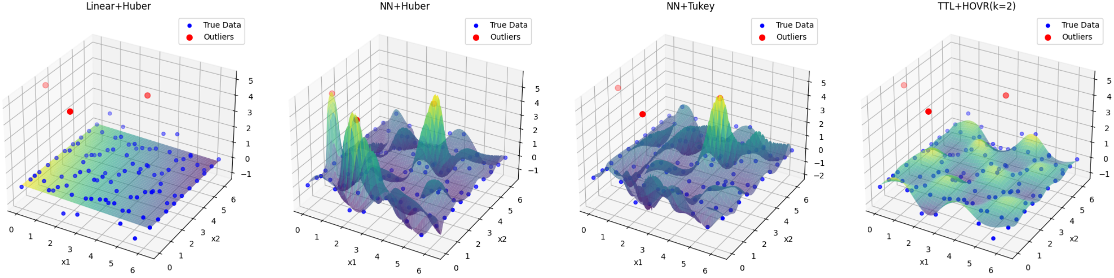
学習率を$\omega_t = \alpha(1+t)^{-1/2}$とすると,

$$\mathbb{E}\left[\inf_{v \in \partial V_h(\xi)} \left\| \frac{\partial U_\lambda(\theta^{(\tau_T)}, \xi^{(\tau_T)})}{\partial(\theta, \xi)} - (0, v) \right\|_2^2 \right] = \mathcal{O}(T^{-1/4}(\log T)^{1/2}).$$

- $\tau_T$はランダムに選択されるStopping Time.
- (劣勾配の自由度を除いて)勾配が0，つまり停留点に収束する.

# というわけで最初の図に戻る.

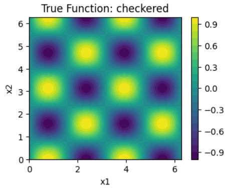
- (入力)2次元 - 100 - 100 - 100 - (出力)1次元の多層パーセプトロン.
- パラメータ数2万超，外れ値3%混入.



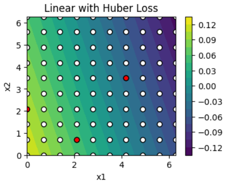

(a) 線形モデル+Huber (b) NN+Huber (c) NN+Tukey (d) 提案法

Figure: $f_*(x) = \sin(2x_1)\cos(2x_2)$.

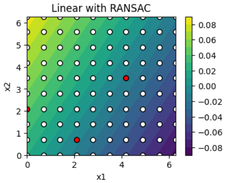
# やってみた 1 (checkered)



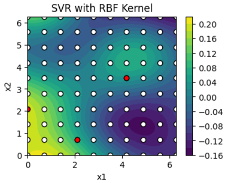

(a) Groundtruth



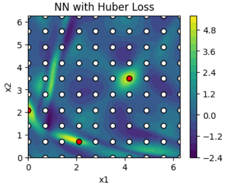

(b) Linear+Huber



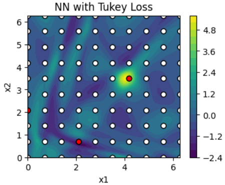

(c) Linear+RANSAC



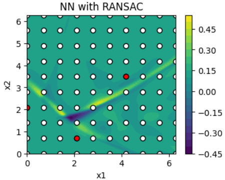

(d) SVR+RBF



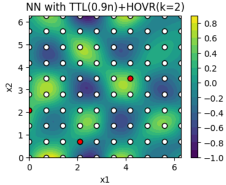

(e) NN+Huber




(f) NN+Tukey

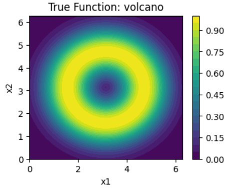



(g) NN+RANSAC

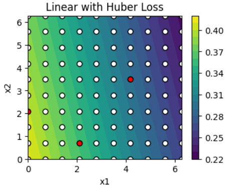



(h) 提案法

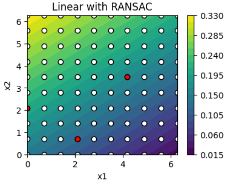
# やってみた 2 (volcano)



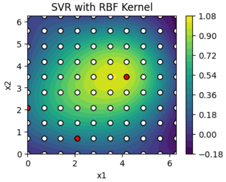

(a) Groundtruth



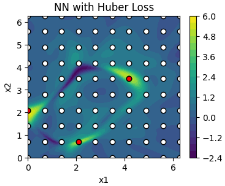

(b) Linear+Huber



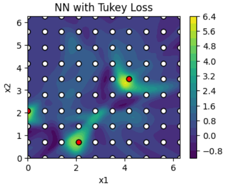

(c) Linear+RANSAC



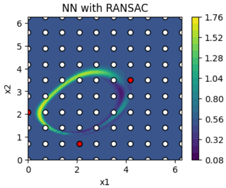

(d) SVR+RBF



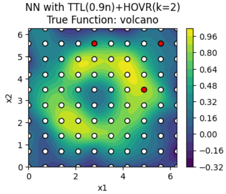

(e) NN+Huber




(f) NN+Tukey

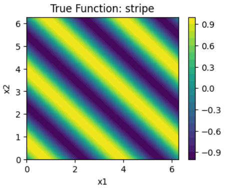



(g) NN+RANSAC

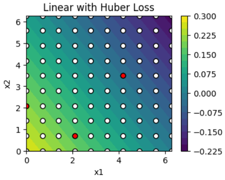



(h) 提案法

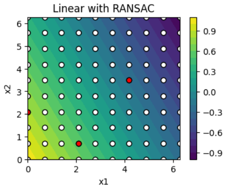
# やってみた 3 (stripe)



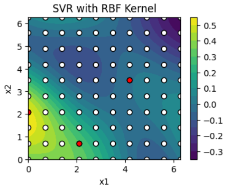

(a) Groundtruth



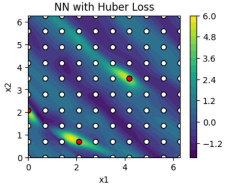

(b) Linear+Huber



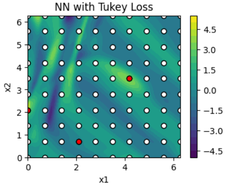

(c) Linear+RANSAC



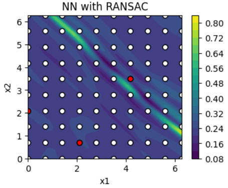

(d) SVR+RBF



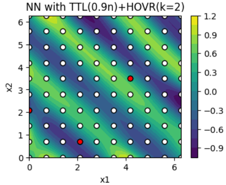

(e) NN+Huber




(f) NN+Tukey

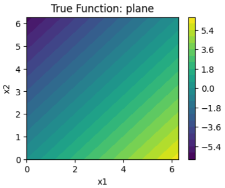



(g) NN+RANSAC

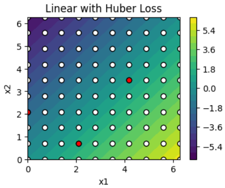



(h) 提案法

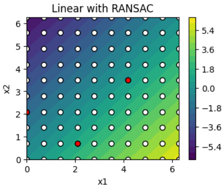
# やってみた 4 (plane)



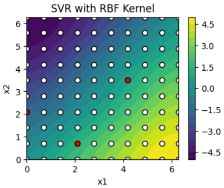

(a) Groundtruth



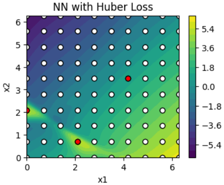

(b) Linear+Huber



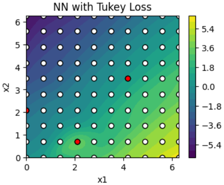

(c) Linear+RANSAC



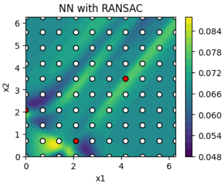

(d) SVR+RBF



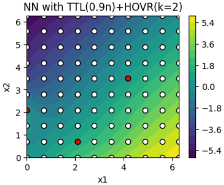

(e) NN+Huber




(f) NN+Tukey




(g) NN+RANSAC

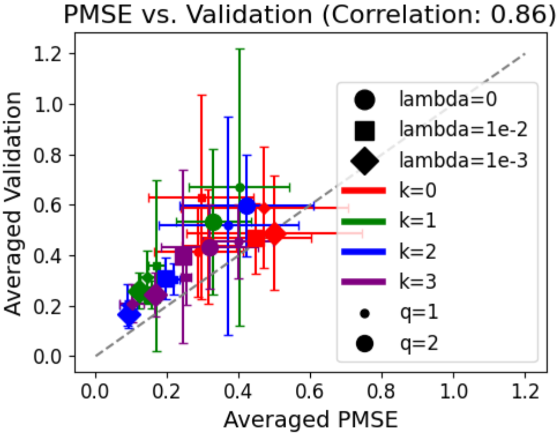



(h) 提案法

| | Non-linear | | | Linear |
|---|---|---|---|---|
| | checkered | volcano | stripe | plane |
| Linear Reg. with Huber's Loss | 0.124 (0.004) | 0.130 (0.003) | 0.498 (0.016) | <u>0.001</u> (0.001) |
| Linear Reg. with RANSAC | 0.140 (0.013) | 0.186 (0.058) | 0.871 (0.277) | **0.001** (0.001) |
| Support Vector Reg. with RBF Kernel | 0.127 (0.008) | 0.113 (0.020) | 0.508 (0.031) | 0.006 (0.002) |
| NN with Huber's Loss | 0.634 (0.608) | 1.031 (0.861) | 0.488 (0.475) | 0.043 (0.025) |
| NN with Tukey's Loss | 0.458 (0.655) | 0.413 (0.630) | 0.304 (0.395) | 0.017 (0.009) |
| NN with Label Noise Reg. | 1.155 (1.068) | 0.872 (0.659) | 0.561 (0.498) | 0.756 (0.945) |
| NN with RANSAC | 0.160 (0.036) | 0.142 (0.009) | 0.527 (0.018) | 0.011 (0.016) |
| NN with ARTL ($h = 0.9n$, $k = 1$) | <u>0.088</u> (0.090) | <u>0.082</u> (0.076) | <u>0.223</u> (0.238) | 0.010 (0.004) |
| NN with ARTL ($h = 0.9n$, $k = 2$) | **0.061** (0.016) | **0.040** (0.029) | **0.119** (0.047) | 0.007 (0.001) |

# パラメータ選択

- ロバスト損失でValidationすればよい.



# おまけ

- Akifumi Okuno. An integrated perspective of robustness in regression through the lens of the bias-variance trade-off. arXiv:2407.10418

では「ロバストネス」の種類を大きく3つに分類した：

1. 外れ値へのロバストネス (統計学)
2. 入力の摂動に対するロバストネス (最適化)
3. モデルの誤特定へのロバストネス (機械学習).

提案法は (1)トリム損失, (2)HOVR, (3)ニューラルネット, で
この3つをすべて網羅した推定になっている.

# まとめ

- ロバスト損失を使っても，ニューラルネットは外れ値に過適合してしまう.
- ニューラルネットを“硬くする”正則化としてHOVRを提案.
- ロバスト損失+正則化と等価なARTLを提案.
- ARTLを最小化する確率的アルゴリズムSGSDを提案.
- 収束証明もできたし，その実験結果もよさそうだった.

コメント・ご質問は okuno@ism.ac.jp まで.