Algebraic Approach to Ridge-Regularized Mean Squared Error Minimization in Minimal ReLU Neural Network (Joint work with R. Fukasaku and Y. Kabata)

Akifumi Okuno[1,2,3]

[1]Inst. Stat. Math., [2]SOKENDAI, [3]RIKEN (AIP/CBS)

https://okuno.net/slides/2025-11-IMI.pdf

1. 自己紹介+講演概要
2. ニューラルネットの基礎とその課題
3. やりたいことと出発点
4. 我々の研究 (Fukasaku, Kabata, and Okuno; arXiv:2508.17783)
5. 今後へ向けて

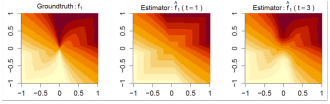
自己紹介+講演概要

奥野彰文 (オクノ アキフミ) 博士 (情報学, 京都大学)[1]

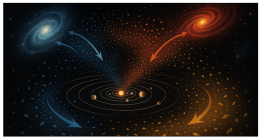
- 専門：統計的機械学習，特に手法開発とその理論解析.
- 普段は統計数理研究所 (東京都立川市) にいます.

[1] https://okuno.net

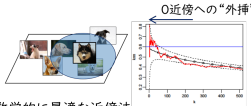
統計〜機械学習の何でも屋です



「可逆」という制約を課した推定の最適性
Okuno and Imaizumi (Elec. J. Stat. 2023)

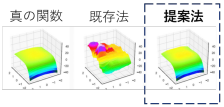


天体の起源を探るクラスタリング：
Hattori, **Okuno** and Roederer (Astrophysical J. 2023)
Okuno and Hattori (Ann. Ins. Stat. Math. 2025)

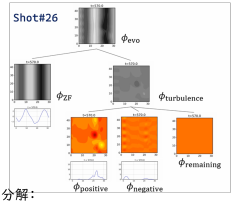


数学的に最適な近傍法
Okuno and Shimodaira (NeurIPS 2020)

組成分析における系統的欠測の補完



類似度ニューラルネットの表現能力とその拡張
Okuno et al. (ICML2018, AISTATS2019)



プラズマ乱流の分解：
Okuno et al. (Plas. Fus. Res. 2024)
Okuno and Sasaki (Physics of Plasmas 2025)

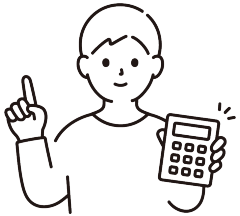
- 理論〜応用まで，統計関係で何かありましたらお気軽にお声がけください．

今日の講演内容

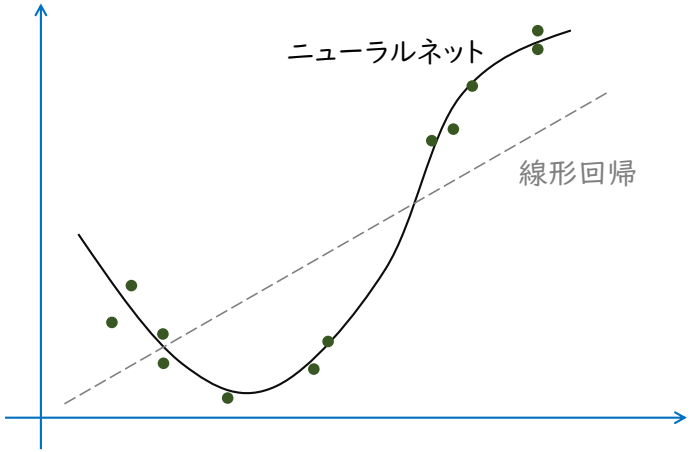
- 活性化関数 $\mathrm{ReLU}(z) = \max\{0, z\}$ を用いるニューラルネット

$$\mathbb{R}^d \ni x \mapsto [\![\, a\,,\, \mathrm{ReLU}(Bx + c)\,]\!] + m \in \mathbb{R},$$

の最適化は非凸で非常に難しいが，計算代数で全ての局所解を列挙できる
(Fukasaku, Kabata, and Okuno; arXiv:2508.17783)



ニューラルネットの基礎とその課題



ニューラルネットはフレキシブルな非線形予測モデル.

ニューラルネットの定義

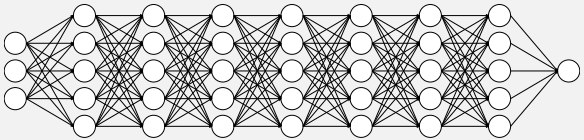
- 線形回帰モデル：

$$f_\theta^{\mathrm{LM}}(x) = Wx + b$$

- ニューラルネット（の特殊形であるパーセプトロン）：

$$f_\theta^{\mathrm{NN}}(x) = W^{(Q+1)}\sigma\left(W^{(Q)}\sigma\left(\cdots\sigma\left(W^{(1)}x + b^{(1)}\right)\cdots\right) + b^{(Q)}\right) + b^{(Q+1)}.$$

  - σを活性化関数という．$1/\{1+\exp(-z)\}$や$\mathrm{ReLU}(z) := \max\{0, z\}$などを要素ごと適用．
  - これ以外にも様々な形のニューラルネットがある．
  - 層の数Qが大きいとき，ディープニューラルネットと呼ぶ（深層学習）．



ニューラルネットは万能近似器である

- fが$I_n = [0,1]^n$上で連続とする．層数$Q = 1$で素子数を無限にとると，
  fを任意の精度で近似できるf^{NN}が存在する[2]．
  - 古典的な結果：Cybenko (1989), Funahashi (1989)など．

- 層数Qを増やすと表現能力が指数的に増大 (Telgarsky, 2016)，
- 層数Qを増やすと高効率な近似を達成 (Yarotsky, 2017)，
- 層数Qを無限に増やせば各層の素子数固定でも万能近似 (Hanin, 2017) 等々…

[2]シグモイド活性化 $\sigma(z) = 1/\{1 + \exp(-z)\}$ で近似誤差は一様に抑えられる

（使うだけなら）実装も非常に容易

```python
# NN Definition
class NeuralNetwork(nn.Module):
    def __init__(self, input_dim=2, activation_func='sigmoid'):
        super(NeuralNetwork, self).__init__()
        self.first = nn.Linear(input_dim, 100)
        self.hidden1 = nn.Linear(100, 100)
        self.hidden2 = nn.Linear(100, 100)
        self.hidden3 = nn.Linear(100, 100)
        self.output = nn.Linear(100, 1)
        self.activation = nn.Sigmoid() if activation_func == 'sigmoid' else nn.ReLU()

    def forward(self, x):
        x = self.activation(self.first(x))
        x = self.activation(self.hidden1(x))
        x = self.activation(self.hidden2(x))
        x = self.activation(self.hidden3(x))
        return self.output(x)
```

構造を記述すると，ボタン一つでいい感じに学習してくれる．

そんなこんなで様々な応用がある

(Generated by ChatGPT)

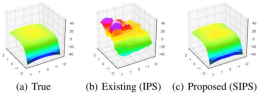
GPUを買いましょう．以上！

ご清聴ありがとうございました！

これでみんながハッピー

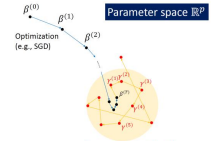
…とはならない

- 統計科学の観点からは，不明なことが多すぎる．



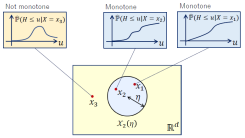
More Expressive Siamese NN

Okuno et al. (AISTATS2019)



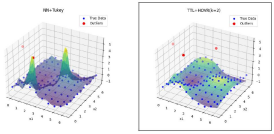
WAIC + Overparameterized NN
+ Langevin dynamics

Okuno and Yano (JCGS2023)



NN + Ordinal Regression

Okuno and Harada (JCGS2024)



NN + Variation Regularization

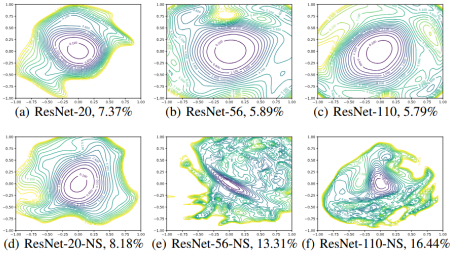
Okuno and Yagishita (in revision)

これまでいろいろやりましたが，結局あまりスッキリしない

難しいポイント

- 非線形である.
- Parametrizationが冗長.
  - 不定性があり，情報行列が退化する $\Rightarrow$ 多くの統計理論が破綻.
  - 学習が非凸最適化問題になる．ヒューリスティックが入りすぎている.

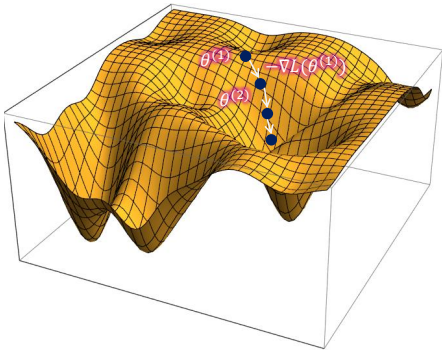
学習に用いる損失関数の例：$L(\theta) = \arg\min_{\theta} \sum_{i=1}^{n} \{y_i - f_\theta(x_i)\}^2$.



Li et al. (NeurIPS2018) Fig.5 より転載

損失地形はボコボコだが，祈りながら谷を下るほかない

- 最小化に用いる，最急降下法（勾配法）：$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma \nabla L(\theta^{(t)})$.



- 凹関数（谷が一つ）では理論保証がたくさんある.
- 多峰の損失関数だと，「どの谷（局所解）に陥るか」は運による….
- 解が孤立点でない場合も $\Rightarrow$ 多くの統計理論が破綻してしまう.

曖昧な気持ちを払拭するための研究プロジェクト

複雑な予測モデルの厳密な事後分布計算手法の確立とその応用

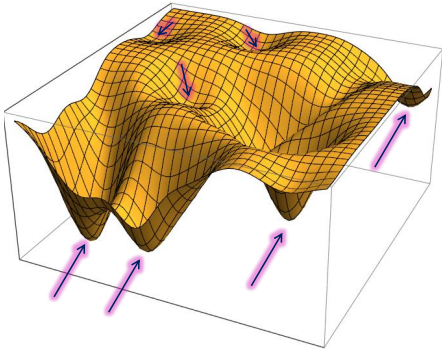
研究課題/領域番号	25K03087
研究種目	基盤研究(B)
配分区分	基金
応募区分	一般
審査区分	小区分60030:統計科学関連 小区分61030:知能情報学関連 合同審査対象区分:小区分60030:統計科学関連、小区分61030:知能情報学関連
研究機関	統計数理研究所
研究代表者	奥野 彰文　統計数理研究所, 統計基盤数理研究系, 助教 (40897972)
研究分担者	深作 亮也　九州大学, 数理学研究院, 助教 (40778924) 高畠 哲也　大阪大学, 大学院基礎工学研究科, 講師 (80846949)
研究期間 (年度)	2025-04-01 – 2029-03-31
研究課題ステータス	交付 (2025年度)

特に計算機援用+数学により，複雑モデルをボトムアップで理解したい.

やりたいことと出発点

本当にやりたいこと

損失関数の局所最小解を全列挙したい.



- 局所解は孤立点ではないかも…(1次元以上の解を持つかも)
- 計算代数で，解の満たすべき方程式が全て求まらないか？

研究の時系列

- 2011年：廣瀬慧先生が下平研究室@阪大で助教に着任された.
- 2013年：奥野が卒論で下平研に配属された.
  （中略）
- 2019年：廣瀬先生+加葉田先生+深作先生が因子分析+計算代数の研究を開始？
- 2022年：廣瀬先生関係で，非常に面白いと思いながら遠目で見ていた.
- 2023年の学術変革領域会議で深作先生・加葉田先生と初対面.
- 2023年のIMI集会に参加して知り合いができる（石原先生，横山先生,…).
- その後で立教大学にお邪魔したり，計良先生をお呼びしたり，etc
- 2024年以降：内輪の集会を何度か開催．研究が始まる.
- 2025年8月：暫定版プレプリントを公開.
- 2025年11月：IMI集会にねじ込んでもらう ← New!

研究の前段となるアイデア

もう一度ニューラルネット（パーセプトロン）の形を見てみよう：

$$f_\theta^{\mathrm{NN}}(x) = W^{(Q+1)}\sigma\left(W^{(Q)}\sigma\left(\cdots\sigma\left(W^{(1)}x + b^{(1)}\right)\cdots\right) + b^{(Q)}\right) + b^{(Q+1)}.$$

- 理論でよくある：活性化関数 $\sigma(z)$ を恒等関数にする.

$$\begin{aligned} f_\theta^{\mathrm{LNN}}(x) &= W^{(Q+1)}\left\{W^{(Q)}\left\{\cdots\left\{W^{(1)}x + b^{(1)}\right\}\cdots\right\} + b^{(Q)}\right\} + b^{(Q+1)} \\ &= \widetilde{W}^{(Q+1)}\widetilde{W}^{(Q)}\cdots\widetilde{W}^{(1)}x + \widetilde{b}. \end{aligned}$$

  - パラメータにやや制約のかかる線形回帰になる.
  - 線形ニューラルネット/縮小ランク回帰などとも呼ばれる.
    (Aoyagi and Watanabe, 2005; Mehta et al., 2022; Aoyagi, 2024 など)

ReLU活性化関数 $\sigma(z) = \max\{0, z\}$ は活性化パターンで記述できる：

ある $W \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m$ とある $x \in \mathbb{R}^d$ を固定すると，
ある $e = e(W, b, x) \in \{0, 1\}^m$ が存在して，

$$\sigma(Wx + b) = \mathrm{diag}(e)(Wx + b).$$

ただし $\mathrm{diag}(e)$ は対角成分が e の対角行列.

- 例えば $Wx + b = (3, -2, 2, 1, -1)$ とすると，$e = e(W, b, x) = (1, 0, 1, 1, 0)$ であり，

$$\mathrm{ReLU}(Wx + b) = (3, 0, 2, 1, 0) = \mathrm{diag}(e)(Wx + b),$$

  というだけの話….
- Arora et al. (2018), Pilanci and Ergen (2020), Mishkin et al. (2022), ….

一般に拡張できる

あるパラメータ $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell=1}^{L}$ と入力 $x \in \mathbb{R}^d$ を固定すると，各層 $\ell = 1, 2, \ldots, L$ の活性化パターン $e_\ell = e_\ell(\theta, x) \in \{0, 1\}^{m_\ell}$ が存在して，

$$f_{\theta,E}^{\mathrm{NN}}(x) = W^{(Q+1)}\mathrm{diag}(e^{(Q)})\Big\{W^{(Q)}\mathrm{diag}(e^{(Q-1)})\{\cdots \cdots \mathrm{diag}(e^{(1)})\Big\{W^{(1)}x + b^{(1)}\Big\}\cdots\Big\} + b^{(Q)}\Big\} + b^{(Q+1)}.$$

- $E = (e^{(\ell)})_{\ell=1}^{L}$ が所与なら，ReLUニューラルネットは単純な行列積になる．
- このとき，損失関数

$$\ell_\lambda(\theta) = \sum_{i=1}^{n}\{y_i - f_{\theta,E}^{\mathrm{NN}}(x_i)\}^2 + \lambda\|\theta\|_2^2$$

は θ に関する多項式になる．

我々のアイデア（出発点）

- 損失関数 $\ell_\lambda(\theta)$ はパラメータ θ に関する多項式.
- その最小解は（たぶん）以下の推定方程式を満たすだろう：

$$\underbrace{\frac{\partial \ell_\lambda(\theta)}{\partial \theta}}_{\text{多項式}} = 0.$$

- これは計算代数が扱っている問題そのものでは?

我々の研究 (Fukasaku, Kabata, and Okuno; arXiv:2508.17783)

というわけで

$$\frac{\partial \ell_\lambda(\theta)}{\partial \theta} = \frac{\partial \{\sum_{i=1}^n \{y_i - f_\theta(x_i)\}^2 + \lambda \|\theta\|_2^2\}}{\partial \theta} = 0$$

を（深作さんに）解いてもらえば万事解決…

とはならない[3].

[3]現実は厳しい.

問題点

- 活性化パターン $E = (e^{(\ell)})$ は パラメータ θ と入力 x に依存.
  - （理想）活性化パターンを決めてから最適パラメータを求めたい.
  - （現実）パラメータが定まると活性化パターンが決まる.
  - 逆！逆！
- もっと言うと，「入力 x への依存」も大変厳しい.
  - 入力によってモデルが変わる，というのは非常に扱い辛い….

深作さんによる非常にパワフルなアイデア

全通りやればよいのでは？

- 各活性化パターンを仮定して推定方程式を解く.
- 出てきた解のうち，仮定した活性化パターンを満たしているものだけ選別.
- 全部の活性化パターンで解を求め，最終的にマージ.

活性化パターンの組み合わせ数に応じた推定方程式を（計算代数で）解けばよい[a].

[a] 言うは易し，実行するは…

詳細な設定：いくらかの仮定

- 以降の議論では，簡単のため $Q = 1$ 層に限定[4]．つまり

$$f_\theta^{\mathrm{NN}}(x) = [\![\, a\,,\, \mathrm{ReLU}(Bx + c)\,]\!], \quad \theta = (a, B, c).$$

ただし素子数を L とする $(a, c \in \mathbb{R}^L, B \in \mathbb{R}^{L \times d})$.

- パラメータ a をあらかじめ消去：$\psi = (B, c)$ について

$$\ell_\lambda(\psi) = \min_a \left\{ \sum_{i=1}^n \{y_i - f_\theta(x_i)\}^2 + \lambda \|\theta\|_2^2 \right\}$$

である[5]．これは有理関数になるので，代数的に $\ell_\lambda(\psi)$ を最小化する．

[4]本質的には一般の層数に対応可能
[5]統計学におけるリッジ線形回帰によって関数が解析的に求まる．

活性化パターンの定義とパラメータ空間の分割

- 回帰に用いるデータセットを $\{(x_i, y_i)\}_{i=1}^n$ とする.
- $\xi_{i\ell}(\psi) = [\![b_\ell \, , \, x_i]\!] + c_\ell$ とし，$e_{i\ell} = e_{i\ell}(\psi) = \begin{cases} 1 & \xi_{i\ell}(\psi) \geq 0 \\ -1 & \xi_{i\ell}(\psi) < 0 \end{cases}$ とする [6]と，

$$\mathrm{ReLU}(\xi_{i\ell}(\psi)) = \frac{e_{i\ell} + 1}{2} \xi_{i\ell}(\psi).$$

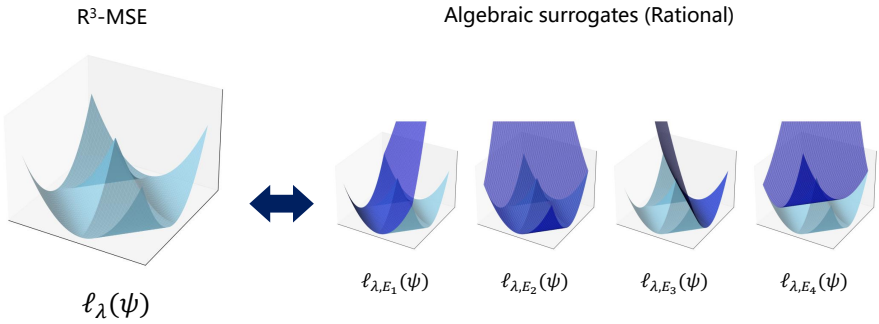
- 逆に，活性化パターンが $E = E(\psi) = (e_{i\ell}) \in \mathbb{R}^{n \times L}$ となるパラメータ集合：

$$\Psi(E) = \{\psi \in \Psi \mid \xi_{i\ell}(\psi) e_{i\ell} \geq 0, \forall i, \ell\}.$$

[6]さっきまでは$\{0, 1\}$でしたが，記号のため ± 1 に変更します.

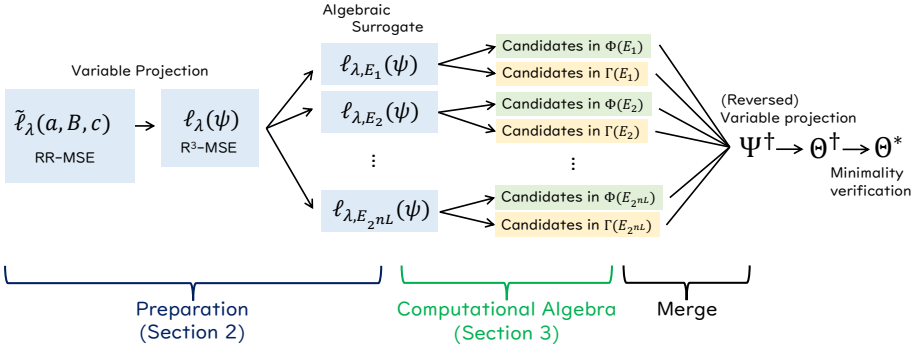
関数の分割と代理関数

- 本来最小化したいのは $\ell_\lambda(\psi)$.
- 各活性化パターン $E_1, E_2, \ldots$ でのパラメータ分割：$\Psi(E_1), \Psi(E_2), \ldots$ において $\ell_{\lambda,E}(\psi)$ と一致する代理関数 $\ell_{\lambda,E_i}(\psi)$ を代わりに最小化する.



- $\partial \ell_{\lambda,E_i}(\psi)/\partial\psi = 0$ の解（のうち領域の内点）は計算代数で求まる.

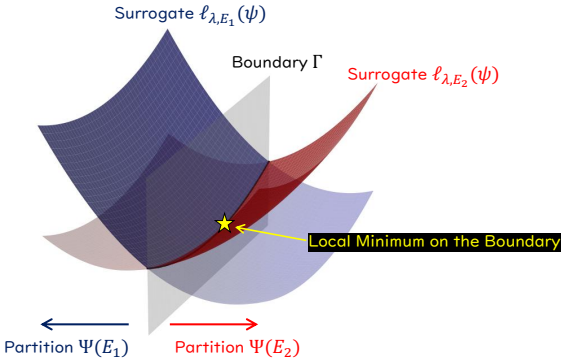
全体の手続き



- パラメータ分割領域の内側の最小解（候補）列挙は比較的単純．解くだけ．
- 境界が面倒臭い…

何が面倒なのか？

- 分割領域が隣接する活性化パターン $E_1, E_2 \in \{-1, +1\}^{n \times L}$ について，
  $\ell_{\lambda, E_1}(\psi)$ と $\ell_{\lambda, E_2}(\psi)$ は境界上に最小解をもちうる．



- Ψ 全域では $\ell_{\lambda, E_1}(\psi), \ell_{\lambda, E_2}(\psi)$ のどちらの（局所）最小解でもないが，
  分割領域で代理関数が切り替わると出現する局所最小解がありうる．

境界上の最小解

- ψ が境界上にある $\Leftrightarrow$ ある i, ℓ について $\xi_{i\ell}(\psi) = [\![b_\ell , x_i]\!] + c_\ell = 0$.
- Lagrange の未定乗数法を解けばよい：

$$\underbrace{\frac{\partial}{\partial \psi}\{\ell_{\lambda,E}(\psi) + \beta \xi_{i\ell}(\psi)\}}_{\text{有理関数}} = 0$$

FKO (arXiv:2508.17783) Theorem 2

ℓ_λの局所最小解は，

(1) パラメータ分割領域 $\psi(E)$ の内側にある，代理関数 $\ell_{\lambda,E}(\psi)$ の局所最小解か，

(2) 境界上にある（境界中での）局所最小解のどちらかしかない.

- したがって，境界の局所解も結局有利多項式（の微分）の零点として出てくる.

代数多様体

$f_1, f_2, \ldots, f_r \in \mathbb{R}[\psi]$ を実係数多項式とするとき，

$$\mathbb{V}(f_1, f_2, \ldots, f_r) = \{\psi \in \Psi \mid f_1(\psi) = 0, f_2(\psi) = 0, \ldots, f_r(\psi) = 0\}$$

を代数多様体と呼ぶ．Gröbner基底を求めて具体的な代数多様体が定まる．

今回での例

分割領域の内側 $\Leftrightarrow \prod_{i,\ell} \xi_{i\ell}(\psi) \neq 0$ での推定方程式 $\frac{\partial \ell_{\lambda,E}(\psi)}{\partial \psi} = 0$ の解集合は，

$$\mathcal{S}_E = \mathbb{V}\left(\mathrm{num}\left(\frac{\partial \ell_{\lambda,E}(\psi)}{\partial \psi}\right)\right) \setminus \mathbb{V}\left(\mathrm{den}\left(\frac{\partial \ell_{\lambda,E}(\psi)}{\partial \psi}\right) \prod_{i,\ell} \xi_{i\ell}(\psi)\right).$$

- 境界の局所解に対応する代数多様体も同様に定まる．

これをたくさんやる

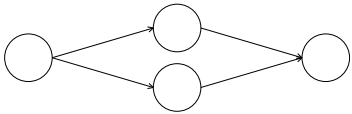
- 分割領域は高々 2^{nL} 個しかないので，
  そのすべての組み合わせで個別に代数多様体を求める．

- 出てきた解は局所最小解の候補[7]なので，
  最小性をチェックして残ったものを局所解として出力する．

[7]停留点のようなもので，最小かどうかは不明だが局所最小解をすべて含んでいる

実際にやってみた

- 入力次元 $d = 1$，素子数 $L = 2$，サンプルサイズ $n = 5$.

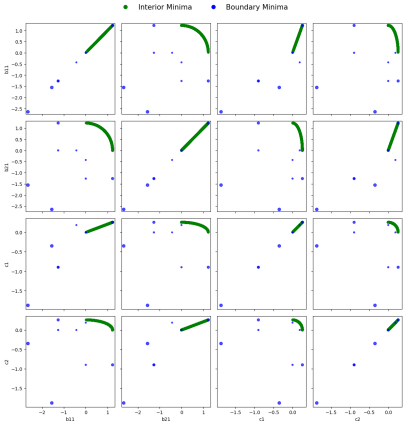


$$(x_1, y_1) = \left(-\frac{17}{100}, \frac{5}{100}\right), \quad (x_2, y_2) = \left(\frac{44}{100}, \frac{102}{100}\right), \quad (x_3, y_3) = \left(-\frac{100}{100}, \frac{61}{100}\right),$$
$$(x_4, y_4) = \left(-\frac{40}{100}, -\frac{36}{100}\right), \quad (x_5, y_5) = \left(-\frac{71}{100}, -\frac{132}{100}\right).$$

- この場合，活性化パターンの組み合わせ数は $2^{nL} = 1024$ ある[8].

[8] つまり1024回Gröbner基底を求めるということ.

計算結果

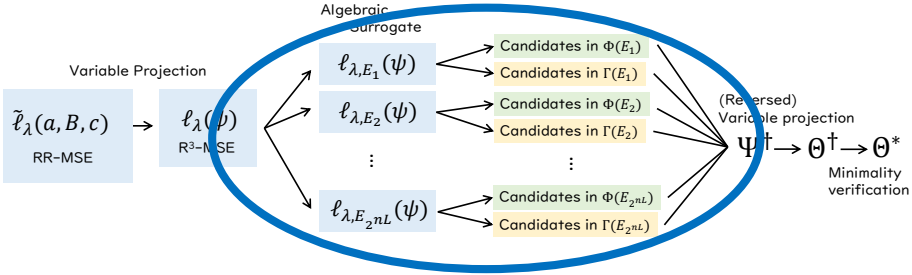


- リッジ正則化を入れたのに連なった解（1次元の解集合）が出てきた.
- 孤立解はすべて境界上だった.

今後へ向けて

残る課題

- 計算量が大きすぎる.
  - パラメータ数を増やす ⇒ 1回あたりの計算量も並列数も増える.
  - サンプルサイズを増やす ⇒ 並列数が増える.



- 並列化，および関連する方程式のGröbner基底の高速計算など…

宣伝

第2回：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，学際的な研究者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）．第1回の開催情報はこちら．参加登録はこちら．

- 日時：2026年2月16日・17日
- 場所：統計数理研究所 D305室（予定）
- オーガナイザ：奥野彬文（統数研/総研大/理研）

10/30追記：1日目の学生発表と2日目のポスター発表の時間を入れ替えました．ポスター発表は両日講演が望ましいですが，両日参加が難しい場合は片方でも大丈夫です．

スケジュール（予定）

(*口頭発表は全て招待講演)

2月16日

13:00-13:30
奥野彬文（統計数理研究所）
統計手法研究と計算技術

13:30-14:30
蓮谷猛（北海道大学）
TBA

14:30-14:45 休憩

14:45-15:45
森岡悠史（道算大学）
TBA

15:45-16:45 ポスターセッション

2月16日-17日@統数研（https://okuno.net/events/ISACT2026）

- どなたでも，ぜひ遊びに来てください！

コメントに限らず統計・機械学習に関する質問・依頼など，お気軽にご連絡ください.

okuno@ism.ac.jp

（今日の資料）https://okuno.net/slides/2025-11-IMI.pdf