深層学習を用いた条件付き確率のノンパラメトリック推定

屋良淳朝[1], 寺田吉壱[1, 2]

[1]大阪大学基礎工学研究科

[2]理研AIP

計算機統計学会 第38回大会

2024/5/24 @やまぎん県民ホール

1. はじめに

1.1 深層学習

▸ 深層学習の発展

- 深層学習は, 現実世界の様々な問題で極めて高い性能を発揮
  - 画像認識：ResNet, U-Net
  - 画像生成：DALL-E, sora
  - 自然言語処理：ChatGPT, Gemini
- 深層学習の性能を説明するための理論も盛んに研究されている
  - **深層学習の汎化性能に関して**

    ✓新しいデータに対する性能の良さ
  - 深層学習の近似性能に関して

    ✓どの程度関数を効率よく近似できるか
  - 深層学習の最適化に関して

    ✓データにfitするパラメータは最適化で得られるか？

1.2 深層学習の汎化誤差解析

- **統計的学習理論の枠組み**
  - 統計的学習理論の目標：機械学習の手法の**正当性**を与える
  - 正当性の評価 = 汎化誤差の評価
    - 汎化誤差：新しいデータに対する推定量の性能の悪さ
    - 汎化誤差が小さい ⇒ ″良い″推定量
  - 汎化誤差評価の観点
    - 訓練データが増えると汎化誤差はゼロに収束するか？
    - 収束するとしたらその速さ（収束レート）は？
    - 得られた収束レートの″最適性″

1.2 深層学習の汎化誤差解析

- **経験誤差最小化（Empirical Risk Minimization, ERM）**
  - 簡単のために教師あり学習の設定で考える
    - $\mathcal{X}$：入力空間, $\mathcal{Y}$：出力空間, P：$\mathcal{X} \times \mathcal{Y}$ 上の確率分布
    - $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \sim_{i.i.d.} P$
    - $\mathcal{D}_n := \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$：観測データ
    - $\ell(\cdot, \cdot) : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$：損失関数
      - どのような問題を考えるかに対応
      - 回帰なら二乗誤差, 分類ならヒンジロスなど
  - 目標：新たなデータ$\boldsymbol{X}$が与えられたときに、Y を**<u>良く予測</u>**する関数 $\hat{f}_n : \mathcal{X} \to \mathbb{R}$ を$\mathcal{D}_n$から構成する
  - **「良く予測する」とは？ ⇒ 汎化誤差が小さいこと！**

1.2 深層学習の汎化誤差解析

- **経験誤差最小化（Empirical Risk Minimization, ERM）**
  - 「良く予測する」とは？ ⇒ 汎化誤差が小さいこと!
  - 汎化誤差（超過リスク）：

$$\mathbb{E}_{(X,Y)}[\ell(Y, \hat{f}(X))] - \inf_{f:\text{measurable}} \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$$

  - **新しいデータに対する損失の期待値は未知** ⇒ **ERM**
  - ERM：観測データ$\mathcal{D}_n$上で損失を最小化
    - 推定量 $\hat{f}$ は仮説空間（モデル）$\mathcal{F}$ から探す
    - 本研究では $\mathcal{F}$ は Deep Neural Networks (DNNs)

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$$

*ERMの解 $\hat{f}$ もしばしばERM (Empirical Risk Minimizer)と呼ぶ

1.2 深層学習の汎化誤差解析

- **汎化誤差のBias-Variance分解**
  - $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$：期待リスク

$$L(\hat{f}) - \inf_{f:\text{measurable}} L(f)$$

$$= \underbrace{L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f)}_{\text{Variance}} + \underbrace{\inf_{f \in \mathcal{F}} L(f) - \inf_{f:\text{measurable}} L(f)}_{\text{Bias}}$$

$\mathcal{G}$：可測関数全体

$\mathcal{F}$：仮説空間（モデル）

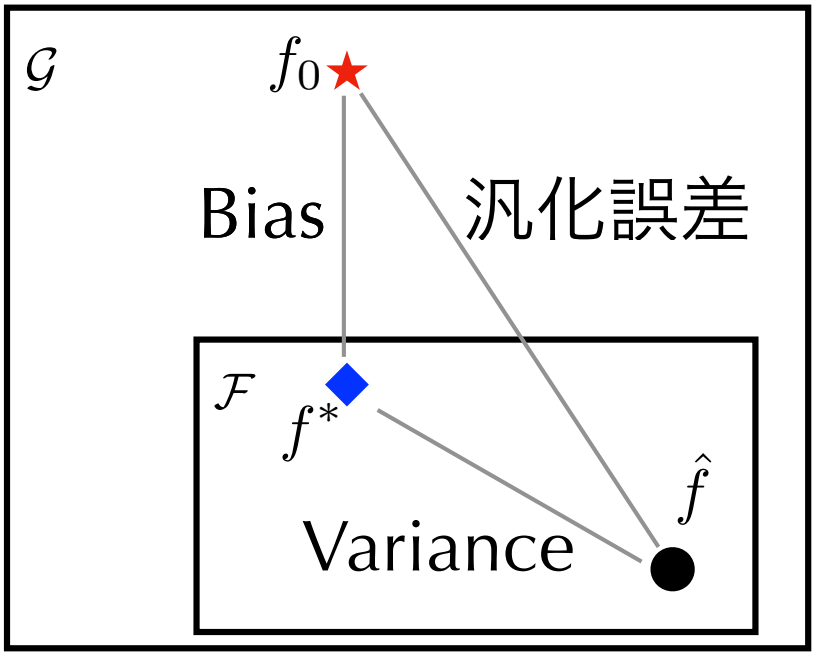
★ $f_0 = \min_{f \in \mathcal{G}} L(f)$

◆ $f^* = \min_{f \in \mathcal{F}} L(f)$

BiasとVarianceは

トレードオフの関係

（モデル$\mathcal{F}$の"大きさ"に依存）



1.3 分類問題と条件付き確率の推定

▸ 分類問題の基本的な設定

- K クラスの分類問題を考える
  - $\mathcal{X} \subset \mathbb{R}^d$：入力空間, $\mathcal{Y} = \{1, \ldots, K\}$：出力空間
  - P：$\mathcal{X} \times \mathcal{Y}$ 上の確率分布, $P_{\boldsymbol{X}}$：$\boldsymbol{X}$ の周辺分布
  - $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \sim_{i.i.d.} P$, $\mathcal{D}_n := \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$
  - $\eta_k(\boldsymbol{x}) := P(Y = k \mid \boldsymbol{X} = \boldsymbol{x})$, $\boldsymbol{\eta}(\boldsymbol{x}) = (\eta_1(\boldsymbol{x}), \ldots, \eta_k(\boldsymbol{x}))^T$
  - $\boldsymbol{X}$ が与えられた下で, Y は以下の多項分布に従う

$$Y | \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{Multi}(\boldsymbol{\eta}(\boldsymbol{x}))$$

- 分類問題の目標：

  新たなデータ $\boldsymbol{X}$ が与えられたときに, Y を良く予測する関数

  $f : \mathcal{X} \to \{1, \ldots, K\}$（決定関数）を見つける

1.3 分類問題と条件付き確率の推定

- **条件付き確率の推定**
  - ヒンジロスを用いた場合は，分類器そのものしか構成できない
    - i.e. ラベルの予測しかできない
  - 条件付き確率を推定できた方が便利な状況もある
    - e.g. AI ベースの病気の診断
  - 条件付き確率を推定できれば，その確率を元に人間の意思決定に役立てることができる！
  - 条件付き確率を推定する方法

    ノンパラメトリックロジスティック回帰

    ⇒ **損失関数として負の対数尤度を利用（最尤推定）**

    ノンパラメトリック最尤推定量（NPMLE） $\hat{\boldsymbol{p}}_n(\boldsymbol{x}) = (\hat{p}_1(\boldsymbol{x}), \ldots, \hat{p}_K(\boldsymbol{x}))$:

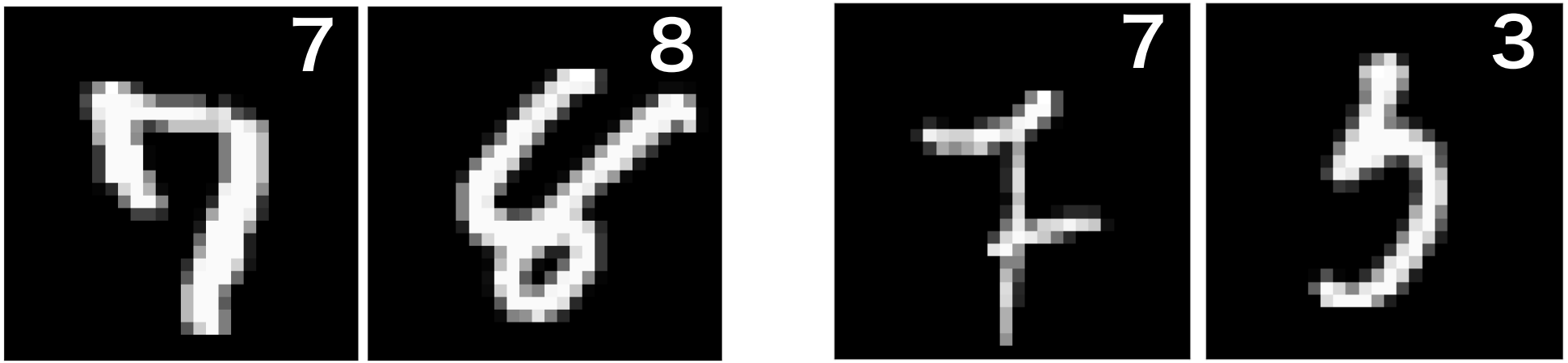
$$\ell(\boldsymbol{p}) := -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i^\top \log \boldsymbol{p}(\boldsymbol{X}_i),\ \hat{\boldsymbol{p}}_n \in \operatorname*{arg\,min}_{\boldsymbol{p}\in\mathcal{F}_n} \ell(\boldsymbol{p})$$

    - $\boldsymbol{Y}_i$ は Y_i の one-hot 表現,

      i.e. $\boldsymbol{Y}_i = (0, \ldots, \underset{k}{1}, \ldots, 0)$ if $Y_i = k$

1.4 深層学習を用いた分類

- **深層学習を用いた分類**
  - 分類問題においては，出力層に softmax 関数を用いた DNNを用いることで条件付き確率の推定が可能
  - MNIST における分類の例（kNNと比較）
    - **判別がつきにくいものは条件付き確率も小さく推定される**



DNN	**4: 10%, 7: 90%**	**4: 10%, 8 : 90%**	**7: 60%, 8 : 40%**	**3: 60%, 4: 10%, 5: 30%**
kNN	7: 80%, 9: 20%	4: 30%, 8 : 20%, 9: 50%	1: 100%	3: 40%, 5: 50%, 9: 10%

1.5 セミパラメトリック推定と条件付き確率の推定

- **Double/Debiased Machine Learning (DML) との関係**
  - 因果推論の文脈において, 傾向スコア（＝共変量を条件付けたときの割り付けの**条件付き確率**）は因果効果を推定する際の局外母数として重要な役割を果たす
  - DML (Chernozhukov et al., 2018)：
    傾向スコアなどの局外母数の推定に機械学習の手法を用いる方法
    ⇒ **局外母数が複雑な状況でも自然な推定量を提供**
  - 局外母数の推定に$o_p\left(n^{-1/4}\right)$の収束レートを示す必要がある

⇒ 例えば, 傾向スコアを深層学習を用いて推定する場合を考えると
深層学習を用いたロジスティック回帰の汎化誤差解析が必須！

- （別の問題として,,,）傾向スコアの推定の応用でよく用いられる
  IPW推定量は傾向スコアがゼロに近いと非常に不安定となる

⇒ ELW推定量 (Liu & Fan, 2023, JRSS B) を用いればOK! (詳細は付録)

2. 先行研究・関連研究

2.2 条件付き確率の収束についての研究

- **条件付き確率の推定における理論解析の困難性**
  - DNN を用いた分類問題では, 最終層にロジスティックシグモイド関数やソフトマックス関数を用いることで各クラスの条件付き確率を出力できる
  - 汎化誤差：

$$\mathbb{E}_{\boldsymbol{X}}\left[\mathrm{KL}\left(\boldsymbol{\eta}(\boldsymbol{X}) \,\|\, \hat{\boldsymbol{p}}(\boldsymbol{X})\right)\right] \qquad (*)$$

    - KL : Kullback-Leibler ダイバージェンス
  - KL は容易に発散するため, (*)を直接評価することは難しい
    - $\mathcal{F}_n$ が"区分的に定数であり，２つ以上の区分をもつ"または,

      "区分的に線形であり，３つ以上の区分をもつ" 関数を全て含む

      ⇒ (*) は発散 (Lemma 2.1 in Bos & Schmidt-Hieber, 2022)
  - **この問題を解決するためのさまざまなアプローチが提案されている**

    (Ohn & Kim, 2022; Bos & Schmidt-Hieber, 2022; Bilodeau et al., 2023)

2.4 本研究の概要

- **ノンパラメトリック最尤推定において, KLの評価は困難**

  例えば, Ohn and Kim (2022) などでは強い仮定が必要に …

  ⇒ van de Geer (2000) のアプローチを用いて

  Hellinger距離を直接評価することで解決 (Theorem 1, 2)

  - **強い仮定が不要, 証明の見通しが良い！**
  - **Remark:** Bos & Schmidt-Hieber (2022)では打ち切り KL を考えて回避
  - **Remark:** Bilodeau et al. (2023)ではMLEを変換して回避
- **深層学習を用いたロジスティック回帰分析の**

  理論解析へ応用 (Theorem 3)

  - **判別問題においても, 次元の呪いを回避できる可能性が示唆**
- **深層学習がミニマックス最適であることを証明 (Theorem 4)**
  - **深層学習はこれ以上改善できない最適な収束レートを達成**

2.4 本研究の概要

▸ 先行研究との比較

	評価指標	損失関数	備考
Hu et al. (2020, arXiv)	誤判別率	0-1損失 ヒンジロス	強い仮定が必要 **ミニマックス最適**
Kim et al. (2021, Neural Networks)	誤判別率	ヒンジロス ロジスティックロス	強い仮定が必要 **ミニマックス最適**
Ohn & Kim (2022, Neural Computation)	**条件付き確率の推定** KL divergence	ロジスティックロス	強い仮定が必要 正則化有り
Bos & Schmidt-Hieber (2022, Electronic Journal of Statistics)	**条件付き確率の推定** 打ち切りKL divergence	負の対数尤度	一般的な仮定 KL より弱く, Hellingerより強い収束
Bildeau et al. (2023, Annals of Statistics)	**条件付き確率** KL divergence	負の対数尤度	MLEを変換した推定量
This work (2024, arXiv)	**条件付き確率の推定** **Hellinger 距離**	**負の対数尤度**	**一般的な仮定** **ミニマックス最適** **証明の見通しが良い**

3. 主結果

3.2 記号と仮定

- **記号**
  - $H\left(P,Q\right)^2 := \frac{1}{2}\int\left(\sqrt{dP}-\sqrt{dQ}\right)^2$: 2つの分布間のHellinger距離の2乗
  - $R(\boldsymbol{\eta}(\boldsymbol{X}),\boldsymbol{p}(\boldsymbol{X})) := \mathbb{E}_{\boldsymbol{X}}\left[H^2(\boldsymbol{\eta}(\boldsymbol{X}),\boldsymbol{p}(\boldsymbol{X}))\right]$
    - 本研究では，$\mathbb{E}_{\boldsymbol{X}}\left[\mathrm{KL}\left(\boldsymbol{\eta}(\boldsymbol{X})\parallel\hat{\boldsymbol{p}}(\boldsymbol{X})\right)\right]$ の代わりに

      $R(\boldsymbol{\eta}(\boldsymbol{X}),\hat{\boldsymbol{p}}(\boldsymbol{X}))$ の収束について議論する
    - $R(\boldsymbol{\eta}(\boldsymbol{X}),\hat{\boldsymbol{p}}_n(\boldsymbol{X})) \leq \mathbb{E}_{\boldsymbol{X}}\left[\mathrm{KL}\left(\boldsymbol{\eta}(\boldsymbol{X})\parallel\hat{\boldsymbol{p}}_n(\boldsymbol{X})\right)\right]$
- **仮定**
  - $\exists\tilde{\boldsymbol{p}}_n\in\mathcal{F}_n,\ \exists c_0>0;\ \dfrac{\eta_k(\mathbf{x})}{\tilde{p}_{n,k}(\mathbf{x})}\leq c_0^2,\ \forall n\in\mathbb{N}, k=1,\ldots,K$
    - 本研究で課す唯一の仮定
    - **<u>DNNによる推定であれば自動的に満たされる</u>**

3.3 ノンパラメトリックロジスティック回帰の一般理論

☆オラクル不等式（推定量の汎化性能を評価）

Theorem 2（オラクル不等式 期待値ver.）

Kクラスの分類問題を考える. $\hat{\boldsymbol{p}}_n$をNPMLEとする.

$\Psi(\delta) \geq J_B(\delta, \bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}_n, \delta), \mu)$を$\Psi(\delta)/\delta^2$が$\delta$の非増加関数となるようにとる. このとき, ある普遍定数cと

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n) \qquad (1)$$

を満たすδ_n, 全ての$\delta \geq \delta_n$に対して以下が成り立つ.

$$\mathbb{E}\left[R\left(\hat{\mathbf{p}}_n, \boldsymbol{\eta}\right)\right] \leq 514(1+c_0^2)\left(\underbrace{\delta_n^2}_{\text{Variance}} + \underbrace{R\left(\tilde{\mathbf{p}}_n, \boldsymbol{\eta}\right)}_{\text{Bias}}\right) + \frac{c}{n},$$

- δがVarianceの評価に対応
- J_Bはモデル$\mathcal{F}_n$の"大きさ" = "複雑さ"を表す

⇒ モデルが複雑なほどVarianceは大きくなってしまう

3.3 ノンパラメトリックロジスティック回帰の一般理論

☆オラクル不等式（推定量の汎化性能を評価）

Theorem 2（オラクル不等式 期待値ver.）

Kクラスの分類問題を考える. $\hat{\boldsymbol{p}}_n$をNPMLEとする.

$\Psi(\delta) \geq J_B(\delta, \bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}_n, \delta), \mu)$を$\Psi(\delta)/\delta^2$が$\delta$の非増加関数となるようにとる. このとき, ある普遍定数cと

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n) \qquad (1)$$

を満たすδ_n, 全ての$\delta \geq \delta_n$に対して以下が成り立つ.

$$\mathbb{E}\left[R\left(\hat{\mathbf{p}}_n, \boldsymbol{\eta}\right)\right] \leq 514(1+c_0^2)\left(\underbrace{\delta_n^2}_{\text{Variance}} + \underbrace{R\left(\tilde{\mathbf{p}}_n, \boldsymbol{\eta}\right)}_{\text{Bias}}\right) + \frac{c}{n},$$

- モデルが複雑なほどVarianceは大きくなってしまう
- 一方, モデルが複雑なほどBias $R(\tilde{\boldsymbol{p}}_n, \boldsymbol{\eta})$は小さくなる

Bias-Variance トレードオフ

（モデルが複雑なほど真の条件付き確率を含みやすくなるため）

3.3 ノンパラメトリックロジスティック回帰の一般理論

☆オラクル不等式（推定量の汎化性能を評価）

Theorem 2（オラクル不等式 期待値ver.）

Kクラスの分類問題を考える. $\hat{\boldsymbol{p}}_n$をNPMLEとする.

$\Psi(\delta) \geq J_B(\delta, \bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}_n, \delta), \mu)$を$\Psi(\delta)/\delta^2$が$\delta$の非増加関数となるようにとる. このとき, ある普遍定数cと

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n) \qquad (1)$$

を満たすδ_n, 全ての$\delta \geq \delta_n$に対して以下が成り立つ.

$$\mathbb{E}\left[R\left(\hat{\mathbf{p}}_n, \boldsymbol{\eta}\right)\right] \leq 514(1+c_0^2)\left(\underbrace{\delta_n^2}_{\text{Variance}} + \underbrace{R\left(\tilde{\mathbf{p}}_n, \boldsymbol{\eta}\right)}_{\text{Bias}}\right) + \frac{c}{n},$$

- モデルの複雑さJ_BとBias $R\left(\tilde{\boldsymbol{p}}_n, \boldsymbol{\eta}\right)$の評価ができれば

  NPMLEの収束レートがわかる！

- Theorem 2は深層学習に限らず一般のNPMLEに対して成立

3.4 深層学習を用いたロジスティック回帰

- **真の条件付き確率が所属する関数クラスの設定**
  - Smoothness $\boldsymbol{\beta}$ のHölder 空間：

$$\mathcal{C}^{\beta}(D,Q) = \Bigg\{ f : D \subset \mathbb{R}^m \to \mathbb{R} : \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta} \|\partial^{\boldsymbol{\alpha}} f\|_{\infty} + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\boldsymbol{x},\boldsymbol{y}\in D,\, \boldsymbol{x}\neq\boldsymbol{y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x}-\mathbf{y}|_{\infty}^{\beta-\lfloor\beta\rfloor}} \leq Q \Bigg\}$$

  - Hölder 空間の元の合成で書ける関数の集合

$$\mathcal{G}_{\text{comp}}(r,\mathbf{d},\mathbf{t},\boldsymbol{\beta},Q) = \{ f = \boldsymbol{g}_r \circ \cdots \circ \boldsymbol{g}_0 : \boldsymbol{g}_i = (g_{ij})_j : [a_i,b_i]^{d_i} \to [a_{i+1},b_{i+1}]^{d_{i+1}}, \\ g_{ij} \in \mathcal{C}^{\beta_i}([a_i,b_i]^{t_i},Q), \text{for some} |a_i|,|b_i| \leq Q \}$$

3.4 深層学習を用いたロジスティック回帰

- **真の条件付き確率が所属する関数クラスの設定**

  - Hölder 空間の元の合成で書ける関数の集合：

$$\mathcal{G}_{\mathrm{comp}}(r, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, Q) = \{f = \boldsymbol{g}_r \circ \cdots \circ \boldsymbol{g}_0 : \boldsymbol{g}_i = (g_{ij})_j : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}},$$
$$g_{ij} \in \mathcal{C}^{\beta_i}([a_i, b_i]^{t_i}, Q), \text{for some} |a_i|, |b_i| \le Q\}$$

  - $f(x_1, x_2, x_3) = g_1 \circ \boldsymbol{g}_2(x_1, x_2, x_3) = g_{11}(g_{01}(x_1, x_3), g_{02}(x_1, x_2))$
    3次元　　2次元
  - は $d_0 = 3, t_0 = 2, d_1 = t_1 = 2, d_2 = 1$ のときの例
  - **背後に低次元構造がある**
  - 条件付き確率として直感的な例も含む（加法モデル）

$$f(x_1, x_2, x_3) = \sigma(f_1(x_1) + f_2(x_2) + f_3(x_3))$$
$$f(x_1, x_2, x_3) = \sigma(f_1(x_1, x_2) + f_2(x_2, x_3))$$

  - 真の条件付き確率は $\eta_k \in \mathcal{G}_{\mathrm{comp}}(r, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, Q), \quad k = 1, \ldots, K$

3.4 深層学習を用いたロジスティック回帰

☆深層学習における収束レート

Theorem 3（深層学習における収束レート）

以下の条件を満たすNNモデルの集合$\mathcal{F}(L, \boldsymbol{m}, s)$ を考える.

$$(\mathrm{i}) n\phi_n \lesssim \min_{i=1,\ldots,L} m_i,\ (\mathrm{ii}) s \asymp n\phi_n \log(n),\ (\mathrm{iii}) L \asymp \log(n)$$

このとき，$r, d, \boldsymbol{t}, \boldsymbol{\beta}, Q, K$ のみに依存する定数 C が存在して,

$$\mathbb{E}\left[R\left(\hat{\boldsymbol{p}}_n, \boldsymbol{\eta}\right)\right] \leq C\phi_n L \log^2(n)$$

が十分大きい n に対して成り立つ．ここで,

$\beta_i^* = \beta_i \prod_{l=i+1}^{r} (\beta_l \wedge 1)$, $\phi_n = \max_{i=0,\ldots,r} n^{-\frac{\beta_i^*}{\beta_i^* + t_i}}$ である.

- Variance の評価 ⇒ Theorem 1, Bias の評価 ⇒ Bos & S-H (2022)
- log-factor を無視すると, 収束レートは ϕ_n
- **収束レートは次元 d に依存していない（次元の呪いの回避）**
- 深層学習は背後の低次元構造（合成関数の構造）をうまく捉える

3.4 深層学習を用いたロジスティック回帰

- **wavelet推定量との比較**
  - ロジスティック回帰による二値分類を考える
  - $f_0(\boldsymbol{x}) := h(\boldsymbol{w}^\top \boldsymbol{x}),\ h \in \mathcal{C}^1([0,d], Q)$：真のロジット関数
  - $\eta(\boldsymbol{x}) := \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1+e^{-f_0(\boldsymbol{x})}}$：真の条件付き確率
  - この場合は推定対象は真のロジット関数
  - Theorem 3 から深層学習の収束レートが, Theorem 3 in Schmidt-Hieber (2020) から, wavelet 推定量の収束レートの下界がわかる

	深層学習	wavelet（下界）
収束レート	$n^{-\frac{1}{2}}$	$n^{-\frac{2}{2+d}}$

*この設定においては深層学習の収束レートは $n^{-2/3}$ まで速くなる（付録参照）

  - 深層学習は次元によらない収束レートだが, **wavelet は次元の呪いの影響を強くうける**

3.4 深層学習を用いたロジスティック回帰

- **wavelet推定量との比較**

	深層学習	wavelet (下界)
収束レート	$n^{-\frac{1}{2}}$	$n^{-\frac{2}{2+d}}$

  - 深層学習は次元によらない収束レートだが, wavelet は次元の呪いの影響を強くうける
  - $d \geq 6$ の状況において, wavelet の収束レートは $o_p(n^{-1/4})$

    を満たさない ⇒ **DMLが適用できない**
  - **深層学習を用いる動機と根拠を提供！**

- Remark: 今回は真の条件付き確率が合成関数の構造をもつと仮定したが, bias の評価ができている関数クラスであれば任意に置き換え可能
  - e.g. Besov 空間, 区分線形関数

3.4 深層学習を用いたロジスティック回帰

⭐ミニマックス最適性

- Theorem 3 で導出したレートは何らかの意味で最適か？

Theorem 4（ミニマックス最適性）

X_j が $[0,1]^d$ 上の Lebesgue 測度に関する密度をもつ分布に従うとし，
その密度が有界であるとする．このとき，任意の非負整数 r，
任意の $\boldsymbol{\beta} \in \mathbb{R}^r$，任意の i に対して，$t_i \leq \min(d_0, \ldots, d_{i-1})$ を満たす $\mathbf{d}, \mathbf{t}$
と任意の十分大きい $Q > 0$ に対して，ある正定数 c が存在して

$$\inf_{\hat{\boldsymbol{p}}_n} \sup_{\eta_k \in \mathcal{G}_{\mathrm{comp}}(r, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, Q),\ k=1,\ldots,K} \mathbb{E}\left[R\left(\hat{\boldsymbol{p}}_n, \boldsymbol{\eta}\right)\right] \geq c\phi_n$$

が成り立つ．下限は全ての推定量についてとる．

- Theorem 4 から，Theorem 3 で導出したレートは
  log-factor を除いてミニマックス最適（これ以上改善できない）

3.5 まとめ

- まとめ
  - ☆ロジスティック回帰において，分散を評価するための汎用的な不等式を示した (Theorem 1, 2)
    - モデルによらず一般的な最尤推定で成立
  - ☆Theorem 2 の適用例として, 深層学習におけるロジスティック回帰の収束レートを導出した (Theorem 3)
  - ☆Theorem 3 で導出したレートのミニマックス最適性を示した (Theorem 4)
  - ☆DMLやELWにおける傾向スコアへの応用が期待される
  - ☆本研究の結果は論文にまとめて投稿済み

    URL: https://arxiv.org/abs/2401.12482

参考文献

Sara A Geer, Sara van de Geer, and D Williams. Empirical Processes in M-estimation, volume 6. Cambridge university press, 2000.

Evarist Giné and Richard Nickl. Mathematical Foundations of Infinite-Dimensional Statistical Models, volume 40. Cambridge University Press, 2016.

Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classifi-cation. Neural Networks, 138:179–197, 2021.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. The Annals of Statistics, 48(4):1875 – 1897, 2020.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In International Conference on Learning Representations, 2019.

Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68

参考文献

Imaizumi, M., & Fukumizu, K. (2019). Deep neural networks learn non-smooth functions effectively. In Proceedings of the International Conference on Artificial Intelligence and Statistics.

Ilsang Ohn, Yongdai Kim; Nonconvex Sparse Regularization for Deep Neural Networks and Its Optimality. Neural Comput 2022; 34 (2): 476–517.

Bilodeau, B., Foster, D. J. and Roy, D. M. (2023). Minimax rates for conditional density estimation via empirical entropy. The Annals of Statistics 51 762 – 790.

Devroye, L. (1982). Any discrimination rule can have arbitrarily bad probability of error for finite sample size. IEEE Transactions on Pattern Analysis and Machine Intelligence 4, 154–157.

Seth Flaxman, Yee Whye Teh, and Dino Sejdinovic. Poisson intensity estimation with reproducing kernels. International Conference on Artificial Intelligence and Statistics, 2017.

Stamatina Lamprinakou, Mauricio Barahona, Seth Flaxman, Sarah Filippi, Axel Gandy, Emma J. McCoy, BART-based inference for Poisson processes, Computational Statistics &
Data Analysis, Volume 180, 2023

Yara, A. and Terada, Y. (2023). Nonparametric logistic regression with deep learning. arXiv.

付録

0. 目次

時間の都合上一部内容を割愛

1.1 深層学習

- **理論研究の重要性**
  - 素朴な疑問：

    Q. あらゆる問題は深層学習で解決するのでは？

    ⇒ 実は, 理論的に**否定的な解**が与えられている！

    (**No-free-lunch theorem**, **Slow rates of convergence**)
  - **No-free-lunch theorem の主張 (informal)：**

    どんな学習アルゴリズムにも苦手な問題（分布）

    が存在して性能が悪化する
  - **すなわち、全ての問題に対して万能な学習法は存在しない**

    ⇒ 様々な手法の開発と, どういうときに各手法がうまくいく

    か調べることが重要！（統計的学習理論）

1.2 深層学習の汎化誤差解析

- **No-free-lunch theorem**
  - 例として, 二値分類問題を考える
    - $\mathcal{X} \subset \mathbb{R}^d$：入力空間, $\mathcal{Y} = \{0, 1\}$：出力空間
    - P：$\mathcal{X} \times \mathcal{Y}$ 上の確率分布、$P_{\boldsymbol{X}}$：X の周辺分布
    - $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \sim_{i.i.d.} P$
    - $\mathcal{D}_n := \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$：観測データ
  - **目標：新たなデータ$\boldsymbol{X}$が与えられたときに、Y を良く予測するような関数 $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ を$\mathcal{D}_n$から構成する（学習）**
    - $R_n = \mathbb{P}(\hat{f}_n(\boldsymbol{X}) \neq Y)$：誤判別率（誤った分類をする確率）
    - $R^* = \inf_{f:\mathcal{X}\to\mathcal{Y} \text{ is measurable}} \mathbb{P}(f(\boldsymbol{X}) \neq Y)$：

      最適な分類に対する誤判別率

1.2 深層学習の汎化誤差解析

- **No-free-lunch theorem**
  - 観測データ$\mathcal{D}_n$から決定関数$\hat{f}_n$を得るアルゴリズムを学習法と呼ぶ

Theorem 0 (No-free-lunch theorem, Devroye, 1982)

任意の学習法に対して

$$\sup_{P\ :\ \text{all distribution of } (\boldsymbol{X},Y),\ R^*=0} R_n \geq \frac{1}{2}$$

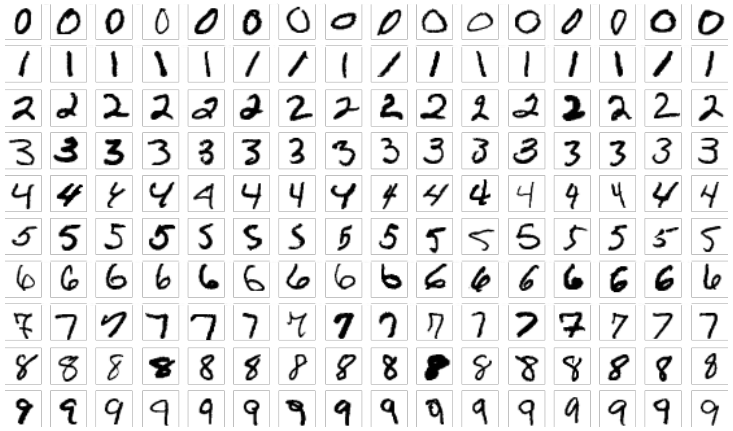
が成り立つ。

  - **「どんな学習法に対しても"苦手な"問題（分布）が存在して性能が悪化する」と言っている**
  - **すなわち, 全ての問題に対して万能な学習法は存在しない**

⇒ 様々な手法の開発と, どういうときに各手法がうまくいくか調べることが重要！（統計的学習理論）

1.3 分類問題と条件付き確率の推定

- **分類問題の基本的な設定**
  - MNISTの例
    - 0~9までの手書き数字の判別問題を考える



MNISTデータセット

- $f($  $) = 7,\ f($  $) = 8$

  となるような"良い" f（分類器）を見つけたい！

- 良さとは？ ⇒ **損失関数**
- どうやって見つける？ ⇒ **経験誤差最小化**

1.3 分類問題と条件付き確率の推定

- **学習の方法**
  - 学習 = 観測データ $\mathcal{D}_n$ から決定関数 f を求めること
  - 学習はERMによって行う
    - 最適化については考えず $\hat{f}$ が得られているものとする

Remark: ERMと推定量の誤差を含んだ形に主結果を拡張することは容易にできる

Remark: 深層学習では local minima と global minima は十分近い

- さまざまな損失関数が用いられる
  - ヒンジロス：$\ell(y, f(x)) := 0 \vee (1 - yf(x))$
  - ロジスティックロス：$\ell(y, f(x)) := \log(1 + e^{-yf(x)})$

1.5 セミパラメトリック推定と条件付き確率の推定

- **Double/Debiased Machine Learning (DML) との関係**
  - 因果推論の文脈において, 傾向スコア（＝共変量を条件付けたときの割り付けの**条件付き確率**）は因果効果を推定する際の局外母数として重要な役割を果たす
  - 局外母数が非常に複雑な場合は, 機械学習 (ML) の方法を用いることで自然な推定量を与えることができる
  - しかし, MLを用いたことによる bias や overfitting は推定量に悪影響を与える
    - 収束が遅くなる（$n^{-1/2}$-一致性が成り立たない）

  ⇒ DML (Chernozhukov et al., 2018) でこの問題を解決！

    - Neyman-orthogonal moments / scores
    - data-splitting の亜種 cross-fitting

1.5 セミパラメトリック推定と条件付き確率の推定

- **Double/Debiased Machine Learning (DML) との関係**
  - DML においては, 局外母数のノンパラメトリック推定
    （この部分を機械学習手法で推定することが想定される）に
    $o_p(n^{-1/4})$の収束レートを示す必要がある

    ⇒ 例えば, 傾向スコアを深層学習を用いて推定する場合を考えると

    深層学習を用いたロジスティック回帰の汎化誤差解析が必須！
  - 本研究では深層学習を用いた条件付き確率の推定における
    収束レートを導出

1.5 セミパラメトリック推定と条件付き確率の推定

- **IPW推定量とELW推定量**
  - 欠測データ問題を考える
  - X：共変量, Y：応答変数, D：Y の欠測の有無（0なら欠測）
  - 観測データ $\{(D_i, D_i Y_i, X_i)\}_{i=1}^N$ から $\mathbb{E}[Y]$ を推定したい
  - 傾向スコア $\pi(X) = \mathbb{P}(D = 1 \mid X)$ が推定できれば,,,

$$\hat{\theta}_{\mathrm{IPW}} = \frac{1}{N} \sum_{i=1}^{N} \frac{D_i Y_i}{\pi(X_i)}$$

として推定できる！（Inverse Probability Weighting (IPW) 推定量）

  - **しかし, 傾向スコアがゼロに近いと非常に不安定**

Q. 傾向スコアを深層学習で推定して問題ないのか？

1.5 セミパラメトリック推定と条件付き確率の推定

- **IPW推定量とELW推定量**

Q. 傾向スコアを深層学習で推定して問題ないのか？

- 実際に深層学習では確率がゼロに推定されることがよくある

⇒ Empirical Likelihood Weighting (ELW) 推定量を用いればOK!

$$\hat{p}_i = \frac{1}{n} \frac{D_i}{1 + \lambda(\hat{\alpha})(\pi(X_i) - \hat{\alpha})}, \quad \hat{\theta}_{\mathrm{ELW}} = \sum_{i=1}^{N} \hat{p}_i Y_i$$

- 実際, 分母 $1 + \lambda(\hat{\alpha})(\pi(X_i) - \hat{\alpha})$ はゼロにならない
- 詳細は Liu & Fan (2023, JRSS B) を参照
- **ELW推定量とDMLを組み合わせることで, 傾向スコアが複雑な場合も深層学習を用いた強力な推定量を構成できる！**

2.1 誤判別率の収束についての研究

- **誤判別率の収束に関する結果が主流**
  - 誤判別率 = 分類器がどれくらい誤った分類をするか

$$\mathbb{P}\left(Y \neq \hat{f}_n(\boldsymbol{X})\right)$$

  - 誤判別率は 0-1 損失の期待値
    - 0-1 損失：$l(y, f(\boldsymbol{x})) = \mathbb{1}\left(y \neq f(\boldsymbol{x})\right)$

  ⇒ 0-1 損失を最小化したい
  - 0-1 損失は不連続で最適化が困難

  ⇒ 凸な損失で代用（代理損失）
    - e.g. ヒンジロス，ロジスティックロス
  - モデルとして深層学習を用いて，さまざまな代理損失を用いた場合の誤判別率の収束に関する研究が行われている

  （Hu et al., 2020; Kim et al., 2021）

2.1 誤判別率の収束についての研究

- **誤判別率の収束に関する結果が主流**
  - Hu et al. (2020, arXive:2001.06892)
    - 評価指標：誤判別率
    - 損失関数：0-1損失，ヒンジロス
    - 母集団分布に関する仮定：
      1. Y を与えたときの $\boldsymbol{X}$ の条件付き密度関数が DNN で表現可能
      2. 決定境界付近にデータが集中する確率が十分小さい
  - Kim et al. (2021, Neural Networks)
    - 評価指標：誤判別率
    - 損失関数：ヒンジロス，ロジスティックロス
    - 母集団分布に関する仮定：
      1. 滑らかな決定境界をもつ
      2. 真の条件付き確率が滑らか
      3. マージン条件：入力 $\boldsymbol{X}$ の密度が決定境界付近で小さい

2.2 条件付き確率の収束についての研究

- **条件付き確率の推定に関する研究**
  - Ohn and Kim (2022, Neural computation)
    - 評価指標：KL ダイバージェンス
    - 損失関数：ロジスティックロス
    - 仮定：真の条件付き確率が0と1から十分離れている

$$0 < \exists c, C < 1;\ c < \eta_k < C, k = 1, \ldots, K$$

  - Bos & Schmidt-Hieber (2022, Electronic Journal of Statistics)
    - 評価指標：打ち切り KL

$$R_B(\boldsymbol{\eta}, \hat{\boldsymbol{p}}_n) := \mathbb{E}\left[KL_B\left(\boldsymbol{\eta}(\boldsymbol{X}), \hat{\boldsymbol{p}}_n(\boldsymbol{X})\right)\right],$$

$$\mathrm{KL}_B\left(\boldsymbol{\eta}(\boldsymbol{X}) \,\|\, \hat{\boldsymbol{p}}_n(\boldsymbol{X})\right) := \boldsymbol{\eta}(\boldsymbol{X})^\top \left(B \wedge \log\left(\frac{\boldsymbol{\eta}(\boldsymbol{X})}{\hat{\boldsymbol{p}}_n(\boldsymbol{X})}\right)\right).$$

    - 損失関数：負の対数尤度
    - 仮定：真の条件付き確率が属するクラス $\mathcal{H}$ が α-SVB

$$\exists C > 0,\ \forall \boldsymbol{p} \in \mathcal{H};\ \mathbb{P}_{\boldsymbol{X}}\left(p_k(\boldsymbol{X}) \geq t\right) \leq Ct^\alpha$$

$$\text{for all } t \in (0, 1] \text{ and all } k \in \{1, \ldots, K\}$$

2.2 条件付き確率の収束についての研究

- **条件付き確率の推定に関する研究**
  - Biloudeau et al. (2023, The Annals of Statistics)
    - この研究は条件付き確率だけでなく，一般の条件付き密度推定に対して議論している
    - 評価指標：KL
    - 損失関数：負の対数尤度
    - 仮定：真の条件付き確率がモデルに含まれる (well specified)
    - 一般の条件付き密度推定問題に対してミニマックス上限，下限を導出した
      - データ数だけでなく，クラス数に関する収束レートを導出
      - モデルが複雑すぎる場合，最尤推定量がsuboptimalであることを示した

2.3 先行研究の問題点

- 推定量が陽な表現をもたない場合において，条件付き確率のノンパラメトリック推定の理論的性質についての議論は少ない
- Ohn & Kim (2022): 限定的な状況下での結果
- Bos & Schmidt-Hieber (2022): KLダイバージェンスの代わりに打ち切りKLダイバージェンスを評価することで仮定を取り除いて少し弱い意味での収束を示した
  - 導出したレートの最適性の議論はされていない
- Bilodeau et al. (2023): MLEを変換した量を用いてKLダイバージェンスに関する収束を示した.
  - ただし，真値がモデルに含まれている (well specified) な状況

本研究

- **分類問題における NPMLE の理論的性質の解明**
- **深層学習における NPMLE の収束レートの導出**
- **ミニマックス最適性の証明**

3.1 主結果の概要

- ☆**Theorem1, 2: 一般のロジスティック回帰における最尤推定の分散を評価する不等式を導出**
  - モデルによらす一般的に成立
  - KLの代わりにHellinger距離を用いることで弱い条件の下証明
  - Van de Geer (2000) のテクニックを用いたシンプルな証明
- ☆**Theorem 3: Theorem 1 を応用して深層学習を用いたロジスティック回帰の収束レート（真値への収束の速さ）を導出**
  - 深層学習が次元の呪いを回避することを示唆
- ☆**Thorem 4: Theorem 3 で導出した収束レートのミニマックス最適性を証明**
  - 深層学習の収束レートはこれ以上改善できない

3.2 記号と仮定

- **記号**
  - $\bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}, \delta) = \left\{ \frac{\boldsymbol{p}+\tilde{\boldsymbol{p}}}{2} : R\left(\frac{\boldsymbol{p}+\tilde{\boldsymbol{p}}}{2}, \tilde{\boldsymbol{p}}\right) \leq \delta^2,\ \boldsymbol{p} \in \mathcal{F}_n \right\}$
  - $N_{p,B}(\delta, \mathcal{F}, Q)$:
    関数クラス $\mathcal{F}$ の $L^p(Q)$ ノルムに関する δ-bracketing number
    - 関数クラス $\mathcal{F}$ のある種の"大きさ"や"複雑度"を表す
  - $J_B(\delta, \bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}_n, \delta), \mu) = \int_{\delta^2/(2^{13}c_0)}^{\delta} \sqrt{\log N_{2,B}\left(u, \bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}_n, \delta), \mu\right)}\, du \vee \delta$
    - μ は $\{1, \ldots, K\}$ 上の数え上げ測度とP_Xの直積測度
    - $J_B(\delta, \bar{\mathcal{F}}_n^{1/2}(\tilde{\mathbf{p}}_n, \delta), \mu)$ は $\tilde{\boldsymbol{p}}_n$ の近くで見たときの局所的な複雑度

    ⇒ 局所的な複雑度を用いることで tight な不等式を導く

    (Bartlett, 2005)

3.3 ノンパラメトリックロジスティック回帰の一般理論

☆オラクル不等式

Theorem 1（オラクル不等式 確率ver.）

Kクラスの分類問題を考える. $\hat{\boldsymbol{p}}_n$をNPMLEとする.

$\Psi(\delta) \geq J_B(\delta, \bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}_n, \delta), \mu)$を$\Psi(\delta)/\delta^2$が$\delta$の非増加関数となるようにとる. δ_n とある普遍定数 c は

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n) \qquad (1)$$

を満たすとする. このとき

$$\mathbb{P}\left(R\left(\hat{\mathbf{p}}_n, \boldsymbol{\eta}\right) > 514(1+c_0^2)(\delta^2 + R\left(\tilde{\mathbf{p}}_n, \boldsymbol{\eta}\right))\right) \leq c\exp\left(-\frac{n\delta^2}{c^2}\right)$$

- Theorem1, 2 は深層学習に限らず任意のNPMLEで成り立つ

3.3 ノンパラメトリックロジスティック回帰の一般理論

☆オラクル不等式

Theorem 1（オラクル不等式 確率ver.）

Kクラスの分類問題を考える. $\hat{\boldsymbol{p}}_n$をNPMLEとする.

$\Psi(\delta) \geq J_B(\delta, \bar{\mathcal{F}}_n^{1/2}(\tilde{\boldsymbol{p}}_n, \delta), \mu)$を$\Psi(\delta)/\delta^2$が$\delta$の非増加関数となるようにとる. δ_n とある普遍定数 c は

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n) \qquad (1)$$

を満たすとする. このとき

$$\mathbb{P}\left(R\left(\hat{\mathbf{p}}_n, \boldsymbol{\eta}\right) > 514(1+c_0^2)(\delta^2 + R\left(\tilde{\mathbf{p}}_n, \boldsymbol{\eta}\right))\right) \leq c\exp\left(-\frac{n\delta^2}{c^2}\right)$$

- モデルの複雑さJ_BとBias $R\left(\tilde{\boldsymbol{p}}_n, \boldsymbol{\eta}\right)$の評価ができれば

  NPMLEの収束レートがわかる！

3.4 深層学習を用いたロジスティック回帰

- **仮説空間（推定モデル）の設定**
  - $\sigma(x) := \max(0, x)$：活性化関数，$\mathbf{v} = (v_1, \ldots, v_r) \in \mathbb{R}$：バイアス
  - $\sigma_{\mathbf{v}}(y_1, \ldots, y_r) = (\sigma(y_1 - v_1), \ldots, \sigma(y_r - v_r))$：

    バイアス込み活性化関数
  - L：隠れ層の数，$\boldsymbol{m} = (m_0, \ldots, m_{L+1})$：各層の幅，
  - $W_j \in \mathbb{R}^{m_{j+1} \times m_j}$：重み行列，$\boldsymbol{v}_i \in \mathbb{R}^{m_i}$：バイアスベクトル，
  - $\boldsymbol{\Phi}$：ソフトマックス関数

3.4 深層学習を用いたロジスティック回帰

- **仮説空間（推定モデル）の設定**
  - 順伝播型ニューラルネットワーク（FFNN）：

$$f : \mathbb{R}^{m_0} \to \mathbb{R}^{m_{L+1}},\ \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \boldsymbol{\Phi} W_L \sigma_{\boldsymbol{v}_L} W_{L-1} \sigma_{\boldsymbol{v}_{L-1}} \cdots W_1 \sigma_{\mathbf{v}_1} W_0 \boldsymbol{x} \quad (\star)$$

$$\mathcal{F}(L, \boldsymbol{m}) := \left\{ \text{f of the form } (\star) : \max_{j=0,\ldots,L} \|W_j\|_\infty \vee |\boldsymbol{v}_j|_\infty \leq 1 \right\}$$

  - $\|W_j\|_0$：W_j の非ゼロの成分の数,

  - $|\boldsymbol{v}_j|_0$：$\boldsymbol{v}_j$ の非ゼロの成分の数,

$$\mathcal{F}(L, \boldsymbol{m}, s) := \left\{ f \in \mathcal{F}(L, \boldsymbol{m}) : \textstyle\sum_{j=0}^{L} \left(\|W_j\|_0 + |\boldsymbol{v}_j|_0 \leq s \right) \right\}$$

3.4 深層学習を用いたロジスティック回帰

- **NNのスパース性に関して**
  - NNのスパース性は本質的な仮定ではなく技術的なもの
  - 仮定が必要な理由：Variance の評価に NN の非ゼロのパラメータの数を用いているため
  - 以下の方法で仮定を除去できる可能性がある
    - Lu et al. (2021): 本研究で用いた近似理論をより洗練させた
    - Kohler & Langer (2019, 2020) の方法
    - Variance の評価をパラメータ数ではなく, ノルムで行う

3.4 深層学習を用いたロジスティック回帰

- **先行研究との比較**
  - 多くの場合クラス数 K は事前に固定されるため，本研究の不等式 Theorem 3, 4ではK は定数とした
  - Bos & Schmidt-Hieber (2022)の比較するためにK を含む収束レートを示す
  - 本研究：$\max_{i=0,\ldots,r} K^{\frac{\beta_i^*+2t_i}{\beta_i^*+t_i}} n^{-\frac{\beta_i^*}{\beta_i^*+t_i}}$
  - Bos & Schmidt-Hieber (2022): $\max_{i=0,\ldots,r} K^{\frac{\beta_i^*+3t_i}{\beta_i^*+t_i}} n^{-\frac{\beta_i^*}{\beta_i^*+t_i}}$

  - **Hellinger距離を評価することによってK に関する収束レートを少し改善している！**

3.4 深層学習を用いたロジスティック回帰

- **ミニマックス最適性に関する補足**
  - 本研究で示したミニマックス下界はデータ数に関するもの
  - Bilodeau et al. (2023) で示されたHölderクラスに対するミニマックスレートは $K^{\frac{2d}{d+\beta}} n^{-\frac{1}{d+\beta}}$
  - 一方で，本研究でHölderクラスになる特殊ケースを考えるとミニマックス下限は $K^{-1} n^{-\frac{1}{d+\beta}}$
  - **クラス数に関するミニマックス下限を示していない**
    - 深層学習もクラス数に関してミニマックス最適性を示せてない
  - **本研究ではクラス数は定数として固定して考えたのでこれ以上詳細な解析は行わなかったが, Bilodeau et al. (2023) の方法を使えばクラス数に関してミニマックス下限が改善できる可能性がある**

3.4 深層学習を用いたロジスティック回帰

- **強い仮定を置くことによる収束レートの改善**
  - 真の条件付き確率が

$$0 < \exists c, C < 1;\ c < \eta_k < C, k = 1, \ldots, K$$

  - を満たすなら, 深層学習の収束レートは$\max_{i=0,\ldots,r} n^{-2\beta_i^*/(2\beta_i^*+t_i)}$まで改善できる
  - しかし, 共変量が与えられたときに, 確実にクラスが分かるケースが存在しないことを仮定している