Deep Learning Theory Team

Taiji Suzuki

深層学習理論チーム　鈴木 大慈

RIKEN Center for Advanced Intelligence Project

概要

チームの目標:

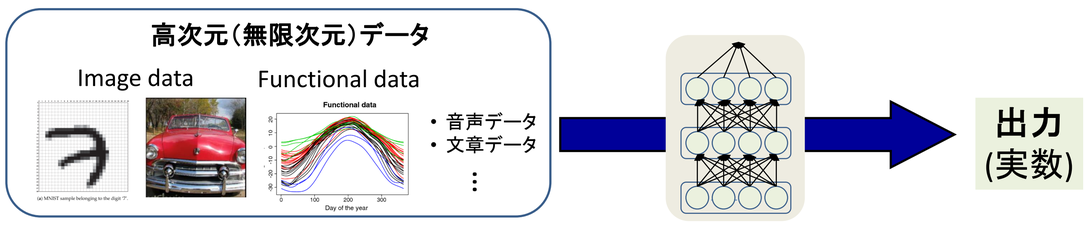
- 深層学習の理論的理解の促進
- 理論に基づいた新しい手法の開発
- 関連する機械学習問題の解法と最適化手法の開発

成果のまとめ

- 深層学習はなぜ高次元データでも動く？→非等方的滑らかさを用いた解析.
- ニューラルネットの理論保証付き最適化. →平均場設定における新しい最適化の枠組み.
- 広い横幅のNNの挙動解析. →NTKを用いたSGDの解析，最適推定誤差の保証.
- 非ユークリッド空間における深層学習の実現→双曲空間におけるNNのリッジレット変換.

深層学習による次元の呪いの回避

- Suzuki&Nitanda: Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. NeurIPS2021, spotlight.
- Okumoto&Suzuki: Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. ICLR2022, spotlight.

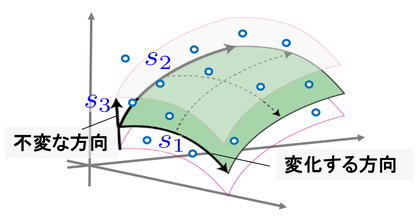


示したこと:

- **真の関数が方向によって異なる滑らかさを持つ状況ではDNNは重要な方向を見つけ，次元の呪いを回避する.**
- **一方で，浅い学習法は次元の呪いを受ける.**

- 教師生徒設定における大域的最適化と次元の呪いの回避: [Suzuki and Akiyama: ICLR2021, spotlight].
- 深層学習の浅層学習への優位性: [Hayakawa and Suzuki: Neural Networks 2020].

s_1, s_2, s_3: 滑らかさ

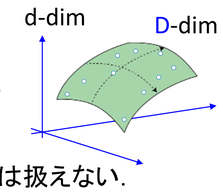


通常のバウンド: 次元がレートに依存（次元の呪い）.

$$n^{-\frac{2s}{2s+d}}$$

多様体回帰:
- データが低次元多様体上に分布することを仮定.
- 低次元多様体の次元のみに依存.
  → 次元の呪いを回避.

$n^{-\frac{2s}{2s+D}}$



問題点: データが低次元多様体からはみ出る場合は扱えない.

解決案:
- 真の関数の滑らかさが方向によって大きく異なることに着目.
- 滑らかでない方向が少なければ次元の呪いが回避できるはず.

真の関数のモデル:

非等方的Besov空間の元の合成関数

$$f^\circ(x) = h_H \circ \cdots \circ h_1(x)$$

$h_\ell : \mathbb{R}^{m_\ell} \to \mathbb{R}^{m_{\ell+1}}$: 非等方的Besov空間 $(B^{s_\ell}_{p,q})$.

➢ 滑らかさが方向によって異なる関数空間: s_iは座標軸i方向への滑らかさ.
➢ 合成することで様々な形状を実現 (多様体上の関数等)
→ さらにその無限次元入力への拡張も考える: γ-平滑関数空間

Def. (非等方的Besov空間)

$$\Delta_h^r(f)(x) := \Delta_h^{r-1}(f)(x+h) - \Delta_h^{r-1}(f)(x), \quad (h \in \mathbb{R}^d)$$
$$\Delta_h^0(f)(x) := f(x)$$
$$w_{r,p}(f,t) = \sup_{h\in\mathbb{R}^d:|h_i|\le t_i} \|\Delta_h^r(f)\|_p$$
$$s = (s_1,\dots,s_d) \in \mathbb{R}_{++}^d$$
$$|f|_{B^s_{p,q}} := \begin{cases} \left(\sum_{k=0}^\infty [2^{ks} w_{r,p}(f,(2^{-k/s_1},\dots,2^{-k/s_d}))]^q\right)^{1/q} & (q<\infty), \\ \sup_{k\ge 0} 2^{ks} w_{r,p}(f,(2^{-k/s_1},\dots,2^{-k/s_d})) & (q=\infty). \end{cases}$$
$$\|f\|_{B^s_{p,q}} := \|f\|_p + |f|_{B^s_{p,q}}$$

直感 $\|f\|_{B^s_{p,q}} = \|f\|_{L^p} + \sum_{i=1}^d \left\|\frac{\partial^{s_i} f}{\partial x_i^{s_i}}\right\|_{L^p}$

深層学習の推定誤差

$$\mathbb{E}[\|\hat{f} - f^\circ\|_{L_2(P_X)}^2] \lesssim n^{-\frac{2\tilde{s}}{2\tilde{s}+1}} \mathrm{poly}(\log(n))$$

ただし, $\tilde{s} = (s_1^{-1} + s_2^{-1} + \cdots)^{-1}$.

次元が現れない！

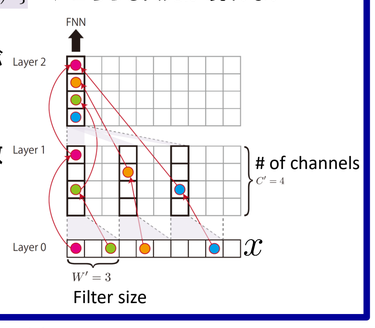
混合平滑Besovの場合: $\mathbb{E}[\|\hat{f} - f^\circ\|_{L_2(P_X)}^2] \lesssim n^{-\frac{2s_1}{2s_1+1}} (\log n)^{\frac{2}{q}+2} \max\{(\log n)^{4/q}, (\log n)^4\}$ → こちらも次元が現れない.

浅い学習方法との比較:

深層	浅層
$n^{-\frac{2\tilde{s}}{2\tilde{s}+1}}$	$n^{-\frac{2s_1}{2s_1+d}}$ （次元の呪いを受ける）

- **特徴抽出能力の重要性を理論的に正当化**
- **浅い学習方法は一番滑らかでない方向の滑らかさ(s_1)が支配的で，次元の呪いを受ける.**
- **既存研究である多様体回帰の解析を用いると浅層学習のレートが出てしまい，深層学習の良さがしっかり解析できていなかった.**

さらに，重要な特徴量が高次元特徴量の中にまばらに存在する状況でもCNNにより重要な特徴量を抽出できることも示した.
→CNNによる特徴量選択.



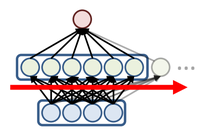
平均場解析による深層NNの最適化

- Nitanda,Wu,Suzuki: Particle Dual Averaging: Optimization of Mean Field Neural Networks with Global Convergence Rate Analysis. NeurIPS2021.
- Oko,Suzuki,Nitanda,Wu: Particle Stochastic Dual Coordinate Ascent: Exponential convergent algorithm for mean field neural network optimization. ICLR2022.
- Nitanda,Wu,Suzuki: Convex Analysis of the Mean Field Langevin Dynamics. AISTATS2021.

2層ニューラルネットワーク:

$$f(x) = \frac{1}{M}\sum_{j=1}^M r_j \sigma(w_j^\top x)$$

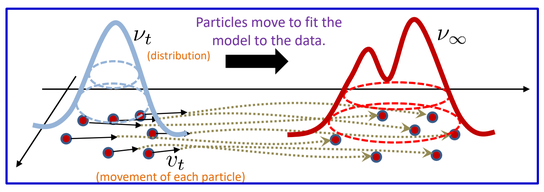
パラメータ $(r_j, w_j)_{j=1}^M$ に関して非線形：非凸最適化



過剰パラメータ化（平均場）:

$$f(x) = \frac{1}{M}\sum_{j=1}^M r_j \eta(w_j^\top x) \xrightarrow{M\to\infty} \int r\sigma(w^\top x)\mathrm{d}\nu(r,w)$$

パラメータの「分布」ν に関する平均：分布に関して線形



正則化学習の再定式化:

$$\min_{\nu:\mathcal{P}(\Theta)} \frac{1}{n}\sum_{i=1}^n \ell\Big(\mathbb{E}_{\theta\sim\nu}[h_\theta(x_i)], y_i\Big) + \lambda\mathbb{E}_\nu[\|\theta\|^2]$$

確率測度 / ℓ:滑らかな損失関数, h_θ: パラメータθのニューロン / L2-正則化

i.e., $h_\theta(x) = r\sigma(w^\top x)$ for $\theta = (r,w)$

負エントロピー正則化を加える：分布は密度を持ち，最適化しやすくなる.

$$\min_{q:\text{prob.density}} \frac{1}{n}\sum_{i=1}^n \ell\Big(\mathbb{E}_q[h_\theta(x_i)], y_i\Big) + \lambda_1\mathbb{E}_q[\|\theta\|^2] + \lambda_2\mathbb{E}_q[\log(q)]$$

$\lambda_2\mathrm{KL}(\nu, N(0, \lambda_2/\lambda_1 I))$ KL-div from a Gaussian distribution.

密度関数に関して凸関数
→ 通常の凸最適化手法を援用できる.

難しさ:
- ➢ 基本的に無限次元最適化.
- ➢ 期待値の陽な記述は得られない.

→ 解決策:
- 粒子を用いた近似.
- 勾配ランジュバン動力学等でサンプリング.

粒子双対平均化法 (Particle Dual Averaging; PDA)

$$\min_{q:\text{prob.density}} \underbrace{\frac{1}{n}\sum_{i=1}^n \ell\Big(\mathbb{E}_q[h_\theta(x_i)], y_i\Big) + \lambda_1\mathbb{E}_q[\|\theta\|^2] + \lambda_2\mathbb{E}_q[\log(q)]}$$

近似 ↓

qに関する線形汎関数で近似 (勾配を用いる)
$\mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)]$ (線形近似; $\bar{g}^{(t)}$は基本的に勾配)
$\bar{g}^{(t)}$の決定に双対平均化法のルールを用いる

$$\min_{q:\text{prob.density}} \mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)] + \lambda_2\mathbb{E}_q[\log(q)]$$

解: $q^{(t+1)}(\theta) \propto \exp(-\bar{g}^{(t)}(\theta)/\lambda_2)$ 具体形が得られる.

→ この分布からは以下の勾配ランジュバン動力学を用いてサンプリング可能:

時間離散化 ↓

$$\mathrm{d}\theta_t = -\nabla(\bar{g}^{(t)}(\theta)/\lambda_2)\mathrm{d}t + \sqrt{2}\mathrm{d}\xi_t.$$
$$\theta_k = \theta_{k-1} - \eta\nabla\bar{g}^{(t)}(\theta)/\lambda_2 + \sqrt{2\eta}\xi_{k-1}$$

計算量解析:
1. **外側ループ:** $\mathcal{L}(\hat{q}^{(t)}) - \mathcal{L}(q^*) \le O(1/t)$
2. **内側ループ:** $T_t = \tilde{O}\left(t^2\exp(8/\lambda_2)/(\lambda_1\lambda_2)\right)$ (GLDによる)

⇒合計: $O(\epsilon^{-3})$の勾配アップデートで十分.

➢ **初の多項式オーダー最適化手法**

粒子確率的双対座標上昇法
(Particle Stochastic Dual Coordinate Ascent; P-SDCA)

主問題

$$\min_p P(p) = \frac{1}{n}\sum_{i=1}^n \ell_i\left(\int p(\theta)h_i(\theta)\right) + \lambda_1\int\|\theta\|^2 p(\theta)\mathrm{d}\theta + \lambda_2\int p(\theta)\log(p(\theta))\mathrm{d}\theta$$

‖ by Fenchelの双対定理

双対問題

$$\ell_i^*(g) := \sup_{u\in\mathbb{R}}\{ug - \ell_i(u)\}$$
$$-\min_{g\in\mathbb{R}^n} D(g) = \frac{1}{n}\sum_{i=1}^n \ell_i^*(g_i) + \lambda_2\log\left(\int q[g](\theta)\mathrm{d}\theta\right)$$

ただし $q[g](\theta) := \exp\left\{-\frac{1}{\lambda_2}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta)g_i + \lambda_1\|\theta\|^2\right)\right\}$

- 双対変数の座標をランダムに選択し，その座標に関して最適化.
→双対座標上昇法

計算量解析:
双対ギャップϵ_Pを達成するのに必要な外側ループ数:

$$t_{\text{end}} = 2\left(n + \frac{1}{\lambda_2\gamma}\right)\log\left(\frac{nC}{\epsilon_P}\right)$$

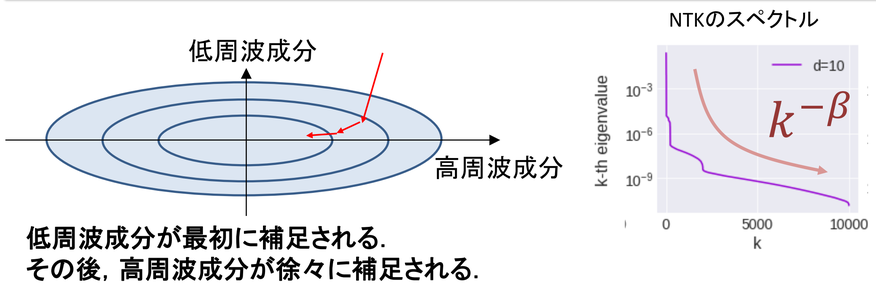
- ➢ 指数オーダーでの収束を達成
- ➢ サンプルサイズnへの依存を緩和

Neural Tangent Kernelの理論

Nitanda&Suzuki: Fast Convergence Rates of Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime, ICLR2021 (oral). **Outstanding paper award**.

ニューラルネットワークの最適化は非凸最適化
→ 横幅の広いネットワークは「凸」的になり大域的最適解へ収束

以下を理論的に示した:
- **確率的最適化により最適な推定レートを達成可能.**
- **ネットワーク固有の周波数成分のスペクトルが学習効率を決める.**



低周波成分が最初に補足される.
その後，高周波成分が徐々に補足される.

$$\mathbb{E}[\|f_T - f^*\|_{L_2}^2] \le \epsilon_M + O(T^{-\frac{2r\beta}{2r\beta+1}})$$

横幅$M\to\infty$で0に収束する項

速い学習レート($O(1/\sqrt{T})$より速い)
→ Minimax最適レート

β: NTK (Neural Tangent Kernel) の固有値の減衰レート
r: 真の関数の“複雑さ” (RKHSとの相対的な位置関係)

➢ 初期値で決まるカーネルの複雑さと真の関数との位置関係でレートが決まる.
➢ 複数層の学習はMultiple Kernel Learningに相当する.

モデル: $Y = f^*(X) + \epsilon$ (ノイズありの観測)

$$f_{a,W}(x) = \frac{1}{\sqrt{M}}\sum_{j=1}^M a_j\eta(w_j{}^\top x)$$ (一層目と二層目を学習)

- 目的関数をオンライン型の確率的最適化で最適化
- Tは更新回数=観測したデータ数

目的関数: 期待損失 / 初期値からのずれ

$$L(a,W) = \mathbb{E}[(Y - f_{a,W}(X))^2] + \frac{\lambda}{2}(\|a - a^{(0)}\|^2 + \|W - W^{(0)}\|_{\mathrm{F}}^2)$$

双曲空間上のリッジレット変換の理論

Sonoda,Ishikawa,Ikeda: Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis. arXiv:2203.01631.

目標：双曲空間上のデータを受け取るNN を設計する.
結果：双曲空間上のリッジレット変換と再構成公式を導出.

ユークリッド空間のNNの各ニューロン:

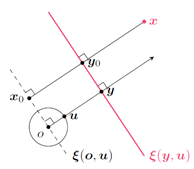
$$\sigma(\langle a,x\rangle - b).$$

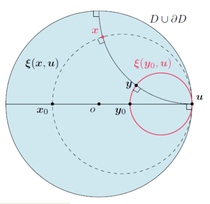
$\boldsymbol{a} = r\boldsymbol{u}, b = r\langle\boldsymbol{u},\boldsymbol{y}\rangle$
とすると，$\langle a,x\rangle - b$はyを通りuに直交した平面とxとの距離$rd(x,\xi(y,u))$に対応.

Poincare円盤上では:

$$\sigma(rd(x,\xi(y,u))).$$

のように“ホロ球”$\xi(y,u)$からの距離に置き換える.





主結果の系: Poincare円盤D^m上の双曲NNの構成

- 連続ホロ球型 HNN（continuous horospherical hyperbolic NN）

$$S[\gamma](\boldsymbol{x}) := \int_{\mathbb{R}\times\partial D^m\times\mathbb{R}} \gamma(a,\boldsymbol{u},b)\sigma(aA(\boldsymbol{x},\boldsymbol{u}) - b)e^{\varrho A(\boldsymbol{x},\boldsymbol{u})}\mathrm{d}a\mathrm{d}\boldsymbol{u}\mathrm{d}b,\ \boldsymbol{x}\in D^m$$

  - $\varrho := (m-1)/2$
  - $A(\boldsymbol{x},\boldsymbol{u}) := d(0,\boldsymbol{\xi}(\boldsymbol{x},\boldsymbol{u})) = \log\left(\frac{1-|\boldsymbol{x}|_E^2}{|\boldsymbol{x}-\boldsymbol{u}|_E^2}\right),\quad (\boldsymbol{x},\boldsymbol{u})\in D^m\times\partial D^m$
  - Note: $A(\boldsymbol{x},\boldsymbol{u}) - b = d(0,\boldsymbol{\xi}(\boldsymbol{x},\boldsymbol{u})) - d(0,\boldsymbol{\xi}(\boldsymbol{y},\boldsymbol{u})) = d(\boldsymbol{x},\boldsymbol{\xi}(\boldsymbol{y},\boldsymbol{u}))$

- リッジレット変換

$$R[f](a,\boldsymbol{u},b) := \int_{D^m} \boldsymbol{c}[f](\boldsymbol{x})\overline{\rho(aA(\boldsymbol{x},\boldsymbol{u}) - b)}e^{\varrho A(\boldsymbol{x},\boldsymbol{u})}\mathrm{d}\boldsymbol{x}$$

  - $\boldsymbol{c}[f](\boldsymbol{x}) := w_m\int_{\mathbb{R}\times\partial D^m}\widehat{f}(\lambda,\boldsymbol{u})e^{(i\lambda+\varrho)A(\boldsymbol{x},\boldsymbol{u})}|\boldsymbol{c}(\lambda)|^{-2}\mathrm{d}\lambda\mathrm{d}\boldsymbol{u}$

- 再構成公式: 任意の $f\in L^2(D^m), \sigma\in\mathcal{S}'(\mathbb{R}), \rho\in\mathcal{S}(\mathbb{R})$ に対し，

$$S[R[f;\rho]] = ((\sigma,\rho))f,$$

  - $((\sigma,\rho)) := \frac{w_m}{2\pi}\int_{\mathbb{R}}\sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-1}\mathrm{d}\omega.$

※非コンパクト対称空間上のデータを受け取るNNまで同様の結果は拡張できる.

意義:
意味埋め込みなど機械学習で有用な双曲空間上のデータを受け取るNNに対し始めて数理的に厳密な“表現”を与えた.
→ 万能近似性，帯域制限した近似誤差といった応用.

その他の話題

- Federated learningのアルゴリズムと収束解析:
バイアス・バリアンスを縮小することでより少ない通信回数で並列分散環境による最適化を達成.
Murata, Suzuki: Escaping Saddle Points with Bias-Variance Reduced Local Perturbed SGD for Communication Efficient Nonconvex Distributed Learning. arXiv:2202.06083.

- 半教師有り学習における重要度重み付きカーネル法による精度改善:
データの重要度を教師なしデータから推定し，重要度に応じた確率で教師ありデータをサンプリングして学習効率を改善.
Murata, Suzuki Gradient Descent in RKHS with Importance Labeling. AISTATS2021.

- 二重降下と前処理付き勾配降下法による最適化との関係および最適な前処理の特徴付け:
過剰パラメータ状況で前処理付き勾配降下法で最小二乗法を解いた時の最適な前処理を特徴づけ.
Amari, Ba, Grosse, Li, Nitanda, Suzuki, Wu, Xu: When Does Preconditioning Help or Hurt Generalization? ICLR2021.

- スパース正則化を用いた過剰パラメータNNの収束解析:
Akiyama, Suzuki: On Learnability via Gradient Method for Two-Layer ReLU Neural Networks in Teacher-Student Setting. ICML2021.

- 深層学習を用いた行列再整列化:
  - ➢ Watanabe, Suzuki: Deep Two-Way Matrix Reordering for Relational Data Analysis. arXiv:2103.14203.
  - ➢ Watanabe, Suzuki: AutoLL: Automatic Linear Layout of Graphs based on Deep Neural Network. SSCI 2021.