FY2024/2024年度

Deep Learning Theory Team

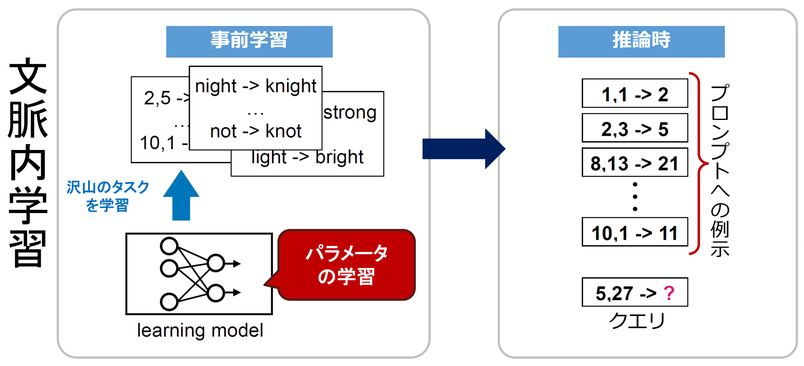
Taiji Suzuki
深層学習理論チーム　鈴木 大慈

RIKEN Center for Advanced Intelligence Project

In-context learning (文脈内学習) の理論

文脈内学習の理論を展開:
なぜTransformerは文脈内学習を実現できるのか？

1. 統計的学習理論：Transformerは文脈内学習でミニマックス最適な誤差を達成
2. 最適化理論：事前学習における勾配法により適切な特徴を学習可能→サンプル複雑度の改善



少数の例をプロンプトを通じて提示する（少数例プロンプトファインチューニング）.
テスト時の例示に対しては<u>パラメータは更新しない</u>.
それでも例示から文脈を読み取って予測できる.
「学習法の学習」

問題設定: $y_{i,t} = F_t^\circ(x_{i,t}) + \epsilon_{i,t} \quad (i = 1, \dots, n)$
$t = 1, \dots, T$: 訓練タスクのインデックス

ミニマックス最適性

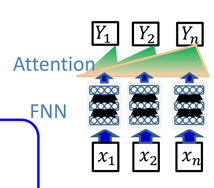
[Kim, Nakamaki, Suzuki: Transformers are Minimax Optimal Nonparametric In-Context Learners. **NeurIPS2024**]

真の関数
$F_t^\circ(x) = \beta_t^\top f^\circ(x)$

- **事前学習 (pretraining):** 特徴量 (表現) を学習 $[f^\circ]$
  - ➢例: Fourier, B-Spline
  - ➢文脈 (t) に非依存
  - ➢データを表現する「最も効率的」な基底を学習
- **文脈内学習 (in-context):** 係数を学習 $[\beta_t]$
  - ➢文脈 (t) に依存
  - ➢例示から現在の文脈β_tを推定
  - → Attentionで獲得

仮定
$\mathrm{Cov}(\beta) = \Sigma = \mathrm{diag}((k^{-2s-1-\epsilon})_{k=1}^\infty)$
(関数空間の複雑さ)

Transformerモデル:
1. 非線形NN層 (N次元特徴量) $\phi : \mathbb{R}^d \to \mathbb{R}^N$
2. Attention層 $\frac{1}{n}\sum_{i=1}^n y_{i,t}\phi(x_{i,t})^\top \Gamma KQ \phi(x_{\mathrm{qr},t})$



定理 (推定誤差)

$$\mathbb{E}[\mathcal{L}(\hat{\phi}, \hat{\Gamma})] \lesssim N^{-2s} + N^2\delta_N^4 + N^{2r+1}\delta_N^2 \quad \text{特徴量の近似誤差}$$
$$+ \frac{N}{n} + \frac{N^{2r}}{n}\log(N) + \frac{N^{4r}}{n^2}\log^2(N) \quad \text{文脈}(\beta_t)\text{を推定する誤差}$$
$$+ \frac{1}{T}\left(N^2\log(\epsilon^{-1}) + \log(\mathcal{N}(\tfrac{\epsilon}{\sqrt{N}}, \mathcal{F}_N, \|\cdot\|_\infty))\right) + \epsilon \quad \text{基底(特徴量)の学習誤差}$$

- **例 (B-spline基底; f_j°がB-spline→Besov/Sobolev空間):**

$$\mathbb{E}[\mathcal{L}(\hat{\phi}, \hat{\Gamma})] \lesssim n^{-\frac{2s}{2s+1}} + \frac{n^{\frac{2}{2s+1}}}{T}$$

「文脈 (β_t)」の推定誤差 ミニマックス最適　「表現 (f°)」の推定誤差

Tが小さい: 記憶中の状態
Tが大きい: 記憶が完了し汎化できる状況
→ テスト時推論のスケーリング則

定理 (ミニマックス下限)

特徴量を知っている時の最適誤差 (β_tの推定誤差)　特徴量を学習する複雑さ (log-covering number)

$$\inf_{\hat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathbb{E}[\mathcal{L}(\hat{f})] \gtrsim n^{-\frac{2s}{2s+1}} + \frac{V(\epsilon_{1,n}, \mathcal{F}^\circ)}{nT}$$
ただし, $\epsilon_{1,n}^2 \simeq \frac{V(\epsilon_{1,n}, \mathcal{F}^\circ)}{nT}$

直感：十分多くの問題例を事前に見ておけば，テスト時に最適な誤差で予測できる．事前知識が足りなければテスト時に失敗する．

平均場NNによる非線形特徴学習の大域的最適性

[Kim, Suzuki: Transformers Learn Nonlinear Features In Context: Nonconvex Mean-field Dynamics on the Attention Landscape. **ICML2024, oral presentation**]

非線形特徴学習(ϕ)に2層NN(平均場NN)を考える：

$$h_\mu(x) = \int h_\theta(x)\mathrm{d}\mu(\theta) \in \mathbb{R}^k$$
$$h_\theta(x) = \mathbf{a}\sigma(\mathbf{w}^\top x) \quad (\theta = (\mathbf{a}, \mathbf{w}) \in \mathbb{R}^k \times \mathbb{R}^d)$$

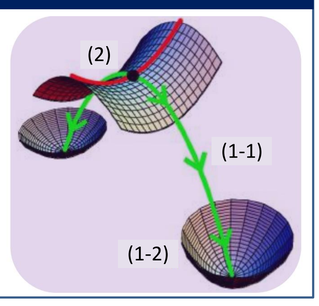
損失関数 $\mathcal{L}(\mu, \Gamma) := \mathbb{E}_{x_{\mathrm{qr}}}\left[\left\|f^\circ(x_{\mathrm{qr}}) - \mathbb{E}_x[f^\circ(x)h_\mu(x)^\top]\Gamma h_\mu(x_{qr})\right\|^2\right]$

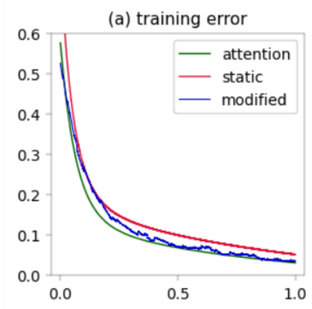
勾配法による最適化 (Wasserstein勾配流)

- $\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla \frac{\delta \mathcal{L}(\mu_t)}{\delta \mu}\right)$
- $\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = -\nabla \frac{\delta \mathcal{L}(\mu_t)}{\delta \mu}(\theta_t) \quad (\mu_t = \mathrm{Law}(\theta_t))$

定理 (損失関数の性質)

損失関数の局所最適解はすべて大域的最適解:
(1) 停留点は鞍点か大域的最適解.
(2) 鞍点には負の曲率方向があり，指数関数的速度で鞍点から脱出可能.





勾配法による特徴学習の最適性

[Jason D. Lee, Kazusato Oko, Taiji Suzuki, Denny Wu: Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. . **NeurIPS2024**]

シングルインデックスモデル:

$$f_*(x) = \sigma_*(\langle x, \theta\rangle) \quad (x \sim N(0, I_d))$$

➢方向$\theta \in \mathbb{R}^d$とリンク関数 σ_*を推定
→ **特徴学習が重要**
➢σ_* は<u>次数p</u>かつ<u>情報指数k</u>.

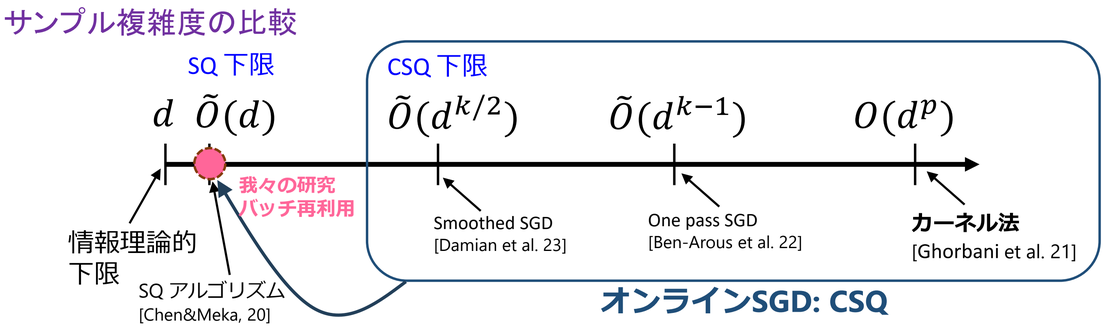
$$\sigma_*(z) = \sum_{i=k}^p \alpha_i^* \mathrm{He}_i(z)$$
He_i: 次数iのエルミート多項式

Q: ニューラルネットワークはどれくらいの計算/サンプル-複雑度で学習できるか？

定義 (情報指数 k [Ben-Arous et al. 2021])
σ_*の情報指数は非ゼロ係数α_i^*をもつ最小次数である:
$$k := \mathrm{IE}(\sigma_*) = \min\{i \in \mathbb{N} \mid \alpha_i^* \neq 0\}$$

情報指数は問題の難しさを計る指標
[kが大きい⇒高次の情報を取り出す必要がある⇒より多くのデータが必要]

サンプル複雑度の比較



- 通常のオンライン確率的勾配降下法 (SGD) はCSQアルゴリズムに含まれる.
- SGDを用いた学習法のサンプル複雑度の下限は$\Omega(d^{k/2})$である．一方で，情報理論的下限は$O(d)$である．⇒ 通常のSGDは<u>情報理論的下限を達成できない.</u>

Q: 通常のSGDを少し修正することでこのギャップを埋めることは可能か？
我々の結果：ミニバッチを二回使いまわす(再利用する)だけで$O(d^{k/2})$を$\tilde{O}(d)$に改善できる.

定理 (バッチ再利用SGDのサンプル複雑度)
バッチ再利用型のSGDは
$$n = \tilde{O}(d\epsilon^{-2}), \quad N = \tilde{O}(\epsilon^{-1}).$$
だけのサンプル複雑度 (更新回数)で平均二乗誤差をϵ以下にできる:
$$\mathbb{E}_x[(f_*(x) - f_{\hat{\Theta}}(x))^2] \leq \epsilon$$

<u>バッチ再利用だけで情報理論的下限を達成できる.</u>
非線形活性化関数を使うことが証明の鍵．非線形活性化関数は最適化の効率化にも寄与している.

- **Correlation statistical query (CSQ):** Algorithm has an access to "noisy" information about correlation between ϕ and y. (τ is the noise level. Noise can be adversarial)
$$|\tilde{q} - \mathbb{E}_{x,y}[\phi(x)y]| \leq \tau \sim n^{-1/2}$$
- **Statistical query (SQ):** General query
$$|\tilde{q} - \mathbb{E}_{x,y}[\phi(x,y)]| \leq \tau \sim n^{-1/2} \quad (\text{with } |\phi| \leq 1)$$

※オンラインSGDはCSQアルゴリズム:
$$\nabla_w \mathbb{E}_{x,y}[(y - f_w(x))^2]$$
$$\propto \underbrace{-\mathbb{E}_{x,y}[y\nabla_w f_w(x)]}_{\text{Correlation query}} + \mathbb{E}_x[f_w(x)\nabla_w f_w(x)]$$
※データを二回使いまわすことで，CSQから外れる.

勾配法による特徴学習：事前学習による学習精度改善

[Oko, Song, Suzuki, Wu: Transformer efficiently learns low-dimensional functions in context. **NeurIPS2024**]

$$F_t^\circ(x) = \sigma_*^t(\langle x, \beta_t\rangle) : \text{タスク}t\text{の関数}$$

リンク関数σ_*^tと方向β_tがランダムに生成される設定 (文脈から推定する)

- β_tはd次元空間のうちr次元部分空間$\mathcal{S}$に分布 $(r \ll d)$:
$$\beta_t \sim \mathrm{Unif}(\mathrm{Unit}(\mathcal{S})) \qquad \dim(\mathcal{S}) = r \ll d$$
- $\sigma_*^t(z) = \sum_{i=k}^P c_i^t \mathrm{He}_i(z)$
  c_i^tはタスクごとにランダムに生成されているとする.

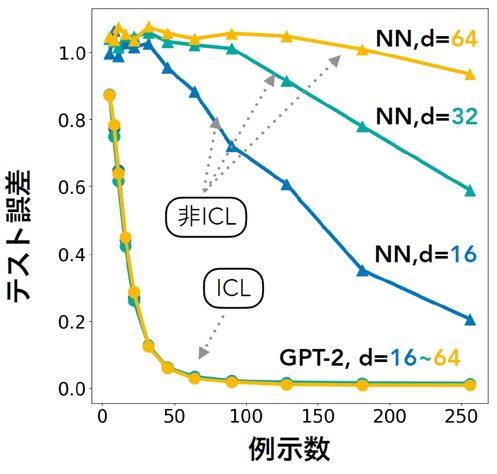
⇒β_tの部分空間を事前学習で取得

	事前学習なし		事前学習あり
手法	カーネル	NN	ICL
テスト時複雑度	d^P	$d^{k/2}$ (d)	r^{5P+2}
事前学習複雑度	---	---	$T_1 = d^{2k}$, $n = d^k$

事前学習後の文脈内学習では次元dへの依存がなくなり少ない例示で文脈を推定できる

定理 (SGDによる学習の複雑度)
1. 事前学習で用いるデータサイズ:
$T = \Theta(d^{2k})$ and $n = \tilde{\Omega}(d^k)$.
2. 一層目のNNの横幅: m
3. テスト時の例示数: n^*
SGD (勾配法) によって**多項式時間で学習可能**で，以下の予測誤差を達成:
$$\mathcal{L}(\widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\Gamma}) \lesssim \underbrace{\frac{r^{3P/2}}{\sqrt{m}}}_{\text{近似誤差}} + \sqrt{r^{5P}}\sqrt{\underbrace{\frac{r^2}{n}}_{\text{汎化誤差}} + \frac{r^2}{n^*}}.$$

必要なテスト時例示数が次元dに依存しない.
➢ 勾配法により不要な特徴量が削れている効果
➢ 事前学習なしでは達成不可能



Koopman作用素の理論によるDNNの汎化誤差解析

[Hashimoto, Sonoda, Ishikawa, Nitanda, Suzuki: Koopman-based generalization bound: New aspect for full-rank weights. **ICLR2024**]

- DNNの一層分をKoopman作用素とみなし，Koopman作用素の作用素ノルムを用いてRademacher複雑度を抑えた.

$$\text{Rademacher complexity} \leq O\left(\frac{\|g\|_{H_L}}{\sqrt{n}}\prod_{j=1}^L \frac{G_j E_j \|W_j\|^{s_{j-1}}}{\det(W_j^* W_j)^{1/4}}\right)$$

これまでの結果とは異なり，中間層の行列の最小特異値が「大きい方が良い」バウンドを与えた.

分布のミニマックス最適化問題

[Kim, Yamamot, Oko, Yang, Suzuki: Symmetric Mean-field Langevin Dynamics for Distributional Minimax Problems, **ICLR2024**]

$$\min_\mu \max_\nu \quad \mathcal{L}(\mu, \nu) + \lambda\mathrm{Ent}(\mu) - \lambda\mathrm{Ent}(\nu) \quad \mu, \nu\text{: 確率分布}$$

応用:
- 敵対的学習
- ロバスト学習
- 強化学習

貢献: 確率分布に対する<u>双対平均加法</u>を提案し，提案手法が$O(1/K)$で収束することを示した.
[離散時間/分布の有限粒子近似誤差も導出]