FY2025/2025年度

High-Dimensional Structure Theory Team

Masaaki Imaizumi

高次元構造理論チーム　今泉 允聡

RIKEN
AIP Center for Advanced Intelligence Project

結果概要：

二側面から学習理論を開発
1. 数学的理論
　→ 解釈性・最適判定
2. 物理学的理論
　→ 精密再現・現象制御

1. 数学的学習理論
- 高い解釈性
- 収束レート解析
- 最適性の判定

2. 物理学的学習理論
- 高い精密性
- 動力学の解析可
- 現象制御技術へ

数学的学習理論：

結果[1]: Transformerは最適な非定常環境適応器

→時間で変化する環境に適応する能力の証明

設定：インコンテキストな非定常強化学習

強化学習 S: 状態, A: 行動, R: 報酬, γ: 割引率
観測 $(s_t^i, a_t^i, r_t^i)_{t=1,\dots,T,i=1,\dots N}$, T: 軌跡長, N: 軌跡数

非定常係数: $\Delta = \sum_t \max_{s,a} |R_t(s,a) - R_{t+1}(s,a)|$

結果：Transformerのリグレット$\mathcal{R}(T)$上界

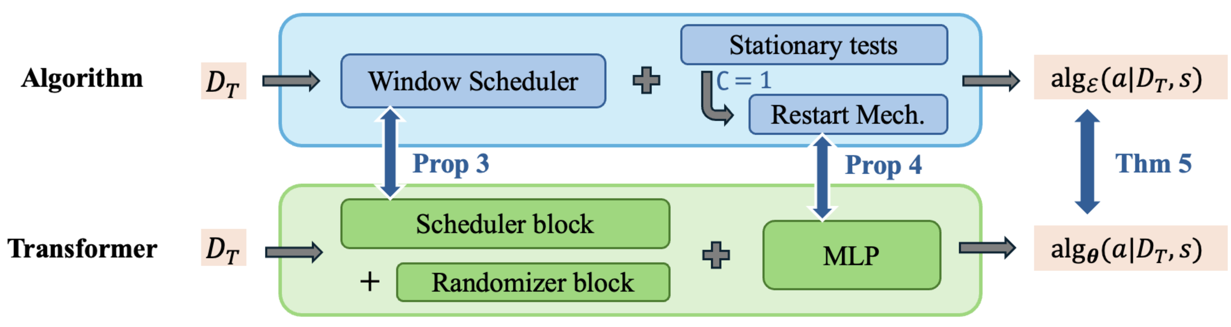
$$\mathcal{R}(T) \le \tilde{O}(\min\{\sqrt{C_T T}, \Delta^{1/3}\ T^{2/3} + \sqrt{T}\})$$

T, Δ に対する最適レートと一致→適応能力の証明

アイディア：適応アルゴリズムを近似するTF構成

内部に部分アルゴリズムを構成して適応的選択

最適アルゴを近似するTF構成



結果[2]: データ非等方性の学習上の有効次元の導出

→学習の難しさを記述する新しい尺度の提案

設定：非等方データのシングルインデックス学習

$w^* \in \mathbb{S}^{d-1}$: シグナルパラメータ, $f: \mathbb{R} \to \mathbb{R}$: 未知関数
$Q: \mathbb{R}^{d \times d}$: 共分散行列 $x_i \sim N(0, Q)$: 非等方入力

$$y_i = f\left(\left\langle x_i, \frac{w^*}{\|Q^{1/2} w^*\|}\right\rangle\right)$$: シングルインデックス出力

2層ニューラルネットを確率的勾配降下法で学習

結果：レート導出と複雑性評価

$$\Theta := \frac{\|Q^{1/2} w^*\|}{\|Q^{1/2}\|}$$:非等方尺度（データの複雑性表現）

定理 (シグナル復元のサンプル複雑性)
w^*の(弱)復元に必要な反復数
$$T = O(\varepsilon^{-2} \Theta^{-2})$$

対応する下限も導出

示唆：NN計算量の非等方への一般化

$Q = I_d$なら既存理論を再現→計算量理論の非等方化

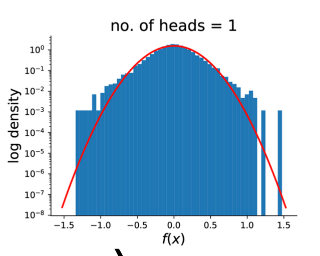
物理学的学習理論：

結果[3]: アテンションの非ガウス極限分布の解明

→Transformer極限分布の未解決問題への一歩

設定：アテンション層の無限幅極限

$X \in \mathbb{R}^{s \times n}, Q, K, V \in \mathbb{R}^{n \times n}$: 重み行列



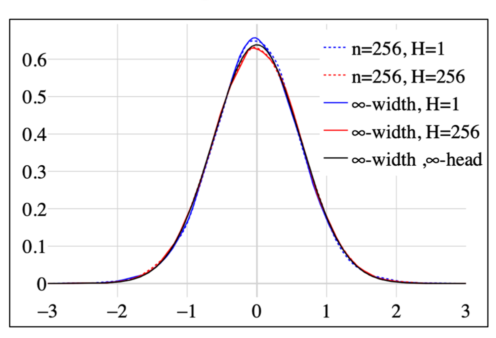
ガウス近似不可モデル

$$\mathrm{Attn}(X) = \mathrm{SoftMax}\left(\frac{1}{\sqrt{n}}(XQ)(XK)^\top\right)(XV)$$

結果：非ガウス極限分布の導出 (Tensor Programs)

定理 (無限幅極限$n \to \infty$)
$p_{i,j}^\circ, Z^j$: ガウス分布（互いに独立）

$$\mathrm{Attn}(X)_i \to^d \sum_{j=1}^{s} \mathrm{SoftMax}(p_{i,1}^\circ, \dots, p_{i,s}^\circ) Z^j$$



実験と理論の整合確認

示唆：理論的なTransformer精密制御への展望
極限分布で学習前に性能予測→ハイパラ選択に応用
ただしTransformerは未導出→本研究の応用可能性

結果[4]: 対角ネットワーク動力学の統一的理論

→既存理論群を包含する一貫した記述

設定：対角線系ネットワークの勾配降下更新

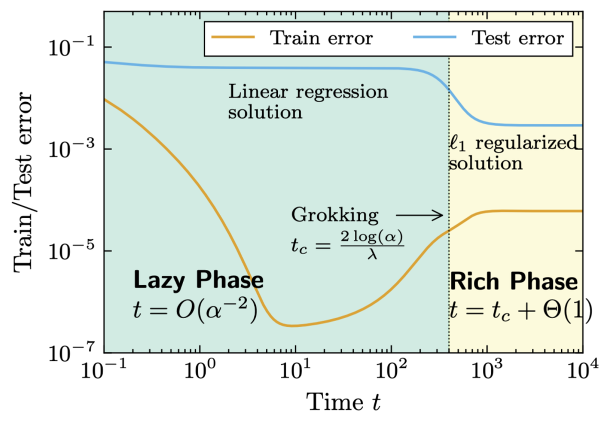
$$f(x; u, v) = w^\top x, w = 2^{-1}(u^2 - v^2)$$

正則化付き勾配流にて更新
（学習ダイナミクスを見る典型トイモデル）

結果：統一記述の方程式を導出

動的平均場理論による方程式

$$C_w(t,t') = \mathbb{E}[(w(t) - w^*)(w(t') - w^*)],$$
$$R_w(t,t') = -\mathbb{E}\left[\frac{\partial w(t)}{\partial z(t')}\right],$$
$$C_f(t,t') = C_w(t,t') + \sigma^2 - \int_0^{t'} R_f(t',s)(C_w(t,s) + \sigma^2)\,\mathrm{d}s - \int_0^{t} R_w(t,s) C_f(t',s)\,\mathrm{d}s,$$
$$R_f(t,t') = R_w(t,t') - \int_{t'}^{t} R_w(t,s) R_f(s,t')\,\mathrm{d}s,$$
$$g(t) = \frac{z(t)}{\delta} + w(t) - w^* - \int_0^t R_f(t,s)(w(s) - w^*)\,\mathrm{d}s,$$
$$\frac{\mathrm{d}}{\mathrm{d}t} w(t) = -\sqrt{w(t)^2 + \alpha^4 \mathrm{e}^{-2\lambda t}}\, g(t) - \lambda w(t),$$



記述可能な現象→
- 収束解の特定
- 学習フェイズ切替
- Grokking記述
- 汎化/収束の二律背反

Selected Papers

[1] B.Chen, S.Ito, M.Imaizumi (2025), "Optimal Dynamic Regret by Transformers for Non-Stationary Reinforcement Learning", Advances in Neural Information Processing Systems.
[2] G.Braun, M.Quang, M.Imaizumi (2025), "Learning a Single Index Model from Anisotropic Data with Vanilla Stochastic Gradient Descent", PMLR: Artificial Intelligence and Statistics.
[3] M.Sakai, R.Karakida, M.Imaizumi (2025), "Infinite-Width Limit of a Single Attention Layer: Analysis via Tensor Programs", Advances in Neural Information Processing Systems.
[4] S.Nishiyama, M.Imaizumi (2026) "Precise Dynamics of Diagonal Linear Networks: A Unifying Analysis by Dynamical Mean-Field Theory", PMLR: Artificial Intelligence and Statistics.