FY2025/2025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


## 生成モデルの新しいアライメント手法

### 拡散モデルのアライメント

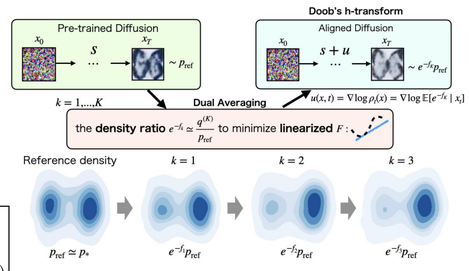
[Kawata, Oko, Nitanda, Suzuki: Direct Distributional Optimization for Provable Alignment of Diffusion Models. ICLR2025]

拡散モデルの生成分布を直接修正することはこれまで困難だった．⇒本研究：直接アライメントを修正する手法を提案

$$\min_{\mu\in\mathcal{P}} \mathcal{L}(\mu) = F(\mu) + \beta \mathrm{KL}(\mu||\mu_{\mathrm{ref}})$$

$\mu_{\mathrm{ref}}$: 事前学習された拡散モデル

例.: DPO, ベイズ, フィルタ



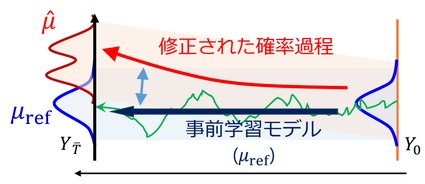

**Phase 1: 最適化.** 双対平均化

For $k = 1, \dots, N-1$:

$$\mu^{(k+1)} = \arg\min_{\mu\in\mathcal{P}} \frac{2}{k(k+1)} \sum_{j=1}^{k} j \left( \mathbb{E}_{\mu}\left[\frac{\delta F(\mu^{(j)})}{\delta \mu}\right] + \beta \mathrm{KL}(\mu||\mu_{\mathrm{ref}}) \right) + \frac{2\beta}{k}\mathrm{KL}(\mu||\mu_{\mathrm{ref}})$$

$$\propto \exp(-\bar{g}^{(k)})\mu_{\mathrm{ref}}$$

where $\bar{g}^{(k)} = \sum_{j=1}^{k} \frac{j}{\beta(k+1)(k+2)/2} \frac{\delta F(\mu^{(j)})}{\delta \mu}$ ← ニューラルネットで近似




$$\Rightarrow \frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\mu_{\mathrm{ref}}} \propto \exp(-\hat{g})$$

ただし $\hat{g} \leftarrow \bar{g}^{(N)}$

$\mu_{\mathrm{ref}}$と最適解$\hat{\mu}$の**密度比**のみが得られる．

**Phase 2: Sampling from $\hat{\mu} \propto \exp(-\hat{g})\,\mu_{\mathrm{ref}}$.**

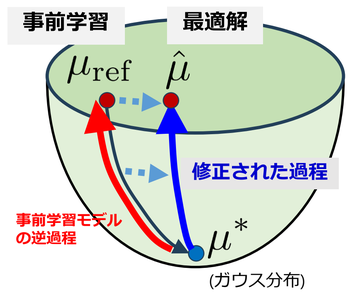
Doob $h$-Transform (Doob, 1957; Rogers & Williams, 2000)

$$\mathrm{d}\bar{Y}_t = (\bar{Y}_t + 2s(\bar{Y}_t, \overline{T}-t) + 2\underbrace{\nabla_x \log(\mathbb{E}[\exp(-\hat{g}(Y_{\overline{T}}))|Y_t = x]|_{x=\bar{Y}_t})}_{\text{修正項}})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$$\bar{Y}_{\overline{T}} \sim \exp(-\hat{g})\mu_{\mathrm{ref}} = \hat{\mu}$$

元の$\mu_{\mathrm{ref}}$の逆過程：

$$\mathrm{d}Y_t = (Y_t + 2s(Y_t, \overline{T}-t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$




最適解$\hat{\mu}$からのサンプリングが可能に！

### LLMのアライメント

人間の選好モデルを特別に仮定せずアライメントさせたい．

各プロンプト$c \sim p(c)$に対し以下のデータを得る：

$y_1 \sim p_w(y|c)$, 高評価データ　　$y_2 \sim p_l(y|c)$ 低評価データ

※ $y_1, y_2$ は必ずしも対になっていない．

**仮定:** $p_{\mathrm{ref}} = \theta p_w + (1-\theta)p_l$

**目標: $p_w$からデータ生成できるようにしたい．**

➢ 特定の選好モデルに依存しない．統計的一致性を保証したい．

**提案法「Density ratio matching estimator」：**

真の密度比からの距離　KL-正則化

$$\min_{p:\mathrm{LLM}} \left\{ \mathrm{B}_f(r||r^*) + \beta^{-1}\mathrm{KL}(p||p_{\mathrm{ref}}) \right\}$$

ただし，$r^* = \frac{p_l}{p_w}$ かつ $r = (1-\theta)^{-1}\left(\frac{p_{\mathrm{ref}}}{p} - \theta\right)$.

$(r = r^* \Leftrightarrow p = p_w)$

$$\mathrm{B}_f(r||r^*) := \int [f(r^*) - f(r) - f'(r)(r^* - r)]\mathrm{d}p_w$$

：凸関数$f$で決まるBregmanダイバージェンス

[Higuchi, Suzuki: Direct Density Ratio Optimization: A Statistically Consistent Approach to Aligning Large Language Models. ICML2025]

$$\mathrm{B}_f(r||r^*) + \beta^{-1}\mathrm{KL}(p||p_{\mathrm{ref}})$$
$$= \int (-f(r) + f'(r)r)\mathrm{d}p_w - \int f'(r)\mathrm{d}p_l$$
$$+ \beta^{-1}\int \hat{r}^{-1}\log(\hat{r})\mathrm{d}p_{\mathrm{ref}} \qquad \hat{r} = \frac{p_{\mathrm{ref}}}{p}$$

→データから推定可能な形

→$r = (1-\theta)^{-1}\left(\frac{p_{\mathrm{ref}}}{p} - \theta\right)$と$\hat{r} = \frac{p_{\mathrm{ref}}}{p}$を代入して，$p$に関して最適化

**例 (logistic regression):**

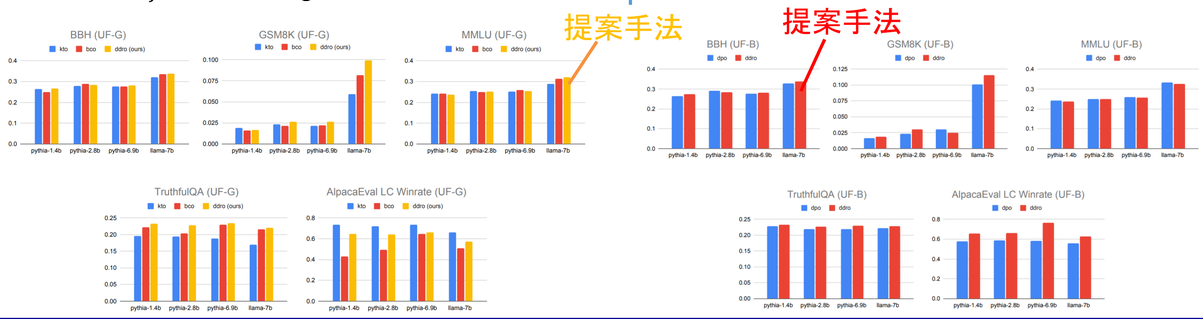
$f(t) = t\log(t) - (1+t)\log(1+t)$

$$\int \underbrace{\log(1+r)\mathrm{d}p_w + \int \log(1+r^{-1})\mathrm{d}p_l}_{\text{Logistic損失}} + \beta^{-1}\int \hat{r}^{-1}\log(\hat{r})\mathrm{d}p_{\mathrm{ref}}$$

KL-正則化

➢ 統計的一致性の理論保証有り

➢ 「相対的密度比」を用いることで安定化可能




## 線形注意手法の理論と最適状態サイズの決定

- 現状のAttentionは本当にそれで十分か？
- 計算量の問題がある．

| Attention | | 代替手法 (線形手法) |
|---|---|---|
| • 計算量 $O(L^2)$ ($L$は入力長) | < 効率性 | • 計算量 $O(L)$ or $O(L\log(L))$ |
| • 基盤モデルのスタンダード | > 能力 | • 能力の限界に関する指摘 言語等の離散データが不得意 |

Q. • 効率的な代替法はあるか？
• どのような タスクなら代替可能？

### 状態空間モデルの表現力

[Naoki Nishikawa, Taiji Suzuki: State Space Models are Comparable to Transformers in Estimating Functions with Dynamic Smoothness. ICLR2025]

- **主結果 1** 既存結果: Copying タスクで 1層のSSM は Transformer を代替できない
  [Jelassi et al.: Repeat After Me: Transformers are Better than State Space Models at Copying. 2024]

  なぜか？⇒入力依存で重要なトークンを抽出可能する必要がある


  **本研究：** 多層の FNN + SSM で Transformer を代替可能

  各トークンの重要度を前の層で計算すれば実は代替できる：

| p | 1 | u | 7 | v | 4 | w | 7 | t | 9 | u | ▶ ? |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 20 | 159 | 2 | 2 | 4 | 3 | 2 | 3 | 24 | ← 重要度（入力依存） e.g. 自分 or 1つ前が最後のトークンと同じ |

- **主結果 2** 区分的$\gamma$-平滑関数の推定において，SSMとTransformerは同じ推定誤差を達成することを証明

### 線形注意機構の特徴次元の決定

[Naoki Nishikawa, Rei Higuchi, Taiji Suzuki: Degrees of Freedom for Linear Attention: Distilling Softmax Attention with Optimal Feature Efficiency. NeurIPS2025]

- Attentionのカーネル関数としての定式化と分解：

$$K(x,y) = \exp(x^\top y/\sqrt{d}) = \mathbb{E}_{z\sim N(0,I)}[\phi(x;z)\phi(y;z)] \quad \text{where } \phi(x;z) = \exp\left(\frac{z^\top x}{d^{1/4}} - \frac{\|x\|^2}{2\sqrt{d}}\right)$$

- 線形注意機構と特徴写像： $\hat{K}(x,y) = \varphi(x)^\top \varphi(y)$ where $\varphi: \mathbb{R}^d \to \mathbb{R}^M$.

**定理 (近似誤差の理論評価)**

$\lambda > 0$に対し，$N_\lambda$をAttentionの**統計的自由度**とする．すると，

$$M \geq \frac{4}{t} N_\lambda \log\left(\frac{32N_\lambda}{t/2}\right),$$

とすることで，Attentionは線形注意機構で以下の誤差で近似できる：

$$\|K - \hat{K}\|^2_{L^2(P_X\otimes P_X)} \leq 2t\left(\lambda^2 + \|K\|^2_{L^2(P_X\otimes P_X)}\right)$$

(ただし，特徴写像$\varphi$は適切に選ぶとする)

統計的自由度: $N_\lambda = \mathrm{Tr}[\Sigma_K(\Sigma_K + \lambda I)^{-1}]$

- Models: GPT-2 ($1.24\times10^8$ parameters) [Radford et al., 2019], Pythia-1B ($1.01\times10^9$ parameters) [Biderman et al., 2023]
- Hyperparameters: $(C,\lambda) = (64, 10^{-4})$ for GPT-2, $(C,\lambda) = (128, 10^{-8})$ for Pythia-1B

✓ Cost $C$ is set to be the same as the head size following prior work (e.g. [Chen et al., 2025])

| Model | Cost | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-2 | 64 | **130** | **182** | 35 | 44 | 42 | 65 | **92** | 33 | 28 | 34 | 46 | 39 | - | - | - | - |
| Pythia-1B | 128 | **277** | 138 | **275** | 132 | 204 | 167 | 115 | 142 | **231** | 64 | 96 | 73 | 40 | 52 | 34 | 9 |

Large in early and middle layers　　Small in latter layers

深いレイヤーの方が少ない特徴量でOK (CNN, FNNと同様の現象 [Arora et al., 2018; Ravichandran et al., 2019; Suzuki et al., 2020])

**Downstream taskの性能** 対抗手法：Performer [Choromanski et al., 2021], DiJiang [Chen et al., 2025]

| Strategy | Method | PiQA | logiQA | ARC-E | ARC-C | Winogrande | MMLU | WSC | Average |
|---|---|---|---|---|---|---|---|---|---|
| **Original GPT-2** | | **0.5985** | 0.3103 | **0.3325** | 0.3003 | 0.5122 | 0.2789 | **0.6538** | **0.4266** |
| — | DiJiang | 0.5065 | 0.2550 | 0.2113 | 0.2244 | 0.4846 | 0.2639 | 0.4615 | 0.3409 |
| | Performer | 0.5468 | 0.2934 | 0.3039 | 0.2747 | 0.4996 | 0.2517 | 0.5962 | 0.3952 |
| Fix | direct | 0.5832 | **0.3195** | 0.2921 | 0.2995 | **0.5335** | 0.2552 | 0.6154 | 0.4141 |
| | softmax | 0.5718 | 0.2673 | 0.2479 | **0.3029** | 0.5020 | 0.2634 | 0.6154 | 0.3958 |
| | $L^2$ | 0.5822 | **0.3195** | 0.2483 | 0.2773 | 0.5107 | 0.2520 | 0.5962 | 0.3980 |
| DoF | direct | 0.5669 | 0.3011 | **0.3241** | **0.3012** | **0.5280** | 0.2712 | **0.6346** | 0.4182 |
| | softmax | 0.5751 | 0.3026 | 0.3224 | 0.2995 | **0.5328** | 0.2564 | **0.6442** | **0.4190** |
| | $L^2$ | 0.5664 | 0.3088 | 0.2736 | 0.2824 | 0.4972 | 0.2608 | 0.5865 | 0.3965 |
| DoF + Clip | direct | **0.5892** | **0.3164** | 0.3136 | 0.2952 | 0.5075 | **0.2993** | **0.6346** | **0.4223** |
| | softmax | **0.5860** | 0.3026 | **0.3401** | 0.2816 | 0.4996 | **0.2832** | **0.6346** | 0.4182 |
| | $L^2$ | 0.5822 | **0.3164** | 0.2942 | **0.3063** | 0.5091 | **0.2799** | 0.5673 | 0.4079 |

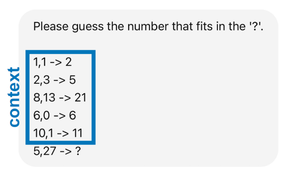
(特徴次元を固定した方法)

(特徴次元を層ごとに選んだ方法 [提案法])

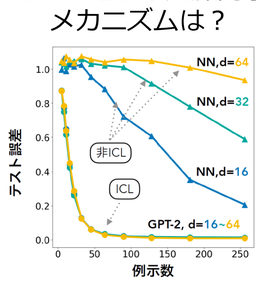
## テスト時推論の理論

### 文脈内学習

[Nishikawa, Song, Oko, Wu, Suzuki: Nonlinear transformers can perform inference-time feature learning.ICML2025]

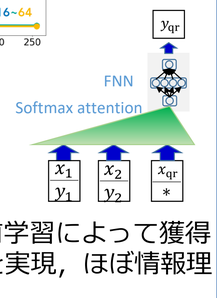



Transformerによる文脈内学習のメカニズムは？




$d$: 入力の次元，$r$: 内在的次元

| | 事前学習なし | | 文脈内学習 | |
|---|---|---|---|---|
| 手法 | Kernel | NN | 線形注意 | 非線形注意 |
| サンプル複雑度 | $d^P$ | $d^{\Theta(\mathrm{ge}(\sigma^*))}$ | $r^{4P}$ | $r^{3\mathrm{ge}(\sigma^*)/2}$ |




- Transformerはデータからの情報抽出の方法を事前学習によって獲得
- ソフトマックス注意が「テスト時の」特徴学習を実現，ほぼ情報理論的下限を達成（生成指数: ge）

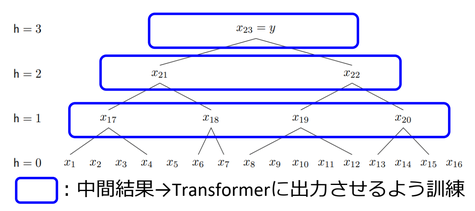
### 思考連鎖

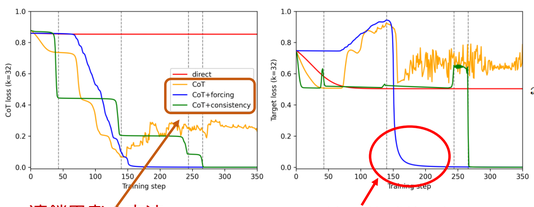
[Kim, Suzuki: Transformers Provably Solve Parity Efficiently with Chain of Thought. ICLR2025]

問題：$k$-パリティ問題

| | 通常学習 | 思考連鎖 |
|---|---|---|
| サンプル/計算複雑度 | $\Omega(d^{k-1})$ | $O(d^{2+\epsilon})$ |

思考連鎖をするように学習することで学習の効率が大幅に改善



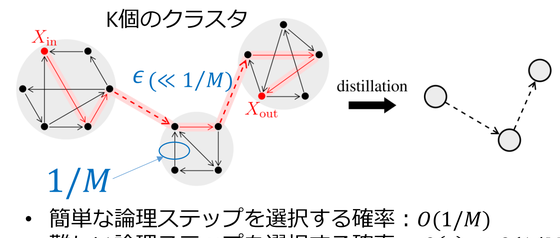




連鎖思考の方法 (いくつかの異なる実装法)

(訓練誤差が落ち切ってからテスト誤差が一気に下がる)

### テスト時スケーリング

[Kim, Wu, Lee, Suzuki: Metastable Dynamics of Chain-of-Thought Reasoning: Provable Benefits of Search, RL and Distillation.. ICML2025]




論理ステップの探索によっていかに推論を改善させることができるか？

- 簡単な論理ステップを選択する確率：$O(1/M)$
- 難しい論理ステップを選択する確率： $O(\epsilon) \ll O(1/M)$

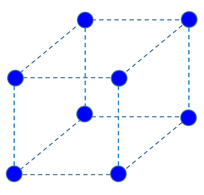
**我々の結果**

- テスト時推論の計算量 $\Theta\left(\frac{KM}{\epsilon}\right)$ PPO前 → $\Theta(KM^2)$ PPO後
- 小さいモデルへ蒸留した場合 $\Theta(K)$

長い時間かけて探索した結果をPPOすることでその後の論理推論が高速化される．

## 離散拡散モデルの理論

[Wakasugi, Suzuki: State Size Independent Statistical Error Bound for Discrete Diffusion Models. NeurIPS2025]



例：$\{0,1\}^D$上の分布を学習

- 順過程： $\frac{\mathrm{d}p_t}{\mathrm{d}t} = Qp_t$, (真の分布) $p_0 = p^*$.
- 逆過程： $\frac{\mathrm{d}q_t}{\mathrm{d}t} = \bar{Q}_{T-t}q_t \quad (t\in[0,T]), \quad q_0 = p_T$.

ただし， $\bar{Q}_t(x,y) = \frac{p_t(y)}{p_t(x)}Q(x,y), \quad \bar{Q}_t(x,x) = -\sum_{y\neq x}\bar{Q}_t(x,y)$

直接多項分布として推定 (拡散モデルを使わず)

$$D_{\mathrm{KL}}(p_0||\hat{q}_T) \lesssim \frac{2^D}{n}$$

**次元の呪いを受ける**

拡散モデルによる改善

深層スコアマッチングによる学習（拡散モデル）

仮定

$p_0 = \sum_{j=1}^{M} c_j u_j(x)$ (固有ベクトルによる分解)

$u_j$: $Q$の$j$番目の固有ベクトル

$\Rightarrow |c_j| \leq |c_1| j^{-s}$ $(s>1)$

$$D_{\mathrm{KL}}(p_0||\hat{q}_T) \lesssim n^{-\frac{2(s-1)}{2(s-1)+1}}$$

次元の呪いを解消

(重要な基底に絞って学習: 特徴学習)

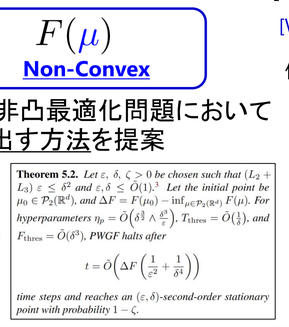
## 確率測度の非凸目的関数の最適化

[Naoya Yamamoto, Juno Kim, Taiji Suzuki: Hessian-guided Perturbed Wasserstein Gradient Flows for Escaping Saddle Points. NeurIPS2025]

$$\min_{\mu\in\mathcal{P}} F(\mu)$$ Non-Convex

確率分布の非凸最適化問題において鞍点を抜け出す方法を提案

**応用:**
・文脈内学習における特徴学習
・推薦システム



**Theorem 5.2.** *Let $\varepsilon, \delta, \zeta > 0$ be chosen such that $(L_2 + L_3)\,\varepsilon \leq \delta^2$ and $\varepsilon, \delta \leq \bar{O}(1)$.[3] Let the initial point be $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\Delta F = F(\mu_0) - \inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} F(\mu)$. For hyperparameters $\eta_p = \bar{O}\left(\delta^{\frac{3}{2}} \wedge \frac{\varepsilon^3}{\delta}\right)$, $T_{\mathrm{thres}} = \bar{O}\left(\frac{1}{\delta}\right)$, and $F_{\mathrm{thres}} = \bar{O}(\delta^3)$, PWGF halts after*

$$t = \bar{O}\left(\Delta F\left(\frac{1}{\varepsilon^2} + \frac{1}{\delta^1}\right)\right)$$

*time steps and reaches an $(\varepsilon,\delta)$-second-order stationary point with probability $1-\zeta$.*

[Wasserstein勾配流] $\partial_t \mu_t = \nabla\cdot\left(\mu_t \nabla \frac{\delta F}{\delta \mu}(\mu_t)\right)$

停留点において以下を実行:

1. 以下のカーネル関数を計算:

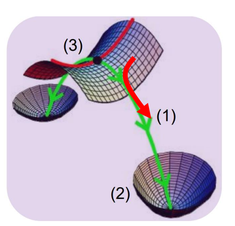
$$K^{(k)}(x,y) = \int \nabla\nabla^\top \frac{\delta^2 F(\mu^{(k)})}{\delta \mu^2}(x,y)\nabla\nabla^\top \frac{\delta^2 F(\mu^{(k)})}{\delta \mu^2}(y,z)\mathrm{d}\mu^{(k)}(y)$$

2. ガウス過程による写像を生成:

$$\xi^{(k)} \sim G(0, K^{(k)})$$

3. 分布を以下のように"摂動":

$$\mu^{(k+1)} = \left(\mathrm{I} + \eta_p \xi^{(k)}\right)\sharp\mu^{(k)}$$




- "Escape" the saddle point.
- Reach a $(\epsilon,\delta)$-second order stationary point