FY2023/2023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

基盤モデルの学習理論

深層生成基盤モデルの統計的学習理論を展開

- 拡散モデルのミニマックス最適性：ICML2023, **oral presentation**.
- Transformerの近似理論と最適性：ICML2023.

生成基盤モデルは現在のAI技術発展の礎である.
しかし，その原理はいまだ不明な点が多い.
⇒ 基盤モデルがいかなるデータを効率的に学習できるかを理論的に解明.

➢ データが高次元でもなぜ学習できる？
➢ モデルのどこがボトルネック？

拡散モデル

[Kazusato Oko, Shunta Akiyama, Taiji Suzuki: Diffusion Models are Minimax Optimal Distribution Estimators. ICML2023, oral presentation]
[IBIS2023最優秀プレゼンテーション賞]

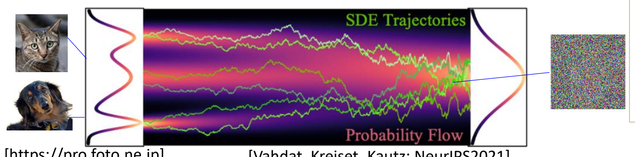
Stable diffusion, 2022

Jason Allen "Théâtre D'opéra Spatial" generated by **Midjourney**. Colorado State Fair's fine art competition, 1st prize in digital art category

・逆過程を深層ニューラルネットワークで推定：スコアマッチング推定 （=デノイジング拡散モデル）
・学習されたスコア関数は訓練データに依存.
➢ どれだけ正確にスコアが学習できるか？
➢ スコアの推定誤差が最終的な生成分布にどう影響するか？

Forward process：ターゲット分布から正規分布へ変換

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$



[https://pro.foto.ne.jp]　[Vahdat, Kreiset, Kautz: NeurIPS2021]

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad (Y_t \sim X_{\overline{T}-t})$$

※ 既存研究は，スコア関数を正確に推定できることを「仮定」していた.

Reverse process：正規分布からターゲット分布へ巻き戻す

経験スコアマッチング損失

$$\min_{s\in \mathrm{DNN}} \frac{1}{n}\sum_{i=1}^{n}\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t|X_0=x_i}\left[\|s(X_t,t) - \nabla \log p_t(X_t|x_i)\|^2\right] \mathrm{d}t$$

疑問：拡散モデルで学習した分布はどれだけ正確？

仮定
真の分布 p_0 は$[-1,1]^d$上の分布で，
$p_0 \in B^s_{p,q}$ (Besov空間)
ただし $s > (1/p - 1/2)_+$.

定理 (推定誤差) (s: 真の分布の滑らかさ，d: xの次元)

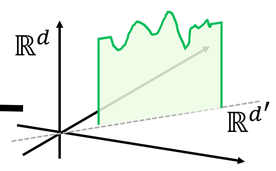
・**TV距離** $\mathbb{E}_{D_n}\left[\mathrm{TV}(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}}\log^9(n).$

これはミニマックス最適：
$n^{-\frac{s}{2s+d}} \lesssim \inf_{\hat{\mu}:\text{estimator}} \sup_{p_0} \mathbb{E}_{D_n}[\mathrm{TV}(\hat{\mu}, X_0)]$

・**W1距離** $\mathbb{E}_{D_n}\left[\mathrm{W}_1(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s+1-\delta}{2s+d'}}.$

これもミニマックス最適 [Niles-Weed & Berthet (2022)].

➢ **真の分布をほぼ最適なレートで推定できる.**



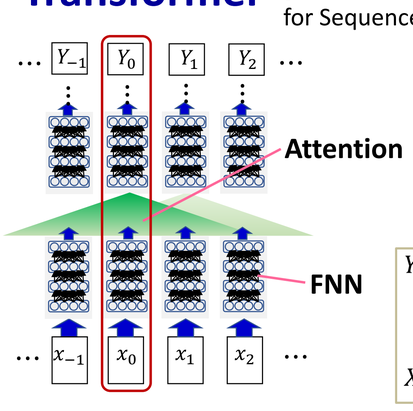
データが低次元多様体上に分布する状況

➢ データの実質的次元に応じて**次元の呪いを回避**.
➢ スコア関数の推定は$t = 0$**付近の方が「難しい」**.
➢ 時刻$t = 0$に近い部分を正確に推定する工夫が必要.

確率過程，関数近似理論，統計的学習理論

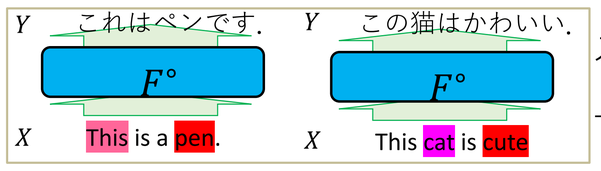
Transformer

[Shokichi Takakura, Taiji Suzuki: Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input. ICML2023]



Transformerは大規模言語モデルの基幹モデル
→ なぜここまでうまくいくのか？

我々の結果：
Transformerは**入力に応じて重要なトークンを選択**し無駄な情報を削ることで次元の呪いを回避できる.



入力応じて重要なトークンは変わる.
→ Attentionで選択.

定理 (推定誤差)

$$\frac{1}{r-l+1}\sum_{j=l}^{r}\mathbb{E}[\|\hat{F}_j - F_j^\circ\|^2_{L_2(P_X)}]$$
$$\lesssim n^{-\frac{2a^\dagger}{2a^\dagger+1}}(\log n)^{2/\alpha+2+\max\{4/\alpha,4\}}$$

(a^+ は関数の滑らかさに関する要約量)

・「重要度」を入出力関係の局所的な滑らかさで特徴付け.
・重要度が局所的に一部のトークンに集中すると仮定.
➢ トークン選択により**次元の呪いを回避**
➢ ほぼ**ミニマックス最適レート**を達成.

平均場ランジュバン動力学の計算・統計理論

平均場ランジュバン動力学の収束解析とNNへの適用

- 平均場ランジュバン動力学の離散化誤差：ICLR2023, NeurIPS2023a.
- 二層NNの最適化と汎化誤差理論：NeurIPS2023b.

機械学習には**確率分布の凸汎関数を最小化**する問題が多数存在.
平均場ランジュバン動力学はこの問題を解く手法である.
(例) 二層ニューラルネットワークの最小化，変分ベイズ法，ノンパラメトリック確率密度推定など

$$\min_{\mu\in\mathcal{P}} \mathcal{L}(\mu) = \underbrace{F(\mu)}_{\text{凸汎関数}} + \lambda_2 \mathrm{Ent}(\mu) \quad (\mu: \text{確率測度})$$

$F(\theta\mu + (1-\theta)\nu) \le \theta F(\mu) + (1-\theta)F(\nu)$
$(\mathrm{Ent}(\mu) = \int \log(\mu)\mathrm{d}\mu)$

➢$\mathcal{L}$を最小化するWasserstein勾配流:

$$\partial_t \mu_t = \nabla\cdot\left[\left(\nabla\frac{\delta F(\mu_t)}{\delta\mu} + \lambda_2\nabla\log(\mu_t)\right)\mu_t\right]$$
$$\mu_t \to \arg\min_{\mu\in\mathcal{P}} \mathcal{L}(\mu)$$

Vanilla GLD:
$\mathcal{L}(\mu) = \int L(x)\mathrm{d}\mu(x) + \lambda_2\mathrm{Ent}(\mu)$ (線形)
$F(\mu) = \int L(x)\mathrm{d}\mu$ (linear)
$\Rightarrow \frac{\delta F(\mu)}{\delta\mu}(\cdot) = L(\cdot)$
$\Rightarrow \mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$

➢これは平均場ランジュバン動力学のFokker-Planck方程式:

$$\mathrm{d}X_t = -\nabla\frac{\delta F(\mu_t)}{\delta\mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t \quad \mu_t = \mathrm{Law}(X_t)$$

時間/空間離散化：

$$X_{k+1}^{(i)} = X_k^{(i)} - \eta_k v_k^i + \sqrt{2\eta_k\lambda_2}\xi_k^{(i)} \quad \xi_k^{(j)} \sim \mathrm{N}(0, I))$$

ただし，$\mathbb{E}[v_k^i] = \nabla\frac{\delta F(\hat{\mu}_k)}{\delta\mu}(X_k^i)$ かつ $\hat{\mu}_k = \frac{1}{M}\sum_{i=1}^{M}\delta_{X_k^{(i)}}$.

定理（離散化アルゴリズムの収束）

時間離散化　空間離散化　確率的勾配

$$\mathscr{L}^{(M)}(\mu_k^{(M)}) - \mathcal{L}(\mu^*) \lesssim \exp(-\lambda_2\eta k\alpha) + \frac{1}{\alpha\lambda}\left(\eta^2 + \lambda\eta + \frac{1}{M} + \eta^{1/2}\lambda^{1/2}\sigma^2\right)$$

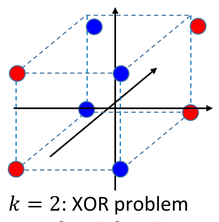
•Suzuki, Wu, Nitanda: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. NeurIPS2023a.
•Taiji Suzuki, Atsushi Nitanda, Denny Wu: Uniform-in-time propagation of chaos for the mean field gradient Langevin dynamics. ICLR2023.

平均場ランジュバン動力学で学習した二層ニューラルネットワークの汎化誤差

[Taiji Suzuki, Denny Wu, Atsushi Nitanda: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. NeurIPS2023b.]

- k-スパースパリティ問題
  - ➢$Z \sim \mathrm{Unif}(\{-1,1\}^d)$　(ただし回転の自由度も許す)
  - ➢$Y = \prod_{j=1}^{k} Z_j$

上位k個の座標だけがラベル Yを決定



$k = 2$: XOR problem
$d = 3, k = 2$

定理 (ニューラルネットワークの判別誤差)

- 設定1: $n > d$
  - ➢テスト判別誤差 = $\boldsymbol{O(d/n)}$
- 設定2: $n > d^2$
  - ➢テスト判別誤差= $\boldsymbol{O(\exp(-n/d^2))}$

※論文内では一般の二値判別問題での判別誤差を導出している

⇒ **NNのサンプル複雑度=** $\boldsymbol{O(d)}$
これはカーネル法を優越する複雑度を達成している：
カーネル法のサンプル複雑度 $\geq \Omega(d^k)$.

Authors	regime/method	k-parity	class error	width	# iterations	
Ji and Telgarsky (2019)	NTK/SGD	No	d^2/n	d^8	d^2/ϵ	カーネル
Telgarsky (2023)	NTK/SGD	No	d^2/n	d^2	d^2/ϵ	
Barak et al. (2022)	Two phase SGD	Yes	$d^{(k+1)/2}/\sqrt{n}$	$O(1)$	d/ϵ^2	NN
Wei et al. (2019)	mean-field/GF	No	d/n	∞	∞	
Telgarsky (2023)	mean-field/GF	No	d/n	d^d	∞	
Ours	mean-field/MFLD	Yes	$\exp(-O(n/d^2))$	$e^{O(d)}$	$e^{O(d)}$	
Ours	mean-field/MFLD	Yes	d/n	$e^{O(d)}$	$e^{O(d)}$	

ニューラルネットワークのサンプル複雑度はkとdが「分離」されている.
⇒ **特徴学習の恩恵**

連合学習の新手法

[Murata, Suzuki: DIFF2: Differential Private Optimization via Gradient Differences for Nonconvex Distributed Learning. ICML2023]

差分プライバシー保証ありの連合学習を実行する効率的な確率的最適化手法Diff2を提案

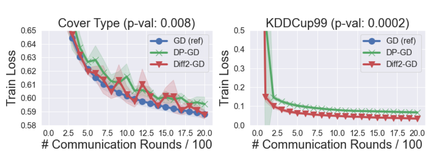
- 理論的貢献: 既存手法より効率性を大きく改善.
- アルゴリズムの性質: Diff2は勾配の差を送る.

$(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$差分プライベートで
$\mathbb{E}[||\nabla f(x_T)||^2] \le ?$

DP-GD (Zhang2017, Wang2018)
$G\left(\frac{\sqrt{d}L}{n\epsilon_{\mathrm{DP}}}\right)$

Diff2 (ours)
$G^{2/3}\left(\frac{\sqrt{d}L}{n\epsilon_{\mathrm{DP}}}\right)^{4/3}$

n: 全クライアントのサンプルサイズ



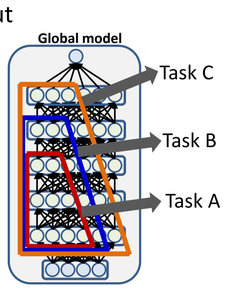
[Nitta, Suzuki, Mulet, Yaguchi, and Hirai: Scalable Federated Learning for Clients with Different Input Image Sizes and Numbers of Output Categories. ICMLA2023]

貢献：各クライアントのタスクが異なる場合でも動く手法を提案
- ➢ **異なる画像サイズ**
- ➢ **異なるラベルの数**

方策：一つの大きな共通モデルから部分モデルを切り出して各クライアントに割り当てる.
→クライアントの問題の複雑さに応じてサイズを自動決定

学習理論で共通モデルを用いることのメリットを証明

$$\hat{L}(\hat{f}) \lesssim L(\hat{f}) + \sum_{j=1}^{M}\left(\sum_{m=j}^{M}\alpha_m\right)\left[\bar{R}_{j,\dot{r}_j} + \dot{r}_j^2 + \frac{\log(M/\delta)}{\bar{n}_j}\right]\log(n)$$



	FedBN [Li *et al.*, 2021]	FedPer [Arivazhagan *et al.*, 2019]	PartialFed-Fix [Sun *et al.*, 2021]	HeteroFL [Diao *et al.*, 2021]	ScalableFL
Local batch normalization layers	√	×	√	√	√
Local output layers	×	√	√	×	√
Adjusting local model widths	×	×	×	√	√
Adjusting local model depths	×	×	×	×	√

その他成果

•入力分布の非等方性の特徴学習への影響：
- ➢ Mousavi-Hosseini, Wu, Suzuki, Erdogdu, NeurIPS2023.
- ➢ Ba, Erdogdu, Suzuki, Wang, Wu, NeurIPS2023.

データの実質的次元が低ければ特徴学習が計算量的にも統計的にも効率的に実現できる.

•分布外汎化に対する特徴学習の理論：
Chen, Huang, Zhou, Bian, Han, Cheng, NeurIPS2023.