Transformers are Minimax Optimal Nonparametric In-Context Learners

ICML 2024 TF2M Workshop Oral, July 27th, Vienna

Juno Kim[1,2] **Tai Nakamaki**[1] **Taiji Suzuki**[1,2]

[1]University of Tokyo [2]RIKEN AIP

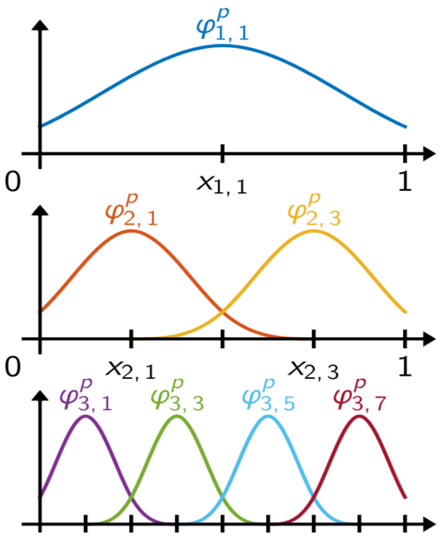
junokim@g.ecc.u-tokyo.ac.jp

Questions on In-Context Learning (ICL)

- Why is few-shot prompting for ICL so effective?
- How does task diversity during pretraining contribute to ICL?
- What is the role of representations learned by MLP layers?
- How optimal is ICL as a learning algorithm?

➡ ***We develop approximation & generalization analyses for ICL from the viewpoint of nonparametric statistical learning theory!***

Setup: Nonparametric Regression



- Input space $\mathcal{X} \subseteq [0,1]^d$
- Basis (e.g. Fourier, wavelet) $(\psi_j^\circ)_{j=1}^\infty \subset L^2(\mathcal{P}_\mathcal{X})$
- **Family** of regression tasks

bounded noise

$$y = F_\beta^\circ(x) + \xi, \quad \mathcal{F}^\circ = \left\{F_\beta^\circ = \sum_{j=1}^\infty \beta_j \psi_j^\circ \,\middle|\, \beta \in \mathbb{R}^\infty, \beta \sim \mathcal{P}_\beta\right\}$$

for hierarchical basis

$\forall N\ \exists \underbrace{\psi_{\underline{N}}^\circ, \cdots, \psi_{\bar{N}}^\circ}_{N}$ s.t.

Assumptions

- partial independence $C_1 \mathbf{I}_N \preceq \left(\mathbb{E}_{x\sim\mathcal{P}_\mathcal{X}}[\psi_j^\circ(x)\psi_k^\circ(x)]\right)_{j,k=\underline{N}}^{\bar{N}} \preceq C_2 \mathbf{I}_N$
- L^∞ growth rate $\left\|\sum_{j=\underline{N}}^{\bar{N}} (\psi_j^\circ)^2\right\|_{L^\infty} \lesssim N^{2r}$
- coefficient decay rate $\|F_\beta^\circ\|_{L^\infty} \le B, \quad \|F_\beta^\circ - F_{\beta,N}^\circ\|_{L^2(\mathcal{P}_\mathcal{X})}^2 \lesssim N^{-2s},$
  $\mathbb{E}_{\beta\sim\mathcal{P}_\beta}[\beta\beta^\top] \lesssim \mathrm{diag}(j^{-2s-1}(\log j)^{-2})$

Setup: MLP+Attn Transformer

$y = F^\circ_\beta(x) + \xi$

$$\mathbf{X}^{(t)} = (x_1^{(t)}, \cdots, x_n^{(t)}), \quad \tilde{x}^{(t)}$$
$$\boldsymbol{y}^{(t)} = (y_1^{(t)}, \cdots, y_n^{(t)})^\top, \tilde{y}^{(t)}$$

prompt query

- Pretraining data: T tasks $F^\circ_{\beta^{(t)}}, t \in [T]$, n i.i.d. samples
- x feeds into N-dim. DNN $\phi \in \mathcal{F}_N$
  - approximation ability $\|\phi\|_{L^\infty} \leq B_N, \quad \|\psi_j^\circ - \phi_j^*\|_{L^\infty} \leq \delta_N$
  - covering entropy $\mathcal{V}(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \epsilon)$
- LSA layer output $f_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}) = \mathrm{clip}_B\left(\frac{1}{n}\sum_{k=1}^{n} y_k \phi(x_k)^\top \underbrace{\Gamma^\top}_{\text{KQ matrices}} \phi(\tilde{x})\right)$, params $\Theta = (\Gamma, \phi)$
- Learn ERM estimator $\widehat{\Theta} = \arg\min_\Theta \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{y}^{(t)} - f_\Theta(\mathbf{X}^{(t)}, \boldsymbol{y}^{(t)}, \tilde{x}^{(t)})\right)^2$

(for optimization dynamics in this setting, see our ICML oral paper)

Upper Bound for In-Context Risk

Theorem

$$\bar{R}(\widehat{\Theta}) := \mathbb{E}\left[\mathbb{E}_{\mathbf{X},\boldsymbol{y},\tilde{x},\beta}\left[(F_\beta^\circ(\tilde{x}) - f_{\widehat{\Theta}}(\mathbf{X},\boldsymbol{y},\tilde{x}))^2\right]\right]$$

$$\lesssim N^{-2s} + N^2\delta_N^4 + N^{2r+1}\delta_N^2$$

▶ DNN approximation error

$$+ \frac{N^{2r}}{n}\log N + \frac{N^{4r}}{n^2}\log^2 N + \frac{N}{n}$$

▶ in-context generalization gap

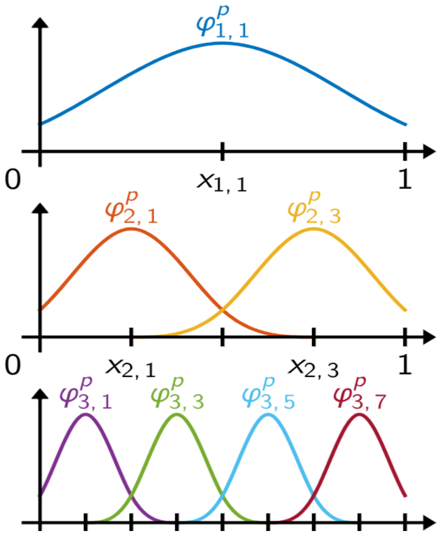
$$+ \frac{N^2}{T}\log\frac{B_N^2}{\epsilon} + \frac{1}{T}\mathcal{V}\left(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \frac{\epsilon}{B_N\sqrt{N}}\right) + \epsilon$$

▶ pretraining generalization gap

ICL is Minimax Optimal in Besov Space

- Task class: unit ball in **Besov space** $\mathcal{F}^\circ = \mathbb{U}(B^\alpha_{p,q}(\mathcal{X})),\ p \in [2,\infty],\ q \in [0,\infty],\ \alpha > d/p$
  - Captures spatial inhomogeneity in smoothness, generalizes Hölder & Sobolev spaces
  - In the supervised setting, DNNs achieve the optimal rate while fixed-kernel methods cannot
- Natural basis: **B-spline wavelets**
  - Multiresolution analysis: forms hierarchy ordered by resolution k

$$\omega^d_{k,\ell}(x) = \prod_{i=1}^{d} \underbrace{\iota_m}_{m\text{-fold convolution of } 1_{[0,1]}}(2^k x_i - \ell_i),\quad k \geq 0,\quad \ell \in \prod_{i=1}^{d}[-m : 2^{k_i}]$$



- Assumptions are verified with $r = 1/2, s = \alpha/d$ if $\mathcal{P}_\mathcal{X}$ has bounded density, coefficients are indep. and $\mathbb{E}_\beta[\beta^2_{k,\ell}] \lesssim 2^{-k(2\alpha+d)}k^{-2}$ (natural decay rate)

ICL is Minimax Optimal in Besov Space

- We apply approximation theory of deep ReLU networks [Suzuki, 19] to obtain:

Theorem

$$\bar{R}(\widehat{\Theta}) \lesssim \boxed{N^{-\frac{2\alpha}{d}}} + \boxed{\frac{N \log N}{n}} + \boxed{\frac{N^2 \log N}{T}}$$

Hence if $T \gtrsim n^{\frac{2\alpha+2d}{2\alpha+d}}$, $N \asymp n^{\frac{d}{2\alpha+d}}$, ICL achieves the optimal rate $n^{-\frac{2\alpha}{2\alpha+d}}$ up to $\log n$.

- LLM pretraining data is nearly infinite in practice: **justifies the effectiveness of ICL at large scales with only few-shot examples**
- Scales suboptimally for small T: *task diversity threshold* [Raventos, 23]
- Curse of dimensionality can be avoided by extending to anisotropic Besov space

Pretraining can Improve ICL Complexity

- Suppose the (unknown) basis is chosen from a wider class $\mathbb{U}(B^{\tau}_{p,q}(\mathcal{X})),\ \tau < \alpha$
- Increased difficulty: complexity of regression is a priori lower bounded by $n^{-\frac{2\tau}{2\tau+d}}$
- For ICL this manifests as increased entropy of $\mathcal{F}_N$ which must be more powerful, however this burden is entirely carried by T

Theorem

$$\bar{R}(\widehat{\Theta}) \lesssim N^{-\frac{2\alpha}{d}} + \frac{N \log N}{n} + \frac{N^{1+\frac{\alpha}{\tau}+\frac{d}{\tau}} \log^3 N}{T}$$

If $T \gtrsim n^{1+\frac{d}{2\alpha+d}\frac{\alpha+d}{\tau}}, N \asymp n^{\frac{d}{2\alpha+d}},$ ICL achieves the rate $n^{-\frac{2\alpha}{2\alpha+d}} \log n,$ improving upon the *a priori* rate by encoding information on the coarser basis during pretraining.

Sequential Input and Transformers

- We also study unbounded sequence inputs $x \in \mathbb{R}^{d\times\infty}$, e.g. entire documents
- Task class: **piecewise γ-smooth class** [Takakura, Suzuki, 23] of smoothness $\alpha \in \mathbb{R}^{d\times\infty}_{>0}$
  - positions of important tokens vary depending on input, requiring dynamical feature extraction
  - γ can be mixed or anisotropic smoothness with $\alpha^\dagger = \max \alpha_{ij}, \ (\sum \alpha_{ij}^{-1})^{-1}$
- DNN class $\mathcal{F}_N$: deep multi-head sliding-window **Transformer networks**

Theorem

Under suitable decay and regularity assumptions, ICL achieves the optimal rate:

$$\bar{R}(\widehat{\Theta}) \lesssim N^{-2\alpha^\dagger} + \frac{N \log N}{n} + \frac{N^{2\vee(1+1/\alpha^\dagger)} \operatorname{polylog}(N)}{T}$$

Information-Theoretic Lower Bound

- We also obtain lower bounds in both n, T by extending the Yang-Barron method and apply to the previous setups
- Holds for any meta-learning scheme for the given regression problem

Theorem

Let Q_1, Q_2 be $\varepsilon_{n1}, \varepsilon_{n2}$-covering numbers of $\mathcal{F}_N, \operatorname{supp} \mathcal{P}_\beta$ and M be the δ_n-packing number of $\mathcal{F}^\circ$ satisfying

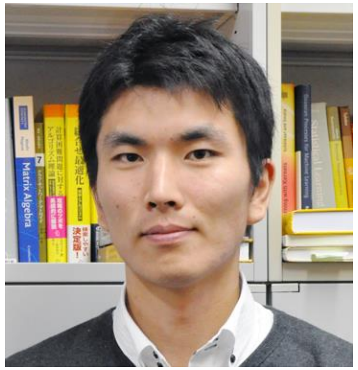
$$\frac{1}{2\sigma^2}\left(n(T+1)\sigma_\beta^2\varepsilon_{n,1}^2 + C_2 n\varepsilon_{n,2}^2\right) \le \log Q_1 + \log Q_2 \le \frac{1}{8}\log M, \quad 4\log 2 \le \log M$$

Then the minimax rate is lower bounded as: $\inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathbb{E}\left[\|\widehat{f} - f^\circ\|_{L^2(\mathcal{P}_\mathcal{X})}^2\right] \ge \frac{1}{4}\delta_n^2$

Q & A

Links

Juno Kim



Taiji Suzuki

contact us!

TF2M oral

ICML oral

References

J. Kim and T. Suzuki. Transformers learn nonlinear features in context: nonconvex mean-field dynamics on the attention landscape. ICML 2024.

A. Raventos, M. Paul, F. Chen, and S. Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. NeurIPS 2023.

T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. ICLR 2019.

S. Takakura and T. Suzuki. Approximation and estimation ability of Transformers for sequence-to-sequence functions with infinite dimensional input. ICML 2023.