How Many Pretraining Tasks Are Needed for In-Context Learning of Linear Regression?

Jingfeng Wu[1], Difan Zou[2], Zixiang Chen[3], Vladimir Braverman[4], Quanquan Gu[3], Peter Bartlett[15]

[1]UC Berkeley, [2]University of Hong Kong, [3]UCLA, [4]Rice University, [5]Google DeepMind

SPOTLIGHT

ICLR

In-context learning

🍓->fruit

🍆->vegetable

🧋->beverage

🧋->?

"solve" new task without updating model

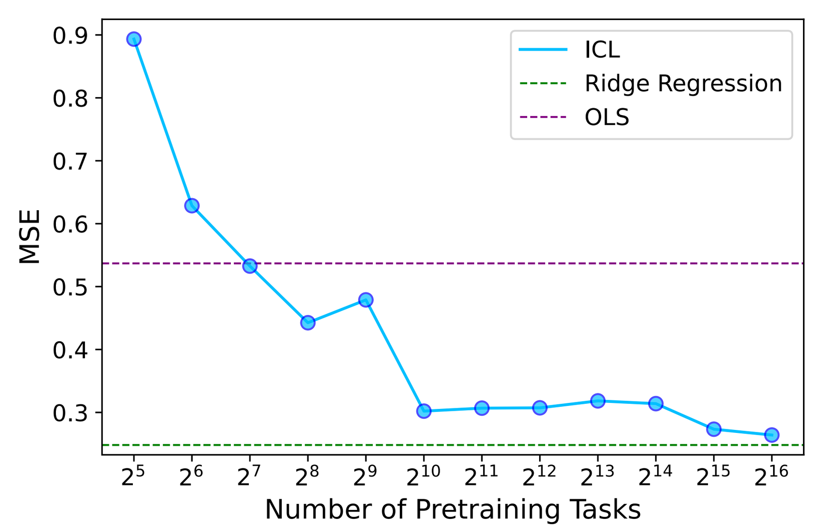
Linear regression

$X \in \mathbb{R}^{*\times d}, \quad Y \in \mathbb{R}^{*}, \quad x \in \mathbb{R}^{d}, \quad y \in \mathbb{R}$

1. task parameter: $\beta \sim \mathcal{N}(0, \psi^2 I_d)$
2. covariate-response: $x \sim \mathcal{N}(0, H)$, $y \sim \mathcal{N}(\beta^\top x, \sigma^2)$
3. context examples: each row of (X, Y) ~iid~ $(x^\top, y)$

(ψ^2, σ^2, H) are fixed (determining the meta-task)

*context length * can vary*



three-layer transformer

An abstract model

$$f: \mathbb{R}^{*\times d} \otimes \mathbb{R}^{*} \otimes \mathbb{R}^{d} \to \mathbb{R}$$
$$(X, Y, x) \mapsto \hat{y}$$

ICL risk at context length n is

$$R_n(f) = \mathbb{E}_{X,Y,x,y}\big(f(X,Y,x) - y\big)^2, \text{ where } \mathtt{dim}(Y) = n$$

Examples

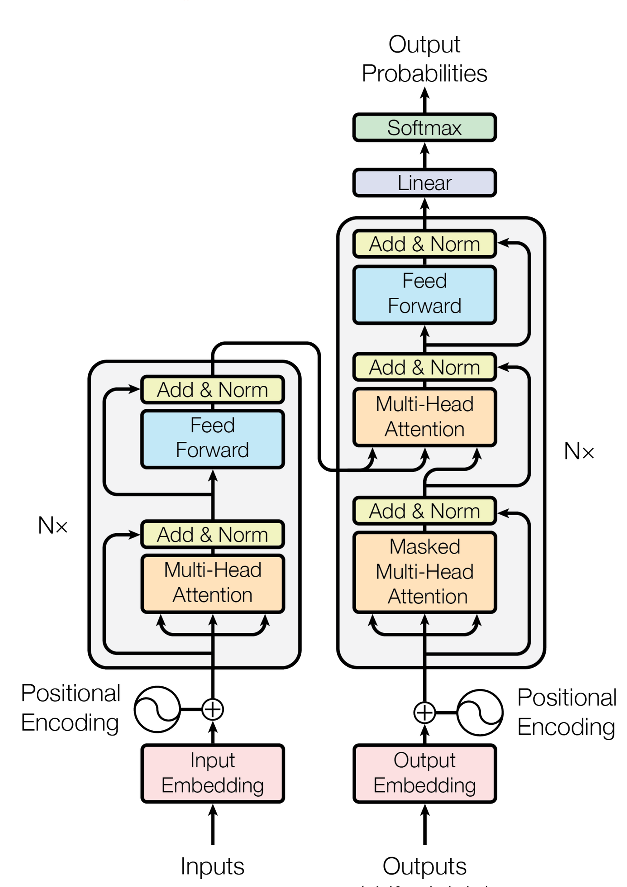
Example 0. Empirical risk minimizer (ERM)

$$\hat{y} := h^*(x),\ h^* := \arg\min_{h \in \mathcal{H}} \|h(X) - Y\|^2$$

zero trainable parameter

Example 1. Transformer

many trainable params



Example 2. Single layer attention

$$Z = \begin{pmatrix} X^\top & x \\ Y^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},$$

$$\hat{y} = \Big(Z + VZ \cdot \mathtt{sfmx}\big((QZ)^\top (KZ)\big)\Big)_{d+1,n+1}$$

V, Q, K are trainable matrix params

Pretraining of an attention model

Simplification 1. Linear attention

$$\hat{y} = \left(Z + VZ \cdot \frac{(QZ)^\top (KZ)}{n}\right)_{d+1,n+1} \qquad Z = \begin{pmatrix} X^\top & x \\ Y^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(n+1)}$$

Simplification 2. Reparameterization

equals to $\hat{y} = \left\langle \boxed{(vW^\top)} \cdot \frac{X^\top Y}{n}, x \right\rangle$ replace by $\Gamma \in \mathbb{R}^{d\times d}$

if the bottom left $1 \times d$ blocks in V and $QK^\top$ are zeros:

$$V = \begin{pmatrix} * & * \\ 0 & v \end{pmatrix} \text{ and } QK^\top = \begin{pmatrix} W & * \\ 0 & * \end{pmatrix}$$

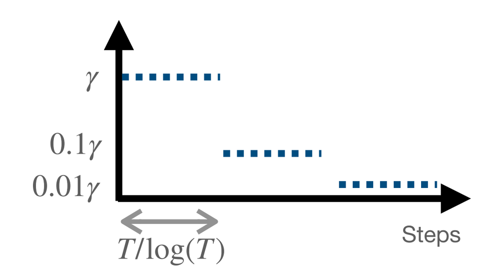
one-step GD with $w_0 = 0$ and trainable matrix stepsize

$$\hat{y} = \langle \hat{w}, x \rangle,\ \hat{w} := w_0 - \Gamma \frac{1}{n} X^\top (X w_0 - Y) = \Gamma \frac{X^\top Y}{n}$$

Simplification 3. Pretraining with fixed context length $n = N$

for $t = 1, \ldots, T$:

1. draw new dataset: $X \in \mathbb{R}^{N\times d}, Y \in \mathbb{R}^{N}, x \in \mathbb{R}^{d}, y \in \mathbb{R}$
2. update: $\Gamma \leftarrow \Gamma - \gamma \nabla_\Gamma (\hat{y} - y)^2$



Pretraining => d^2-dim linear fitting

$$R_N(\Gamma) = \mathbb{E}_{X,Y,x,y}(\hat{y} - y)^2, \text{ where } \mathtt{dim}(Y) = N \quad \hat{y} = \left\langle \Gamma \frac{X^\top Y}{N}, x \right\rangle$$

linearly fit $\left(\frac{X^\top Y}{N} \otimes x,\ y\right)$ with a matrix parameter $\Gamma \in \mathbb{R}^{d\times d}$

Task complexity

Theorem 1. For T steps of pretraining, we have

$$\mathbb{E}R_N(\Gamma_T) - \min R_N \lesssim \left\langle H\tilde{H}, \left(\prod_{t=1}^{T}(I - \gamma_t H\tilde{H})\Gamma^*\right)^2 \right\rangle + \big(\psi^2 \mathtt{tr}(H) + \sigma^2\big) \cdot \frac{D_{\mathtt{eff}}}{T_{\mathtt{eff}}}$$

- $\Gamma^* = \left(\frac{N+1}{N} H + \frac{\mathtt{tr}(H) + \sigma^2/\psi^2}{N} I\right)^{-1} \approx \left(H + \frac{1}{N} I\right)^{-1}$
- $\tilde{H} = \psi^2 \cdot H \left(\frac{N+1}{N} H + \frac{\mathtt{tr}(H) + \sigma^2/\psi^2}{N} I\right) \approx \psi^2 H \left(H + \frac{1}{N} I\right)$
- $T_{\mathtt{eff}} = T/\log(T), \quad D_{\mathtt{eff}} = \sum_{1 \le i,j \le d} \min\left\{1, \gamma^2 T_{\mathtt{eff}}^2 \lambda_i^2 \tilde{\lambda}_j^2\right\}$

λ_i and $\tilde{\lambda}_j$ are eigenvalues of H and $\tilde{H}$

Optimality of ICL

For a model f, its average risk (conditional on X) **at length M** is

$$R_M(f; X) = \mathbb{E}_{Y,x,y}\big(f(X,Y,x) - y\big)^2, \text{ where } \mathtt{dim}(Y) = M$$

Proposition. Tuned **ridge** is Bayesian optimal

$$\hat{y} = \left\langle (X^\top X + \sigma^2/\psi^2 I)^{-1} X^\top Y, x \right\rangle$$

Moreover, if $\psi^2 \mathtt{tr}(H) \lesssim \sigma^2$, then the average risk (w.h.p.) is

$$R_M(\mathtt{ridge}; X) - \sigma^2 \approx \psi^2 \cdot \sum_i \min\{\lambda_i, \mu_M\}, \text{ where } \mu_M \approx \frac{\sigma^2/\psi^2}{M}$$

Near Bayes optimality

$$\hat{y} = \left\langle \Gamma^* \frac{X^\top Y}{N}, x \right\rangle$$

Theorem 2. If $\psi^2 \mathtt{tr}(H) \lesssim \sigma^2$, then the average risk **at length M** is

$$\mathbb{E}R_M(f; X) - \sigma^2 \approx \boxed{\psi^2 \cdot \sum_i \min\{\lambda_i, \mu_M\}} \text{ opt ridge risk}$$
$$+ \underline{\psi^2 (\mu_M - \mu_N)^2} \cdot \sum_i \min\left\{\frac{\lambda_i}{\mu_N^2}, \frac{1}{\lambda_i}\right\} \cdot \min\left\{\frac{\lambda_i}{\mu_M}, 1\right\}$$

where $\mu_M \approx \frac{\sigma^2/\psi^2}{M}$, $\mu_N \approx \frac{\sigma^2/\psi^2}{N}$. small when $M \approx N$

Contributions

- statistical task complexity of pretraining
- optimality of ICL achieved by an attention model
- techniques for analyzing high-order tensors

Open problems

- varying context length?
- non-linearities?

- Garg, Shivam, Dimitris Tsipras, Percy S. Liang, and Gregory Valiant. "What can transformers learn in-context? a case study of simple function classes." NeurIPS 2022
- Akyürek, Ekin, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. "What learning algorithm is in-context learning? Investigations with linear models." ICLR 2022
- Ahn, Kwangjun, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. "Transformers learn to implement preconditioned gradient descent for in-context learning." NeurIPS 2023
- Tsigler, Alexander, and Peter L. Bartlett. "Benign overfitting in ridge regression." JMLR 2024
- Zhang, Ruiqi, Spencer Frei, and Peter L. Bartlett. "Trained transformers learn linear models in-context." JMLR 2024