Minimax Rates for Learning Pairwise Interactions in Attention-Style Models

Shai Zucker[1] Xiong Wang[2] Fei Lu[3] Inbar Seroussi[1,4]

[1]Dept. of Applied Mathematics, Tel Aviv University

[2]School of Mathematics, Sun Yat-sen University

[3]Dept. of Mathematics, Johns Hopkins University

[4]School of Computer Science, Tel Aviv University

Attention as an Interacting Particle System

- We study the following single-layer attention-style Interacting Particle System (IPS) model:

$$Y_i = \frac{1}{N-1} \sum_{j \neq i} \phi_\star(X_i^\top A_\star X_j) + \eta_i$$

- This model captures the core structure of a self-attention layer: interactions are determined by pairwise bilinear scores $X_i^\top A_\star X_j$, where $A_\star$ plays the role of a query-key matrix:

$$A_\star = \frac{1}{\sqrt{d_k}} W_Q W_K^\top.$$

- We aim to learn the induced pairwise interaction

$$g_\star(x, y) = \phi_\star(x^\top A_\star y).$$

- Main question: **what is the minimax convergence rate in sample size and how does the rate depend on d?**

Upper Bound

Estimator: ERM over low-rank matrices and piecewise-polynomial activations, given M i.i.d. samples of N tokens

$$\widehat{g}_M = \operatorname*{argmin}_g \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\left|Y_i^m - \frac{1}{N-1}\sum_{j\neq i} g(X_i^m, X_j^m)\right|^2$$

Theorem (Informal)

If there are enough samples such that,

$$rd \leq \left(\frac{M}{\log M}\right)^{\frac{1}{2\beta+1}},$$

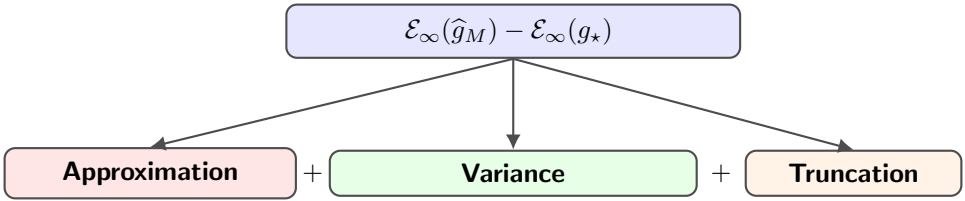
where $\beta > 0$ is the smoothness index of the scalar interaction function $\phi_\star$ and r is the rank of the matrix $A_\star$, then the error behaves as a **scalar nonparametric rate** up to logarithmic factors.

$$\sup_{g_\star} \mathbb{E}\|\widehat{g}_M - g_\star\|_{L^2_\rho}^2 \lesssim M^{-\frac{2\beta}{2\beta+1}}$$

Upper Bound Proof Strategy

- **Nonlocal dependency.** Y_i is defined over a sum of the target function $g_\star$, so we first prove this aggregation can be decomposed by a **coercivity condition** bounding the local error by the aggregated population error defined as $\mathcal{E}_\infty(g_\star)$.

$$\frac{1}{N-1}\mathbb{E}\left[\|\widehat{g}_M - g_\star\|_{L^2_\rho}^2\right] \leq \mathcal{E}_\infty(\widehat{g}_M) - \mathcal{E}_\infty(g_\star).$$



Minimax Lower Bound

Lemma (Reduction to single-index risk)

Let $A_\star$ follow the low rank assumption. For every estimator of the form $\hat{g}(x,y) = \hat{\phi}(x^\top \hat{A} y)$, we have:

$$\inf_{\hat{A},\hat{\phi}} \left\| \hat{g} - \phi_\star(x^\top A_\star y) \right\|_{L^2_\rho}^2 \geq \inf_{\psi_M \in L^2(P_U)} \int_{\mathbb{R}} \left| \psi_M(u) - \phi_\star(u) \right|^2 dP_U(u)$$

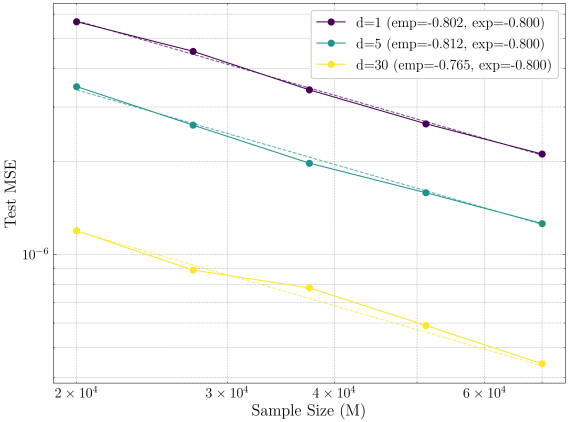
where $U = x^\top A_\star y$ and P_U is its induced law.

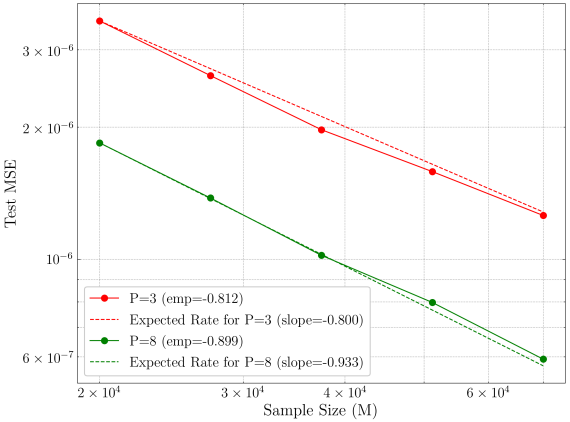
Theorem (Informal)

Fix a smoothness index $\beta > 0$ and radius $L > 0$. There exists a constant $C_{\text{lower}} > 0$, such that:

$$\inf_{\substack{\widehat{A} \in \mathbb{R}^{d \times d} \\ \widehat{\phi} \in \mathcal{L}^2_\rho}} \sup_{\phi_\star \in \mathcal{C}(\beta, L)} \left[\left\| \widehat{\phi}(x^\top \widehat{A} y) - \phi_\star(x^\top A_\star y) \right\|_{L^2_\rho}^2 \right] \geq C_{\text{lower}} N^{-\frac{2\beta}{(2\beta+1)}} M^{-\frac{2\beta}{(2\beta+1)}}$$

Numerical experiments





Varying $d \in \{1, 5, 30\}$**:** nearly parallel slopes.

Varying smoothness: smoother $\phi_\star$ gives steeper slope.

Observed slopes match $M^{-2\beta/(2\beta+1)}$ **and depend on** β **, not on** d.

Take-home message

- We show that attention models can learn high-dimensional interactions at a scalar rate independent of the embedding dimension.

$$\|\widehat{g}_M - g_\star\|_{L^2_\rho}^2 \asymp M^{-\frac{2\beta}{2\beta+1}} \qquad \text{(up to logarithmic factors)}$$

- This rate is achieved under the condition $rd \leq (M/\log M)^{1/(2\beta+1)}$.
- **This result highlights that the attention model is able to avoid the curse of dimensionality.** In other words, our analysis provides insight into why attention models maintain strong performance in high-dimensional settings.

Thank you!

Paper on OpenReview