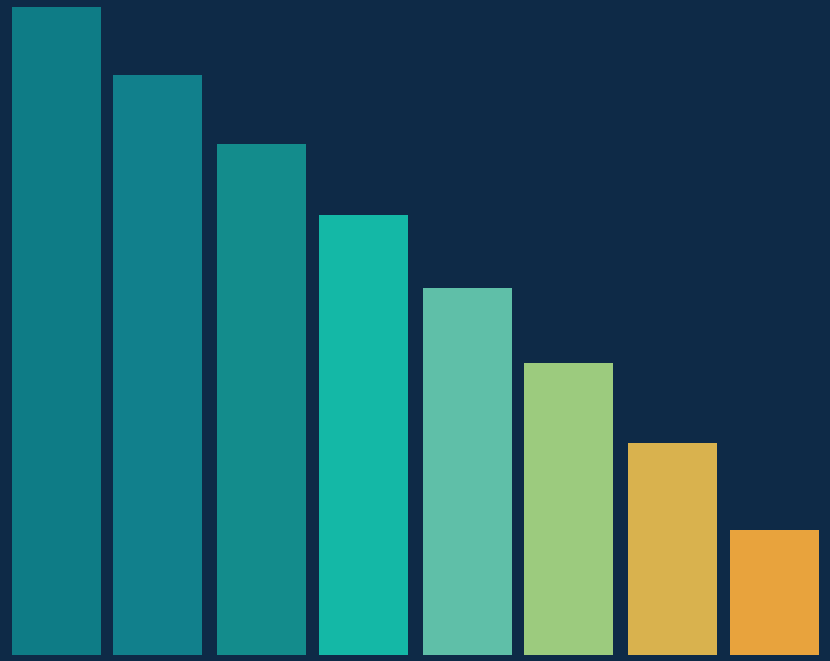
ICML 2026

Alignment-Sensitive Minimax Rates for Spectral Algorithms with Learned Kernels

A new complexity measure that bridges fixed-kernel theory and adaptive learning.



Dongming Huang (National University of Singapore)

Zhifan Li · Yicheng Li · Qian Lin

Adaptive learning breaks the classical assumptions

Classical fixed-kernel theory

- The kernel is fixed before seeing the target.
- Need
  - a polynomial decay condition for the eigenvalues of the kernel $\lambda_j \asymp j^{-\gamma}$
  - a source (smoothness) condition on the signal $\sum_j \langle f^*, \phi_j \rangle^2 / \lambda_j^s \leq R_s$

These assumptions describe one prescribed spectrum and cannot speak to a spectrum that changes.

vs

Modern adaptive learning

- Training reshapes the representation, so the induced kernel evolves.
- Generalization improves whenever this raises the alignment between the kernel and the target.

The adapted spectrum rarely satisfies eigen-decay, and the source exponent is unknown.

Goal: a complexity measure that stays valid and informative if the kernel is learned from data.

The Effective Span Dimension (ESD)

Start with a sequence model:

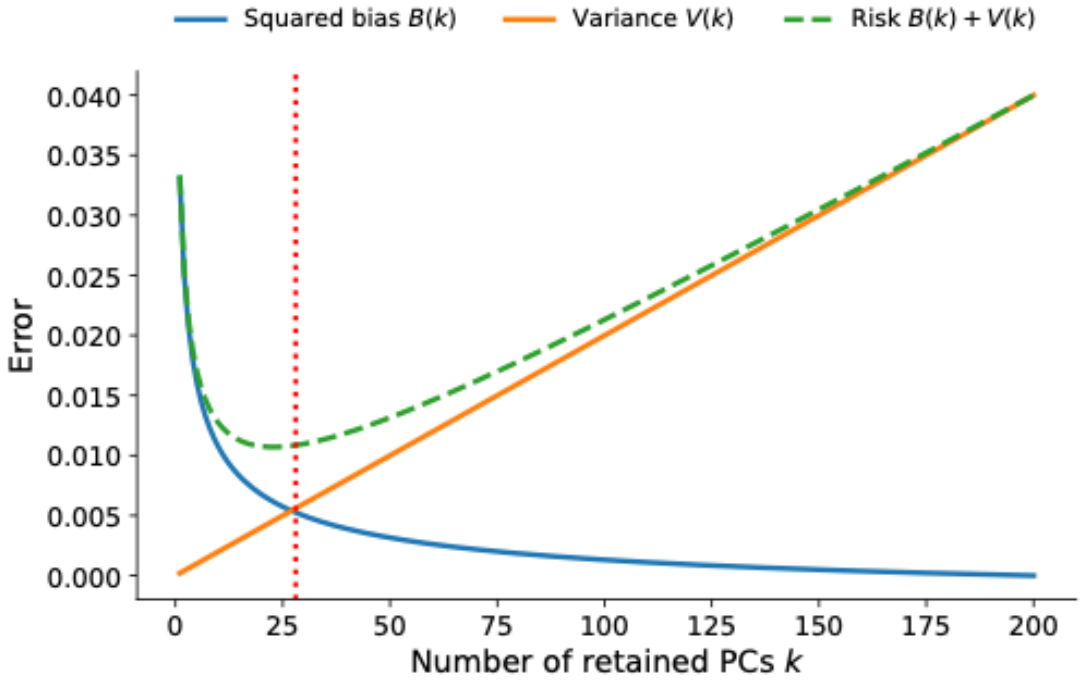
$$Z_j = \theta_j^* + \xi_j, \quad \mathrm{Var}(\xi_j) = \sigma^2, \quad j = 1, \ldots, d$$

obs signal noise

Suppose the spectrum is $\lambda = \{\lambda_1, \lambda_2, \ldots\}$ and $\{\pi_i\}$ be its decreasing ordering.

Effective Span Dimension

$$d^{\dagger}(\sigma^2; \theta^*, \lambda) = \min\{ k : \frac{1}{k} \sum_{i=k+1}^{d} (\theta^*_{\pi_i})^2 \le \sigma^2 \}$$



- The fewest (k) leading directions to keep so that the leftover signal energy drops below $k \cdot \sigma^2$.
- This is the bias–variance crossing of the principal-component estimator.

Different from other measures

- It depends jointly on the signal θ^*, the spectrum λ, and the noise variance σ^2.
- Signal-agnostic measures (such as effective dimension) ignore where the signal sits, so they cannot see alignment.

Minimax rates for ESD-bounded classes

$\mathcal{F}_{K,\boldsymbol{\lambda}}^{(\sigma^2)}$ *: a class of signals with ESD at most K*

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \mathcal{F}_{K,\boldsymbol{\lambda}}^{(\sigma^2)}} \mathcal{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \asymp K\sigma^2$$

1 **No classical assumptions**

Neither a source condition nor an eigenvalue-decay law is required.

2 **Recovers the classics**

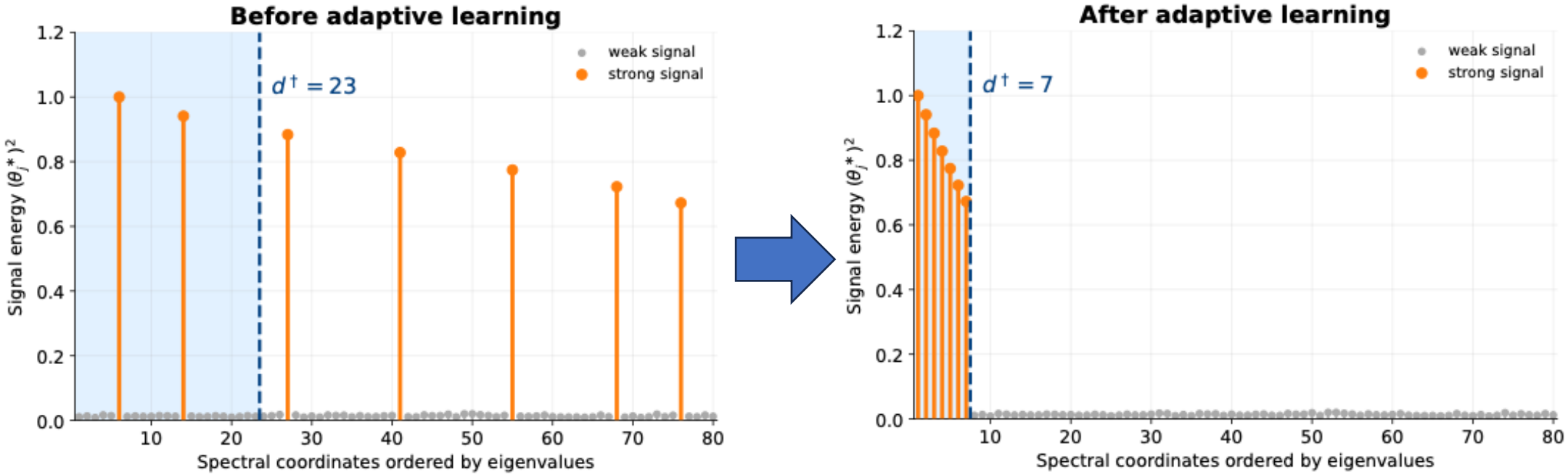
When the s-source condition and β-power eigenvalue-decay condition do hold, this recovers the classical source-condition rate $n^{-s\,\beta/(1+s\,\beta)}$

3 **Broadly applicable**

The framework extends to linear regression and RKHS kernel regression.

Feature learning via ESD reduction

Adaptation reshapes the kernel



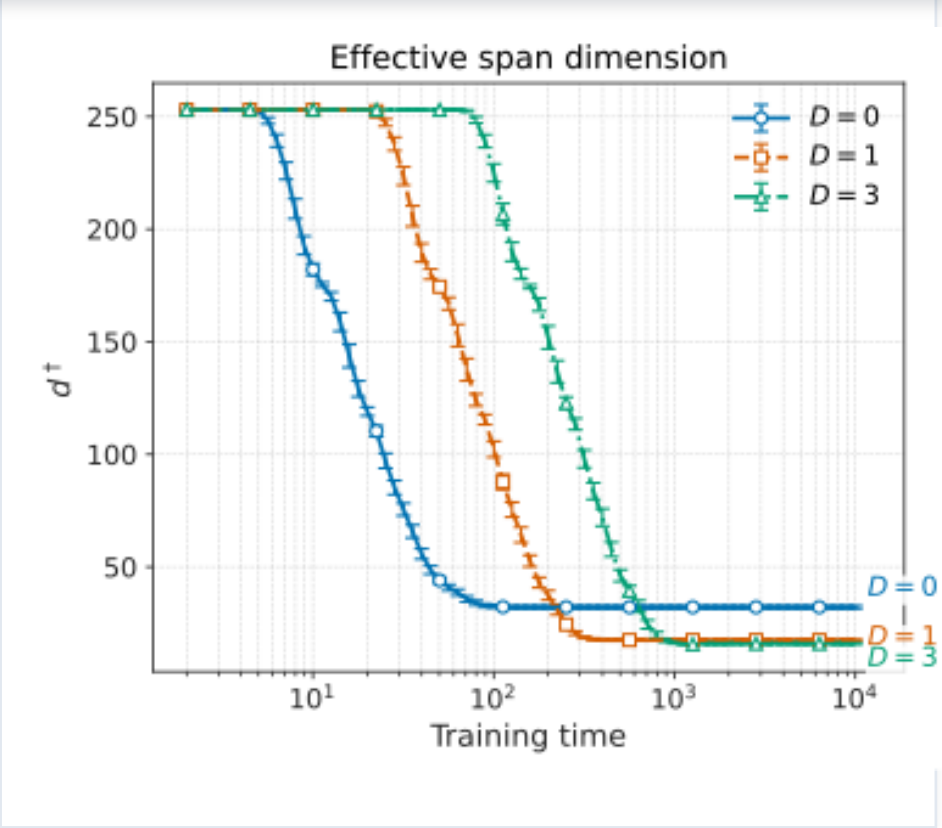
- Badly-aligned kernel → heavy tail → large ESD → **minimax-hard** class.

- Better-aligned kernel → light tail → ESD drops → the *same* target lands in an **easier** class.

Effective span dimension vs. training time

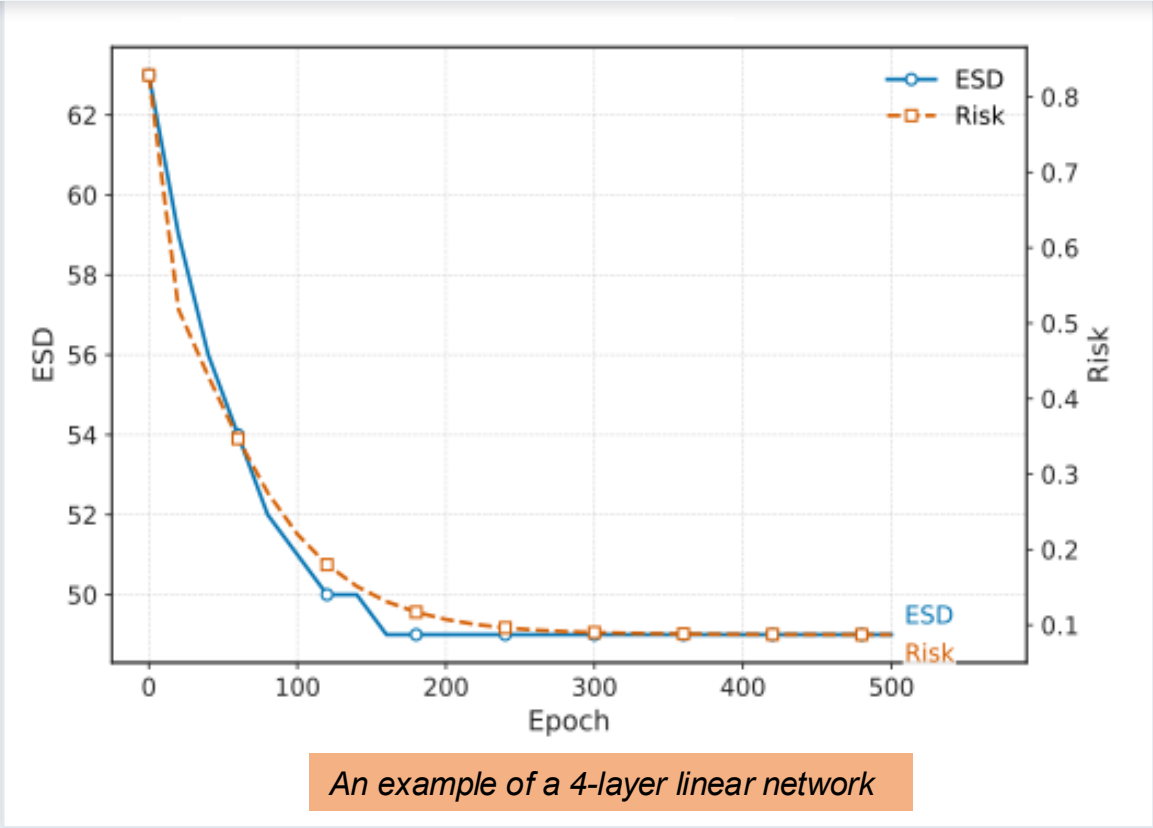
Proven result under a fixed eigenbasis

- **Over-parameterized gradient flow** learns the eigenvalues and reduces the ESD $\theta_j = a_j b_j^D \beta_j$



Pathwise ESD for evolving eigenfunctions

- Freeze the learned kernel at each time (t), read its eigensystem, get a pathwise ESD $d^\dagger(\sigma^2; f^*, \mathbf{k}_t)$



An example of a 4-layer linear network

Takeaway

1

An alignment-sensitive complexity measure

ESD reads statistical difficulty from how the signal, the spectrum, and the noise level.

2

Simple rates that applied to learned kernels

It pins the minimax excess risk at $K \cdot \sigma^2$, without source or eigen-decay conditions, while still recovering the classical rates.

3

Reinterpreting adaptation

It reframes successful feature learning as ESD reduction.

Future directions: *evolving eigenfunctions such as in learned neural networks.*