Alignment-Sensitive Minimax Rates for Spectral Algorithms with Learned Kernels

Dongming Huang, Zhifan Li, Yicheng Li, and Qian Lin

1. Motivation

Question: ***How do we quantify statistical difficulty or target–kernel alignment after the kernel has been learned?***

Classical rates $n^{-s\beta/(1+s\beta)}$ for spectral algorithms with kernel $\mathbf{k}(x,x') = \sum_j \lambda_j \phi_j(x)\phi_j(x')$ usually assume the polynomial eigen-decay and source conditions:

$$\lambda_j \asymp j^{-\beta}, \qquad \sum_j \frac{\langle f^*, \phi_j\rangle^2}{\lambda_j^s} \le R_s.$$

Classical assumptions may fail when the kernel evolves over time: the learned spectrum may not decay polynomially and the index s may be unclear.

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

2. Effective Span Dimension

A sequence model with zero mean & uncorrelated noise:

$$\underbrace{Z_j}_{\text{obs.}} = \underbrace{\theta_j^*}_{\text{signal}} + \underbrace{\xi_j}_{\text{noise}}, \quad \mathrm{Var}(\xi_j) = \sigma^2, \quad j = 1, \dots, d \ (d \text{ can be } \infty).$$

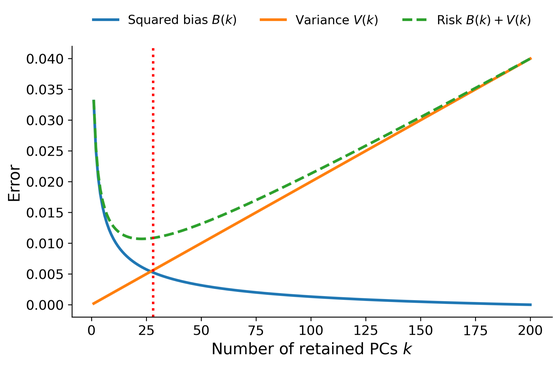
Given a spectrum $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots)$ and a filter $\psi_\nu(\cdot)$, spectral estimator: $\hat\theta_j = (1 - \psi_\nu(\lambda_j))Z_j$, $j \in [d]$; e.g., a spectral-cutoff estimator keeps coordinates with larger λ_j.

Definition

Let $\{\pi_j\}_1^d$ order the eigenvalues decreasingly. Define the Effective Span Dimension (ESD) as

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

$d^\dagger$ is the cutoff where the squared bias of the spectral-cutoff estimator $B(k) = \sum_{i>k}(\theta^*_{\pi_i})^2$ is no larger than the accumulated variance $V(k) = k\sigma^2$, indicated by the dotted line in the bias–variance plot below.



Unlike signal-agnostic measures such as effective dimension, ESD captures target-kernel alignment.

3. Minimax Risk

Given a quota $K \in [d]$, define the ESD-bounded class

$$\mathcal{F}^{(\sigma^2)}_{K,\boldsymbol{\lambda}} = \{\boldsymbol{\theta} : d^\dagger(\sigma^2; \boldsymbol{\theta}, \boldsymbol{\lambda}) \le K\}.$$

Minimax characterization

Consider risk $\mathcal{R}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{\theta}^*}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$. The minimax risk over this class grows linearly with K:

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

$d^\dagger\sigma^2$ captures the statistical difficulty of the problem.

4. Examples

- **Classical source ellipsoid.** If $\lambda_i = i^{-\beta}$, the source condition is the ellipsoid

$$\Theta_s(R) = \left\{\boldsymbol{\theta} : \sum_{i=1}^d \lambda_i^{-s}\theta_i^2 \le R\right\} = \left\{\boldsymbol{\theta} : \sum_{i=1}^d i^{s\beta}\theta_i^2 \le R\right\}.$$

  It embeds into an ESD class with

$$K_{\mathrm{src}} = \min\left\{d, \left\lceil (R/\sigma^2)^{1/(1+s\beta)}\right\rceil\right\}, \quad \Theta_s(R) \subseteq \mathcal{F}^{(\sigma^2)}_{K_{\mathrm{src}},\boldsymbol{\lambda}}.$$

  Taking $\sigma^2 = \sigma_0^2/n$ gives the rate

$$\sigma^2 K_{\mathrm{src}} \asymp \sigma_0^2 \min\{n^{-s\beta/(1+s\beta)}, d/n\}.$$

- **Near-parametric rates beyond classical assumptions.** Let $\boldsymbol{\lambda}$ be decreasing. For $b \ge 1$, set $g(x) = \sigma_0^2 x e^{-x^b}$, $\theta_1^* = 0$, and $(\theta^*_{j+1})^2 = g(j) - g(j+1)$. Then with $K_n = \lceil (\log n)^{1/b} \rceil$,

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathcal{F}^{(\sigma_0^2/n)}_{K_n,\boldsymbol{\lambda}}} \mathcal{R}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) \asymp \sigma_0^2 K_n/n \asymp \sigma_0^2 (\log n)^{1/b}/n.$$

  The classical rate does not yield the $(\log n)^{1/b}/n$ term.

- **Same eigenvalues, different difficulty.** Let $S = \mathrm{supp}(\boldsymbol{\theta}^*)$ and suppose $|S| = s \ll d$. Compare two spectra with the same eigenvalue multiset.
  (1) [Well-aligned] If the largest s eigenvalues of $\boldsymbol{\lambda}^{(1)}$ are on S, then $d^\dagger(\sigma^2; \boldsymbol{\theta}^*, \boldsymbol{\lambda}^{(1)}) \le s$.
  (2) [Misaligned] If the largest $d - s$ eigenvalues of $\boldsymbol{\lambda}^{(2)}$ are on S^c, then $d^\dagger(\sigma^2; \boldsymbol{\theta}^*, \boldsymbol{\lambda}^{(2)}) \ge \min\{d - s, \|\boldsymbol{\theta}^*\|^2/\sigma^2\}$.
  Spectrum-only measures cannot distinguish (1) & (2).

5. ESD in Regression Models with Noise Variance σ_0^2

- **Fixed-design linear model** $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$. Given SVD $n^{-1/2}\mathbf{X} = \mathbf{U}_{n\times r}\mathbf{S}_{r\times r}\mathbf{V}^\top_{r\times p}$, transform

$$\mathbf{Z} = n^{-1/2}\mathbf{U}^\top\mathbf{Y}, \quad \boldsymbol{\theta}^* = \mathbf{S}\mathbf{V}^\top\boldsymbol{\beta}^*, \quad \lambda_j = s_j^2, \quad \sigma^2 = \sigma_0^2/n.$$

  Define $d^\dagger(\sigma_0^2/n; \boldsymbol{\beta}^*, \mathbf{X}) := d^\dagger(\sigma_0^2/n; \boldsymbol{\theta}^*, \boldsymbol{\lambda})$.

- **RKHS** $y = f^*(x) + \varepsilon$. For $\mathbf{k}(x,x')$, set

$$\theta_j^* = \langle f^*, \phi_j\rangle_{L^2(\mu)}, \qquad \bar\sigma^2 = \frac{\sigma_0^2 + \|f^*\|_\infty^2}{n}.$$

  Define $d^\dagger(\bar\sigma^2; f^*, \mathbf{k}) := d^\dagger(\bar\sigma^2; \boldsymbol{\theta}^*, \boldsymbol{\lambda})$.

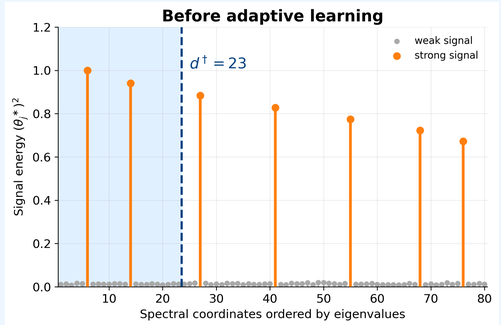
Minimax characterization can also be established.

6. Feature Learning

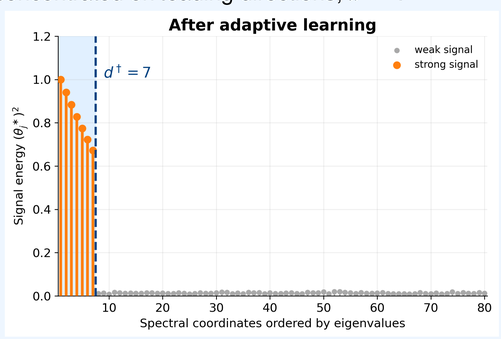
If the kernel is changed during training, one may compare the ESD before and after feature learning.

Example: one signal, two spectral orderings

Before adaptation: a badly-aligned kernel, strong signal coordinates appear late in the spectral order, $d^\dagger = 23$



After adaptation: a better-aligned kernel, strong signal is concentrated on leading directions, $d^\dagger = 7$



Risk interpretation

If adaptation changes $\boldsymbol{\lambda}^{(0)}$ to $\boldsymbol{\lambda}^{(a)}$ and

$$d^\dagger(\sigma^2; \boldsymbol{\theta}^*, \boldsymbol{\lambda}^{(a)}) < d^\dagger(\sigma^2; \boldsymbol{\theta}^*, \boldsymbol{\lambda}^{(0)}),$$

the same signal moves from a **minimax-hard class** with a larger ESD quota to an **easier class** with a smaller quota.

General mechanism. Feature learning reshapes the kernel so the same target has a smaller ESD at the relevant noise level, which means it moves into an easier class.

Trade-off function and span profile. Define

$$\mathbf{H}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}}(k) = \frac{1}{k}\sum_{i=k+1}^{d}(\theta^*_{\pi_i})^2, \ \mathbf{D}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}}(\tau) = \min\{k : \mathbf{H}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}}(k) \le \tau\}.$$

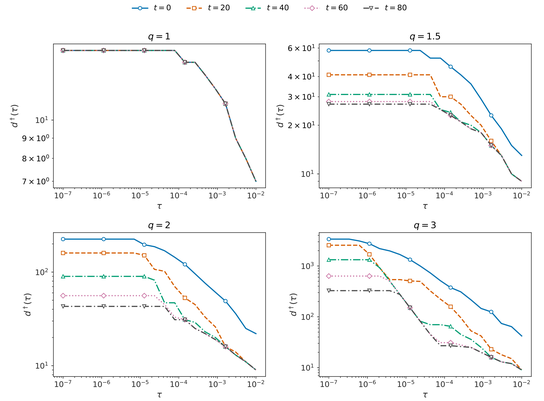
$\sigma^{-2}\mathbf{H}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}}(k)$ is the spectral-cutoff bias–variance ratio, and $\mathbf{D}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}}(\tau) = d^\dagger(\tau; \boldsymbol{\theta}^*, \boldsymbol{\lambda})$ is the span profile.
We expect $\mathbf{D}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}}(\cdot)$ moves down during feature learning.

7. ESD reduction in over-parameterized sequence models

Estimate: $\hat\theta_j(t) = \tilde\lambda_j(t)^{1/2}\beta_j(t)$, where $\tilde\lambda_j(t) = (a_j(t)b_j(t)^D)^2$ and (a_j, b_j, β_j) are learned using gradient flow. This learns the eigenvalues while fixing the eigencoordinates. We analyze

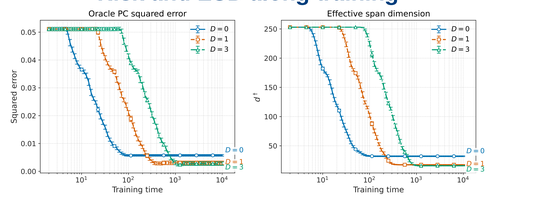
$$d^\dagger(t) = d^\dagger(\sigma^2; \boldsymbol{\theta}^*, \tilde{\boldsymbol{\lambda}}(t)).$$

Span-profile evolution under OP-GF



Setup: $\lambda_j = j^{-\gamma}$; nonzeros at $\ell(j) = \lfloor j^q \rfloor$, $\theta^*_{\ell(j)} = Cj^{-(p+1)/2}$; others zero. Here $q \in (1, 1.5, 2, 3)$, $n = 10^4, \sigma_0 = 1, d = 5000, J = 15, p = 2.5, \gamma = 1$ and depth $D = 0$. Larger q: worse alignment.

Risk and ESD along training



Here $q = 2.5$, with the same $n, \sigma_0, d, J, p, \gamma$ as above; compare $D \in (0, 1, 3)$. Mean over 20 Monte Carlo runs; bars: one standard error. The plotted risk is the squared error of the spectral-cutoff estimator tuned at the current ESD.

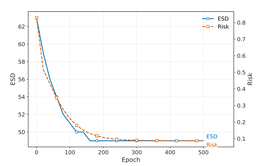
8. Beyond fixed eigenbasis: pathwise ESD

For a learned kernel $\mathbf{k}_t$ at time t, we track the pathwise ESD

$$d^\dagger(t) = d^\dagger(\bar\sigma^2; f^*, \mathbf{k}_t).$$

Example: training a 4-layer linear neural network

- $\mathbf{A}(t)$ is the learned representation matrix in $f_t(x) = w(t)^\top\mathbf{A}(t)x$
- $\mathbf{k}_t(x,x') = x^\top\mathbf{A}(t)^\top\mathbf{A}(t)x'$



Both $d^\dagger(t)$ and risk decrease during training. $p = 900, n = 1000, \beta_j^* = j^{-1.1}$ for $j \le 200$. Risk is $\|\mathbf{A}(t)^\top w(t) - \beta^*\|_2^2$.

9. Take-Home Message

1. ESD depends jointly on the target, kernel, and noise level.
2. For an ESD quota K, minimax risk $\asymp K\sigma^2$.
3. Successful adaptation should reduce the ESD of the same signal.

Scan for paper

arXiv:2509.20294