Risk Phase Transitions in Spiked Regression

Alignment-Driven Benign and Catastrophic Overfitting

Jiping Li (UCLA) Rishi Sonthalia (Boston College)

ICLR 2026

UCLA

BOSTON COLLEGE

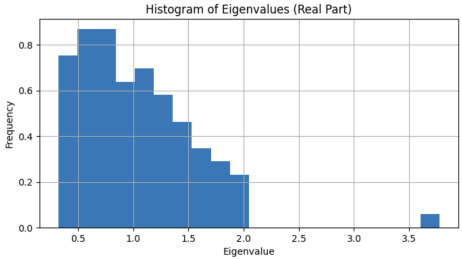
ICLR

Why Study Spiked Models?

- Many learned representations are **anisotropic**:
  - **Signal** (spike that captures key information)
  - **Noise** (bulk that "buries" the signal)

Motivation

We want a precise map of when spike alignment helps generalization and when it hurts.



Problem Setup

Data model

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{A} \in \mathbb{R}^{d\times n}, \quad \boldsymbol{Z} = \theta \boldsymbol{u}\boldsymbol{v}^\top \text{ (signal)}, \quad \boldsymbol{\Sigma} = \theta^2 \boldsymbol{u}\boldsymbol{u}^\top + \tau^2 \boldsymbol{I}_d.$$

- Noise Assumption: $\mathbb{E}[A_{ij}^2] = \tau^2$ with zero mean

Target model

For the i-th label: $y_i = \alpha_Z \boldsymbol{z}_i^\top \boldsymbol{\beta}_* + \alpha_A \boldsymbol{a}_i^\top \boldsymbol{\beta}_* + \varepsilon_i, \quad \|\boldsymbol{\beta}_*\|^2 = 1$

- $\mathbb{E}[\varepsilon_i^2] = \tau_\varepsilon^2$ with mean zero.
- $(\boldsymbol{\beta}_*^\top \boldsymbol{u})^2$: target alignment.

Minimum-norm interpolator

$$\boldsymbol{\beta}_{\text{int}} = \boldsymbol{X}^\dagger \boldsymbol{y}.$$

Asymptotic Proportional Regime

$$c = d/n.$$

Generalization Error

Model Specification

- $\alpha_Z = \alpha_A$: well-specified
- $\alpha_Z \neq \alpha_A$: mis-specified.

For a new test data point $\{\tilde{\boldsymbol{x}}, \tilde{y}\}$, with potentially different parameters $\tilde{\theta}, \tilde{\tau}, \tilde{\tau_\varepsilon}, \tilde{\alpha}_A, \tilde{\alpha}_Z$

Generalization error

$$\mathcal{R}(\boldsymbol{\beta}_{int}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon},\{\tilde{\boldsymbol{x}},\tilde{\varepsilon}\}}\left[(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{y}})^2\right] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon},\tilde{\boldsymbol{x}},\tilde{\varepsilon}}\left[(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{x}}^T\boldsymbol{\beta}_{int})^2\right]$$

What Do We Measure?

Taxonomy of asymptotic overfitting behavior

$$\mathcal{R}_c = \lim_{n,d\to\infty,\, d/n\to c} \mathcal{R}(\boldsymbol{\beta}_{int}) - \tilde{\tau}_\varepsilon^2.$$

- **Benign**: $\lim_{c\to\infty} \mathcal{R}_c = 0$
- **Tempered**: $0 < \lim_{c\to\infty} \mathcal{R}_c < \infty$
- **Catastrophic**: $\lim_{c\to\infty} \mathcal{R}_c = \infty$

1. When does alignment between $\boldsymbol{\beta}_*$ and the spike $\boldsymbol{u}$ improve generalization?
2. How does spike size change benign, tempered, and catastrophic overfitting?

Two Spike-Size Regimes

Operator Norm Scaling

$$\theta^2 = \gamma\tau^2.$$

- γ tunes the spike strength relative to the bulk (noise).
- BBP transition ("visible" spike) occurs at $\gamma = (1+\sqrt{c})^2$.

Frobenius Norm Scaling

$$\theta^2 = d\tau^2.$$

- $\mathbb{E}[\|\boldsymbol{Z}\|_F^2] = \mathbb{E}[\|\boldsymbol{A}\|_F^2]$.
- "Stronger" spike signal (scale with d).

Takeaway

These two regimes give qualitatively different generalization behavior.

Well-Specified Case: Exact Risk Theorem

In this case ($\alpha_A = \alpha_Z = \tilde{\alpha}_A = \tilde{\alpha}_Z = \alpha > 0$), the target $\boldsymbol{y}$ is a direct linear function of $\boldsymbol{X}$.

Theorem (Well-Specified Risk)

Given data $(\boldsymbol{X}, \boldsymbol{y})$ and $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$ generated according to our setting. If the well-specification condition $\alpha_Z = \alpha_A = \tilde{\alpha}_Z = \tilde{\alpha}_A = \alpha > 0$ holds, the asymptotic excess risk $\mathcal{R}_c$ is:

$$\mathcal{R}_c = \begin{cases} \tau_\varepsilon^2 \frac{c}{1-c} & \text{if } c < 1 \\ \tau_\varepsilon^2 \frac{1}{c-1} + \alpha^2 \tau^2 \left(1 - \frac{1}{c}\right) \left[\|\boldsymbol{\beta}_*\|^2 + (\boldsymbol{\beta}_*^\top \boldsymbol{u})^2 \frac{\theta^2 \tau^2 c^2 - 2\theta^2 \tau^2 c - \theta^4}{(\theta^2 + \tau^2 c)^2} \right] & \text{if } c > 1 \end{cases}.$$

Well-Specified Case: Operator Norm Scaling

In **operator-norm scaling**, for overparameterized models ($c > 1$),

$$\mathcal{R}_c = \alpha^2\tau^2\left(1 - \frac{1}{c}\right)\left(\|\boldsymbol{\beta}_*\|^2 + \frac{\gamma c^2 - 2\gamma c - \gamma^2}{(\gamma + c)^2}(\boldsymbol{\beta}_*^\top \boldsymbol{u})^2\right) + \tau_\varepsilon^2 \frac{1}{c-1}.$$

$\lim_{c\to\infty} \mathcal{R}_c = \alpha^2\,\tau^2\left(\|\boldsymbol{\beta}_*\|^2 + \gamma(\boldsymbol{\beta}_*^\top \boldsymbol{u})^2\right) \implies$ always **tempered overfitting**

Representative Conclusion

For operator norm scaling, alignment is beneficial *if and only if*

$$\gamma > c(c-2).$$

- If $\gamma = \Theta_c(1)$ (constant with respect to c), this threshold is **not the same** as the BBP threshold $\gamma > (1 + \sqrt{c})^2$, so a visible spike does *not* automatically mean alignment helps.

Well-Specified Case: Operator Norm Scaling (Continued)

If instead $\gamma = \omega_c(1)$ (γ grows with c),

- Anti-aligned ($\boldsymbol{\beta}_*^\top \boldsymbol{u} = 0$), always **tempered overfitting**
- Aligned ($\boldsymbol{\beta}_*^\top \boldsymbol{u} \neq 0$ (say $= \|\boldsymbol{\beta}_*\|^2$)), mixed behaviors:

$$\lim_{c\to\infty} \mathcal{R}_c = \alpha^2\tau^2 \cdot \begin{cases} \infty & \text{if } \omega_c(1) \le \gamma \le o_c(c^2) \\ \|\boldsymbol{\beta}_*\|^2 + (\frac{1}{\phi} - 1)(\boldsymbol{\beta}_*^\top \boldsymbol{u})^2 & \text{if } \gamma = \phi c^2 \text{ for const. } \phi > 0 \\ \|\boldsymbol{\beta}_*\|^2 - (\boldsymbol{\beta}_*^\top \boldsymbol{u})^2 & \text{if } \gamma = \omega_c(c^2) \end{cases}$$

As $\gamma \uparrow$, the overfitting regime: **catastrophic** $\to$ **tempered** $\to$ **benign**.

Well-Specified Case: Frobenius Norm Scaling

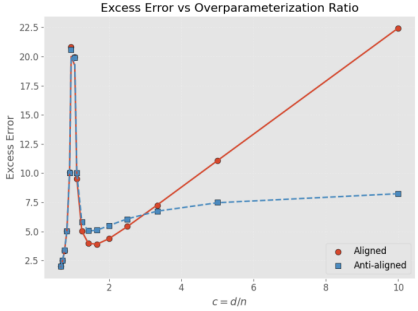
For Frobenius norm scaling, the excess risk for $c > 1$ simplifies to:

$$\mathcal{R}_{c>1} = \alpha^2 \tau^2 \left(1 - \frac{1}{c}\right) \left(\|\boldsymbol{\beta}_*\|^2 - (\boldsymbol{\beta}_*^\top \boldsymbol{u})^2\right) + \tau_\varepsilon^2 \frac{1}{c-1}.$$

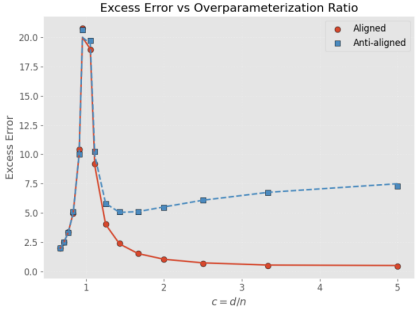
Representative Conclusion

Alignment is always beneficial *when overparameterized* ($c > 1$).

Well-Specified Case: Phase Transitions in Practice



(a) Operator norm scaling ($\theta^2 = c\tau^2$). Alignment initially improves generalization, but have catastrophic risk as $c \to \infty$. Anti-alignment yields tempered risk.



(b) Equal Frobenius norm scaling ($\theta^2 = d\tau^2$). Alignment leads to benign overfitting, while anti-alignment results in tempered risk.

Most surprising result

Increasing spike strength can induce catastrophic overfitting *before* benign overfitting appears.

Misspecified Case: Alignment Has Beneficial Region

In this case ($\alpha_A = \tilde{\alpha}_A \neq \alpha_Z = \tilde{\alpha}_Z$), the target $\boldsymbol{y}$ is no longer a direct linear function of $\boldsymbol{X}$.

Theorem (Mis-specified Risk)

Given data $(\boldsymbol{X}, \boldsymbol{y})$ and $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$ generated according to our setting. If the mis-specification condition holds with $\alpha_Z = \tilde{\alpha}_Z$, $\alpha_A = \tilde{\alpha}_A$, define $\Delta_c := \alpha_Z - \frac{\alpha_A}{c}$, $\Delta_1 := \alpha_Z - \alpha_A$, then

$$\mathcal{R}_c = \begin{cases} \tau_\varepsilon^2 \frac{c}{1-c} + \tau^2 \left(\boldsymbol{\beta}_*^\top \boldsymbol{u}\right)^2 \frac{\Delta_1^2}{1-c} \frac{\theta^2}{\theta^2+\tau^2} & c < 1 \\ \tau_\varepsilon^2 \frac{1}{c-1} + \alpha_A^2 \tau^2 \|\boldsymbol{\beta}_*\|^2 \left(1 - \frac{1}{c}\right) + \tau^2 \left(\boldsymbol{\beta}_*^\top \boldsymbol{u}\right)^2 \Delta_c^2 \frac{\theta^2}{\theta^2+\tau^2 c} \left[\frac{c}{c-1} \frac{\theta^2+\tau^2 c^2}{\theta^2+\tau^2 c} - 2\frac{\alpha_A}{\Delta_c}\right] & c > 1 \end{cases}$$

- Alignment is always detrimental for $c < 1$.
- The ratio α_Z/α_A defines a region where alignment is beneficial for $c > 1$.

Misspecified Case: Beneficial Regions

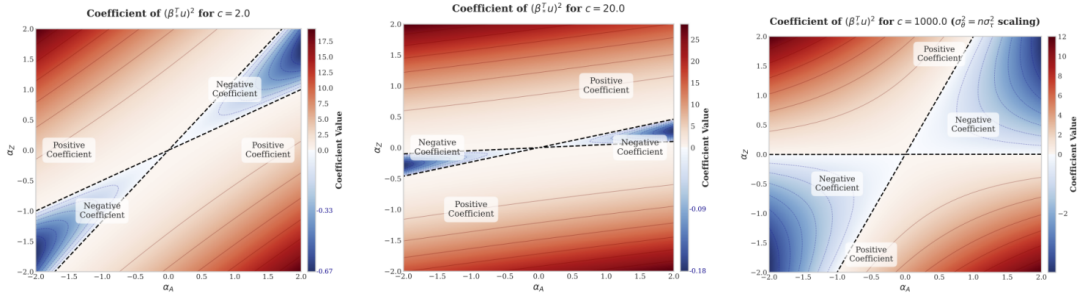
- For **operator norm scaling**, the beneficial region **shrinks as** $c \to \infty$:

$$\frac{1}{c} \leq \frac{\alpha_Z}{\alpha_A} \leq \frac{1}{c}\left(\frac{3c^2 - \gamma + 2c\gamma - 2c}{(c^2 + \gamma)}\right).$$

- For **Frobenius norm scaling**, the beneficial region **persists as** $c \to \infty$:

$$\frac{1}{c} \leq \frac{\alpha_Z}{\alpha_A} \leq 2 - \frac{1}{c}.$$

Beneficial Region Plots



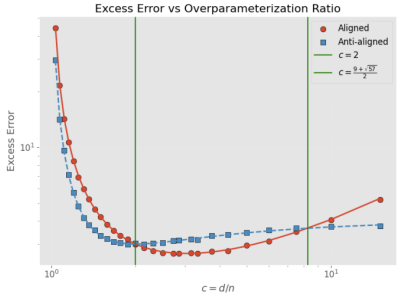
(a) Operator norm scaling, $c = 2$. Large beneficial region.

(b) Operator norm scaling, $c = 20$. Smaller beneficial region

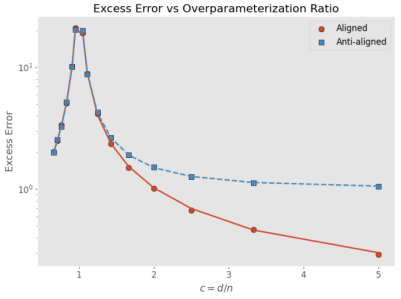
(c) Frobenius norm scaling, $c = 1000$. The beneficial region persists at extreme overparameterization.

Figure 3: Phase boundaries for spike alignment impact. Coefficient of $(\boldsymbol{\beta}_*^\top \boldsymbol{u})^2$ as a function of α_Z/α_A, indicating whether alignment improves or harms generalization.

Misspecified Case: Phase Transitions in Practice



(a) Under operator norm scaling ($\theta^2 = c\tau^2$) with $\alpha_Z = 1$, $\alpha_A = 2$, alignment initially improves generalization for small c, but becomes harmful beyond a critical point, leading to catastrophic overfitting.



(b) Under Frobenius norm scaling ($\theta = \sqrt{d}\tau$) with $\alpha_A = 1$ and $\alpha_Z = 1.1$, alignment remains better than anti-alignment across all c, but benign overfitting is not achieved unless $\alpha_Z = \alpha_A$.

Most surprising result

The same alignment can become harmful beyond a critical point (see operator norm scaling).

Risk Decomposition Theorem

General Theorem (Informal)

The risk $\mathcal{R}$ can be decomposed into four terms.

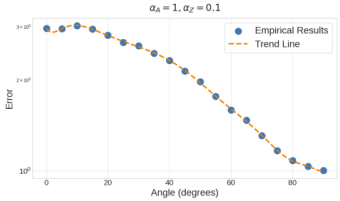
- **Bias**: error along the spike.
- **Variance**: norm growth of interpolator.
- **Noise floor**: irreducible error.
- **Target alignment**: cross-term that can amplify or dampen risk!

Intuition

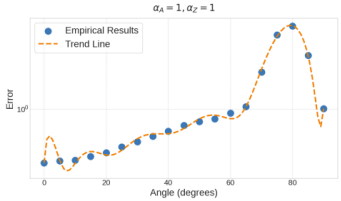
The phase transitions come from competition among these four terms.

These Phase Transitions Persist in Nonlinear Models

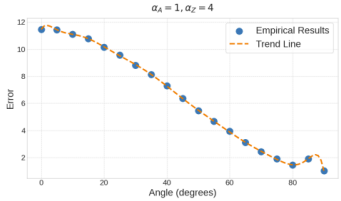
For 3-layer ReLU networks, we have



(a) $\alpha_Z = 0.1$, alignment helps.



(b) $\alpha_Z = 1$, mixed behavior.



(c) $\alpha_Z = 4$, alignment hurts.

Why this matters

Alignment-driven phase transitions are not just an artifact of our linear model.

We also have results for covariate shifts ($\alpha_Z \neq \tilde{\alpha}_Z$ or $\alpha_A \neq \tilde{\alpha}_A$). Check out the paper!

Scaling	Benign	Tempered	Catastrophic
Well-Specified, No Covariate Shift: $\alpha_A = \tilde{\alpha}_A = \alpha_Z = \tilde{\alpha}_Z = \alpha > 0$			
$\theta^2 = \gamma\tau^2$	$\gamma = \omega_c(c^2), \boldsymbol{\beta}_* \parallel \boldsymbol{u}$	All other cases	$o_c(c^2) \geq \gamma \geq \omega_c(1), \boldsymbol{\beta}_* \not\perp \boldsymbol{u}$
$\theta^2 = d\tau^2$	$\boldsymbol{\beta}_* \parallel \boldsymbol{u}$	$\boldsymbol{\beta}_* \nparallel \boldsymbol{u}$	Never
Misspecified, No Covariate Shift: $\alpha_A = \tilde{\alpha}_A, \alpha_Z = \tilde{\alpha}_Z, \alpha_A \neq \alpha_Z$			
$\theta^2 = \gamma\tau^2$	Never	All other cases	$o_c(c^2) \geq \gamma \geq \omega_c(1), \boldsymbol{\beta}_* \not\perp \boldsymbol{u}$
$\theta^2 = d\tau^2$	Never	Always	Never
Misspecified with Covariate Shift: $\alpha_A \neq \tilde{\alpha}_A$ or $\alpha_Z \neq \tilde{\alpha}_Z$			
$\theta^2 = \gamma\tau^2$	Never	All other cases	$\alpha_Z \neq \tilde{\alpha}_Z, \boldsymbol{\beta}_* \not\perp \boldsymbol{u}, \gamma = \omega_c(1)$ or $\alpha_Z = \tilde{\alpha}_Z, \boldsymbol{\beta}_* \not\perp \boldsymbol{u}, \omega_c(1) \leq \gamma \leq o_c(c^2)$
$\theta^2 = d\tau^2$	$\alpha_Z = \tilde{\alpha}_Z = \tilde{\alpha}_A, \boldsymbol{\beta}_* \parallel \boldsymbol{u}$	All other cases	$\alpha_Z \neq \tilde{\alpha}_Z$ and $\boldsymbol{\beta}_* \not\perp \boldsymbol{u}$
Spike Recovery: $\alpha_A = \tilde{\alpha}_A = 0, \alpha_Z = \tilde{\alpha}_Z$			(Appendix D)
$\theta^2 = \gamma\tau^2$	$\gamma\tau^2 = o_c(1)$	$\gamma\tau^2 = \Theta_c(1)$	$\gamma\tau^2 = \omega_c(1)$
$\theta^2 = d\tau^2$	$\tau^2 = o_c(1)$	$\tau^2 = \Theta_c(1)$	Never

Please visit our poster session.

Thank you!