PhD Defense:

Random Matrix Theory for Linearized Neural Networks



Collin Cranston

University of California, San Diego

June 9, 2026

Acknowledgments

- Chapter 2 is extracted from "Collin Cranston, Zhichao Wang, Todd Kemp. NTK Regression under dual power-law model: Deterministic Equivalents via SDE and PDE Methods." Currently under review. Currently under review.
- Chapter 3 is extracted from "Collin Cranston, Zhichao Wang, Todd Kemp, Michael W. Mahoney. Eigen-Spike Emergence and Quadratic Equivalents for Conjugate Kernels on Nonlinearly Separable Data, Proceedings of the 39th Annual Conference on Learning Theory, *Proceedings of Machine Learning Research (PMLR)*. arXiv:2605.29669 " To appear.



Todd Kemp



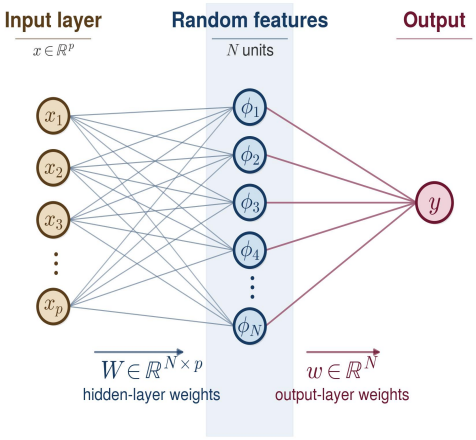
Zhichao Wang



Michael Mahoney

Abstract

- Nonlinear Random Matrix Theory (RMT) and Free Probability in Deep Learning.
- Universality result, called *Gaussian Equivalence*, allows for the behavior of complex nonlinear neural networks to be understood through tractable linear kernel models.
- This motivates the question whether this idealized linear equivalence remains meaningful beyond isotropic Gaussian setting.



Single hidden-layer Neural Network

Outline

Chapter 2: NTK Regression under dual power-law model: Deterministic Equivalents via SDE and PDE Methods

- Model settings and motivation
- Outline of approach
- Main results
- Experiments

Chapter 3: Eigen-Spike Emergence and Quadratic Equivalents for Conjugate Kernels on Nonlinearly Separable Data

- Background and motivation
- Main results
- Experiments

Conclusion

Chapter 2: NTK Regression under dual power-law model: Deterministic Equivalents via SDE and PDE Methods

Two-Layer Neural Network

Consider a two-layer Neural Network

$$f_\theta(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \boldsymbol{w}^\top \sigma(\boldsymbol{W}_1 \boldsymbol{x}), \quad \boldsymbol{\theta} := \left(\boldsymbol{w}_1^{(1)}, \ldots, \boldsymbol{w}_N^{(1)}, w_1, \ldots, w_N\right) \in \mathbb{R}^{N(p+1)}$$

- $\boldsymbol{x} \in \mathbb{R}^p$ is the input data, p is the ambient feature dimension, N is the network width.
- $\sigma : \mathbb{R} \to \mathbb{R}$ is the entry-wise nonlinear activation function (e.g ReLU, tanh, sigmoid).
- First layer weights $\boldsymbol{W}_1 \in \mathbb{R}^{N \times p}$ are initialized as $(\boldsymbol{W}_1)_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, p^{-1})$.
- Second layer weights $\boldsymbol{w} \in \mathbb{R}^N$ are initialized as $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_W^2)$

Lazy Training Regime

When the width $N \to \infty$, the weights $\boldsymbol{\theta}$ stay close to their initialization $\boldsymbol{\theta}_0$ throughout training via gradient descent. Moreover, the model output $f_\theta(\boldsymbol{x})$ is well-approximated by a first-order Taylor expansion around $\boldsymbol{\theta}_0$. [JGH18]

Linearize the neural network, i.e perform a first-order Taylor expansion around $\boldsymbol{\theta}_0$.

$$f_\theta(\boldsymbol{x}) \approx f(\boldsymbol{x};\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta}_0)$$

- $f(\boldsymbol{x};\boldsymbol{\theta}_0)$ denotes the network output at initialization.
- Perform centering by setting $f(\boldsymbol{x};\boldsymbol{\theta}_0) = 0$ [AP20].

Neural Tangent Kernel (NTK)

The (finite-width) NTK is the gram matrix of the Jacobian of the network output respect to weights $\boldsymbol{\theta}$:

$$\boldsymbol{K}^{\mathrm{NTK}}(\boldsymbol{x},\boldsymbol{x}') := \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}';\boldsymbol{\theta}_0)$$

For data matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$,

$$\boldsymbol{K}^{\mathrm{NTK}} = \boldsymbol{J}^\top \boldsymbol{J}, \quad \boldsymbol{J} := \nabla_{\boldsymbol{\theta}} f(\boldsymbol{X};\boldsymbol{\theta}_0) \in \mathbb{R}^{N(p+1)\times n}$$

Why study the NTK?

Introduced to study training dynamics of ultra-wide Neural Networks (NNs) through tractable kernel models [JGH18].

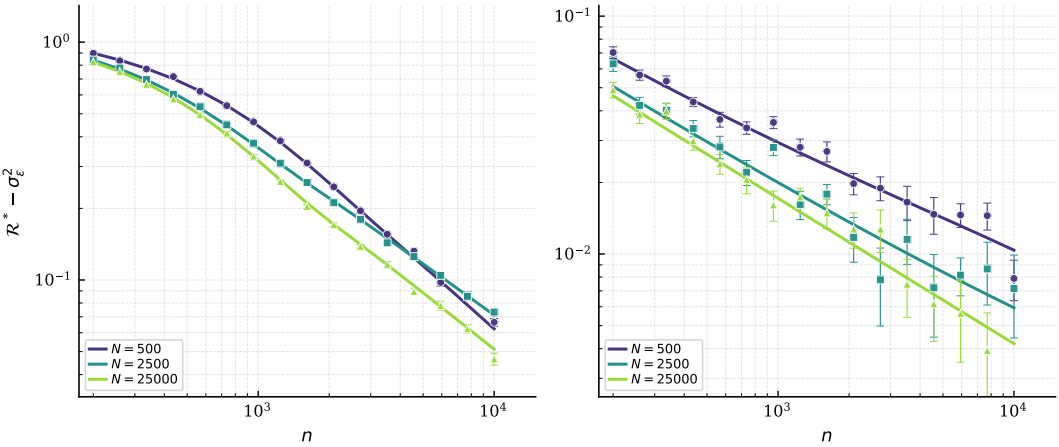
- Under gradient-flow dynamics, the NTK is approximately constant and training reduces to linear regression in tangent features.
- Spectral properties of NTK govern generalization performance of NNs.
- We focus on the finite-width realization of NTK for its connection to NNs of practical width.
- Scaling laws[1] in finite-width NTK remain largely open theoretically.

$$\mathcal{L}(n, N) = \left[\left(\frac{N_c}{N} \right)^{\alpha_N/\alpha_n} + \left(\frac{n_c}{n} \right) \right]^{\alpha_n}$$

Key references: Jacot et al. (2018), Chizat et al. (2019), Kaplan et al. (2020), Anwar et al. (2024)

[1] Model performance scales as a power-law in samples n and parameters N. (Kaplan et al. 2020)

Scaling Law



Optimally-tuned excess NTK risk: Plot versus sample size n with log-log axis. **(Left)** slowly-decaying spectrum $\alpha = 0.2$, with visible curvature on log-log plot. **(Right)** rapidly-decaying spectrum $\alpha = 0.8$, exhibiting a power-law decay in n. Empirical test loss reported as markers $\pm$SE over 20 trials. Plots generated with $r = 1.5, p = 1200, \sigma_\varepsilon^2 = 0.3, \sigma_W^2 = 1.0$

Learning Problem:

- Goal: learn unknown target function $f^* : \mathbb{R} \to \mathbb{R}$ from n training samples $\{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$.
- Student-teacher model: $y_i = f^*(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$
- Consider model defined by the linearization of $f_\theta(\cdot)$ around $\boldsymbol{\theta}_0$.
- Perform (approximate) kernel ridge regression (KRR) via the finite-width NTK with ridge parameter $\lambda > 0$.

NTK Predictor:

$$\hat{f}_\theta^\lambda(\boldsymbol{x}) = \boldsymbol{k}_{\mathrm{NTK}}(\boldsymbol{x})^\top \left(\boldsymbol{K}^{\mathrm{NTK}} + \lambda \mathbf{I}_n\right)^{-1} \boldsymbol{y}$$

- $\boldsymbol{k}_{\mathrm{NTK}}(\boldsymbol{x}) \in \mathbb{R}^n$ has entries $\boldsymbol{K}^{\mathrm{NTK}}(\boldsymbol{x}, \boldsymbol{x}_\alpha)$ for $\alpha \in [n]$.
- Ridge parameter $\lambda > 0$ corresponds to ℓ^2-regularization of parameter weights towards initialization.
- As $N \to \infty$, converges to deterministic kernel that describes network trajectory during training under gradient descent.

Training:

- **Empirical Risk minimization (ERM)** of ℓ^2-regularized Mean Squared Error (MSE):

$$L_\lambda(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_\theta^\lambda(\boldsymbol{x}_i)\right)^2 + \frac{\lambda}{n}||\boldsymbol{\theta}||_2^2$$

- **Closed-form solution**: Optimal kernel ridge predictor:

$$\widehat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L_\lambda(\boldsymbol{\beta}) = \boldsymbol{k}_{\mathrm{NTK}}(\boldsymbol{x})^\top \left(\boldsymbol{K}^{\mathrm{NTK}} + \lambda \mathbf{I}_n\right)^{-1}\boldsymbol{y}$$

- **Training error** under optimal predictor $\widehat{\boldsymbol{\beta}}_\lambda$:

$$\mathrm{E}_{\mathrm{train}} = \frac{1}{n}||\boldsymbol{y} - \hat{f}_\theta^\lambda(\boldsymbol{X})||^2 = \frac{\lambda^2}{n}\boldsymbol{y}^T\left(\boldsymbol{K}^{\mathrm{NTK}} + \lambda \mathbf{I}_n\right)^{-2}\boldsymbol{y}$$

Testing:

- Prediction risk:

$$\mathcal{R}(\lambda) = \mathbb{E}_{\boldsymbol{x}}\left[\left(\hat{f}_\theta^\lambda(\boldsymbol{x}) - y\right)^2\right]$$

  - Expectation is over an independent test point $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$ conditional on training dataset and random weights.
  - Random variable, but will concentrate around its mean (over $\boldsymbol{X}, \varepsilon$) in the proportional asymptotic regime.

Main goal

Derive deterministic equivalents[2] for the random variables

$$\mathrm{E}_{\text{train}} = \frac{1}{n}||\boldsymbol{y} - \hat{f}_\theta^\lambda(\boldsymbol{X})||_2^2, \qquad \mathcal{R}(\lambda) = \mathbb{E}_{\boldsymbol{x}}\left[\left(y - \hat{f}_\theta^\lambda(\boldsymbol{x})\right)^2\right]$$

[2] finite-dimensional deterministic surrogates that approximate the random matrix under scalar functionals (traces, bilinear forms)

Focus on dataset that allows for scaling laws in the excess prediction risk $\mathcal{R}(\lambda) - \sigma_\varepsilon^2$.

(A0): Dual Power-Law Model

The training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are drawn i.i.d from $\mathcal{N}(0, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \cdots, \sigma_p), \qquad \sigma_j = j^{-\alpha}, \quad j \in [p] \tag{1.1}$$

for some fixed $\alpha \in (0, \infty)$. The labels are generated as a linear power-teacher.

$$\boldsymbol{y}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \boldsymbol{\beta}_j = j^{-r}, \quad j \in [p] \tag{1.2}$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ and $r > 0$ is a fixed constant.

- Capacity exponent α controls effective dimensionality of data, with larger values causing lower effective dimensionality.
- Source exponent $r > 0$ controls difficulty of learning problem, measures alignment of signal $\boldsymbol{\beta}$ onto top eigenvectors of $\boldsymbol{\Sigma}$.

Focus on dataset that allows for scaling laws in the excess prediction risk $\mathcal{R}(\lambda) - \sigma_\varepsilon^2$.

(A0): Dual Power-Law Model

The training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are drawn i.i.d from $\mathcal{N}(0, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \cdots, \sigma_p), \qquad \sigma_j = j^{-\alpha}, \quad j \in [p]$$

for some fixed $\alpha \in (0, \infty)$. The labels are generated as a linear power-teacher.

$$\boldsymbol{y}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \boldsymbol{\beta}_j = j^{-r}, \quad j \in [p]$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ and $r > 0$ is a fixed constant.

- *Signal captured by coordinate j* is given by $\beta_j^2 \sigma_j \sim j^{-(2r+\alpha)}$; the exponent $2r + \alpha$ dictates the scaling rate.
- Entries of $\boldsymbol{W}_1 \boldsymbol{X}$ are marginally $\mathcal{N}(0, \tau^2)$, where $\tau^2 := \frac{1}{p} \operatorname{Tr} \boldsymbol{\Sigma}$.

Main Assumptions

We make the following general assumptions.

Assumptions

(A1) Proportional Regime. $n, p, N \to \infty$ at rates $\frac{n}{p} = \phi$ and $\frac{N}{p} = \psi$ for some $\phi, \psi \in (0, \infty)$.

(A2) Capacity Exponent. We take $\mathbf{\Sigma}_{jj} = j^{-\alpha}$ with $\alpha > 0$.

(A3) Source Exponent. The signal vector satisfies $\boldsymbol{\beta}_j = j^{-r}$ with $r > \frac{1}{2}$.

(A4) Nonlinearity. The non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$ is Lipschitz, centered with respect to $\mathcal{N}(0, \tau^2)$, and has finite Gaussian moments of all order.

- Assumption (A3) ensures $||\boldsymbol{\beta}||_2 < \infty$ as $p \to \infty$.
- Assumption (A4) allows for most common activation functions (after centering), e.g ReLU, sigmoid, tanh, softplus.

Simplifying Principles

We apply three **simplifying principles**

Simplifying Principles

(**C1**) **Gaussian Equivalence I**: Replace nonlinear matrices with **linear** Gaussian surrogates with matching first and second order moments.

(**C2**) **Gaussian Equivalence II**: Asymptotic training and test error of NTK regression are asymptotically equivalent to that of regression on **linearized** kernel.

(**C3**) **Generalized Cross Validation (GCV)**: GCV is a uniformly consistent estimator of prediction risk.

We conjecture (C1)-(C3) extend to the anisotropic power-law setting.

Key references: Adlam et al. (2019), Hu–Lu (2022), Montanari–Saeed (2022), Peche et al. (2019)

Universality principle from RMT in proportional asymptotic ($n \asymp N \asymp p$) analysis of nonlinear kernel models.

Gaussian Equivalence I

$$\sigma(\boldsymbol{W}_1\boldsymbol{X}) \approx b_1\boldsymbol{W}_1\boldsymbol{X} + b_2\boldsymbol{\Theta}, \qquad \boldsymbol{\Theta}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

when $\mathbb{E}[\sigma(g)] = 0$, where $b_1 = \mathbb{E}[\sigma'(g)]$, $\quad b_2^2 = \mathbb{E}[\sigma(g)^2] - \tau^2 b_1^2$ and $g \sim \mathcal{N}(0, \tau^2)$.

Gaussian Equivalence II

The following holds pointwise for each fixed $\lambda \geq 0$.

$$\left|\mathcal{R}(\lambda) - \mathcal{R}_{\mathsf{GE}}(\lambda)\right| = o_{n,\mathbb{P}}(1)$$

After further simplification, we define the **linearized** NTK. Let $c_1 = \mathbb{E}[\sigma'(g)^2]$

$$\boldsymbol{K}_{\text{lin}} = \boldsymbol{F}_{\text{lin}}^\top \boldsymbol{F}_{\text{lin}} + \sigma_W^2 \left(p^{-1} b_1^2 \boldsymbol{X}^\top \boldsymbol{X} + \tau^2 (c_1 - b_1^2) \mathbf{I}_n \right) + z\mathbf{I}_n, \quad \boldsymbol{F}_{\text{lin}} = \frac{1}{\sqrt{N}} \left(b_1 \boldsymbol{W}_1 \boldsymbol{X} + b_2 \boldsymbol{\Theta} \right)$$

Generalized Cross Validation (GCV): Introduced as rotationally-invariant approximation of LOOCV [GHW79].

GCV

$$\mathrm{GCV}(\lambda) := \frac{\mathrm{E}_{\mathrm{train}}(\lambda)}{\Big(\lambda \,\mathrm{tr}\big(\boldsymbol{K}^{\mathrm{NTK}} + \lambda \mathbf{I}_n\big)\Big)^2}$$

- In-sample estimator: agnostic to true data distribution.
- Consistent estimator of prediction risk for wide class of kernel models and their associated learning problems.
- Shown to approximate prediction risk of pretrained Residual Neural Network (RNN) on dataset of images (CIFAR-10) [WHS22].

Key references: Golub et al. (1979), Jacot et al. (2020), Wei et al. (2022)

Uniform Consistency of GCV

$$\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \left| \mathcal{R}(\lambda) - \mathrm{GCV}(\lambda) \right| = o_{n,\mathbb{P}}(1)$$

- Proved for KRR under general concentration assumptions on features, which covers our dual power-law model, where a rate of $\mathcal{O}_{n,\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ was achieved [Mis22].
- Extends to $\lambda_{\min} = 0$ in benign overfitting regime, which allows for optimal ridge-tuning via GCV.
- Conjecture GCV is a uniformly consistent estimator for prediction risk of finite-width NTK regression on power-law data.

Key references: Adlam–Pennington (2020), Misiakiewicz–Saeed (2024), Patil et al. (2021)

Asymptotic Training Error

$$\mathrm{E}_{\text{train}}(\lambda) = \lambda^2\Big(\sigma_\varepsilon^2 \operatorname{tr}\big(\boldsymbol{K} + \lambda \mathbf{I}_n\big)^{-2} + \operatorname{tr}\Big(\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{K} + \lambda \mathbf{I}_n)^{-2}\Big)\Big) + o_{n,\mathbb{P}}(1)$$

Define

$$m_1^{(n)}(\lambda) = \operatorname{tr}\big(\boldsymbol{K} + \lambda \mathbf{I}_n\big)^{-1}, \qquad m_2^{(n)}(\lambda) = \operatorname{tr}\big(p^{-1}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{K} + \lambda \mathbf{I}_n)^{-1}\big)$$

$$m_\beta^{(n)}(\lambda) = \frac{1}{n}\big(\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{X}\boldsymbol{\beta}\big)$$

Asymptotic Training Error corr.

$$\mathrm{E}_{\text{train}}(\lambda) = -\lambda^2\Big(\sigma_\varepsilon^2 \partial_\lambda m_1^{(n)}(\lambda) + \partial_\lambda m_\beta^{(n)}(\lambda)\Big) + o_{n,\mathbb{P}}(1)$$

GCV corr.

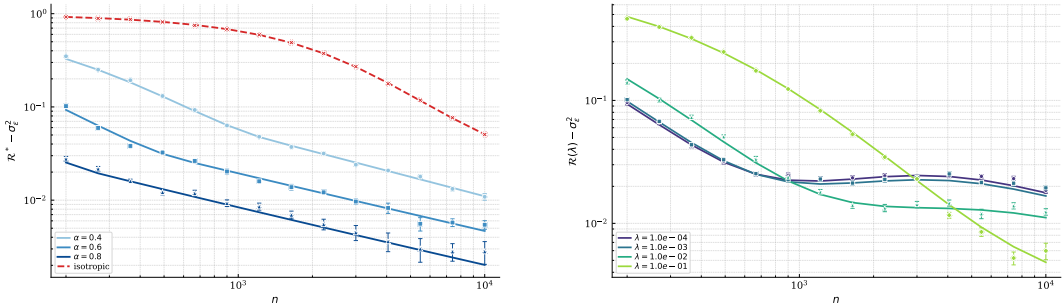
$$\mathrm{GCV}(\lambda) = \frac{-\left(\sigma_\varepsilon^2 \partial_\lambda m_1^{(n)}(\lambda) + \partial_\lambda m_\beta^{(n)}(\lambda)\right)}{\left(m_1^{(n)}(\lambda)\right)^2}$$

- GCV can be understood as correction to train error to account for model complexity.
- $m_1^{(n)}(\lambda)$ accounts for train-test gap, measures effective degrees of freedom of linear estimator onto $\boldsymbol{K}^{\mathrm{NTK}}$.
- Signal resolvent $m_\beta^{(n)}$ measures alignment of signal $\boldsymbol{\beta}$ with resolvent of $\boldsymbol{K}^{\mathrm{NTK}}$.
- $m_\beta^{(n)}(\lambda) \approx m_2^{(n)}(\lambda)$ concentrates under isotropic random signal $\boldsymbol{\beta} \sim \mathcal{N}(0, p^{-1}\mathbf{I}_p)$.

New Goal

Derive deterministic equivalents to $m_1^{(n)}(\lambda), m_2^{(n)}(\lambda)$ and $m_\beta^{(n)}(\lambda)$

Isotropic vs power law



Isotropic vs power-law: **(Left)** Optimally-tuned excess NTK risk plotted against n for varying α with isotropic baseline in red. Isotropic case exhibits curve in log-log axes plot, while power law curve straightens as α increases. **(Right)** excess risk against n for various ridge λ with fixed $\alpha = 0.6$. Both plots generated with $p = 2000, \sigma_\varepsilon^2 = 0.3, r = 1.5, N = 2500, \sigma_W^2 = 1.0$. Error bars are plotted with marker $\pm$ SE over 20 trials.

Main Technical Tool: Log-determinant of Brownian Motion embedded NTK matrix.

Log-Determinant Potential

$$\mathcal{F}_n(t, \kappa, \gamma) = \frac{1}{n} \log\det \boldsymbol{K}(t, \kappa, \gamma)$$

Recall linearized NTK matrix

$$\boldsymbol{K} = \boldsymbol{F}^\top \boldsymbol{F} + \sigma_W^2\big(p^{-1} b_1^2 \boldsymbol{X}^\top \boldsymbol{X} + \tau^2(c_1 - b_1^2)\mathbf{I}_n\big) + z\mathbf{I}_n, \quad \boldsymbol{F} = \frac{1}{\sqrt{N}}\big(b_1 \boldsymbol{W}_1 \boldsymbol{X} + b_2 \boldsymbol{\Theta}\big)$$

Approach idea: Condition on train data $\boldsymbol{X}$, embed each Gaussian-entry matrix with i.i.d Brownian motions

$$\boldsymbol{W}_1 \leftrightarrow \boldsymbol{W}_1(t), \quad \boldsymbol{\Theta} \leftrightarrow \boldsymbol{\Theta}(t), \quad \boldsymbol{F} \leftrightarrow \boldsymbol{F}(t)$$

where

$$\boldsymbol{A}(t) \Leftrightarrow \boldsymbol{A}_{ij}(t) \sim \mathrm{BM}(0, 1)$$

Define

Time-parametrized NTK

$$\boldsymbol{K}(t,\kappa,\gamma) = \boldsymbol{F}(t)^\top \boldsymbol{F}(t) + p^{-1}\kappa \boldsymbol{X}^\top \boldsymbol{X} + \gamma \mathbf{I}_n, \quad \boldsymbol{F}(t) = \frac{1}{\sqrt{N}}\big(b_1 \boldsymbol{W}_1(t)\boldsymbol{X} + b_2 \boldsymbol{\Theta}(t)\big)$$

- $\kappa = \kappa(t)$ and $\gamma = \gamma(t)$ are parameters, with target values

$$\kappa_{\text{target}} := \sigma_W^2 b_1^2, \qquad \gamma_{\text{target}} = \lambda + \sigma_W^2 \tau^2 (c_1 - b_1^2)$$

- Motivation:

$$\frac{\partial \mathcal{F}_n}{\partial \gamma} = \underbrace{\operatorname{tr}\big(\boldsymbol{K}(t,\kappa,\gamma)^{-1}\big)}_{m_1^{(n)}(t,\kappa,\gamma)}, \quad \frac{\partial \mathcal{F}_n}{\partial \kappa} = \underbrace{\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top \boldsymbol{X}\big(\boldsymbol{K}(t,\kappa,\gamma)^{-1}\big)\Big)}_{m_2^{(n)}(t,\kappa,\gamma)}$$

Step 1: Apply entry-wise Itô-calculus to $\mathcal{F}_n$ to derive a finite-dimensional Stochastic Differential Equation (SDE) of the form[3]

$$d\mathcal{F}_n(t, \boldsymbol{u}) = d\mathcal{M}_n(t, \boldsymbol{u}) + \mathcal{D}_n(t, \boldsymbol{u})dt + E_n(t)$$

Step 2: Control the Martingale terms $\mathcal{M}_n(t, \boldsymbol{u})$ and remainder terms $E_n(t)$ uniformly in $t \in [0, 1]$, recognize drift terms $\mathcal{D}_n(t, \boldsymbol{u})$ as partial derivatives to derive an asymptotic deterministic partial differential equation (PDE) of the form

$$\partial_t \mathcal{F}_\infty(t, \boldsymbol{u}) + H(\boldsymbol{u}, \partial_{\boldsymbol{u}} \mathcal{F}_\infty) = 0 \tag{1.3}$$

[3] $\boldsymbol{u} := (\kappa, \gamma)$

Step 3: Recognize PDE (1.3) as a first-order nonlinear PDE of Hamilton-Jacobi (HJ) type.

$$\partial_t \mathcal{F}_\infty(t, \boldsymbol{u}) + H(\boldsymbol{u}, \boldsymbol{p_u}) = 0$$

where the Hamiltonian H is obtained from replacing partial derivatives ∂_{u_j} with corresponding momentum variables p_j.

Step 4: Solve HJ PDE using Method of Characteristics. Hamilton's equations translate PDE into system of ordinary differential equations (ODEs)

$$\dot{\boldsymbol{u}}_j = \partial_{p_j} H, \qquad \dot{p}_j = -\partial_{\boldsymbol{u}_j} H$$

with initial conditions defined by $\mathcal{F}_\infty\big(0, \boldsymbol{u}(0)\big)$. Momentum variables p_j will correspond to partial derivatives of $\mathcal{F}_\infty(t)$ along curves $\boldsymbol{u}_j$.

Proposition: Asymptotic PDE

Under Assumptions (A1) and (A4), $\mathcal{F}_n(t, \kappa, \gamma)$ *approximately* satisfies the following PDE

$$\partial_t \mathcal{F} = \left(b_1^2 \partial_\kappa \mathcal{F} + b_2^2 \partial_\gamma \mathcal{F}\right)\left[(1 - \frac{\phi}{\psi}) + \frac{\phi}{\psi}\left(\kappa \partial_\kappa \mathcal{F} + \gamma \partial_\gamma \mathcal{F}\right)\right] \tag{1.4}$$

Next Steps:

- Derive unique analytic solution to (1.4) with initial condition

$$\mathcal{F}_\infty(0, \kappa_0, \gamma_0) = \lim_{n \to \infty} \frac{1}{n} \log\det\left(p^{-1}\kappa_0 \boldsymbol{X}^\top \boldsymbol{X} + \gamma_0 \mathbf{I}_n\right)$$

- κ_0, γ_0 are initial values of $(\kappa(t), \gamma(t))$ characteristic curves, chosen to reach target values $(\kappa_{\text{target}}, \gamma_{\text{target}})$ at time $t = 1$.
- Deterministic equivalents for initial momentum variables $p_\kappa(0), p_\gamma(0)$ are given by the general Marchenko-Pastur Theorem [BS10].

Main Theorem Setup:

Denote the following, $D_\lambda = \lambda + \tau^2\sigma_W^2(c_1 - b_1^2)$. Then for $\eta > 0$, define the rational functions

$$s(\eta) = \frac{b_1^2(\eta + \sigma_W^2)}{\eta(D_\lambda - b_2^2\sigma_W^2)}\big(\phi\eta - \psi(1-\eta)(\eta + \sigma_W^2)\big), \quad \zeta(\eta) = \frac{b_2^2\eta + D_\lambda}{b_1^2(\eta + \sigma_W^2)}$$

Let $\eta^* > 0$ be the unique positive solution to the fixed point equation

$$\Psi(\eta^*) = F_\alpha\big(s(\eta^*)\big) + \zeta(\eta^*)s(\eta^*) - \phi = 0, \quad F_\alpha(s) = \frac{1}{p}\sum_{j=1}^{p}\frac{s}{s + j^\alpha} \tag{1.5}$$

– The fixed point equation (1.5) has a unique physical solution $\eta^* > 0$ corresponding to $s(\eta^*) > 0$, and can be found via standard numerical methods (e.g bracketing via brentq [Bre73]).

Theorem: Power-Law Resolvents

Under Assumptions (A0)-(A4), we have

$$m_1(\lambda) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} s(\eta^*), \qquad m_2(\lambda) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} F_\alpha\big(s(\eta^*)\big)$$

where $F_\alpha(s) = \frac{1}{p}\sum_{j=1}^{p} \frac{s}{s+j^\alpha}$. Moreover,

$$m_\beta(\lambda) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} F_r(s(\eta^*)), \qquad F_r(s) = \sum_{j=1}^{p} \frac{s j^{-(\alpha+2r)}}{s j^{-\alpha} + 1}$$

Remarks

- Resolvent traces are characterized as momentum curves of HJ PDE $p_\kappa, p_\gamma, p_\xi$. The full rigorous justification of the identifications $m_1^\infty = m_1, m_2^\infty = m_2$, and $m_\beta^\infty = m_\beta$ is beyond scope of this work.

Remarks

— Signal resolvent m_β is the momentum curve p_ξ of **augmented** HJ PDE:

$$\widetilde{\mathcal{F}}_n(t,\lambda,\kappa,\xi) = n^{-1}\log\det\Big(\boldsymbol{K}(t,\kappa,\gamma) + \xi \boldsymbol{X}^\top \beta\beta^\top \boldsymbol{X}\Big)$$

This is a rank-1 perturbation of original log-determinant potential, recover original HJ along $\xi = 0$ slice.

— Resolvent traces take the form

$$m_1 = \frac{p_\gamma(0)}{\eta^*}, \quad m_2 = \frac{p_\kappa(0)}{\eta^*}, \quad m_\beta = \frac{p_\xi(0)}{\eta^*}$$

Reduces coupled polynomial system of [AP20] and fixed-point equations of [MM19] to single scalar-root solver.

Remarks

— Signal resolvent m_β is the momentum curve p_ξ of **augmented** HJ PDE:

$$\widetilde{\mathcal{F}}_n(t,\kappa,\gamma) = \mathcal{F}_n(t,\kappa,\gamma) + \frac{1}{n}\log\big(1+\xi \boldsymbol{u}^\top \boldsymbol{K}(t,\kappa,\gamma)^{-1}\boldsymbol{u}\big)$$

This is a rank-1 perturbation of original log-determinant potential, recover original HJ along $\xi = 0$ slice.

— Resolvent traces take the form

$$m_1 = \frac{p_\gamma(0)}{\eta^*}, \quad m_2 = \frac{p_\kappa(0)}{\eta^*}, \quad m_\beta = \frac{p_\xi(0)}{\eta^*}$$

Reduces coupled polynomial system of [AP20] and fixed-point equations of [MM19] to single scalar-root solver.

Proposition: Prediction Risk

Under Assumptions (A0)-(A4), the prediction risk admits a bias-variance decomposition

$$\mathcal{R}(\lambda) \asymp \mathcal{B}(\lambda; \alpha, r) + \mathcal{V}(\lambda; \alpha, r) \tag{1.6}$$

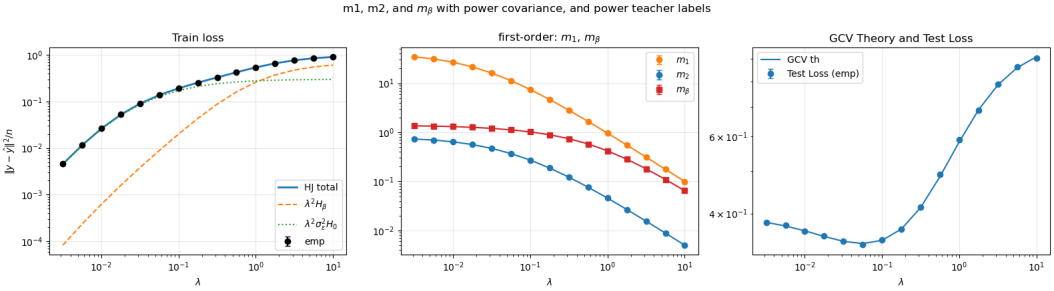
where the bias and variance terms are defined as

$$\mathcal{B}(\lambda; \alpha, r) = -\frac{\partial_\lambda m_\beta(\lambda)}{m_1(\lambda)^2}, \qquad \mathcal{V}(\lambda; \alpha, r) = -\frac{\sigma_\varepsilon^2 \partial_\lambda m_1(\lambda)}{m_1(\lambda)^2}$$

Remarks:

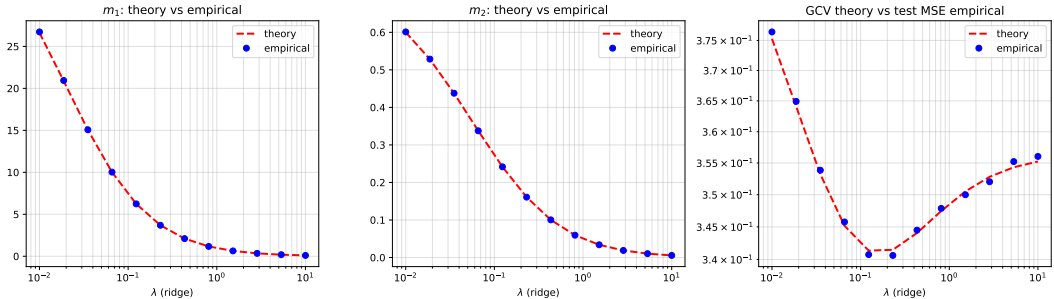
- Follows from implicit differentiation of resolvent traces from main theorem and the GCV estimator, full formulas omitted.
- Bias term depends on source strength r through m_β, while variance term depends *only* on capacity α through m_1.

Experiments



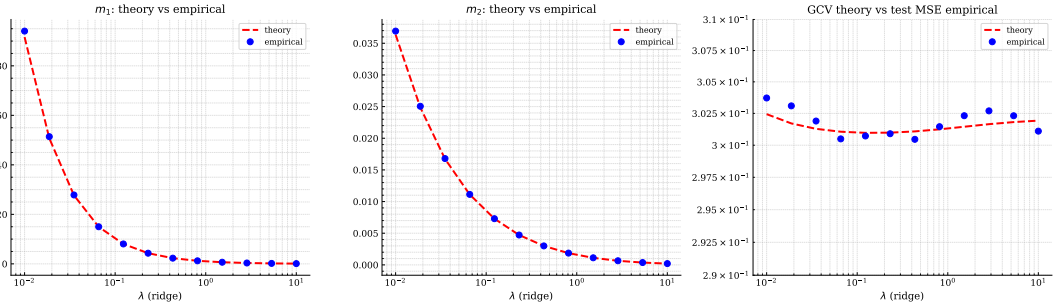
Power-law resolvents: All three trace resolvents m_1, m_2, m_β plotted against ridge λ (middle). m_β shows divergence from m_2. Theoretical curves for all three trace resolvents track empirical plots well. Right figure depicts theoretical GCV prediction and with empirical test MSE. Plots generated with $n = 2250, N = 3000, p = 1500, \sigma_\varepsilon^2 = 0.3, \sigma_W^2 = 1.0, \alpha = 0.5, r = 0.8$ and error bars over 20 trials

Experiments



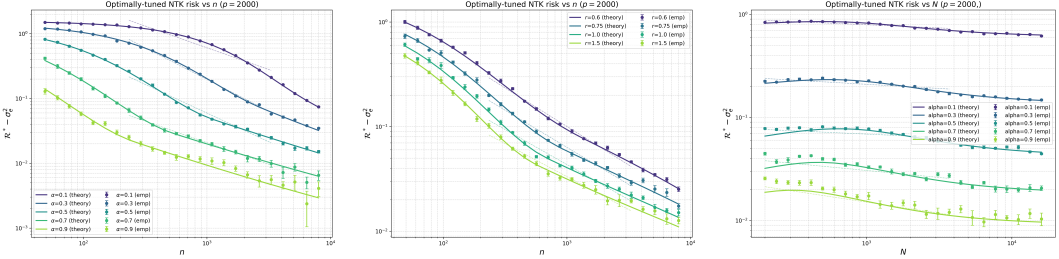
Theory vs empirical $\alpha = 0.5$: lots for m_1 (left), m_2 (middle), and GCV_λ (right) against ridge λ. All three show plots strong agreement across. Plots generated with $n = 2400, N = 1800, p = 1200, \sigma_\varepsilon^2 = 0.3, \sigma_W^2 = 1.0$ and isotropic random labels.

Experiments



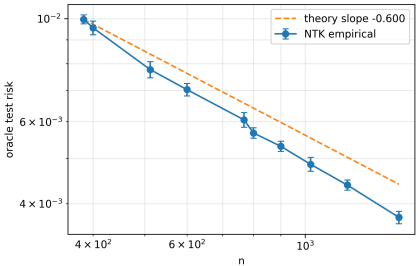
Theory vs empirical $\alpha = 1.5$: Same setup as figure on preceding slide, with larger capacity exponent $\alpha > 1$. Test loss (right) is noisier, but still tracked by theoretical curve.

Experiments



Optimally-tuned Risk: Excess NTK risk of optimally-tuned ridge $\lambda = \lambda^*$ (markers with $\pm$SE over 10 trials) on log-log axes. Theoretical curves plotted in solid lines, showing close alignment across all regimes.**Left**: excess risk vs training samples n with varying capacity exponent α and fixed $N = 5000, r = 1.0$. As α increases, curves straighten out to follow power-law decay. **Middle**: Varying values of source exponent r. **Right:** risk against width N for sweeping α, and fixed $n = 2000, r = 0.7$. Throughout $\sigma_W = 1, \sigma_\varepsilon^2 = 0.01$, $p = 2000$.

Scaling law of NTK regression Under a further assumption on the GCV- NTK Risk deterministic equivalence, we derive a scaling law for finite-width NTK. The simulation uses a capacity strength of $\alpha = 1.5$ and a source complexity of $r = 0.5$. The resulting empirical risk exponent of $\theta_{\mathrm{NTK}} \approx 0.687$ aligns closely with the theoretical prediction of 0.6, confirming that the Gaussian-equivalent model accurately captures the scaling behavior of finite-width neural networks in this regime. Moreover, classical KRR *source-capacity* scaling exponents are recovered in the $N \to \infty$ limit.



Scaling Law: Test risk scaling of a finite-width Gaussian-equivalent NTK on unnormalized power-law data ($\alpha = 1.5, r = 0.5$).

Chapter 3: Eigen-Spike Emergence and Quadratic Equivalents for Conjugate Kernels on Nonlinearly Separable Data

Define the Conjugate Kernel (CK):

Conjugate Kernel

$$\boldsymbol{K}^{\mathrm{CK}} := \boldsymbol{Y}^\top \boldsymbol{Y}, \qquad \boldsymbol{Y} = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{X})$$

Notation:

$$\boldsymbol{X} \in \mathbb{R}^{d\times n}, \quad \boldsymbol{W} \in \mathbb{R}^{N\times d}, \quad \boldsymbol{W}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

Motivations:

- CK captures the *similarity* structure of data in random feature space of NN.
- Spectral properties of CK provide insight into
  - **Representation quality**: CK eigenvalues describe strength of learned similarities.
  - **Feature extraction**: CK Eigenvectors indicate direction of task-model alignment.
  - **Generalization**: CK spectra dictates generalization ability of NTK and RF models.

Motivating Question:

What are the machine learning implications of the Linear Equivalence (LE):

$$\sigma(\boldsymbol{WX}) \approx b_1 \boldsymbol{WX} + b_2 \boldsymbol{\Theta}, \qquad \boldsymbol{\Theta}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

LE says CK behaves as a linear kernel plus isotropic noise

Problem

The LE **hides important nonlinear task structure**

- For **nonlinearly separable data**, LE implies spectral methods on CK fail.
- **Neural Networks**, in contrast, have been shown to learn complex nonlinear structure in data
- LE fails to explain *how* or *when* nonlinearity σ helps learning task.

Motivating Question:

What are the machine learning implications of the Linear Equivalence (LE):

$$\sigma(\boldsymbol{WX}) \approx b_1 \boldsymbol{WX} + b_2 \boldsymbol{\Theta}, \qquad \boldsymbol{\Theta}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

To answer this, we consider the central spectral question of this chapter.

Central Spectral Question

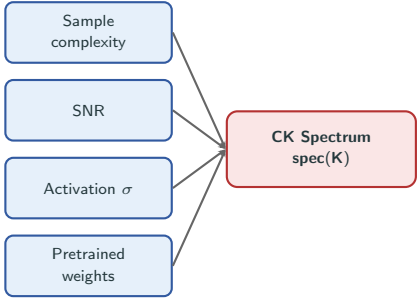
When can a neural network transfer a **nonlinear pattern** in data to a **linearly accessible direction**[4] that allows for linear classification?

Key mechanism: BBP (Baik–Ben Arous–Péché) transitions[Baik-Ben Arous-Péché, 2005]– when an informative outlier eigenvalue seperates from the bulk spectrum, and its corresponding eigenvector aligns with the signal.

[4] aligned eigen-spike

Practical Implications: ML practitioners tune various "knobs" to improve model performance:

- **Sample complexity** : n vs. feature dimension d vs. width N
- **Signal-to-noise ratio (SNR)**: ratio of meaningful signal to random noise.
- **Activation function**: Choice of nonlinearity σ
- **Training**: Random initialization vs. pretrained weights $\boldsymbol{W}$



Question:

How do each of these knobs affect the spectral structure of $\boldsymbol{K}^{\mathrm{CK}}$, and in turn the *learnability* of nonlinear patterns?

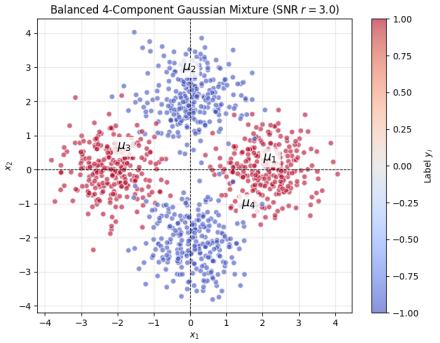
Four-Component Gaussian Mixture Model (GMM): $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \frac{1}{4}\sum_{a=1}^{4} \mathcal{N}\Big(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_a,\ \frac{1}{d}\mathbf{I}_d\Big)$

Cluster means: for SNR parameter $r \geq 0$,

$$\boldsymbol{\mu}_1 = r\,\boldsymbol{u}_1,\quad \boldsymbol{\mu}_2 = r\,\boldsymbol{u}_2,\quad \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_1,\quad \boldsymbol{\mu}_4 = -\boldsymbol{\mu}_2$$

with $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$ orthonormal.

Labels: $y_i = +1$ for clusters $\{1, 3\}$, $y_i = -1$ for clusters $\{2, 4\}$, making the classification problem nonlinearly separable.



Key Property: Nonlinear Separability

- **Not linearly separable**: class means cancel $\Rightarrow$ no linear method works
- **Quadratically separable**: second-order features (e.g., $\boldsymbol{x}_i^\top \boldsymbol{A}\, \boldsymbol{x}_i$) can separate

Why study XOR?

- **Canonical nonlinearly separable dataset**: widely studied in Deep Learning theory (Refinetti et al. 2021, Glasgow et al. 2023)
- **Quadratic feature channel:** isolates quadratic features, which can generalize to more complex nonlinear tasks.

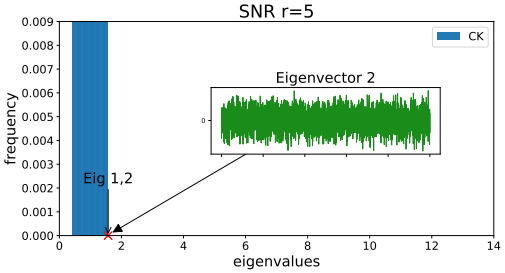
Proxy for nonlinear learnability

When do **spike eigenvalues** emerge from the bulk of $\boldsymbol{K}^{\mathrm{CK}}$ whose eigenvectors **asymptotically align** with XOR-labels?

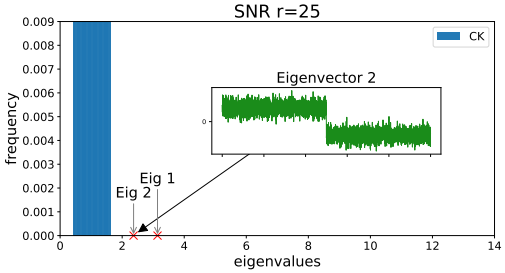
When yes $\Rightarrow$ linear spectral classification via $\boldsymbol{K}^{\mathrm{CK}}$ is possible.

When no $\Rightarrow$ nonlinear structure remains hidden from linear readouts.

Phase Transition in SNR



$r = 5$: no eigenvalues separate from bulk



$r = 25$: two eigenvalues emerge

Four Practical "Knobs"

Knob 1: Sample complexity n vs d vs N

Knob 2: Signal-to-noise ratio (SNR)

Knob 3: Pretrained/spiked weights $\boldsymbol{W}$

Knob 4: Activation choice σ

For each configuration of these knobs, we ask:

1. Do **outlier eigenvalues** emerge in the spectrum of $\boldsymbol{K}^{\mathrm{NTK}}$?
2. Do their **eigenvectors align with XOR labels $\boldsymbol{y}$**?
3. Can we perform **linear spectral classification**?

We identify BBP-type phase transitions for each knob, and consider their practical implications.

Write XOR dataset as a signal-plus-noise low rank perturbation:

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{M}, \qquad \boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

where $\boldsymbol{M} = r\sqrt{\frac{n}{2d}}\left(\boldsymbol{u}_1\boldsymbol{v}_1^\top + \boldsymbol{u}_2\boldsymbol{v}_2^\top\right)$ is a **rank-2** signal, and $\boldsymbol{v}_1 \perp \boldsymbol{v}_2 \in \mathbb{R}^n$ encode cluster membership, and $\mathrm{SNR} := r^2$

Assumptions

The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies

- σ is centered and normalized with respect to the Gaussian: $\mathbb{E}[\sigma(\xi)] = 0, \mathbb{E}[\sigma(\xi)^2] = 1$.
- σ is three-times differentiable with bounded derivatives.

Hermite Coefficients of σ:

$$\boxed{b_\sigma := \mathbb{E}[\sigma'(\xi)] = \mathbb{E}[\xi\,\sigma(\xi)]} \qquad \boxed{c_\sigma := \mathbb{E}[\sigma''(\xi)] = \mathbb{E}[(\xi^2 - 1)\sigma(\xi)]}$$

Write XOR dataset as a signal-plus-noise low rank perturbation:

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{M}, \qquad \boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

where $\boldsymbol{M} = r\sqrt{\frac{n}{2d}}\left(\boldsymbol{u}_1\boldsymbol{v}_1^\top + \boldsymbol{u}_2\boldsymbol{v}_2^\top\right)$ is a **rank-2** signal, and $\boldsymbol{v}_1 \perp \boldsymbol{v}_2 \in \mathbb{R}^n$ encode cluster membership, and $\mathrm{SNR} := r^2$

Assumptions

The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies

- σ is centered and normalized with respect to the Gaussian: $\mathbb{E}[\sigma(\xi)] = 0, \mathbb{E}[\sigma(\xi)^2] = 1$.
- σ is three-times differentiable with bounded derivatives.

Main Takeaway:

- b_σ: controls the **linear channel**
- c_σ: controls the **quadratic channel** $\longleftarrow$ key for XOR!

Summary of Main Results

Regime	SNR	Sample	Weights	Label Align?
Case 1 (Baseline)	$\Theta(1)$	$n \asymp d$	Random	**No**
Case 2 (Large SNR)	$\Theta(n^{1/4})$	$n \asymp d$	Random	**Yes**
Case 3 (Trained $\boldsymbol{W}$)	$\Theta(1)$	$n \asymp d$	Spiked	**Yes**
Case 4 (Quad. size)	$\Theta(1)$	$n \asymp d^2$	Random	**Yes**

Common thread: In each successful case, the quadratic Hermite component c_σ creates a feature channel that makes the nonlinear XOR structure spectrally visible.

Activation requirement: $c_\sigma \neq 0$ is **necessary** in all cases. If $c_\sigma = 0$ (e.g., tanh after centering/normalization), linear classification via CK is impossible.

Core idea: Expand $\sigma(\boldsymbol{WX})$ around null-model $\sigma(\boldsymbol{WZ})$ and retain **second-order Hermite coefficients**.

Proposition: Quadratic Equivalent (QE)

Let $\boldsymbol{Y}_0 = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{WZ})$, $\boldsymbol{g}_k = \boldsymbol{W}\boldsymbol{u}_k$, $\theta_{\mathrm{SNR}} = r\sqrt{n/(2d)}$. Then:

$$\boldsymbol{Y}_{\mathrm{QE}} := \underbrace{\boldsymbol{Y}_0}_{\text{bulk}} + \underbrace{\frac{\theta_{\mathrm{SNR}} b_\sigma}{\sqrt{N}}\left(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top\right)}_{T_1 \text{ (linear spikes)}} + \underbrace{\frac{\theta_{\mathrm{SNR}}^2 c_\sigma}{\sqrt{N}}\left(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}\right)}_{T_2 \text{ (quadratic spikes)}}$$

satisfies $\|\boldsymbol{Y} - \boldsymbol{Y}_{\mathrm{QE}}\| \to 0$ in probability.

- Y_0: determines bulk spectrum μ (may also create uninformative outliers)
- $T_1 \sim \theta_{\mathrm{SNR}} b_\sigma$: linear spikes, eigenvectors align with v_1, v_2 but $\perp \boldsymbol{y}$
- $T_2 \sim \theta_{\mathrm{SNR}}^2 c_\sigma/\sqrt{n}$: quadratic spikes, can align with $\boldsymbol{y}$ appears when $r = \Theta(n^{1/4})$

Proportional Regime: $n, d, N \to \infty$

$$n/N \to \varphi \in (0, \infty), \quad n/d \to \psi \in (0, \infty), \quad r = \Theta(1)$$

Bulk law: Empirical spectral distribution (ESD) of $\boldsymbol{K}^{\mathrm{CK}}$ converges to

$$\mu := \rho_{\varphi}^{\mathrm{MP}} \boxtimes \nu, \quad \nu = (1 - b_\sigma^2)\delta_0 + b_\sigma^2 \, \rho_{\psi}^{\mathrm{MP}}$$

(free multiplicative convolution of Marchenko–Pastur laws) [FW20].

Key transforms:

$$z(s) = -\frac{1}{s} + \phi \int \frac{x}{1 + xs} \nu(dx), \qquad \varphi(s) = -\frac{sz'(s)}{z(s)},$$

$$T(s) = \frac{z(s)s^2 - (\phi - 1)s}{\phi}, \qquad T^{-1}(t) = -\frac{t\big(1 - \psi b_\sigma^2 \, t\big)}{1 + \big(1 - \psi b_\sigma^2\big)t - \psi b_\sigma^2 (1 - b_\sigma^2)t^2},$$

Theorem (Finite-SNR, Proportional Limit)

Under the above setting with $r = \Theta(1)$:

(i) **Bulk**: *ESD of $\boldsymbol{K}^{\mathrm{CK}} \to \mu$.*

(ii) **BBP for covariance spike**: *If $\tau > \tau_{\mathrm{crit}}$ and $z'(-1/\Lambda_\tau) > 0$, then $\boldsymbol{K}^{\mathrm{CK}}$ has an outlier at $\hat{\lambda}_\tau = z(-1/\Lambda_\tau) + o_{n,\mathbb{P}}(1)$. This spike is uninformative (present even for $r = 0$).*

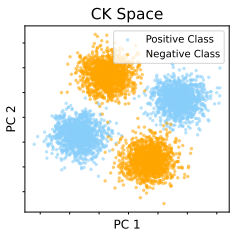
(iii) **BBP for mean spike**: *Similar BBP transition for the mean direction.*

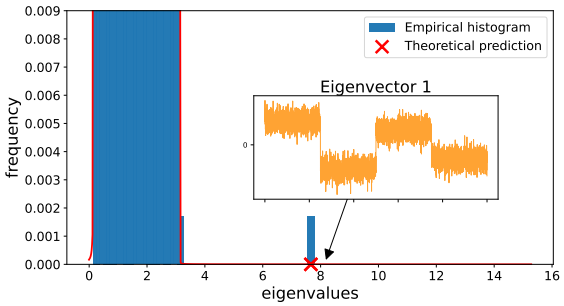
(iv) **Two linear spikes from XOR**: *If $z'(s_{\mathrm{lin}}) > 0$, two outlier eigenvalues at $z(s_{\mathrm{lin}})$, with eigenvectors aligning with $\boldsymbol{v}_1, \boldsymbol{v}_2$ (within-class structure).*

(v) **No label alignment**: **Every** *outlier eigenvector of $\boldsymbol{K}^{\mathrm{CK}}$ is asymptotically* **orthogonal** *to $\boldsymbol{y}$:*

$$\|\hat{V}\,\boldsymbol{y}\|^2 \to 0 \quad a.s.$$

Failure case: Finite-SNR proportional regime is the baseline where linear classification fails.





Observations:

- XOR visible in PC space **but not linearly separable**
- Theory (red) matches
- Leading eigenvec $\perp \boldsymbol{y}$

$n = 5000$, $N = d = 150000$,

$r = 6$, $\sigma \propto \mathrm{ReLU}$

Key Takeaway

The **linear spikes** (from b_σ) carry within-class information $(\boldsymbol{v}_1, \boldsymbol{v}_2)$, but $\boldsymbol{y} \perp \{v_1, v_2\}$.

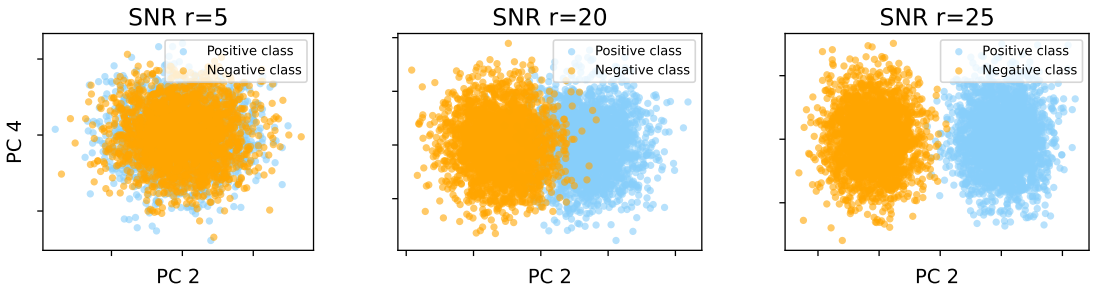
In the proportional regime with finite SNR, the CK exhibits outliers, but **linear classification of XOR fails**. The LE correctly predicts this failure.

Additional negative result: Kernel spectral clustering [Couillet-Florent Benaych–Georges 2016] (normalized Laplacian with Euclidean distance kernel) **also fails** for XOR in this regime.

Remark: The uninformative spike can be removed if $\tau \leq \tau_{\mathrm{crit}}$, i.e., $\sqrt{\psi}\left(\frac{c_\sigma^2}{2} - b_\sigma^2\right) \leq b_\sigma^2$. If $c_\sigma/\sqrt{2}$ is small relative to b_σ (e.g., centered normalized GELU).

Knob: Increase SNR — PCA Phase Transition

Phase Transition in Principal Component (PC): PC 2 vs PC 4 of CK over increasing SNR. Yellow: -1 labels, Blue: $+1$ labels.



$r = 5$ (no separation)

$r = 20$ (transition)

$r = 25$ (linearly separable)

Clear phase transition: Around $r = 20$–25, XOR becomes linearly separable in CK principal component space.

Idea: Increase the SNR so that the quadratic Hermite component becomes non-negligible.

Scaling: Same proportional limit, but now

$$r = r_0\, n^{1/4}, \qquad r_0 \in (0, \infty) \text{ fixed.}$$

New constants:

$$\kappa_0 = \frac{r_0^4\, \psi^2}{4}, \quad \beta_{\text{quad}} = c_\sigma^2\, \kappa_0, \quad \tilde{\beta}_{\text{quad}} = \beta_{\text{quad}} + \tau, \quad s_{\text{quad}} = T^{-1}(\beta_{\text{quad}}).$$

Key observation: With $\theta_{\text{SNR}} = r\sqrt{n/(2d)} = \Theta(n^{1/4})$, the quadratic term

$$T_2 = \frac{\theta_{\text{SNR}}^2 c_\sigma}{2\sqrt{N}} \left(\boldsymbol{g}_1^{\odot 2}\, \boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\, \boldsymbol{v}_2^{\odot 2\top} \right)$$

becomes $O(1)$ and comparable to the bulk.

Theorem (Large-SNR Proportional Limit)

Under $r = r_0 n^{1/4}$, let $\boldsymbol{P}_{\text{lin}} = v_1 v_1^\top + v_2 v_2^\top$ and $\boldsymbol{K}_\perp^{\text{CK}} = P_\perp \boldsymbol{K}^{\text{CK}} P_\perp$:

(i) **Bulk**: *ESD of $\boldsymbol{K}^{\text{CK}} \to \mu$.*

(ii) **Two diverging linear outliers**: *If $b_\sigma \neq 0$, top two eigenvalues diverge as $\hat{\lambda}_{1,2} = b_\sigma^2 r_0^2 \psi \cdot n^{1/2}(1 + o_P(1))$, but their eigenvectors are $\perp \boldsymbol{y}$.*

(iii) **Quadratic 1_n-spike**: *BBP transition for 1_n-direction at effective strength $\tilde{\beta}_{\text{quad}}$.*

(iv) **Quadratic y-spike (label alignment!)**: *If $z'(s_{\text{quad}}) > 0$, then $\boldsymbol{K}_\perp^{\text{CK}}$ has an outlier whose eigenvector satisfies:*

$$\boxed{\frac{1}{n}\left|\hat{v}_y^\top \boldsymbol{y}\right|^2 \xrightarrow{P} \frac{s_{\text{quad}}\, z'(s_{\text{quad}})}{\beta_{\text{quad}}\, T'(s_{\text{quad}})} > 0}$$

Key Takeaway

By increasing SNR to $r = \Theta(n^{1/4})$, the quadratic nonlinearity creates **label-informative outliers $\Rightarrow$ linear spectral classification of XOR becomes possible!**

Why does this work? The identity

$$\boldsymbol{v}_1^{\odot 2} = \frac{1}{\sqrt{n}}(1_n + \boldsymbol{y}), \quad \boldsymbol{v}_2^{\odot 2} = \frac{1}{\sqrt{n}}(1_n - \boldsymbol{y})$$

shows that element-wise squaring converts within-class directions $\boldsymbol{v}_1, \boldsymbol{v}_2$ into label $\boldsymbol{y}$.

Role of activation choice:

- If $b_\sigma = 0$: diverging linear outliers vanish $\Rightarrow$ **simpler** spectral classification
- If $c_\sigma = 0$: quadratic channel absent $\Rightarrow$ **no label alignment possible**
- The activation σ must have $c_\sigma \neq 0$ for nonlinear classification!

Conclusion

Conclusion:

- In Chapter 2, we utilized a universality result (**Gaussian Equivalence**) to study finite-width NTK regression under a dual-power law model.
  - We introduced a novel technique to derive deterministic equivalents for trace resolvents that characterize the prediction risk of NTK, extending previous results beyond isotropic setting.
  - As an application, we derived a *source-capacity* scaling law for finite-width NTK, contributing to the fast growing literature on neural scaling laws in kernel methods.
- In Chapter 3, we study the practical implications of the **Gaussian Equivalence** using spike eigenvalues of the CK and their task-alignment as a proxy for nonlinear learnability.
  - We develop a Quadratic Equivalent (QE) framework that allows for the precise BBP-type phase transitions in four regimes of practical interest.

References

[AP20] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[Bre73] Richard P. Brent. *Algorithms for Minimization Without Derivatives*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, NJ, 1973. Reprinted by Dover Publications, 2002.

[BS10] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[FW20] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020.

[GHW79] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8580–8589, 2018.

[Mis22] Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

[RR07] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1177–1184, 2007.

[WHS22] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR, 17–23 Jul 2022.

Thank you!

Appendix

Chapter 2 Appendix

Decompose Jacobian over $\boldsymbol{\theta}$ into first and second layer weights,

$$\boldsymbol{K}^{\mathrm{NTK}}(\boldsymbol{x}, \boldsymbol{x}') = \underbrace{\nabla_{\boldsymbol{W}_1} f_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{W}_1} f_{\boldsymbol{\theta}}(\boldsymbol{x}'; \boldsymbol{\theta}_0)}_{\boldsymbol{K}_1(\boldsymbol{x}, \boldsymbol{x}')} + \underbrace{\nabla_{\boldsymbol{w}} f_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{w}} f_{\boldsymbol{\theta}}(\boldsymbol{x}'; \boldsymbol{\theta}_0)}_{\boldsymbol{K}_2(\boldsymbol{x}, \boldsymbol{x}')}$$

- Define the second layer Kernel $\boldsymbol{K}_2(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{F}^\top \boldsymbol{F}$, where $\boldsymbol{F} = \frac{1}{\sqrt{N}} \sigma(\boldsymbol{W}_1 \boldsymbol{X})$.
- A quick computation gives the first layer kernel[5]
  $\boldsymbol{K}_1(\boldsymbol{X}, \boldsymbol{X}) = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \frac{(\boldsymbol{F}')^\top \mathrm{diag}(\boldsymbol{w})^2 \boldsymbol{F}'}{N}$, where $\boldsymbol{F}' = \frac{1}{\sqrt{N}} \sigma'(\boldsymbol{W}_1 \boldsymbol{X})$.
  - $\boldsymbol{K}_2$ is also called the Conjugate Kernel (CK).
  - As $\sigma_W \to 0$, $\boldsymbol{K}^{\mathrm{NTK}}$ degenerates to the CK and the learning problem reduces to Random Feature (RF) regression [RR07].

[5] $\odot$ denotes Hadamard (entry-wise) product

Standard tool from RMT in proportional asymptotic ($n \asymp N \asymp p$) analysis of nonlinear kernel models.

Linear Equivalence (LE)[6]

$$\sigma(\boldsymbol{W}_1\boldsymbol{X}) \approx b_1\boldsymbol{W}_1\boldsymbol{X} + b_2\boldsymbol{\Theta}, \qquad \boldsymbol{\Theta}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

when $\mathbb{E}[\sigma(g)] = 0$, where $b_1 = \mathbb{E}[\sigma'(g)]$, $\quad b_2^2 = \mathbb{E}[\sigma(g)^2] - \tau^2 b_1^2$ and $g \sim \mathcal{N}(0, \tau^2)$.

- Under LE, second layer Kernel simplifies $\boldsymbol{K}_2 \approx \boldsymbol{F}_{\text{lin}}^\top \boldsymbol{F}_{\text{lin}}$, where $\boldsymbol{F}_{\text{lin}} = \frac{1}{\sqrt{N}}(b_1\boldsymbol{W}_1\boldsymbol{X} + b_2\boldsymbol{\Theta})$.
- By centering and controlling the off-diagonal terms of $\boldsymbol{K}_1$, we have

$$\boldsymbol{K}_1 \approx b_1^2\sigma_W^2 p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \tau^2\sigma_W^2(c_1 - b_1^2)\mathbf{I}_n, \quad c_1 := \mathbb{E}[\sigma'(g)^2]$$

Key references: Mei–Montanari (2022), Hu–Lu (2022), Goldt et al. (2021), Adlam–Pennington (2020)

[6] approximation $\approx$ holds on order of bulk statistics

Standard tool from RMT in proportional asymptotic ($n \asymp N \asymp p$) analysis of nonlinear kernel models.

Linear Equivalence (LE)

$$\sigma(\boldsymbol{W}_1\boldsymbol{X}) \approx b_1\boldsymbol{W}_1\boldsymbol{X} + b_2\boldsymbol{\Theta}, \qquad \boldsymbol{\Theta}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

when $\mathbb{E}[\sigma(g)] = 0$, where $b_1 = \mathbb{E}[\sigma'(g)]$, $\quad b_2^2 = \mathbb{E}[\sigma(g)^2] - \tau^2 b_1^2$ and $g \sim \mathcal{N}(0, \tau^2)$.

— Collecting both substitutions, we define the *linearized NTK*

$$\boldsymbol{K}_{\text{lin}} = \boldsymbol{F}_{\text{lin}}^\top \boldsymbol{F}_{\text{lin}} + \sigma_W^2 \left(p^{-1} b_1^2 \boldsymbol{X}^\top \boldsymbol{X} + \tau^2 (c_1 - b_1^2) \mathbf{I}_n \right) + z\mathbf{I}_n, \quad \boldsymbol{F}_{\text{lin}} = \frac{1}{\sqrt{N}} \left(b_1 \boldsymbol{W}_1 \boldsymbol{X} + b_2 \boldsymbol{\Theta} \right)$$

We will drop the the *lin* subscript, and study KRR on the *linearized* kernel $\boldsymbol{K}$ directly.

Universality Principle: Training and test error of kernel ridge regression on $\boldsymbol{K}^{\mathrm{NTK}}$ are asymptotically equivalent to that of regression on *linearized* kernel $\boldsymbol{K}$.

Gaussian Equivalence Theorem (GET)

The following holds pointwise for each fixed $\lambda \geq 0$.

$$\left|\mathcal{R}(\lambda) - \mathcal{R}_{\mathsf{GE}}(\lambda)\right| = o_{n,\mathbb{P}}(1)$$

- $\mathcal{R}_{\mathsf{GE}}(\lambda)$ denotes prediction risk of KRR on the linearized NTK kernel.
- Conjecture GET extends to power-law input (A0).

Key references: Adlam et al. (2019), Hu–Lu (2022), Montanari–Saeed (2022), Peche et al. (2019)

NTK Scaling Law

Assumption 1: Unnormalized power-law

Replace Lecun factor $p^{-1/2}$ on sample covariance with trace-normalization $T_p = \mathrm{Tr}\,\boldsymbol{\Sigma} = \sum_{j=1}^{p} j^{-\alpha}$, where

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p), \qquad \sigma_j = j^{-\alpha}, \quad \alpha > 1.$$

Normalize signal vector $\boldsymbol{\beta}_j = c_{\alpha,r} j^{-r}$, where $c_{\alpha,r}$ is chosen to ensure $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$, and assume $\alpha + 2r > 1$.

Remarks

- Trace-normalization ensures pre-activations are non-degenerate as $p \to \infty$. A similar normalization convention was taken in Loureiro et al (2025).
- Condition $\alpha + 2r > 1$ ensures $||f^*||_{L^2} < \infty$ as $p \to \infty$.
- For simplicity, we assume noiseless test label $y(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$, although this can be dropped by subtracting the variance of test noise from GCV estimator.

Define constants

$$\bar{\sigma}_{p,j} := \frac{\sigma_j}{T_p}, \qquad c_{\mathrm{NTK}} := b_1^2(1+\sigma_W^2) > 0,$$

For effective ridge $\mu > 0$, define **effective dimension** and **source-bias** profiles

$$\mathcal{N}_p(\mu) := \sum_{j=1}^{p} \frac{c_{\mathrm{NTK}}\bar{\sigma}_{p,j}}{c_{\mathrm{NTK}}\bar{\sigma}_{p,j}+\mu}, \quad \mathcal{B}_p(\mu) := \sum_{j=1}^{p} \sigma_j \beta_{p,j}^2 \left(\frac{\mu}{c_{\mathrm{NTK}}\bar{\sigma}_{p,j}+\mu}\right)^2.$$

Assumption 2: Local GCV Deterministic Equivalent

Let $\mu := \lambda + \lambda_{\mathrm{diag}}$ be the effective ridge appearing in the HJ deterministic equivalent, and assume either $\lambda_{\mathrm{diag}} = 0$ or $\lambda_{\mathrm{diag}} = o(\mu_{\lambda^*})$. Further assume there exists constants $C_B, C_V \in (0,\infty)$ such that uniformly for $\mu_\lambda \downarrow 0$ with $\mathcal{N}_p(\mu) = o(n \wedge N)$ and $\mu_\lambda^{-1/\alpha} = o(p)$,

$$R_{n,N,p}(\lambda) = C_B\mathcal{B}_p(\mu_\lambda) + C_V\left(\frac{1}{n}+\frac{1}{N}\right)\mathcal{N}_p(\mu_\lambda) + o\left(\mathcal{B}_p(\mu_\lambda) + \left(\frac{1}{n}+\frac{1}{N}\right)\mathcal{N}_p(\mu_\lambda)\right).$$

Theorem (Finite-width NTK Scaling Law)

Under Assumptions 1 and 2, define

$$q_{\alpha,r} := \min\Big\{2, \frac{\alpha + 2r - 1}{\alpha}\Big\}$$

Suppose $\alpha + 2r \neq 2\alpha + 1$, *and* $\frac{1}{n} + \frac{1}{N} \to 0$ *with p sufficiently large, then the optimally tuned NTK risk satisfies the following* **scaling law**

$$\inf_{\lambda \geq 0} \mathcal{R}_{n,N,p}(\lambda) \asymp \Big(\frac{1}{n} + \frac{1}{N}\Big)^{\theta}, \qquad \theta := \frac{\alpha q_{\alpha,r}}{\alpha q_{\alpha,r} + 1} < 1$$

Moreover, the **optimal ridge** *satisfies* $\mu_{\lambda^*} \asymp \Big(\frac{1}{n} + \frac{1}{N}\Big)^{\frac{\alpha}{\alpha q_{\alpha,r}+1}}$

Recover classical KRR *source-capacity* exponent Caponnetto–De Vito (2007) in the $N \to \infty$ limit, under the renormalization of source parameter $r_{\mathrm{CdV}} = \frac{q}{2}$.

Two Regimes

Non-saturated source regime: if $\alpha + 2r < 2\alpha + 1$, then $q_{\alpha,r} = \frac{\alpha+2r-1}{\alpha}$ and

$$\inf_{\lambda \geq 0} \mathcal{R}_{n,N,p}(\lambda) \asymp \left(\frac{1}{n} + \frac{1}{N}\right)^{\frac{\alpha+2r-1}{\alpha+2r}}, \qquad \mu_{\lambda^*} \asymp \left(\frac{1}{n} + \frac{1}{N}\right)^{\frac{\alpha}{\alpha+2r}}$$

Saturated source regime: $\alpha + 2r > 2\alpha + 1$, then $q_{\alpha,r} = 2$

$$\inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp \left(\frac{1}{n} + \frac{1}{N}\right)^{\frac{2\alpha}{2\alpha+1}}, \qquad \mu_{\lambda_\star} \asymp \left(\frac{1}{n} + \frac{1}{N}\right)^{\frac{\alpha}{2\alpha+1}}.$$

Remarks:

- **Non-saturated source :** $r < \frac{\alpha+1}{2}$ rate depends on both (α, r).
- **Saturated source:** $r > \frac{\alpha+1}{2}$ rate set by capacity α only, source parameter r no longer contributes.

Corollary

Under the setup of the previous Theorem, if $N \asymp n^{\ell}$ for some $\ell > 0$, then

$$\left(\frac{1}{n} + \frac{1}{N}\right) \asymp n^{-\min(1,\ell)}.$$

Consequently,

$$\inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp n^{-\theta_{\mathrm{NTK}}(\alpha,r,\ell)}, \qquad \theta_{\mathrm{NTK}}(\alpha, r, \ell) := \min\{1, \ell\} \frac{\alpha q_{\alpha,r}}{\alpha q_{\alpha,r} + 1}$$

Data vs Width Bottleneck: When $\ell < 1$, width is the limiting factor, and θ_{NTK} is degraded by the same factor. When $\ell > 1$, data binds and finite-width NTK gets full kernel exponent θ_{NTK}.

Core idea: Split modes at resolution-scale $J_\mu := \lfloor \mu^{-1/\alpha} \rfloor$.

$$\frac{c_{\mathrm{NTK}}\bar{\sigma}_{p,j}}{c_{\mathrm{NTK}}\bar{\sigma}_{p,j}+\mu} \asymp \begin{cases} 1, & j \le J_\mu \\ j^{-\alpha}/\mu, & j > J_\mu \end{cases}, \qquad \frac{\mu}{c_{\mathrm{NTK}}\bar{\sigma}_{p,j}+\mu} \asymp \begin{cases} \mu j^{\alpha}, & j \le J_\mu \\ 1, & j > J_\mu \end{cases}$$

Step one: Estimate *effective dimension* $\mathcal{N}_p(\mu)$.

$$\mathcal{N}_p(\mu) \asymp \sum_{j \le J_\mu} 1 + \mu^{-1} \sum_{J_\mu < j \le p} j^{-\alpha} \asymp \boxed{\mu^{-1/\alpha}.}$$

Step two: Estimate the *source-bias profile* $\mathcal{B}_p(\mu) = B_{\le}(\mu) + B_{\ge}(\mu)$, where

$$B_{\le}(\mu) \asymp \mu^2 \sum_{j \le J_\mu} j^{\alpha-2r}, \qquad B_{\ge}(\mu) \asymp \sum_{j > J_\mu} j^{-(\alpha+2r)}$$

When $\alpha + 2r > 2\alpha + 1$, the below-resolution sum is $O(1)$ and $\mathcal{B}_p(\mu) \asymp \mu^2$. Otherwise, $\mathcal{B}_p(\mu) \asymp \mu^{(\alpha+2r-1)/\alpha}$. In either case,

$$\boxed{\mathcal{B}_p(\mu) \asymp \mu^{q_{\alpha,r}}}$$

Step three: Optimize the deterministic equivalent.

$$\mathcal{R}_{n,N,p}(\lambda) \asymp \mu_\lambda^{q_\alpha,r} + \left(\frac{1}{n}+\frac{1}{N}\right)\mu_\lambda^{-1/\alpha}$$

Balance the two terms when $\mu_{\lambda^*} \asymp \left(\frac{1}{n}+\frac{1}{N}\right)\mu_\lambda^{-1/\alpha}$. At this optimal ridge,

$$\inf_{\lambda\geq 0} \mathcal{R}_{n,N,p}(\lambda) \asymp \left(\frac{1}{n}+\frac{1}{N}\right)^{\frac{\alpha q_{\alpha,r}}{\alpha q_{\alpha,r}+1}}$$

Remark: At logarithmic boundary $\alpha + 2r = 2\alpha + 1$,

$$\inf_{\lambda\geq 0} \mathcal{R}_{n,N,p}(\lambda) \asymp \left(\frac{1}{n}+\frac{1}{N}\right)^{\frac{2\alpha}{2\alpha+1}} \left(\log \frac{1}{n^{-1}+N^{-1}}\right)^{\frac{1}{2\alpha+1}}$$

Chapter 3 Appendix

Case 3: Setup — Spiked Weights

Motivation: Pretrained NNs have weight matrices with spectral spikes (heavy-tailed ESDs observed empirically). Model this as a rank-one perturbation:

$$\boldsymbol{W}_1 = \boldsymbol{W} + \theta\, a\, b^\top, \qquad \theta = \theta_0\, N^{1/4}$$

- $\boldsymbol{W}$: random initialization ($W_{ij} \sim \mathcal{N}(0,1)$)
- $a \sim \mathcal{N}(0, I_N/N)$: random direction
- $b = \mathbf{1}_d$: deterministic direction
- SNR $r = \Theta(1)$ (finite!)

Idea: Even with finite SNR, the **large spike in $\boldsymbol{W}_1$** amplifies signals via the quadratic channel.

Related to: one-step gradient descent (Ba et al., 2022), LoRA (Hu et al., 2021).

Theorem (Spiked Weights)

Define $\kappa = 1 + r^2/2$, $\kappa_\perp = 2 + 2r^2 + r^4/4$, *and* $\beta_y = \frac{3c_\sigma^2}{4}\theta_0^4\,\varphi\,\kappa_\perp$.

(i) **Bulk**: *ESD* $\to \mu$ *(same as random weights).*

(ii) **One giant outlier**: $\hat{\lambda}_1 = b_\sigma^2\,\theta_0^2\,\varphi\,\kappa\,\sqrt{N}\,(1 + o_P(1))$, *eigenvector* $\perp \boldsymbol{y}$.

(iii) **Inherited** $O(1)$ **spikes**: *All outliers from finite SNR case persist* ($\perp \boldsymbol{y}$).

(iv) **Additional** $O(1)$ **spike**: *If* $z'(s_y) > 0$ *where* $\beta_y\,T(s_y) = 1$, *a new outlier emerges.*

(v) **Label alignment**: *In the supercritical case,*

$$\boxed{\frac{1}{n}\left|\hat{v}_y^\top \boldsymbol{y}\right|^2 \xrightarrow{P} \frac{s_y\,z'(s_y)}{\beta_y\,T'(s_y)} \cdot \frac{r^4/4}{\kappa_\perp} > 0}$$

Key Takeaway

Pretrained weights create a **quadratic feature channel** even at finite SNR. The large spike in $\boldsymbol{W}_1$ amplifies the XOR signal through the second-order Hermite component c_σ.

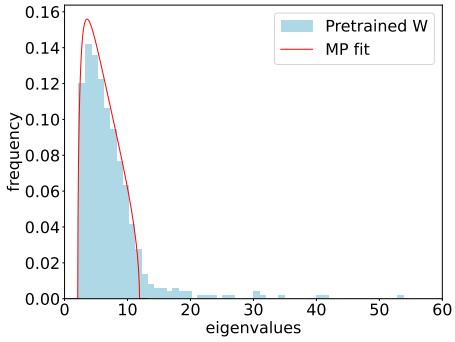
Interpretation:

- The LE is no longer sufficient: a Quadratic Equivalent is needed
- The spike in $\boldsymbol{W}_1$ "boosts" the quadratic channel to $O(1)$ strength
- If $c_\sigma = 0$, no label alignment occurs (quadratic channel absent)

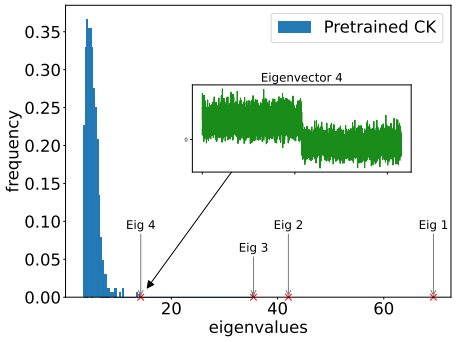
Empirical verification: Using weights from a 20-layer NN pretrained on CIFAR-2:

- After sufficient training, 4th eigenvector of CK aligns with XOR labels

Knob: Pretrained Weights — CIFAR-2 Experiments



ESD of FC1 weights (40 epochs). MP overlaid in red.

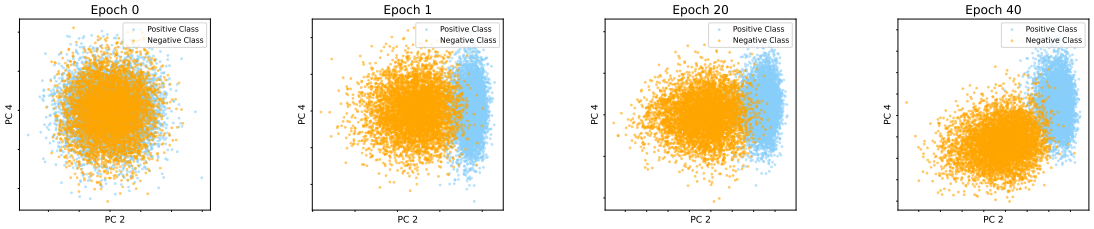


ESD of CK with pretrained weights.

20-layer NN trained on CIFAR-2 (88% accuracy, 40 epochs). CK tested on XOR data. $n = 10000$, $d = 3072$, $N = 512$, $r = 5.25$, $\sigma \propto \mathrm{ReLU}$.

Key: Pretrained CK exhibits 4 spikes, with 4th eigenvector aligning with XOR labels.

Knob: Pretrained Weights — Phase Transition Over Training

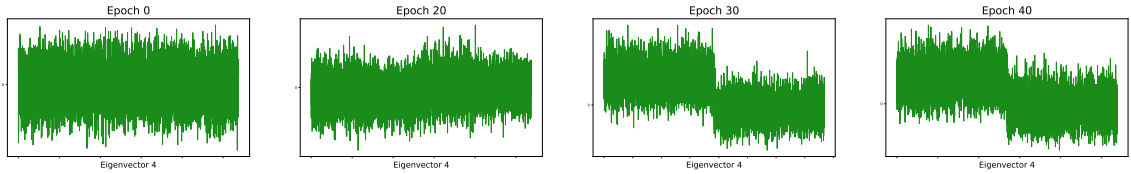


Epoch 0 Epoch 1 Epoch 20 Epoch 40

PC 2 vs PC 4 of CK with trained weights, tracked from Epoch 0 to 40. Yellow: -1, Blue: $+1$.

Phase transition: 4th eigenvector transitions from **pure noise** $\rightarrow$ **near-perfect XOR alignment**.

Knob: Pretrained Weights — Eigenvector Alignment



Epoch 0 Epoch 20 Epoch 30 Epoch 40

4th leading eigenvector of CK with pre-trained weights over training.

Sharp phase transition between Epoch 20 and 30: eigenvector aligns perfectly with XOR labels $\boldsymbol{y}$.

$\Rightarrow$ Spectral transition in the CK parallels the feature learning dynamics in the NN.

Motivation: In proportional regime $n = \Theta(d)$, the LE washes out higher-order structure. What if we have more data?

Regime:

$$\frac{2n}{d^2} \to \gamma \in (0, \infty), \qquad \frac{n^2}{N} \to 0 \quad \text{(ultra-wide)}$$

Key assumptions: $b_\sigma = 0$ (no linear channel), $c_\sigma \neq 0$, SNR $r = O(1)$.

Intuition: With $n \asymp d^2$ samples, the CK concentrates around a quadratic polynomial kernel:

$$\boldsymbol{K}^{\mathrm{CK}} \approx a_\sigma^2 P + b_\sigma^2 \boldsymbol{X}^\top \boldsymbol{X} + \frac{c_\sigma^2}{2}\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \odot \left(\boldsymbol{X}^\top \boldsymbol{X}\right)$$

The Hadamard square $(\boldsymbol{X}^\top \boldsymbol{X})^{\odot 2}$ provides the quadratic features needed for XOR.

Theorem (Quadratic Sample Size)

Let $\ell = r^4/4$, $\boldsymbol{K}_c^{\mathrm{CK}} = P\boldsymbol{K}^{\mathrm{CK}}P$ (centered CK), and $\lambda_+(\gamma) = (1+\sqrt{\gamma})^2$.

(i) **Bulk**: *ESD of $\boldsymbol{K}_c^{\mathrm{CK}} \to \rho_\gamma^{MP}$.*

(ii) **BBP for informative outlier**: *If $\ell > 1/\sqrt{\gamma}$, then*

$$\hat{\lambda}_1(\boldsymbol{K}_c^{\mathrm{CK}}) \xrightarrow{P} (1 - c_\sigma^2/2) + c_\sigma^2\left(1 + \gamma + \gamma\ell + \frac{1}{\ell}\right).$$

If $\ell \le 1/\sqrt{\gamma}$, then $\hat{\lambda}_1(\boldsymbol{K}_c^{\mathrm{CK}}) \xrightarrow{P} \lambda_+(\gamma)$ (no outlier).

(iii) **Eigenvector alignment with label**: *If $\ell > 1/\sqrt{\gamma}$,*

$$\boxed{\frac{1}{n}\left|\langle \hat{v}_1(\boldsymbol{K}_c^{\mathrm{CK}}), \boldsymbol{y}\rangle\right|^2 \xrightarrow{P} \frac{\gamma\ell^2 - 1}{\gamma\ell(\ell+1)} \in (0,1)}$$

Key Takeaway

With quadratic sample complexity $n \asymp d^2$, the CK reduces to a quadratic polynomial kernel. Under the BBP threshold $\ell > 1/\sqrt{\gamma}$, an informative outlier emerges with **positive label alignment**.

Phase transition: The threshold $\ell > 1/\sqrt{\gamma}$, i.e., $r^4/4 > 1/\sqrt{\gamma}$, is a sharp condition on the SNR for the kernel to "see" the XOR structure.

Additional assumptions:

- Ultra-wide network $N \gg n^2$: random kernel concentrates around a deterministic kernel
- $b_\sigma = 0$: removes rank-d linear spikes from $\boldsymbol{X}^\top \boldsymbol{X}$
- Extending to $N \asymp n$ (linear width) is future work