NOISE PROPAGATION OPERATORS AND A TWO-PARAMETER CHARACTERIZATION OF BENIGN OVERFITTING

CΛ / LIGHTMAN

ABSTRACT. Overparameterized models that interpolate noisy training data can nonetheless generalize well, a phenomenon known as benign overfitting. Existing taxonomies, notably that of Mallinar et al. (2022), characterize the benign–tempered–catastrophic trichotomy through a single eigenvalue sequence of the kernel matrix. We introduce the *noise propagation operator* $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1}$, where $\mathbf{\Sigma}$ collects the singular values of the training Jacobian and $\mathbf{\Gamma}$ measures the test–train cross-correlation of gradient features. This operator yields a unified generalization bound whose excess risk decomposes into a signal-bias term, a noise-variance term $\sigma^2 \operatorname{tr}(\mathbf{M})$, and a concentration tail controlled by the Frobenius norm of $\mathbf{M}$. Under power-law decay assumptions $\sigma_j \sim j^{-\alpha}$ and $\rho_j \sim j^{-\beta}$, we prove a sharp dichotomy: the noise-variance term $\sigma^2 \operatorname{tr}(\mathbf{M})$ is finite if and only if $\beta > \alpha + \frac{1}{2}$, diverges logarithmically when $\beta = \alpha + \frac{1}{2}$, and diverges polynomially otherwise. The two-parameter (α, β) framework separates the mechanism by which noise enters the model (governed by α) from the mechanism by which noise is expressed at test points (governed by β), thereby explaining why feature learning can convert catastrophic overfitting into benign overfitting by increasing β while α remains fixed. We illustrate the framework with a spectral-separation result for two-layer ReLU teacher–student networks.

1. INTRODUCTION

A central puzzle in the theory of overparameterized learning is that models with far more parameters than training samples can interpolate noisy data—fitting every label exactly—yet still predict accurately on unseen inputs. Classical bias–variance trade-off reasoning predicts that such interpolation should amplify noise and destroy generalization, but modern neural networks routinely defy this prediction. Understanding

Date: May 5, 2026.

2020 *Mathematics Subject Classification.* 62J05, 68T07, 62G20.

Key words and phrases. Benign overfitting, overparameterized models, generalization bounds, noise propagation, Jacobian spectrum, neural tangent kernel.

when and why interpolation is compatible with generalization is one of the most active questions in statistical learning theory.

What problem do we address? We seek a quantitative criterion that determines whether an interpolating predictor generalizes well (benign overfitting), moderately (tempered overfitting), or poorly (catastrophic overfitting), in terms of the spectral geometry of the model's Jacobian and its interaction with the test distribution.

Why is this important? Existing criteria for benign overfitting, while powerful, rely on a single spectral sequence. In fixed-kernel regression the kernel eigenvalues simultaneously control how noise is absorbed during training and how it is expressed at test time, so a single sequence suffices. However, after feature learning the training-time and test-time spectral behaviors can decouple: the learned features may suppress noise in directions that are invisible at test points even if those directions carry substantial noise energy in the training fit. A single-parameter criterion cannot capture this decoupling.

What have others done? Bartlett, Long, Lugosi, and Tsigler [2] established the first rigorous benign-overfitting result for linear regression, showing that the minimum-norm interpolant generalizes when the effective rank of the covariance tail is large relative to the sample size. Tsigler and Bartlett [10] extended these results to ridge regression with sharp non-asymptotic bounds. Mallinar et al. [7] introduced a taxonomy—benign, tempered, and catastrophic—based on the eigenvalue decay rate of the kernel matrix, providing a unified qualitative picture across kernel methods and neural networks. On the optimization side, Jacot, Gabriel, and Hongler [5] introduced the Neural Tangent Kernel (NTK), which linearizes the network around initialization and connects overparameterized training dynamics to kernel regression. Hastie, Montanari, Rosset, and Tibshirani [4] gave precise asymptotic characterizations of the double-descent phenomenon in least-squares regression. Boursier and Flammarion [3] analyzed gradient flow in two-layer ReLU teacher–student networks, proving that the learned weights align with the teacher directions. Belkin, Hsu, Ma, and Mandal [1] empirically demonstrated the double-descent curve. Mei and Montanari [6] analyzed generalization in random feature models, providing sharp asymptotics that connect kernel regression to neural network behaviour. Rakhlin and Zhai [8] gave minimax-optimal risk bounds for kernel regression that highlight the role of the effective dimension.

What is new here? We introduce the *noise propagation operator* $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1}$, which factors the test-time noise amplification into two independent components:

- $\mathbf{\Sigma}$: the singular values of the training Jacobian, governing how noise enters the interpolating solution (decay rate α);
- $\mathbf{\Gamma}$: the test–train cross-correlation of gradient features, governing how noise is expressed at test points (decay rate β).

In fixed-kernel regression, $\mathbf{\Gamma}$ is determined by $\mathbf{\Sigma}$ (both are controlled by the same kernel eigenvalues), so the framework reduces to a single-parameter condition. After feature learning, $\mathbf{\Gamma}$ can change independently of $\mathbf{\Sigma}$: the learned representation can increase β (making noise directions invisible at test time) while α remains fixed. This separation provides a mechanistic explanation for why feature learning improves generalization beyond what any fixed-kernel theory can predict.

Our main contributions are:

(1) A **unified generalization bound** (Theorem 3.1) expressing the test mean squared error as the sum of a signal-bias term, the noise-variance term $\sigma^2 \operatorname{tr}(\mathbf{M})$, and a concentration tail controlled by $\|\mathbf{M}\|_F$.
(2) A **benign-overfitting dichotomy** (Theorem 3.4) giving a sharp two-parameter condition: overfitting is benign if and only if $\beta > \alpha + \frac{1}{2}$.
(3) An **application to two-layer ReLU networks** (Proposition 4.1) demonstrating that gradient-flow training achieves spectral separation in the Jacobian, yielding benign generalization.

2. Preliminaries

2.1. **Setting and notation.** Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a training set with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We write $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ for the label vector. We assume a standard regression model

$$y_i = g^*(\mathbf{x}_i) + \xi_i, \qquad i = 1, \dots, n, \tag{1}$$

where $g^* \colon \mathbb{R}^d \to \mathbb{R}$ is the unknown ground-truth function and $\xi_1, \dots, \xi_n$ are independent, mean-zero, sub-Gaussian random variables with variance proxy σ^2. That is, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[e^{\lambda \xi_i}\big] \le e^{\lambda^2 \sigma^2 / 2}. \tag{2}$$

Remark 2.1 (Notation convention). Throughout this paper, σ (without subscript) denotes the noise standard deviation, while σ_j (with subscript) denotes the j-th singular value of the Jacobian $\mathbf{J}$. The diagonal matrix of singular values is $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$.

Let $f(\cdot\,;\theta)\colon \mathbb{R}^d \to \mathbb{R}$ be a parametric predictor with parameter vector $\theta \in \mathbb{R}^p$. We assume that θ^* is the *minimum-norm* interpolating parameter vector relative to a reference θ_0:

$$f(\mathbf{x}_i;\theta^*) = y_i, \qquad i = 1,\dots,n, \tag{3}$$

with $\theta^* - \theta_0 = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top(\mathbf{y} - f(\mathbf{X};\theta_0))$ (the minimum-norm linearized solution). For simplicity we take θ_0 such that $f(\mathbf{x}_i;\theta_0) = 0$ for all i, which can always be achieved by redefining f.

2.2. **Jacobian and its singular value decomposition.**

Definition 2.2 (Training Jacobian)**.** The *training Jacobian* at θ^* is the matrix $\mathbf{J} \in \mathbb{R}^{n\times p}$ whose i-th row is the gradient of the prediction at the i-th training point:

$$\mathbf{J}_{i,:} = \nabla_\theta f(\mathbf{x}_i;\theta^*)^\top, \qquad i = 1,\dots,n. \tag{4}$$

Let the singular value decomposition (SVD) of $\mathbf{J}$ be

$$\mathbf{J} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \tag{5}$$

where $\mathbf{U} = [\mathbf{u}_1,\dots,\mathbf{u}_n] \in \mathbb{R}^{n\times n}$ is orthogonal, $\mathbf{V} = [\mathbf{v}_1,\dots,\mathbf{v}_p] \in \mathbb{R}^{p\times p}$ is orthogonal (with only the first n columns relevant when $p > n$), and $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1,\dots,\sigma_n)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n > 0$. The assumption $\sigma_n > 0$ is equivalent to requiring that $\mathbf{J}$ has full row rank, which is a necessary condition for interpolation via linearized perturbation of the parameters.

2.3. **Gradient feature functions and cross-correlation.**

Definition 2.3 (Gradient feature functions)**.** For each $j = 1,\dots,n$, the *j-th gradient feature function* is

$$\psi_j(\mathbf{x}) = \nabla_\theta f(\mathbf{x};\theta^*)^\top \mathbf{v}_j, \tag{6}$$

where $\mathbf{v}_j$ is the j-th right singular vector of $\mathbf{J}$. Evaluated at a test point $\mathbf{x}$, the scalar $\psi_j(\mathbf{x})$ measures the projection of the test gradient onto the j-th training principal direction.

Definition 2.4 (Cross-correlation matrix)**.** The *cross-correlation matrix* $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times n}$ is defined by

$$\boldsymbol{\Gamma}_{jl} = \mathbb{E}_{\mathbf{x}}\big[\psi_j(\mathbf{x})\,\psi_l(\mathbf{x})\big], \qquad j,l = 1,\dots,n, \tag{7}$$

where the expectation is over the test distribution. The matrix $\boldsymbol{\Gamma}$ is symmetric and positive semidefinite. We write $\rho_j = \sqrt{\boldsymbol{\Gamma}_{jj}}$ for the root-mean-square magnitude of the j-th gradient feature at test points.

2.4. **The noise propagation operator.**

Definition 2.5 (Noise propagation operator)**.** The *noise propagation operator* is the matrix

$$\mathbf{M} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\,\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{n\times n}. \tag{8}$$

Since $\boldsymbol{\Gamma}$ is positive semidefinite and $\boldsymbol{\Sigma}^{-1}$ is positive diagonal, $\mathbf{M}$ is symmetric and positive semidefinite.

Remark 2.6. The operator $\mathbf{M}$ captures the full pipeline of noise amplification: the noise vector $\boldsymbol{\xi} \in \mathbb{R}^n$ is first projected onto the training singular directions (the columns of $\mathbf{U}$), scaled by σ_j^{-1} (the inverse singular values of $\mathbf{J}$), and then the resulting coefficients are propagated to test points through the cross-correlation $\boldsymbol{\Gamma}$. The factor $\boldsymbol{\Sigma}^{-1}$ appears twice because the interpolation condition forces the noise coefficient in direction j to be proportional to σ_j^{-1}, and the test-time variance of that coefficient depends on its squared magnitude σ_j^{-2}.

2.5. **Linearization residual.** Our analysis relies on a local linearization of the predictor around a reference parameter θ_0 (for example, the initialization).

Assumption 2.7 (Linearization)**.** There exists a constant $\delta_{\text{lin}} \geq 0$ such that for every test point $\mathbf{x}$ in the support of the test distribution,

$$\left|f(\mathbf{x};\theta^*) - f(\mathbf{x};\theta_0) - \nabla_\theta f(\mathbf{x};\theta_0)^\top(\theta^* - \theta_0)\right| \leq \delta_{\text{lin}}. \tag{9}$$

When $\delta_{\text{lin}} = 0$ the predictor is exactly linear in the parameters (as in kernel regression or the NTK regime). For finite-width networks, δ_{lin} quantifies the quality of the first-order Taylor approximation.

2.6. **Signal bias.**

Definition 2.8 (Signal bias)**.** The *signal bias* is defined as

$$B^2_{\text{signal}} = \mathbb{E}_{\mathbf{x}}\left[\left(\sum_{j=1}^{n} \frac{\mathbf{u}_j^\top \mathbf{g}^*}{\sigma_j}\,\psi_j(\mathbf{x}) - g^*(\mathbf{x})\right)^2\right], \tag{10}$$

where $\mathbf{g}^* = (g^*(\mathbf{x}_1), \ldots, g^*(\mathbf{x}_n))^\top \in \mathbb{R}^n$ is the vector of noiseless labels. This term measures the approximation error incurred by the linearized interpolant in reconstructing the true function, and depends on the richness of the feature space but not on the noise.

3. Main Results

3.1. Unified generalization bound.

Theorem 3.1 (Unified Generalization Bound). *Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be generated according to the model (1) with sub-Gaussian noise satisfying (2). Let θ^* be an interpolating solution satisfying (3), and let $\mathbf{J}$, $\mathbf{\Sigma}$, $\mathbf{V}$, ψ_j, $\mathbf{\Gamma}$, and $\mathbf{M}$ be as in Definitions 2.2–2.5. Suppose Assumption 2.7 holds with residual δ_{lin}. Then there exists an absolute constant $C > 0$ (depending only on the sub-Gaussian norm of ξ_i) such that for every $\delta \in (0,1)$, with probability at least $1-\delta$ over the noise $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$,*

$$\text{MSE}_{\text{test}} \le C\big(B_{\text{signal}}^2 + \sigma^2 \operatorname{tr}(\mathbf{M}) + \sigma^2 \|\mathbf{M}\|_F \sqrt{\log(1/\delta)} + \delta_{\text{lin}}^2\big), \tag{11}$$

where $\text{MSE}_{\text{test}} = \mathbb{E}_{\mathbf{x}}\big[(f(\mathbf{x};\theta^*) - g^*(\mathbf{x}))^2\big]$ is the test mean squared error (conditional on the training data and noise realization), and $\|\mathbf{M}\|_F = (\sum_{j,l} \mathbf{M}_{jl}^2)^{1/2}$ is the Frobenius norm.

Proof. The proof proceeds in four steps.

Step 1: Decomposition of the test prediction error. By the interpolation condition (3) and the data model (1), the parameter perturbation $\theta^* - \theta_0$ is determined by the system $\mathbf{J}(\theta^* - \theta_0) \approx \mathbf{y} - f(\mathbf{X};\theta_0)$, where the approximation is controlled by the linearization residual. Under Assumption 2.7, the test prediction error admits the decomposition

$$f(\mathbf{x};\theta^*) - g^*(\mathbf{x}) = S(\mathbf{x}) + N(\mathbf{x}) + R(\mathbf{x}), \tag{12}$$

where the three terms are defined as follows.

The *signal reconstruction term* is

$$S(\mathbf{x}) = \sum_{j=1}^{n} \frac{\mathbf{u}_j^\top \mathbf{g}^*}{\sigma_j} \psi_j(\mathbf{x}) - g^*(\mathbf{x}). \tag{13}$$

This term captures how well the linearized interpolant reconstructs the noiseless function g^*.

The *noise leakage term* is

$$N(\mathbf{x}) = \sum_{j=1}^{n} \frac{\mathbf{u}_j^\top \boldsymbol{\xi}}{\sigma_j} \psi_j(\mathbf{x}). \tag{14}$$

This term captures how training noise propagates to the test prediction.

The *linearization residual term* satisfies $|R(\mathbf{x})| \le \delta_{\text{lin}}$ by Assumption 2.7.

To verify this decomposition, note that the interpolation condition gives $\mathbf{J}(\theta^* - \theta_0) = \mathbf{g}^* + \boldsymbol{\xi} - f(\mathbf{X};\theta_0) + \mathbf{r}$ where $\mathbf{r}$ collects training-point residuals. Using the SVD $\mathbf{J} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, the minimum-norm solution is

$\theta^* - \theta_0 = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top(\mathbf{g}^* + \boldsymbol{\xi} - f(\mathbf{X};\theta_0) + \mathbf{r})$. The linearized test prediction is

$$\begin{aligned} f(\mathbf{x};\theta^*) &\approx f(\mathbf{x};\theta_0) + \nabla_\theta f(\mathbf{x};\theta_0)^\top(\theta^* - \theta_0) \\ &= f(\mathbf{x};\theta_0) + \sum_{j=1}^{n} \frac{\mathbf{u}_j^\top(\mathbf{g}^* + \boldsymbol{\xi} - f(\mathbf{X};\theta_0) + \mathbf{r})}{\sigma_j}\,\psi_j(\mathbf{x}). \end{aligned}$$

Subtracting $g^*(\mathbf{x})$ and rearranging yields (12), where the signal reconstruction involves $\mathbf{u}_j^\top \mathbf{g}^*/\sigma_j$, the noise leakage involves $\mathbf{u}_j^\top \boldsymbol{\xi}/\sigma_j$, and all remaining terms (including those involving $f(\mathbf{X};\theta_0)$ and $\mathbf{r}$) are absorbed into $S(\mathbf{x})$ and $R(\mathbf{x})$.

Step 2: Expected noise variance equals $\sigma^2\,\mathrm{tr}(\mathbf{M})$. We compute the expected test MSE contribution from the noise term. Defining the random variable

$$Z = \mathbb{E}_{\mathbf{x}}\big[N(\mathbf{x})^2\big] = \mathbb{E}_{\mathbf{x}}\left[\left(\sum_{j=1}^{n} \frac{\mathbf{u}_j^\top \boldsymbol{\xi}}{\sigma_j}\,\psi_j(\mathbf{x})\right)^2\right], \tag{15}$$

we expand the square and exchange the expectation over $\mathbf{x}$ with the sum:

$$\begin{aligned} Z &= \sum_{j=1}^{n}\sum_{l=1}^{n} \frac{\mathbf{u}_j^\top \boldsymbol{\xi}}{\sigma_j}\,\frac{\mathbf{u}_l^\top \boldsymbol{\xi}}{\sigma_l}\,\mathbb{E}_{\mathbf{x}}\big[\psi_j(\mathbf{x})\,\psi_l(\mathbf{x})\big] \\ &= \sum_{j=1}^{n}\sum_{l=1}^{n} \frac{(\mathbf{u}_j^\top \boldsymbol{\xi})\,(\mathbf{u}_l^\top \boldsymbol{\xi})}{\sigma_j\,\sigma_l}\,\boldsymbol{\Gamma}_{jl} \\ &= \boldsymbol{\xi}^\top \mathbf{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\,\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top \boldsymbol{\xi} \\ &= \boldsymbol{\xi}^\top \mathbf{U}\mathbf{M}\mathbf{U}^\top \boldsymbol{\xi}. \end{aligned} \tag{16}$$

Taking the expectation over $\boldsymbol{\xi}$, using $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] = \sigma^2\mathbf{I}_n$ and the orthogonality of $\mathbf{U}$:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\xi}}[Z] &= \sigma^2\ \mathrm{tr}\big(\mathbf{U}\mathbf{M}\mathbf{U}^\top\big) \\ &= \sigma^2\ \mathrm{tr}\big(\mathbf{U}^\top\mathbf{U}\mathbf{M}\big) \\ &= \sigma^2\ \mathrm{tr}(\mathbf{M}). \end{aligned} \tag{17}$$

This establishes that the expected noise contribution to the test MSE is $\sigma^2\,\mathrm{tr}(\mathbf{M})$.

Step 3: Concentration via the Hanson–Wright inequality. The quadratic form $Z = \boldsymbol{\xi}^\top\mathbf{U}\mathbf{M}\mathbf{U}^\top\boldsymbol{\xi}$ from (16) involves a sub-Gaussian random vector $\boldsymbol{\xi}$ and the positive semidefinite matrix $\mathbf{A} = \mathbf{U}\mathbf{M}\mathbf{U}^\top$. Note that $\|\mathbf{A}\|_F = \|\mathbf{M}\|_F$ and $\|\mathbf{A}\|_{\mathrm{op}} = \|\mathbf{M}\|_{\mathrm{op}}$ since $\mathbf{U}$ is orthogonal.

By the Hanson–Wright inequality (see, e.g., Rudelson and Vershynin [9]), there exists an absolute constant $c > 0$ such that for all $t > 0$,

$$\mathbb{P}\big[|Z - \mathbb{E}[Z]| > t\big] \le 2\exp\left(-c\,\min\left(\frac{t^2}{\sigma^4\,\|\mathbf{M}\|_F^2},\ \frac{t}{\sigma^2\,\|\mathbf{M}\|_{\mathrm{op}}}\right)\right). \tag{18}$$

Setting the right-hand side equal to δ and solving for t in the Frobenius-norm regime (which dominates when t is not too large), we obtain that with probability at least $1 - \delta$,

$$Z \le \mathbb{E}[Z] + C_1\,\sigma^2\,\|\mathbf{M}\|_F\,\sqrt{\log(1/\delta)}, \tag{19}$$

where $C_1 > 0$ is an absolute constant depending only on c and the sub-Gaussian norm of ξ_i.

Step 4: Combining the three contributions. The test MSE satisfies

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{test}} &= \mathbb{E}_{\mathbf{x}}\big[(f(\mathbf{x};\theta^*) - g^*(\mathbf{x}))^2\big] \\
&= \mathbb{E}_{\mathbf{x}}\big[(S(\mathbf{x}) + N(\mathbf{x}) + R(\mathbf{x}))^2\big] \\
&\le \mathbb{E}_{\mathbf{x}}\big[\big(|S(\mathbf{x})| + |N(\mathbf{x})| + |R(\mathbf{x})|\big)^2\big] \\
&\le 3\,\mathbb{E}_{\mathbf{x}}\big[S(\mathbf{x})^2\big] + 3\,\mathbb{E}_{\mathbf{x}}\big[N(\mathbf{x})^2\big] + 3\,\delta_{\mathrm{lin}}^2, \qquad (20)
\end{aligned}$$

where the last inequality uses $(a+b+c)^2 \le 3(a^2+b^2+c^2)$.

By the definition of signal bias (Definition 2.8), $\mathbb{E}_{\mathbf{x}}[S(\mathbf{x})^2] = B_{\mathrm{signal}}^2$. By (17) and (19), $\mathbb{E}_{\mathbf{x}}[N(\mathbf{x})^2] \le \sigma^2\,\mathrm{tr}(\mathbf{M}) + C_1\sigma^2\,\|\mathbf{M}\|_F\,\sqrt{\log(1/\delta)}$ with probability at least $1 - \delta$. Substituting into (20) and absorbing the factor of 3 into the constant C yields the bound (11). □

Remark 3.2. In kernel regression with a fixed kernel K, the Jacobian is $\mathbf{J} = \mathbf{X}$ (the data matrix in feature space), and the test–train cross-correlation $\mathbf{\Gamma}_{jl}$ reduces to the kernel eigenvalue structure. In this case, $\mathbf{M}$ is determined entirely by the kernel eigenvalues, and $\mathrm{tr}(\mathbf{M})$ reduces to the effective-rank-based criteria studied in [2, 10, 7]. The two-parameter framework thus strictly generalizes the single-parameter taxonomy.

3.2. **Benign overfitting dichotomy.** We now specialize to the setting where the singular values and cross-correlation exhibit power-law decay, which is common in practice and connects directly to the taxonomy of Mallinar et al. [7].

Assumption 3.3 (Power-law spectral decay)**.** There exist constants $\alpha > 0$, $\beta > 0$, $c_1, c_2 > 0$ such that for all $j \ge 1$:

(1) The singular values of $\mathbf{J}$ satisfy $c_1\,j^{-\alpha} \le \sigma_j \le c_1^{-1}\,j^{-\alpha}$.

(2) The diagonal entries of $\mathbf{\Gamma}$ satisfy $c_2\, j^{-\beta} \le \rho_j \le c_2^{-1}\, j^{-\beta}$, where $\rho_j = \sqrt{\mathbf{\Gamma}_{jj}}$.

Theorem 3.4 (Benign Overfitting Dichotomy)**.** *Under Assumption 3.3, the noise-variance term $\sigma^2 \operatorname{tr}(\mathbf{M})$ in the bound of Theorem 3.1 exhibits the following trichotomy:*

(1) ***Benign regime*** *($\beta > \alpha + \frac{1}{2}$): $\operatorname{tr}(\mathbf{M}) < \infty$, so the noise contribution to the test MSE is bounded.*
(2) ***Catastrophic regime*** *($\beta < \alpha + \frac{1}{2}$): $\operatorname{tr}(\mathbf{M}) = \infty$; equivalently, the tail sum $\sum_{j>k} \rho_j^2/\sigma_j^2$ diverges for every finite k, and the noise contribution is unbounded.*
(3) ***Tempered regime*** *($\beta = \alpha + \frac{1}{2}$): $\operatorname{tr}(\mathbf{M})$ diverges logarithmically; specifically, $\sum_{j=1}^{J} \rho_j^2/\sigma_j^2 = \Theta(\log J)$ as $J \to \infty$.*

Proof. We analyze the trace of $\mathbf{M}$ by computing the diagonal entries. Since $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$ is diagonal, the diagonal entries of $\mathbf{M}$ are

$$\mathbf{M}_{jj} = \frac{\mathbf{\Gamma}_{jj}}{\sigma_j^2} = \frac{\rho_j^2}{\sigma_j^2}. \tag{21}$$

Therefore,

$$\operatorname{tr}(\mathbf{M}) = \sum_{j=1}^{n} \frac{\rho_j^2}{\sigma_j^2}. \tag{22}$$

Since $\mathbf{\Sigma}^{-1}$ is diagonal, we have $\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1}) = \sum_j \sigma_j^{-2}\mathbf{\Gamma}_{jj} = \sum_j \rho_j^2/\sigma_j^2$, using the identity $\operatorname{tr}(\mathbf{DAD}) = \sum_j D_{jj}^2 A_{jj}$ for diagonal $\mathbf{D}$. Note that the off-diagonal entries of $\mathbf{\Gamma}$ do not contribute to $\operatorname{tr}(\mathbf{M})$.

Under Assumption 3.3, each summand satisfies

$$\frac{\rho_j^2}{\sigma_j^2} \asymp \frac{j^{-2\beta}}{j^{-2\alpha}} = j^{2(\alpha-\beta)}, \tag{23}$$

where $\asymp$ denotes equality up to positive multiplicative constants depending on c_1 and c_2. The tail sum starting from any fixed index k is therefore

$$L_{\text{tail}} = \sum_{j>k} \frac{\rho_j^2}{\sigma_j^2} \asymp \sum_{j>k} j^{2(\alpha-\beta)}. \tag{24}$$

We now apply the integral comparison test. The series $\sum_{j=1}^{\infty} j^{2(\alpha-\beta)}$ converges if and only if the exponent satisfies $2(\alpha-\beta) < -1$, that is, $\beta > \alpha + \frac{1}{2}$.

Case 1: $\beta > \alpha + \frac{1}{2}$. Then $2(\alpha - \beta) < -1$, so the series $\sum_j j^{2(\alpha-\beta)}$ converges. By the comparison bounds from Assumption 3.3, $\mathrm{tr}(\mathbf{M}) = \sum_j \rho_j^2/\sigma_j^2$ is bounded above by $c_2^{-2} c_1^2 \sum_j j^{2(\alpha-\beta)} < \infty$.

Case 2: $\beta < \alpha + \frac{1}{2}$. Then $2(\alpha - \beta) > -1$, and the series diverges. By the lower bound in Assumption 3.3, $\mathrm{tr}(\mathbf{M}) \geq c_2^2 c_1^{-2} \sum_j j^{2(\alpha-\beta)} = \infty$.

Case 3: $\beta = \alpha + \frac{1}{2}$. Then $2(\alpha - \beta) = -1$, and the series becomes the harmonic series:

$$\sum_{j=1}^{J} j^{-1} = \Theta(\log J). \tag{25}$$

Therefore $\sum_{j=1}^{J} \rho_j^2/\sigma_j^2 = \Theta(\log J)$, confirming logarithmic divergence.

This completes the proof of the trichotomy. □

Remark 3.5. The key distinction from the single-parameter taxonomy of Mallinar et al. [7] is the following. In that framework, both the noise absorption and noise expression are controlled by the same eigenvalue sequence $\{\lambda_j\}$ of the kernel, leading to a condition that depends only on the decay rate of $\{\lambda_j\}$. Our condition $\beta > \alpha + \frac{1}{2}$ involves *two independent* rates: α controls how efficiently the interpolation absorbs noise (via σ_j), and β controls how visible the noise is at test points (via ρ_j). The condition $\beta > \alpha + \frac{1}{2}$ captures the requirement that noise directions must be *invisible at test points*, not merely that noise be small in magnitude. After feature learning, β can increase (the learned representation suppresses noise visibility at test points) while α remains fixed, potentially converting catastrophic overfitting into benign overfitting—a transition that cannot be explained by any single-parameter criterion.

4. Application to Teacher–Student Networks

We demonstrate the utility of the (α, β) framework by analyzing a concrete setting: two-layer ReLU teacher–student networks trained by gradient flow. The following result shows that gradient flow achieves spectral separation in the Jacobian, which in turn guarantees benign generalization through our Theorem 3.1.

4.1. Setting.

Consider the following teacher–student configuration.

- **Teacher network.** The ground-truth function is a two-layer ReLU network with k hidden neurons:

$$f^*(\mathbf{x}) = \sum_{j=1}^{k} a_j^* \, \mathrm{ReLU}(\mathbf{w}_j^{*\top} \mathbf{x}), \tag{26}$$

where $\mathbf{w}_1^*, \ldots, \mathbf{w}_k^* \in \mathbb{R}^d$ are orthonormal vectors and $a_1^*, \ldots, a_k^* \in \mathbb{R}$ are the second-layer weights.

- **Student network.** The student is a two-layer ReLU network with $p \gg k$ hidden neurons:

$$f_\theta(\mathbf{x}) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} a_j \ \mathrm{ReLU}(\mathbf{w}_j^\top \mathbf{x}), \tag{27}$$

where $\theta = (a_1, \mathbf{w}_1, \ldots, a_p, \mathbf{w}_p)$ collects all trainable parameters.

- **Training.** The student is trained by population gradient flow (gradient flow on the expected loss) starting from an initialization scale $\alpha_{\text{init}} \to 0$ (the "rich" or "feature-learning" regime). The training inputs $\mathbf{x}_i$ are drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

4.2. **Spectral separation result.**

Proposition 4.1 (Jacobian Spectral Separation)**.** *In the setting described above, suppose n is sufficiently large relative to k and d, and $p \gg k(d+1)$. Then at a global minimizer θ^* of the population risk reached by gradient flow, the following spectral separation holds for the training Jacobian $\mathbf{J} = \mathbf{J}(\theta^*)$:*

1. *For $i \le k(d+1)$: $\sigma_i(\mathbf{J}) \ge c\sqrt{n}$, where $c > 0$ is a constant depending on the teacher weights.*
2. *For $i > k(d+1)$: $\sigma_i(\mathbf{J}) \le C\alpha_{\text{init}}$, where $C > 0$ is an absolute constant.*

Consequently, the effective rank satisfies $r_{\text{eff}} \le k(d+1)$, and

$$\mathrm{MSE}_{\text{test}} \le O\left(\sqrt{\frac{k\,d\,\log n}{n}}\right). \tag{28}$$

Proof outline. The argument combines results from Boursier and Flammarion [3] with our Theorem 3.1. We proceed in four steps.

Step 1: Gradient flow alignment (cited result). By the main theorem of Boursier and Flammarion [3], in the rich regime ($\alpha_{\text{init}} \to 0$), population gradient flow drives the student network to a k-sparse aligned solution: at the global minimizer, exactly k student neurons have their first-layer weights aligned with the teacher directions $\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*$, with second-layer weights matching $a_1^*, \ldots, a_k^*$. The remaining $p - k$ neurons have weights of order $O(\alpha_{\text{init}})$.

Step 2: Active block of the Jacobian. The k aligned neurons contribute an "active block" $\mathbf{J}_{\text{act}}$ to the training Jacobian. For each aligned neuron j ($j = 1, \ldots, k$), the gradient $\nabla_{\mathbf{w}_j} f_{\theta^*}(\mathbf{x}_i)$ involves the indicator $\mathbf{1}[\mathbf{w}_j^{*\top}\mathbf{x}_i > 0]$ and the input $\mathbf{x}_i$. Since $\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*$ are orthonormal and

the inputs are Gaussian, by a general-position argument (the probability that any collection of n Gaussian vectors lies in a degenerate configuration with respect to k hyperplanes is zero), the active block has rank $k(d+1)$: each neuron contributes d degrees of freedom from the first-layer weight gradient and 1 from the second-layer weight, and the k neurons contribute independently due to orthonormality. For n sufficiently large relative to $k(d+1)$, the minimum singular value of $\mathbf{J}_{\text{act}}$ is at least $c\sqrt{n}$ by standard random matrix concentration results for Gaussian designs.

Step 3: Dead block bound. The remaining $p-k$ neurons have weights of order $O(\alpha_{\text{init}})$ by Step 1. Their contribution to the Jacobian, denoted $\mathbf{J}_{\text{dead}}$, satisfies $\|\mathbf{J}_{\text{dead}}\|_{\text{op}} = O(\alpha_{\text{init}})$. Each dead neuron has weights of order $O(\alpha_{\text{init}})$, and the operator norm of its rank-$(d+1)$ Jacobian block scales linearly with α_{init}. This means the dead neurons contribute negligibly to the Jacobian spectrum.

Step 4: Spectral separation via Weyl's inequality. The full Jacobian is $\mathbf{J} = \mathbf{J}_{\text{act}} + \mathbf{J}_{\text{dead}}$. By Weyl's inequality for singular values,

$$\sigma_i(\mathbf{J}) \geq \sigma_i(\mathbf{J}_{\text{act}}) - \|\mathbf{J}_{\text{dead}}\|_{\text{op}}\,. \tag{29}$$

For $i \leq k(d+1)$: $\sigma_i(\mathbf{J}) \geq c\sqrt{n} - O(\alpha_{\text{init}}) \geq c\sqrt{n}/2$ for α_{init} sufficiently small. For $i > k(d+1)$: $\sigma_i(\mathbf{J}_{\text{act}}) = 0$ (since $\mathbf{J}_{\text{act}}$ has rank at most $k(d+1)$), so $\sigma_i(\mathbf{J}) \leq \|\mathbf{J}_{\text{dead}}\|_{\text{op}} = O(\alpha_{\text{init}})$ by the reverse Weyl inequality.

This establishes the spectral gap. The generalization bound (28) follows from Theorem 3.1: with effective rank $r_{\text{eff}} \leq k(d+1)$, the trace of $\mathbf{M}$ is dominated by the $k(d+1)$ large singular values, giving $\text{tr}(\mathbf{M}) \leq k(d+1) \cdot O(1/n)$, and the standard $\sqrt{\log n/n}$ rate follows from the concentration term. □

Remark 4.2. This result illustrates the mechanism described in Remark 3.5. Before feature learning (i.e., at initialization with a random kernel), the singular values of $\mathbf{J}$ would not exhibit spectral separation: all n singular values would be of comparable magnitude, and $\text{tr}(\mathbf{M})$ would grow with n. After gradient-flow training, the learned features concentrate the Jacobian spectrum on $k(d+1)$ directions aligned with the teacher, effectively increasing β (the noise directions become invisible at test points) while α for the active directions remains unchanged. This is the (α, β) decoupling in action.

Remark 4.3. Proposition 4.1 is labeled as a proposition rather than a theorem because its proof relies on the gradient-flow alignment result of [3], which is established under specific conditions (population gradient flow, Gaussian inputs, orthonormal teacher weights, rich regime).

Extending this to finite-sample gradient descent with non-Gaussian inputs remains an open problem.

5. Discussion

5.1. **Connection to double descent.** The noise propagation operator $\mathbf{M}$ provides a natural lens through which to view the double-descent phenomenon [4]. Consider least-squares regression with n samples and p features, and let $\gamma = p/n$ denote the overparameterization ratio.

The trace $\operatorname{tr}(\mathbf{M})$ is dominated by the smallest singular values of the Jacobian:

$$\operatorname{tr}(\mathbf{M}) = \sum_{j=1}^{n} \frac{\rho_j^2}{\sigma_j^2}, \tag{30}$$

and the sum is most sensitive to the indices j where σ_j is smallest.

In the underparameterized regime ($\gamma < 1$), the Marchenko–Pastur law for random design matrices predicts that the smallest singular value of $\mathbf{J}$ is of order $\sqrt{n}(1 - \sqrt{\gamma})$, and a calculation using the Marchenko–Pastur density (see, e.g., [4, Section 4]) yields

$$\frac{\operatorname{tr}(\mathbf{M})}{n} \to \frac{(1+\gamma)}{(1-\gamma)^3} \qquad \text{as } n, p \to \infty \text{ with } p/n \to \gamma < 1. \tag{31}$$

This expression diverges as $\gamma \to 1^-$, reflecting the peak of double descent.

At the interpolation threshold ($\gamma = 1$), the smallest singular value $\sigma_{\min}(\mathbf{J}) \to 0$ by the Marchenko–Pastur edge behavior, so $\operatorname{tr}(\mathbf{M}) \to \infty$. This is the exact point where the interpolating solution first exists, and it amplifies noise maximally.

In the overparameterized regime ($\gamma > 1$), the minimum-norm interpolant selects the solution with the smallest parameter norm. This implicit regularization improves the effective conditioning of the problem: the relevant singular values are those of the $n \times n$ matrix $\mathbf{J}\mathbf{J}^\top/p$, whose smallest eigenvalue is of order $(\sqrt{\gamma} - 1)^2$, bounded away from zero. Consequently, $\operatorname{tr}(\mathbf{M})$ decreases as γ increases beyond 1, completing the descent after the peak.

This analysis does not constitute a formal theorem, as it relies on asymptotic random matrix theory results applied in a specific (linear, Gaussian design) setting. Extending it to the nonlinear setting covered by Theorem 3.1 requires controlling $\mathbf{\Gamma}$ in addition to $\mathbf{\Sigma}$, which is an interesting direction for future work.

5.2. **Limitations.** We identify the following limitations of our framework, which we state explicitly to delineate the boundaries of our results.

(1) **Linearization residual is not quantified.** The bound in Theorem 3.1 includes the term δ_{lin}^2, but we do not provide an explicit estimate of δ_{lin} in terms of the network architecture or training procedure. Our framework applies rigorously only when local linearization is a valid approximation (e.g., in the NTK regime or for wide networks). Quantifying δ_{lin} for finite-width networks trained by gradient descent is a significant open problem.
(2) **Signal bias is not analyzed.** The term B_{signal}^2 in (11) is left as a given quantity. Our analysis focuses entirely on the noise amplification mechanism; understanding B_{signal}^2 requires a separate investigation of the approximation-theoretic properties of the learned features.
(3) **Population-level cross-correlation.** The matrix $\mathbf{\Gamma}$ is defined via a population expectation over test points (Definition 2.4). In practice, one observes only finite test samples, and bounding the deviation between the empirical and population versions of $\mathbf{\Gamma}$ requires matrix concentration inequalities (e.g., matrix Bernstein bounds). We defer this finite-sample analysis to future work.
(4) **Sub-Gaussian noise assumption.** The concentration step (Step 3 of the proof of Theorem 3.1) relies on the Hanson–Wright inequality for sub-Gaussian random vectors. Extending the framework to heavy-tailed noise distributions would require alternative concentration tools, such as truncation arguments or the recently developed heavy-tailed Hanson–Wright inequalities.
(5) **Scope of the teacher–student result.** Proposition 4.1 is established only for two-layer ReLU networks with orthonormal teacher weights, Gaussian inputs, and population gradient flow in the rich regime. Extending the spectral separation result to deeper architectures, non-Gaussian inputs, or finite-sample stochastic gradient descent remains open.

5.3. **Open problems.** We conclude by highlighting several directions for future investigation.

(1) **Quantifying δ_{lin} for finite-width networks.** Can one bound the linearization residual for networks of moderate width trained

by gradient descent, potentially using higher-order Taylor expansions or mean-field approximations?

(2) **Finite-sample $\mathbf{\Gamma}$ estimation.** What is the minimax rate for estimating $\mathrm{tr}(\mathbf{M})$ from a finite test sample, and can one construct practical estimators that predict the benign/tempered/catastrophic regime from data?

(3) **Feature-learning dynamics of (α, β).** Can one track the evolution of α and β during gradient-descent training, and characterize the conditions under which training drives the system from the catastrophic to the benign regime?

(4) **Beyond power-law decay.** The dichotomy in Theorem 3.4 assumes power-law decay. Extending the characterization to more general spectral profiles (e.g., exponential decay, mixed regimes) would broaden the applicability of the framework.

(5) **Heavy-tailed and dependent noise.** Extending Theorem 3.1 to accommodate heavy-tailed label noise or dependent noise structures (as arise in time-series or spatially correlated data) would significantly increase the practical relevance of the bounds.

Acknowledgements. The exploratory analysis and iterative refinement of the arguments in this paper were conducted with the assistance of Claude (Anthropic). All mathematical statements and proofs have been verified by the author.

References

[1] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[2] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[3] E. Boursier and N. Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35, 2022.

[4] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

[5] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[6] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[7] N. Mallinar, J. B. Simon, A. Abedsoltan, P. Pandit, M. Belkin, and P. Nakkiran. Benign, tempered, or catastrophic: toward a refined taxonomy of overfitting. *Advances in Neural Information Processing Systems*, 35, 2022.

[8] A. Rakhlin and X. Zhai. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory (COLT)*, pages 2595–2623, 2019.

[9] M. Rudelson and R. Vershynin. Hanson–Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[10] A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

Email address: Lightman.chang@gmail.com