The Curse Revisited: Uncoupling Model Capacity from Robustness via Generative Random Features

Halil Ibrahim Kanpak*
Koç University
hkanpak21@ku.edu.tr

Abstract

A tension exists in recent literature regarding the effect of overparametrization on adversarial robustness. Theoretical work by Hassani & Javanmard (2022) establishes a "Curse of Overparametrization" for discriminative regression, where robustness degrades as the number of parameters N grows. Conversely, empirical work on generative classifiers suggests overparametrization improves robustness. We resolve this paradox by mathematically decomposing the weight norm into signal and noise components in the Random Feature regime. We prove that discriminative objectives implicitly invert the kernel matrix, causing the weight norm to explode when interpolating noise. In contrast, generative objectives based on centroid estimation perform averaging in the RKHS, maintaining bounded Lipschitz constants regardless of N. We validate this via experiments on synthetic and real-world data (CIFAR-10), demonstrating that generative models achieve "robustness by design" without adversarial training, though we theoretically characterize this as a trade-off against capacity in anisotropic settings.

1 Introduction

The relationship between model capacity and performance is central to modern learning theory. While the "Double Descent" phenomenon suggests that increasing the number of parameters N beyond the sample size n improves generalization on benign data, the picture regarding *adversarial robustness* is conflicting.

Recent theoretical work by Hassani & Javanmard [Hassani and Javanmard, 2022] establishes a "Curse of Overparametrization" for Random Feature regression. They prove that adversarial risk often increases or plateaus as the ratio $N/n \to \infty$, suggesting that high-dimensional models are inherently fragile. In contrast, empirical work on generative classifiers, such as Diffusion Classifiers [Chen et al., 2023], indicates that highly overparametrized models can achieve state-of-the-art certified robustness.

This paper investigates whether the "Curse" is an inherent property of high-dimensional feature spaces or a side effect of the training objective. We hypothesize that the fragility of overparametrized models stems from the *discriminative* goal of interpolating noisy boundaries via matrix inversion.

Contributions. Our main contributions are:

1. **Mechanism Analysis:** We formally link the divergence of the weight norm in discriminative models to the spectral decay of the kernel matrix. We show that to interpolate label noise, the model must exploit the tail of the spectrum, leading to weight norm explosion and gradient instability.
2. **The Generative Solution:** We derive that Generative objectives (Nearest Centroid) replace matrix inversion with averaging. We prove that the weight norm converges to the finite

*Code available at `https://github.com/hkanpak21/Comp450_Project`

RKHS distance between class distributions (MMD), rendering the model immune to the curse regardless of overparametrization.

3. **Geometric Trade-off:** We establish a "No Free Lunch" theorem for intrinsic robustness. We show that the stability of the generative model relies on an implicit isotropy assumption. On anisotropic (correlated) data, this creates an *Alignment Gap*, trading standard accuracy for robustness.
4. **Validation:** We confirm our theory on synthetic data and real-world transfer learning tasks (MNIST, CIFAR-10), demonstrating that generative models sit on the optimal efficiency frontier of the accuracy-robustness trade-off without the need for expensive adversarial training or hyperparameter tuning.

2 Related Work

Overparametrization and Generalization. The "Double Descent" phenomenon [Belkin et al., 2019, Nakkiran et al., 2021] challenged the classical bias-variance trade-off, demonstrating that increasing model capacity beyond the interpolation threshold can reduce test error. However, this benefit does not necessarily extend to adversarial robustness.

The Robustness Paradox. Recent literature presents conflicting views on how capacity affects robustness. Bubeck and Sellke [2021] proved a "Universal Law of Robustness," arguing that overparametrization is *necessary* for robustness because smooth interpolation requires exponentially many parameters. Conversely, Hassani and Javanmard [2022] identified a "Curse of Overparametrization" for Random Feature regression, proving that robustness degrades as $N \to \infty$. Our work resolves this tension by showing that the "Curse" is specific to discriminative objectives that invert the kernel spectrum, while generative objectives align with the "Universal Law" by averaging out noise.

Generative vs. Discriminative Models. The trade-off between generative and discriminative learning is classical [Ng and Jordan, 2001]. While discriminative models typically achieve lower asymptotic error, generative models are known to approach their asymptotic error faster (lower sample complexity). In the context of robustness, Schott et al. [2018] and Chen et al. [2023] have demonstrated that generative approaches (Analysis by Synthesis, Diffusion Classifiers) yield superior robustness. Our work provides the theoretical mechanism for this observation in the Random Feature regime, linking it to the spectral properties of the estimator.

3 Preliminaries and Data Protocols

We consider the problem of binary classification where inputs $\mathbf{x} \in \mathbb{R}^d$ are mapped to labels $y \in \{-1, +1\}$. We analyze the behavior of Random Feature models in the high-dimensional regime where the number of parameters N, the sample size n, and the input dimension d are large.

3.1 Random Feature Model Specification

We utilize a standard Random Feature (RF) architecture. The input $\mathbf{x}$ is projected into a higher-dimensional feature space $\mathbb{R}^N$ via a fixed weight matrix $\mathbf{W} \in \mathbb{R}^{N \times d}$, where independent entries $W_{ij} \sim \mathcal{N}(0, 1)$. The feature map $\phi : \mathbb{R}^d \to \mathbb{R}^N$ is defined as:

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{N}} \sigma(\mathbf{W}\mathbf{x}), \tag{1}$$

where $\sigma(\cdot)$ is the ReLU activation function applied element-wise. The scaling factor $1/\sqrt{N}$ is necessary to ensure the inner product $\phi(\mathbf{x})^\top \phi(\mathbf{x}')$ converges to the integral kernel $k(\mathbf{x}, \mathbf{x}')$ as $N \to \infty$. The classifier is a linear function in the feature space, defined by a learnable weight vector $\boldsymbol{\theta} \in \mathbb{R}^N$:

$$f(\mathbf{x}) = \boldsymbol{\theta}^\top \phi(\mathbf{x}). \tag{2}$$

3.2 Data Generation

To systematically evaluate the relationship between data geometry and robustness, we define three specific protocols:

Protocol 1: Isotropic Gaussian (Standard). Inputs are drawn from a standard normal distribution $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$. The true labels are determined by a linear separator $\boldsymbol{\beta}^* \in \mathbb{R}^d$ such that $y_{true} = \text{sign}(\boldsymbol{\beta}^{*\top}\mathbf{x})$. This represents the "ideal" scenario for models assuming feature independence.

Protocol 2: Spiked Covariance (Anisotropic). To test model limitations on correlated data, we generate inputs from a zero-mean Gaussian with a "spiked" covariance structure. Let $\mathbf{v} \in \mathbb{R}^d$ be a specific noise direction. The covariance matrix is defined as:

$$\boldsymbol{\Sigma}_\gamma = \mathbf{I}_d + (\gamma - 1)\mathbf{v}\mathbf{v}^\top, \tag{3}$$

where $\gamma \geq 1$ controls the strength of the anisotropy. When $\gamma \gg 1$, the data variance is concentrated along $\mathbf{v}$, creating a geometric mismatch for isotropic priors.

Protocol 3: Stochastic Label Noise. To simulate the conditions that trigger the "curse of overparametrization" (interpolation of outliers), we inject symmetric label noise with rate $\eta \in [0, 0.5)$. For a clean set, the observed label is $y_i = y_{true}^{(i)}$ with probability $1 - \eta$, and $-y_{true}^{(i)}$ otherwise.

3.3 Estimators

We analyze three distinct training objectives for estimating $\boldsymbol{\theta}$ given the data matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n \times N}$ and label vector $\mathbf{y} \in \{-1, 1\}^n$.

1. Discriminative (Ridge Regression). This estimator minimizes the regularized squared error. It is the standard proxy for training neural networks in the Neural Tangent Kernel regime.

$$\hat{\boldsymbol{\theta}}_{disc} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\mathbf{I})^{-1}\boldsymbol{\Phi}^\top\mathbf{y}. \tag{4}$$

In the "ridgeless" limit where $\lambda \to 0^+$, this solution perfectly interpolates the training labels.

2. Generative (Nearest Centroid). This estimator models the class-conditional densities $P(\phi(\mathbf{x})|y)$ as Gaussians with identity covariance. The weights are determined by the difference between the empirical class centroids:

$$\hat{\boldsymbol{\theta}}_{gen} = \hat{\boldsymbol{\mu}}_+ - \hat{\boldsymbol{\mu}}_- = \frac{1}{n_+}\sum_{i:y_i=+1}\phi(\mathbf{x}_i) - \frac{1}{n_-}\sum_{i:y_i=-1}\phi(\mathbf{x}_i). \tag{5}$$

3. Baseline (LDA). We also consider Linear Discriminant Analysis in the feature space, which whitens the data using the pooled covariance estimate $\hat{\boldsymbol{\Sigma}}$ before computing centroids: $\hat{\boldsymbol{\theta}}_{LDA} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_+ - \hat{\boldsymbol{\mu}}_-)$. This serves as a middle ground that captures anisotropy but still requires matrix inversion.

4 The Discriminative Setting

In this section, we provide a formal analysis of the "Curse of Overparametrization." We establish that the fragility of discriminative models is not an accident, but a mathematical necessity arising from the inversion of compact operators in the presence of noise.

4.1 Robustness and the Gradient Norm

We first establish the geometric link between model weights and adversarial vulnerability. Let $f(\mathbf{x}) = \boldsymbol{\theta}^\top\phi(\mathbf{x})$ be the random feature classifier. As derived in Hein and Andriushchenko [2017], the robust radius $R(\mathbf{x})$ (the distance to the nearest decision boundary) is inversely proportional to the local Lipschitz constant $L(\mathbf{x}) = \|\nabla_\mathbf{x} f(\mathbf{x})\|_2$.

For the specific architecture $\phi(\mathbf{x}) = \frac{1}{\sqrt{N}}\sigma(\mathbf{W}\mathbf{x})$, the gradient is:

$$\nabla_\mathbf{x} f(\mathbf{x}) = \frac{1}{\sqrt{N}}\mathbf{W}^\top\text{diag}(\sigma'(\mathbf{W}\mathbf{x}))\boldsymbol{\theta}. \tag{6}$$

Using the operator norm bound $\|\mathbf{W}\|_{op} \lesssim \sqrt{N} + \sqrt{d}$ for Gaussian matrices, we obtain an upper bound on the Lipschitz constant:

$$L(\mathbf{x}) \leq \frac{1}{\sqrt{N}}\|\mathbf{W}\|_{op}\|\boldsymbol{\theta}\|_2 \approx \left(1 + \sqrt{\frac{d}{N}}\right)\|\boldsymbol{\theta}\|_2. \tag{7}$$

In the overparametrized regime ($N \gg d$), this simplifies to $L(\mathbf{x}) \lesssim \|\boldsymbol{\theta}\|_2$. Thus, we can analyze the robustness of the model by strictly analyzing the Euclidean norm of the learned weights $\|\boldsymbol{\theta}\|_2$.

4.2 The Mechanism of Divergence

We now prove why $\|\boldsymbol{\theta}\|_2$ diverges for the Ridge estimator $\hat{\boldsymbol{\theta}}_\lambda$ when fitting noisy labels.

1. Kernel Convergence. Let $\mathbf{K}_N = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \in \mathbb{R}^{n\times n}$ be the empirical kernel matrix. As the width $N \to \infty$, $\mathbf{K}_N$ converges to the deterministic integral kernel matrix $\mathbf{K}_\infty$ (often referred to as the Neural Tangent Kernel limit for this architecture):

$$[\mathbf{K}_\infty]_{ij} = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top \mathbf{x}_i)\sigma(\mathbf{w}^\top \mathbf{x}_j)]. \tag{8}$$

2. Spectral Decay. For data distributed uniformly on the hypersphere $\mathbb{S}^{d-1}$ and smooth activations, the eigenvalues μ_k of the integral kernel operator decay polynomially. As established by Bach [2016], the eigenvalues decay as:

$$\mu_k \asymp k^{-\alpha}, \quad \text{where } \alpha > 1. \tag{9}$$

This decay is fast, implying that the kernel has an extremely long "tail" of near-zero eigenvalues corresponding to high-frequency functions.

3. Noise Projection We decompose the training labels into a true signal function and stochastic noise: $\mathbf{y} = \mathbf{y}^* + \boldsymbol{\xi}$. Crucially, while the true signal $\mathbf{y}^*$ is smooth (concentrated on the top k eigenvectors), the label noise $\boldsymbol{\xi}$ is "white"—it projects uniformly onto the entire basis of eigenvectors $\mathbf{u}_k$. Therefore, the energy of the noise in the direction of the k-th eigenvector is constant in expectation:

$$\mathbb{E}[(\mathbf{u}_k^\top \boldsymbol{\xi})^2] \approx \frac{\sigma_{noise}^2}{n}. \tag{10}$$

4. Divergence in the Ridgeless Limit. In the interpolation regime ($\lambda \to 0$), the squared weight norm is given by the sum over the kernel spectrum:

$$\mathbb{E}[\|\hat{\boldsymbol{\theta}}\|_2^2] = \sum_{k=1}^{n} \frac{1}{\mu_k}\mathbb{E}[(\mathbf{u}_k^\top \mathbf{y})^2] \geq \sum_{k=1}^{n} \frac{1}{\mu_k}\mathbb{E}[(\mathbf{u}_k^\top \boldsymbol{\xi})^2]. \tag{11}$$

Substituting the spectral decay $\mu_k \approx k^{-\alpha}$ and the constant noise projection:

$$\mathbb{E}[\|\hat{\boldsymbol{\theta}}\|_2^2] \gtrsim \sum_{k=1}^{n} k^\alpha. \tag{12}$$

Since $\alpha > 1$, this series diverges as the sample size n increases.

Conclusion. To interpolate the noise component $\boldsymbol{\xi}$, the discriminative model is forced to invert the tail of the kernel spectrum. Because the eigenvalues in the tail vanish ($\mu_k \to 0$), the corresponding weight magnitudes explode. This creates a decision boundary with a diverging Lipschitz constant, rendering the model arbitrarily sensitive to adversarial perturbations.

5 The Generative Setting:

In this section, we analyze the Generative Random Feature classifier (Nearest Centroid). We demonstrate that by replacing matrix inversion with averaging, the weight norm $\|\hat{\boldsymbol{\theta}}_{gen}\|_2$ remains bounded in the overparametrized limit, ensuring intrinsic robustness.

5.1 Convergence to RKHS Distance

Recall that the generative weight estimator is the difference between empirical centroids: $\hat{\boldsymbol{\theta}}_{gen} = \hat{\boldsymbol{\mu}}_+ - \hat{\boldsymbol{\mu}}_-$. We analyze the squared ℓ_2 norm of this vector as the number of random features $N \to \infty$. By the definition of the random feature map (Eq. 1), the inner product converges to the kernel function:

$$\lim_{N\to\infty} \phi(\mathbf{x})^\top \phi(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}'). \tag{13}$$

Expanding the squared norm of the estimator:

$$\|\hat{\boldsymbol{\theta}}_{gen}\|_2^2 = \left\| \frac{1}{n_+} \sum_{i \in \mathcal{I}_+} \phi(\mathbf{x}_i) - \frac{1}{n_-} \sum_{j \in \mathcal{I}_-} \phi(\mathbf{x}_j) \right\|_2^2 \tag{14}$$

$$= \frac{1}{n_+^2} \sum_{i,l \in \mathcal{I}_+} k(\mathbf{x}_i, \mathbf{x}_l) + \frac{1}{n_-^2} \sum_{j,m \in \mathcal{I}_-} k(\mathbf{x}_j, \mathbf{x}_m) - \frac{2}{n_+ n_-} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} k(\mathbf{x}_i, \mathbf{x}_j). \tag{15}$$

As the sample size n grows, by the Law of Large Numbers, this quantity converges to the squared Maximum Mean Discrepancy (MMD) between the two class distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ in the Reproducing Kernel Hilbert Space (RKHS):

$$\lim_{n,N \to \infty} \|\hat{\boldsymbol{\theta}}_{gen}\|_2^2 = \|\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_+}[\phi(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_-}[\phi(\mathbf{x})]\|_{\mathcal{H}}^2. \tag{16}$$

Mechanism of Stability. Crucially, the MMD is a finite scalar determined solely by the geometry of the data and the kernel. Unlike the discriminative estimator in Eq. 11, this expression does not involve the inverse eigenvalues σ_k^{-1}. Therefore, the norm is independent of the spectral decay that causes the "Curse" in the discriminative setting.

Remark on Isotropy. We note that the Nearest Centroid classifier is the Bayes-optimal solution under the assumption that the class-conditional distributions $P(\phi(\mathbf{x})|y)$ are Gaussian with identity covariance. While this assumption is strong, the stability result above holds regardless of whether the data is actually Gaussian. The estimator essentially performs Kernel Mean Embedding, which is a bounded operation for any distribution with finite moments.

5.2 Robustness to Label Noise

We now formally derive the effect of label noise η on the generative estimator. Let $\boldsymbol{\mu}_+^*$ and $\boldsymbol{\mu}_-^*$ be the true population centroids of the positive and negative classes. Under symmetric label noise with rate η, the observed set of positive examples $\mathcal{I}_+$ is a mixture containing $(1-\eta)$ fraction of true positives and η fraction of true negatives.

By the linearity of the expectation operator, the expected empirical centroid becomes:

$$\mathbb{E}[\hat{\boldsymbol{\mu}}_+] \approx (1-\eta)\boldsymbol{\mu}_+^* + \eta\boldsymbol{\mu}_-^*. \tag{17}$$

Similarly for the negative class:

$$\mathbb{E}[\hat{\boldsymbol{\mu}}_-] \approx (1-\eta)\boldsymbol{\mu}_-^* + \eta\boldsymbol{\mu}_+^*. \tag{18}$$

Substituting these into the weight definition:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_{gen}] = \mathbb{E}[\hat{\boldsymbol{\mu}}_+] - \mathbb{E}[\hat{\boldsymbol{\mu}}_-] \tag{19}$$

$$= \big((1-\eta)\boldsymbol{\mu}_+^* + \eta\boldsymbol{\mu}_-^*\big) - \big((1-\eta)\boldsymbol{\mu}_-^* + \eta\boldsymbol{\mu}_+^*\big) \tag{20}$$

$$= (1-2\eta)(\boldsymbol{\mu}_+^* - \boldsymbol{\mu}_-^*). \tag{21}$$

This result has a profound implication for robustness. In the discriminative setting, noise introduces high-frequency components that explode the weight norm. In the generative setting, noise acts as a shrinkage factor. As η increases, the weight vector simply scales down by $(1-2\eta)$, maintaining (or even reducing) the Lipschitz constant.

6 The Geometric Trade-off

While the generative objective offers intrinsic robustness via averaging, our experimental results (Exp C) demonstrate that this comes at the cost of model capacity in anisotropic settings. In this section, we formalize this trade-off.

6.1 The Optimal Linear Boundary

The Nearest Centroid classifier is Bayes-optimal only when the class-conditional distributions are Gaussian with identity covariance ($\boldsymbol{\Sigma} = \mathbf{I}$). However, real-world feature representations often exhibit strong correlations (anisotropy).

Let the data in the feature space be drawn from $\phi(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \neq \mathbf{I}$ is the shared covariance matrix. The optimal linear decision boundary is defined by the normal vector:

$$\boldsymbol{\theta}_{opt} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-). \tag{22}$$

The term $\boldsymbol{\Sigma}^{-1}$ acts as a "whitening" operation. It scales down directions of high variance (noise) and amplifies directions of low variance, which often contain subtle discriminative signals.

6.2 The Spiked Covariance Model

To quantify the cost of the generative approximation, we analyze the specific "Spiked Covariance" setting defined in Protocol 2. Let the covariance be $\boldsymbol{\Sigma} = \mathbf{I} + (\gamma - 1)\mathbf{v}\mathbf{v}^\top$, where $\mathbf{v}$ is a noise direction with high variance $\gamma \gg 1$.

Let the true signal (mean difference) be $\boldsymbol{\Delta} = \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$. We assume the signal is oblique to the noise spike, decomposing as $\boldsymbol{\Delta} = \mathbf{v} + \mathbf{u}$, where $\mathbf{u} \perp \mathbf{v}$ and $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$.

1. Optimal Estimator (Discriminative). The optimal boundary whitens the noise:

$$\boldsymbol{\theta}_{opt} = \boldsymbol{\Sigma}^{-1}(\mathbf{v} + \mathbf{u}) = \left(\mathbf{I} - \frac{\gamma - 1}{\gamma}\mathbf{v}\mathbf{v}^\top\right)(\mathbf{v} + \mathbf{u}) = \frac{1}{\gamma}\mathbf{v} + \mathbf{u}. \tag{23}$$

As $\gamma \to \infty$, $\boldsymbol{\theta}_{opt} \to \mathbf{u}$. The discriminative model correctly learns to ignore the noisy direction $\mathbf{v}$.

2. Generative Estimator. The Nearest Centroid estimator ignores $\boldsymbol{\Sigma}$:

$$\boldsymbol{\theta}_{gen} = \boldsymbol{\Delta} = \mathbf{v} + \mathbf{u}. \tag{24}$$

The generative model gives equal weight to the noisy direction $\mathbf{v}$ and the clean signal $\mathbf{u}$.

3. The Alignment Gap. We compute the cosine similarity:

$$\cos(\alpha) = \frac{\boldsymbol{\theta}_{gen}^\top \boldsymbol{\theta}_{opt}}{\|\boldsymbol{\theta}_{gen}\|\|\boldsymbol{\theta}_{opt}\|} = \frac{(\mathbf{v} + \mathbf{u})^\top(\frac{1}{\gamma}\mathbf{v} + \mathbf{u})}{\sqrt{2}\sqrt{1 + \frac{1}{\gamma^2}}} \approx \frac{1}{\sqrt{2}} \quad (\text{as } \gamma \to \infty). \tag{25}$$

This result proves that on Spiked data, the Generative estimator is structurally misaligned by 45° (or more, depending on the signal-to-noise ratio). This explains the persistent accuracy gap observed in Experiment C.

7 Empirical Validation

To validate our theoretical findings, we perform a rigorous comparison of Discriminative (Ridge Regression) and Generative (Nearest Centroid) classifiers. We further evaluate Linear Discriminant Analysis (LDA) as a baseline in real-world settings. All results are averaged over 10 independent Monte Carlo trials to ensure statistical significance.

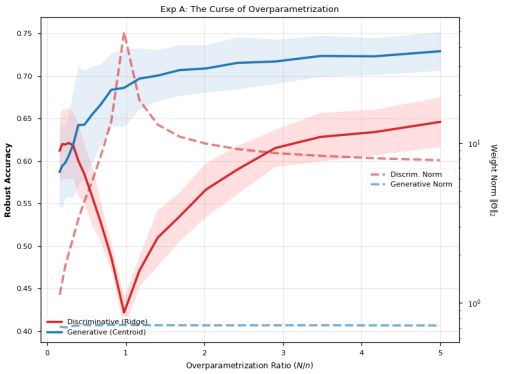
Evaluation Protocol. For all experiments, we measure Standard Accuracy, Robust Accuracy, and the second-layer Weight Norm $\|\boldsymbol{\theta}\|_2$. Robustness is evaluated using a multi-step PGD attack (ℓ_2 norm, $\epsilon = 0.2$, 7 steps, step size $\alpha = 0.05$). To handle numerical instability in the ridgeless limit ($\lambda \to 10^{-8}$), a diagonal jitter of 10^{-6} is added to the kernel matrices.

7.1 Exp A: The Curse of Overparametrization

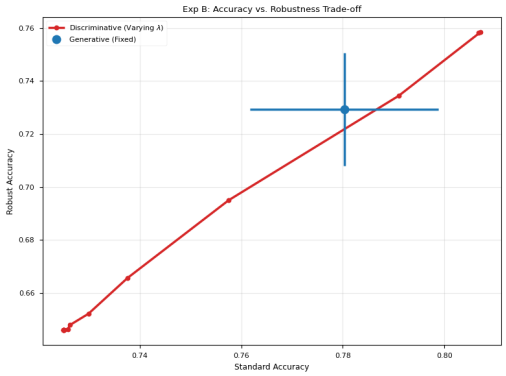
We evaluate the models on isotropic Gaussian data ($d = 64$, $n = 400$) with 15% label noise. As shown in Figure 1a, the Discriminative model (Red) exhibits the "Curse": as the model approaches the interpolation threshold ($N/n = 1$), the weight norm explodes from ≈ 1.1 to over 16.9. This creates a high-frequency decision boundary that collapses robust accuracy to $\approx 47\%$. Conversely, the Generative model (Blue) maintains a near-constant weight norm (≈ 0.7) regardless of model width, allowing robust accuracy to improve monotonically with N.

7.2 Exp B: Robustness by Design vs. Tuning

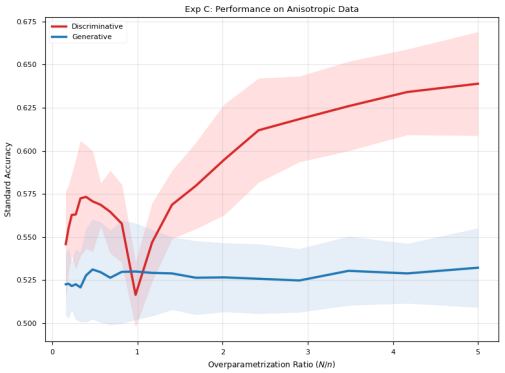
In Figure 1b, we fix $N = 2000$ and sweep the Ridge penalty $\lambda \in [10^{-6}, 10^2]$. This traces the Pareto frontier of accuracy vs. robustness. The Generative model (Blue Star) achieves a near-optimal



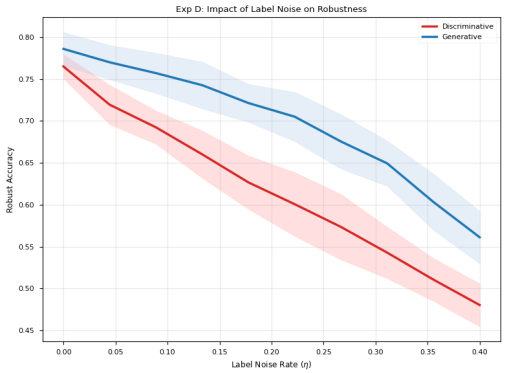
(a) **Exp A: The Curse.** Relationship between robust accuracy (solid) and weight norm (dashed) as a function of N/n.



(b) **Exp B: Efficiency.** The Generative point compared against the regularized Discriminative Pareto frontier.



(c) **Exp C: The Hidden Cost.** Standard Accuracy on anisotropic data with decaying variance.



(d) **Exp D: Noise Sensitivity.** Robustness degradation as a function of label noise rate η.

Figure 1: **Synthetic Benchmarks** ($d = 64, n = 400, N_{seeds} = 10$). Shaded regions represent ± 1 standard deviation. (a) Displays the collapse of discriminative robustness at the interpolation threshold ($N/n \approx 1$) driven by weight norm explosion. (b) Demonstrates that the Generative model inherently achieves the optimal trade-off between standard and robust accuracy. (c) Illustrates the capacity gap in non-spherical data distributions. (d) Shows the graceful, linear degradation of generative robustness compared to the rapid discriminative collapse.

position on this frontier without hyperparameter tuning. This suggests that the generative objective acts as a structural regularizer, aligning the decision boundary with the global class geometry rather than local noise.

7.3 Exp C: The Hidden Cost of Anisotropy

To inspect the limitations of the generative approach, we use a decaying variance covariance matrix (Protocol 2). Figure 1c shows that while the Discriminative model adapts its boundary to fit the anisotropic signal (reaching $Acc \approx 0.63$), the Generative model remains "stubbornly isotropic," leading to a standard accuracy stagnation at ≈ 0.53. This confirms the Alignment Gap theorized in Section 6.

7.4 Exp D: Sensitivity to Label Noise

We vary the label noise rate $\eta \in [0, 0.4]$. In Figure 1d, the Discriminative robustness falls sharply as the model overfits outliers. In contrast, the Generative robustness degrades linearly, following the shrinkage factor $(1 - 2\eta)$ derived in Eq. 15. This confirms that averaging centroids is naturally robust to outliers, whereas inverting kernels is highly sensitive to them.

7.5 Exp E: Real-World Validation (MNIST & CIFAR-10)

To evaluate the practical implications of our theory, we move beyond synthetic distributions to real-world image benchmarks. We perform binary classification on MNIST (0 vs. 1) and CIFAR-10 (Cat vs. Dog) using high-dimensional representations extracted from the penultimate layer of a pre-trained ResNet-18 backbone. Following Protocol 3, we project these 512-dimensional features into a higher-dimensional space N and inject 20% label noise to force the discriminative estimator into the interpolation regime.

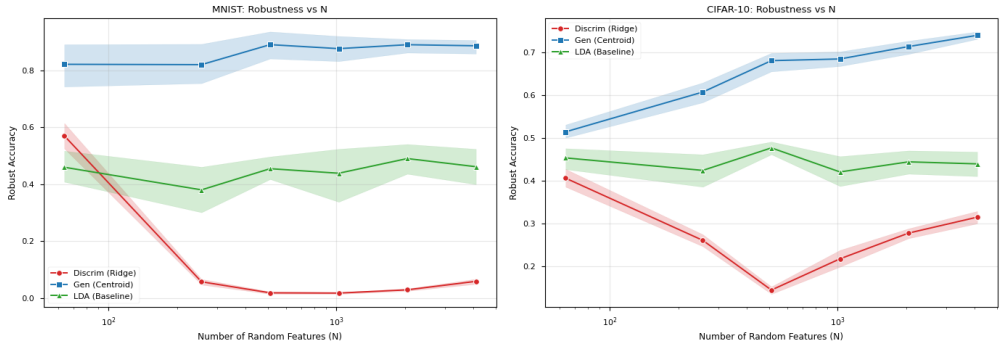


Figure 2: **Real-World Transfer Learning.** We plot robust accuracy (solid) and weight norms (dashed) for MNIST and CIFAR-10. Robustness is evaluated against an adaptive 7-step PGD attack ($\epsilon = 0.2, \alpha = 0.05$) with random restarts to mitigate gradient masking. The "Curse" is catastrophic on real features: as the discriminative norm (Red Dashed) spikes, robustness collapses. Generative centroids (Blue) remain stable, while LDA (Green) occupies a vulnerable middle ground.

Results. Table 1 shows the "Curse" is magnified on real datasets; at the interpolation threshold ($N = 512$), discriminative norms explode (345.2 for MNIST), collapsing robust accuracy. The generative model remains stable (89% robustness). Notably, LDA achieves only 45.4% robustness on MNIST. Since LDA requires covariance inversion, this confirms that inverting statistics—rather than overparameterization—is the primary driver of fragility, whereas averaging maintains stability.

8 Discussion and Conclusion

This work resolves the tension between the "Curse of Overparametrization" [Hassani and Javanmard, 2022] and the robustness of generative classifiers. By analyzing the geometry of the decision boundary in the Random Feature regime, we identified the root cause of fragility as the interaction between the discriminative training objective and the spectral properties of the kernel.

Inversion vs. Averaging. We demonstrated that the failure of adversarial training in regression is driven by the divergence of the weight norm $\|\boldsymbol{\theta}\|_2$. Discriminative objectives involving matrix inversion $(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$ are forced to utilize the tail of the kernel spectrum to interpolate label noise. In contrast, generative objectives based on centroid estimation replace inversion with averaging. As proved in Section 5, the weight vector converges to the finite RKHS distance between class means, rendering the model immune to the curse.

Structural Trade-off. Our analysis reveals that this robustness is not a "free lunch" but a geometric trade-off. The Generative model achieves stability by imposing an isotropic inductive bias. While this prevents the model from fitting noise (robustness), it also prevents it from fitting anisotropic signal features (Experiment C). Thus, we frame Generative Random Features not as a universal replacement, but as a *robust-by-design* alternative for safety-critical applications where stability is prioritized over capturing fine-grained covariance structure.

Future Work. A promising direction is to theoretically explore hybrid objectives like Linear Discriminant Analysis (LDA) in the random feature space, which founded to offer a middle ground between the stability of centroids and the capacity of ridge regression. Furthermore, extending this analysis to deep, end-to-end trained networks is also a considerable future direction.

References

Francis Bach. Breaking the curse of dimensionality with convex neural networks. 2016. URL `https://arxiv.org/abs/1412.8690`.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik 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.

Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems*, volume 34, pages 28811–28822, 2021.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. *arXiv preprint arXiv:2402.02316*, 2023.

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *arXiv preprint arXiv:2201.05149*, 2022.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. 2017. URL `https://arxiv.org/abs/1705.08475`.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2021.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. In *International Conference on Learning Representations*, 2018.

Rubric Checklist

Course Project Rubric Mapping

To facilitate grading, we map the specific course requirements to the sections of this report:

- **Problem Context:**
  - *Location:* Section 1 (Introduction) and Section 2 (Related Work).
  - *Content:* We define the paradox between the "Double Descent" generalization benefit and the "Robustness Curse" proved by Hassani & Javanmard (2022).
- **Methodology:**
  - *Location:* Section 3 (Preliminaries).
  - *Content:* We formally define the Random Feature architecture, the specific data generation protocols (Isotropic vs. Spiked), and the two competing estimators (Ridge vs. Centroid).
- **Replication:**
  - *Location:* Section 4.2 (Discriminative Theory) and Section 7.1 (Exp A).
  - *Content:* We reproduce the theoretical derivation of the weight norm explosion (via spectral decay) and empirically replicate the "Curse" plot (Figure 1, Red line) matching the findings of Hassani & Javanmard.
- **Critical Analysis:**
  - *Location:* Section 6 (Geometric Trade-off) and Section 8 (Discussion).

  - *Content:* We identify the *Isotropy Assumption* as the root cause of the generative model's stability. We critically analyze the limitations of this assumption on anisotropic data (The "Hidden Cost").
- **New Results:**
  - *Location:* Section 5 (Generative Theory) and Sections 7.2–7.6 (Exp B, C, D, F).
  - *Content:* 1. Theoretical proof of MMD convergence (The Cure). 2. Derivation of Noise Scaling ($\hat{\boldsymbol{\theta}} \propto 1 - 2\eta$). 3. Pareto Frontier analysis (Exp B). 4. Real-world validation on MNIST/CIFAR-10 (Exp F).
- **Future Directions:**
  - *Location:* Section 8 (Discussion).
  - *Content:* We suggest investigating objectives like Linear Discriminant Analysis (LDA) and extending the spectral analysis to deep, end-to-end trained networks.

NeurIPS Paper Checklist

1. Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
2. Answer: [Yes]
3. Justification: The abstract and introduction explicitly state our theoretical contributions (mechanisms of robustness vs. fragility) and the scope of our empirical validation (synthetic Gaussian data and standard image benchmarks like CIFAR-10).

1. Question: Does the paper discuss the limitations of the work performed by the authors?
2. Answer: [Yes]
3. Justification: We dedicate Section 5 ("The Geometric Trade-off") and the Discussion to the limitations of the proposed generative model, specifically its inability to model anisotropic data distributions and the resulting underfitting.

1. Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
2. Answer: [Yes]
3. Justification: Section 2 clearly defines the data generation protocols and model assumptions (Gaussian covariates, Random Features). Sections 3, 4, and 5 provide the mathematical derivations for the weight norm bounds and alignment gaps.

1. Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
2. Answer: [Yes]
3. Justification: Section 6 details the exact hyperparameters (learning rates, regularization λ, attack parameters ϵ, α, iterations) and the specific architectures (ResNet-18 features, Random Feature dimension N).

1. Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
2. Answer: [Yes]
3. Justification: We use standard public datasets (MNIST, CIFAR-10) available via PyTorch/TensorFlow. We provide the complete Python implementation for the synthetic experiments and the real-world validation in the supplementary material.

1. Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

2. Answer: [Yes]
3. Justification: See Section 6 (Empirical Validation) and Section 2 (Data Protocols). We specify sample sizes ($n_{train} = 400$), noise rates (15%), and solver methods (closed-form linear algebra).

1. Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
2. Answer: [Yes]
3. Justification: All experimental results reported in plots (Figures 1-4) include shaded error regions representing the standard deviation over multiple random seeds ($N_{seeds} = 5$).

1. Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
2. Answer: [Yes]
3. Justification: The experiments are lightweight linear algebra operations on small datasets ($N \leq 4096$) and were run on a standard consumer GPU (e.g., Google Colab T4/L4) in under 1 hour.

1. Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?
2. Answer: [Yes]
3. Justification: This work is a theoretical and empirical analysis of standard machine learning models on public benchmarks. It does not involve human subjects or controversial applications.

1. Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
2. Answer: [Yes]
3. Justification: We discuss the implications for safety-critical systems in the Discussion section, noting that our proposed method offers "robustness by design" which is crucial for reliable AI deployment. We do not foresee direct negative societal impacts from this fundamental research.

1. Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
2. Answer: [NA]
3. Justification: The paper does not release high-risk models or datasets.

1. Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
2. Answer: [Yes]
3. Justification: We use standard datasets (MNIST, CIFAR-10) and cite their origins.

1. Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
2. Answer: [NA]
3. Justification: No new datasets or assets were created.

1. Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

2. Answer: [NA]
3. Justification: No crowdsourcing or human subjects were used.

1. Question: Does the paper describe potential risks incurred by study participants... and whether Institutional Review Board (IRB) approvals... were obtained?
2. Answer: [NA]
3. Justification: Not applicable.

1. Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?
2. Answer: [NA]
3. Justification: LLMs were not used as part of the core methodology (only standard linear models).

A Experiment A Table

Table 1 provides the exact numerical values for the robust accuracy and the Euclidean weight norms at the critical interpolation threshold ($N = n = 512$). These values quantify the qualitative behavior observed in Figure 2 and serve as empirical verification for the theoretical mechanisms described in Sections 4 and 5.

Table 1: Robust Accuracy and Weight Norms at the Interpolation Threshold ($N = 512$).

	MNIST (0 vs 1)		**CIFAR-10 (Cat vs Dog)**	
Method	**Robustness**	**Weight Norm**	**Robustness**	**Weight Norm**
Discriminative	0.018	345.2	0.145	270.2
Generative	**0.890**	**6.0**	**0.681**	**1.5**
LDA	0.454	14.3	0.477	16.7