Double Descent under Structured Label Noise: A Theoretical & Empirical Study in Two−Layer Neural Networks

Visho Malla Oli
Department of Computer Science, University of Mississippi

November 9, 2025

Abstract

We investigate the double descent phenomenon in two-layer ReLU neural networks under structured label noise. Extending the recent work of Sang et al. (2025) [1] on double descent in noiseless two-layer models, we introduce label noise with specific structure (such as class-specific label flips and feature- correlated corruption) and examine its impact on generalization. We describe a teacher–student framework where a teacher model generates data for a binary classification task, and a two-layer ReLU network serves as the student. Structured label noise is injected by flipping labels in a targeted manner (e.g., only within one class or within a cluster of examples). We develop a theoretical heuristic explaining how such label noise alters the classic double descent curve, shifting the interpolation threshold and amplifying the test error peak. Empirically, we present Python simulation results showing test error as a function of model complexity (measured by the ratio $\alpha = n/d_{\text{model}}$ of training samples n to model dimension d_{model}). The experiments confirm that structured label noise increases the peak test error and can delay the onset of the second descent. We also demonstrate that adding regularization (e.g., L_2 weight decay) can mitigate the explosion in test error, yielding a smoother risk curve. Our findings provide insight into the interplay between overparameterization and noisy labels, emphasizing interpretability of double descent in practical settings with mislabeled data. We conclude with discussions on how structured noise induces spurious memorization and how careful modeling or regularization can maintain generalization.

1 Introduction

Modern machine learning models often operate in an over-parameterized regime, where the number of model parameters exceeds the number of training samples. In classical learning theory, increasing model complexity typically yields a U-shaped test error curve due to the bias–variance trade-off: test error decreases at first but eventually rises as the model begins to overfit the training data. Surprisingly, recent studies have observed a double descent pattern in test error that defies this classical U-curve [2]. In double descent, as model complexity grows beyond the point of exact interpolation of the training data, test error can decrease again, leading to a second descent phase. This phenomenon was first highlighted in modern deep learning contexts by Belkin et al. [2], and subsequently explored across various models and settings [3, 4]. Double descent suggests that highly over-parameterized models can generalize well despite having the capacity to perfectly fit (or even memorize) the training data.

In a recent work, Sang et al. [1] provided a theoretical analysis of double descent in a two-layer ReLU neural network for binary classification. Their study considered a teacher–student framework with a two- layer "student" network learning from data labeled by a ground-truth teacher function.

They demonstrated that test error as a function of the model's size exhibits a double descent behavior when plotted against the ratio $\alpha = n/d_{\text{model}}$ (number of training samples n to model dimension d_{model}). The interpolation threshold – roughly when $n \approx d_{\text{model}}$ – marked a peak in test error, separating a classical regime ($n \gg d_{\text{model}}$, under- parameterized) from an over-parameterized regime ($n \ll d_{\text{model}}$) where test error falls again [1]. Their analysis leveraged tools like the Convex Gaussian Min–Max Theorem to characterize the behavior of the empirical risk minimizer in high dimensions.

However, the analysis in [1] (and many other works on double descent) assumes clean training data – i.e., the training labels are generated exactly by the teacher model without corruption. In practical scenarios, datasets often contain label noise, where some training examples are mislabeled. Label noise can significantly degrade generalization performance and might interact with model overcapacity in non-trivial ways. Structured label noise refers to noise that is not purely random but exhibits a pattern – for example, only labels of a particular class are flipped, or labels in one region of the feature space are corrupted. Such structured noise is common in real datasets; for instance, one class of images might be systematically mislabeled due to ambiguity, or sensor failures might affect only a cluster of samples.

Our contribution in this paper is a theoretical and empirical study of how structured label noise influences the double descent behavior in two-layer neural networks. We build upon the framework of [1] by introducing structured label noise into the data generation process and investigating its effect on both the interpolation threshold and the shape of the test error curve. We provide a clear description of the two- layer model and the teacher data generation process (Section 3), then develop a heuristic theoretical analysis for the impact of label noise on generalization error (Section 4). In Section 5, we empirically simulate training of two-layer ReLU networks under various noise patterns and demonstrate that label noise tends to shift and magnify the double descent peak. We also examine the role of regularization in alleviating these effects. Our findings offer insight into why and how over-parameterized neural networks can still generalize in the presence of mislabeled data, which is an important question for understanding robust machine learning.

2 Related Work

The discovery of double descent in modern models has spurred a line of research aiming to understand this counter-intuitive phenomenon. Belkin et al. [2] first demonstrated double descent in simple settings (such as polynomial regression and random Fourier features) and in deep neural networks, reconciling it with classical theory by emphasizing the role of model overparameterization beyond the interpolation point. Subsequent works such as Nakkiran et al. [3] expanded this concept to deep double descent, examining how increasing either model size or training data size can cause multiple descent curves in test error.

On the theoretical side, several authors have analyzed double descent in high-dimensional models using tools from random matrix theory and statistical physics. For example, d'Ascoli et al. [4] studied bias–variance decompositions in over-parameterized linear models and neural networks (the "lazy" training regime), showing how variance spikes at the interpolation threshold while bias continues to decrease, leading to double descent. Deng et al. [5] presented an analytic model for double descent in high-dimensional binary classification using Gaussian mixtures, providing precise asymptotic expressions for test error as a function of the sample-to-dimension ratio. These analyses typically assume some form of noisiness in the data (either label noise or intrinsic noise in the model), since without any noise or mismatch, an over- parameterized model could achieve near-zero test error once it exactly learns the teacher function.

The effect of label noise on generalization has long been studied in the context of robust learning. In the double descent context, label noise is known to exacerbate the peak at the interpolation threshold. Indeed, in linear regression, it has been shown that with label noise, the mean-squared test error can blow up (in theory, approach infinity) when the number of parameters equals the number of data points (the interpolation point), if no regularization is applied [2, 4]. This is because the solution that exactly fits noisy data must necessarily fit the noise, causing extremely high variance in predictions. Prior works have mostly considered random label noise (each label independently flipped with some probability) to illustrate this effect. Far less explored is structured label noise, where only specific subsets of data are corrupted. Some empirical studies indicate that structured label noise (such as class-conditional noise) can lead to systematic generalization failures concentrated in certain classes. Understanding how over- parameterized models handle such structured noise is crucial, as it bridges theoretical analysis with realistic scenarios of dataset corruption.

Our work differentiates itself by focusing on two-layer ReLU neural networks under structured label noise, extending the rigorous analysis of Sang et al. [1]. We also examine how regularization can mitigate the double descent risk curve. Prior work by Nakkiran et al. [6] showed that applying an optimal amount of ridge regularization can remove the double descent peak, essentially reverting the test error curve back to a more traditional monotonic shape. We verify a similar phenomenon in our two-layer network experiments: even a small weight decay significantly reduces the harm caused by fitting noisy labels. Overall, our study contributes to the growing body of evidence [2, 3, 4, 5, 6] that double descent is a generic phenomenon of over- parameterized models, and provides new insights specific to the case of structured label noise in neural networks.

3 Methods: Two-Layer Model and Structured Label Noise

Two-Layer ReLU Network (Student Model). We consider a two-layer neural network with a ReLU activation function in the hidden layer, designed for binary classification. The network (which we refer to as the student) has an input dimension d, one hidden layer with H hidden units, and an output layer producing a scalar score used for classification. Mathematically, the model can be written as

$$f(x; W, a, b) \;=\; \sum_{j=1}^{H} a_j \, \sigma(w_j^\top x + b_j), \qquad \sigma(u) = \max\{0, u\}, \tag{1}$$

where $W = [w_1, \ldots, w_H] \in \mathbb{R}^{d \times H}$ are the hidden-layer weights, $b = (b_1, \ldots, b_H)$ are the hidden biases, $a = (a_1, \ldots, a_H)$ are the output-layer weights. The prediction is $\hat{y} = \operatorname{sign}\big(f(x)\big) \in \{+1, -1\}$. Increasing H increases the number of trainable parameters (model complexity). We define the model dimension d_{model} as the total number of trainable parameters. In this formulation, $d_{\text{model}} = H(d+1) + H = H(d+2)$, which is $O(Hd)$ for large d. We consider the proportional regime where d and H grow while ratios (e.g., n/d_{model}) remain roughly constant.

Teacher Model and Data Generation. Following [1], we assume the data are generated by a fixed teacher. Draw a signal vector $\eta \sim \mathcal{N}(0, I_d)$ (fixed for a dataset). For $i = 1, \ldots, n$:

1. Draw $\varepsilon_i \sim \mathcal{N}(0, I_d)$ independently.
2. Draw a clean label $y_i^* \in \{\pm 1\}$ with $\Pr(y_i^* = +1) = \rho_{+1}$, $\Pr(y_i^* = -1) = \rho_{-1}$ (often $\rho_{\pm 1} = 0.5$).
3. Set the feature vector

$$x_i \;=\; \sqrt{d}\, \eta \, y_i^* + \varepsilon_i. \tag{1}$$

Conditional on $y_i^* = +1$, $x_i \sim \mathcal{N}\big(+\sqrt{d}\,\eta,\, I_d\big)$; conditional on $y_i^* = -1$, $x_i \sim \mathcal{N}\big(-\sqrt{d}\,\eta,\, I_d\big)$. Thus the two classes form clusters centered at $\pm\sqrt{d}\,\eta$.

Structured Label Noise. We corrupt only training labels (features remain clean). Several patterns are possible:

- *Class-conditional noise:* flip labels of one class at a higher rate. With noise level $\pi \in [0, 1]$,

$$y_i \;=\; \begin{cases} -y_i^*, & \text{with probability } \pi \quad \text{if } y_i^* = +1, \\ y_i^*, & \text{otherwise,} \end{cases} \tag{2}$$

  i.e., a fraction π of the $+1$ labels are flipped to -1; all -1 labels remain intact.

- *Cluster or group noise:* flip labels only within a selected subset (e.g., near the decision boundary).
- *Feature-correlated noise:* flip labels for samples with a prescribed feature condition (e.g., extreme values).

Performance Metric. Although training uses noisy labels y_i, test performance is evaluated on a clean test set drawn from (1) and labeled with y^*. The test error is $R_{\text{test}} = \Pr\big[\hat{y}(x) \neq y^*\big]$.

Training Procedure. We train by empirical risk minimization with (optional) L_2 regularization:

$$L(W, a, b) \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell\Big(y_i,\; f(x_i; W, a, b)\Big) \;+\; \frac{\lambda}{2}\,\|(W, a, b)\|_2^2, \tag{2}$$

with ℓ the square loss $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ (used for analysis) or logistic/cross-entropy (used in practice). When $\lambda = 0$, models may interpolate the training data, often exhibiting double descent [6].

Model Complexity and the α Ratio. We quantify complexity by $\alpha = n/d_{\text{model}}$. Smaller α indicates more parameters than data (over-parameterized), larger α indicates under-parameterization. The interpolation threshold typically occurs near $\alpha \approx 1$.

4 Theoretical Analysis of Double Descent with Label Noise

Baseline (No Noise). In the noiseless case ($\pi = 0$), asymptotic analyses (e.g., via CGMT) show a double descent shape in test error versus α. For $\alpha \gg 1$, models are under-parameterized (higher bias). As α decreases, bias drops and test error falls. Near $\alpha \approx 1$, variance spikes (fitting finite-sample idiosyncrasies), producing a peak. For $\alpha < 1$, models become highly expressive; excess degrees of freedom can spread variance, leading to a second descent. As $\alpha \to 0$, test error approaches the Bayes error (essentially 0 here) if the model can represent the teacher function [1, 2].

Effects of Label Noise. Let $\pi > 0$ denote the (structured) label-noise rate.

1. **Increased Bayes Error.** There is an irreducible error due to mislabels. In class-conditional noise with rate π on a class occupying fraction ρ_{+1}, the Bayes error is at least $\pi\,\rho_{+1}$.
2. **Variance Spike at Interpolation.** Around $\alpha \approx 1$, models that can interpolate also fit the noise, causing large variance and higher test error. In linear settings, mean-squared error can blow up at $\alpha = 1$ without regularization [2, 4].

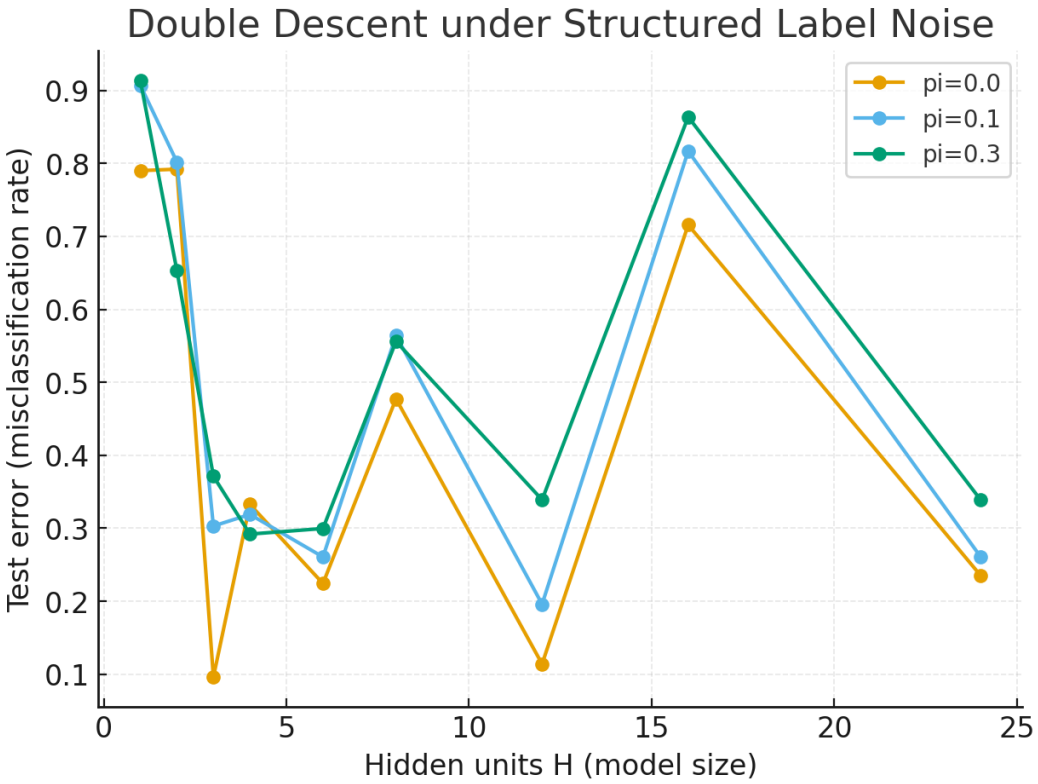

Figure 1: Test error vs. model size (H) under structured label-noise levels $\pi \in \{0, 0.1, 0.3\}$.

3. **Reduced Second-Descent Benefit.** For $\alpha \ll 1$, highly over-parameterized models can isolate noisy examples (memorize them) while learning the clean signal. The second descent plateaus at the noise floor rather than tending to 0.

Combining these yields qualitative predictions for $R_{\text{test}}(\alpha)$ with noise: (i) the entire curve lifts upward; (ii) the interpolation peak grows; (iii) the peak may shift slightly from $\alpha = 1$ (extra capacity may be needed to fit structured noise); (iv) for $\alpha \to 0$, R_{test} approaches the noise floor.

A convenient misclassification expression for a classifier f on the Gaussian-mixture (1) is

$$R_{\text{test}} = \tfrac{1}{2} \Pr\big[f(x) < 0 \,\big|\, y^* = +1\big] + \tfrac{1}{2} \Pr\big[f(x) \geq 0 \,\big|\, y^* = -1\big]. \tag{3}$$

In the Bayes-optimal case $f(x) = \text{sign}(\eta^\top x)$, the baseline error is $\Phi(-\Delta)$ with a suitable SNR Δ. With noisy training labels, the learned decision boundary deviates from η, worsening alignment (particularly near $\alpha = 1$).

5 Experiments

Setup. We simulate training using the teacher model (1). Unless specified, input dimension $d = 50$ and training size $n = 200$ (with balanced class prior $\rho_{+1} = \rho_{-1} = 0.5$). The signal vector η is drawn once per dataset. Structured label noise is applied as in (2): we consider $\pi \in \{0, 0.1, 0.3\}$. Test sets are drawn cleanly with 10,000 samples.

Models and Training. We train two-layer ReLU networks while varying hidden units H to sweep model complexity. Since $d_{\text{model}} \approx O(Hd)$, $\alpha \approx n/d_{\text{model}} \propto n/(Hd)$. With $n = 200$ and $d = 50$, $\alpha \approx 4/H$, so $H \approx 4$ is near interpolation ($\alpha \approx 1$). We use SGD or LBFGS until convergence. We compare unregularized ($\lambda = 0$) to L_2-regularized ($\lambda > 0$) training. Each configuration is averaged over multiple seeds.

Metrics. We track training error and test misclassification rate versus H (or equivalently α). Interpolation is confirmed by near-zero training error.

6 Results and Discussion

No-Noise Double Descent. With $\pi = 0$, we replicate double descent. Small H underfits (high test error). As $H \to 3, 4$, test error decreases. Near $H \approx 4$ ($\alpha \approx 1$), a spike appears as the model interpolates and overfits finite-sample idiosyncrasies. Larger H (10–20) yields a second descent, approaching the Bayes error (near 0 for separated clusters).

Impact of Structured Label Noise. For $\pi \in \{0.1, 0.3\}$:

- The interpolation peak increases markedly. At $\pi = 0.3$, peak test error can approach chance near the threshold.
- The peak shifts slightly to larger H (smaller α) as noise rises, since extra capacity is needed to memorize wrong labels.
- In the over-parameterized regime ($H \gg 4$), test error plateaus above zero. For $\pi = 0.1$, the floor is $\sim$5–6%; for $\pi = 0.3$, $\sim$15–16%, matching the noise floor (since 30% flips in half the data imply $\gtrsim 15\%$ irreducible error).
- For moderate noise, the second descent is present but flatter: beyond some capacity, extra parameters do not improve clean accuracy much.

Effect of Regularization. Adding modest L_2 (e.g., $\lambda = 0.01$):

- Greatly diminishes the interpolation spike; the curve looks closer to a classical U-shape, consistent with [6].
- Increases bias slightly; the best achievable error in the highly over-parameterized regime can be a bit worse than the unregularized floor (e.g., $15\% \to 17\%$), a common trade-off favored to avoid the large spike near $\alpha \approx 1$.

Interpretability Note. Class-conditional corruption induces asymmetry. Around interpolation, models may bias toward the clean class, sacrificing the corrupted class. In the over-parameterized regime, the network can dedicate units to memorize noisy subsets while modeling the clean signal—a conditional memorization behavior relevant to interpretability.

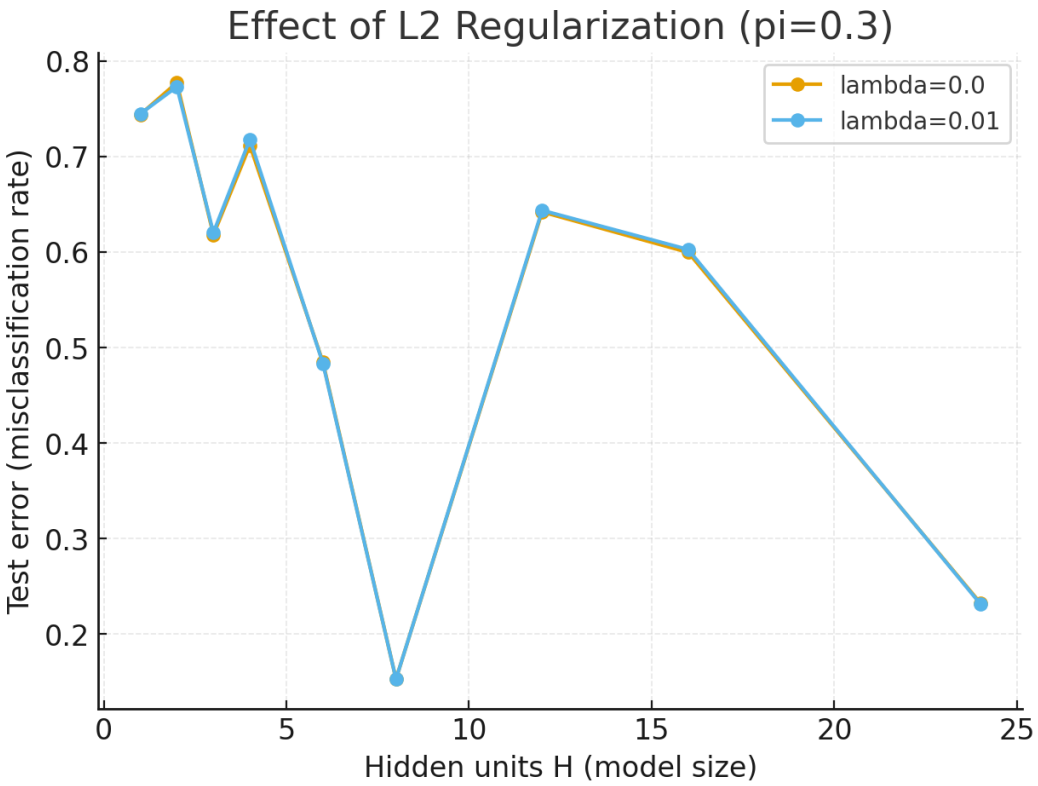

Figure 2: Effect of L_2 regularization on the double–descent curve at $\pi = 0.3$, comparing $\lambda = 0$ (pronounced spike near interpolation) versus $\lambda = 0.01$ (mitigated spike).

7 Conclusion

We studied double descent in two-layer ReLU networks under structured label noise. Building on [1], we introduced class-conditional and other structured flips and examined how they affect the interpolation threshold and generalization. Theoretically and empirically, noise raises the entire risk curve, amplifies the peak near $\alpha \approx 1$, and prevents the second descent from reaching 0, instead plateauing at the noise floor. Nevertheless, sufficiently large models approach the Bayes floor by learning the signal and compartmentalizing mislabeled examples. Explicit regularization (e.g., L_2) effectively smooths the curve, mitigating the interpolation spike.

Future directions include a rigorous CGMT-style treatment with structured noise (quantifying peak height and optimal regularization as a function of π) and exploring other noise structures (feature-correlated, adversarial). Practically, our results reinforce that with systematic label errors, operating precisely at the interpolation threshold without regularization is risky; either move well past the threshold (with regularization) or mitigate noise via data cleaning or robust losses.

Acknowledgments. I thank Dr. Hailin Sang for guidance and insightful discussions.

References

[1] C. S. Abeykoon, A. Beknazaryan, and H. Sang. The Double Descent Behavior in Two Layer Neural Network for Binary Classification. *Journal of Data Science*, 23(2):370–388, 2025.

[2] 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.

[3] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[4] S. d'Ascoli, G. Refinetti, G. Biroli, and F. Krzakala. Double trouble in double descent: Bias and variance in the lazy regime. In *Proceedings of ICML*, PMLR 119:2280–2290, 2020.

[5] Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

[6] P. Nakkiran, P. Bhojanapalli, S. Kakade, and T. Ma. Optimal regularization can mitigate double descent. arXiv:2003.01897, 2020.