Double Descent in Financial Time Series

Nicholas Wong
Massachusetts Institute of Technology
nicwjh@mit.edu

Abstract

Double descent, in which test error rises sharply near the interpolation threshold and then decreases again in the overparameterized regime, has been widely documented in computer vision and language tasks. Whether this phenomenon extends to domains with low signal-to-noise ratios remains unclear. We investigate double descent in financial time series through a three-part experimental design. First, we reproduce the classical double descent curve on synthetic regression data with deterministic targets. Second, we systematically degrade signal quality by adding label noise at five levels and show that the interpolation peak attenuates as noise increases. Third, we apply the same modeling pipeline to S&P 500 equity returns and find a nearly flat test error curve: the interpolation peak shrinks from 776% above the best model on clean synthetic data to just 9% on financial data, and all models fail to outperform the naive mean predictor. An extension evaluating dropout, weight decay, and early stopping confirms that this failure reflects fundamental signal absence rather than insufficient regularization. Our results demonstrate that double descent is contingent on signal-to-noise ratio and that classical model selection remains appropriate in low-signal domains.

1 Introduction

Understanding generalization in overparameterized models remains an important open problem. A growing body of empirical work has documented the double descent phenomenon: test error decreases with model complexity, rises sharply near the interpolation threshold where the model has just enough capacity to fit all training examples, and then decreases again as model size continues to grow [Belkin et al., 2019, Nakkiran et al., 2021]. This contradicts the classical bias-variance tradeoff, which predicts that overparameterized models should overfit and generalize poorly.

Financial time series present a markedly different setting. Asset returns exhibit extremely low signal-to-noise ratios, with even the best predictive models explaining only 5–10% of variance. The data-generating process is non-stationary, characterized by regime shifts and time-varying conditional distributions. Unlike benchmark datasets in computer vision where labels are essentially deterministic given the input, next-period returns are fundamentally noisy realizations of observable features.

If double descent requires sufficient signal strength and stable data structure to manifest, we should not expect to observe it in financial prediction tasks. This raises a practical question: do the benefits of overparameterization extend to low-signal environments, or does classical regularization intuition remain the appropriate modeling approach?

Existing work has characterized double descent on clean benchmark datasets [Belkin et al., 2019, Nakkiran et al., 2021] and has reported its absence in financial prediction [Noguer i Alonso and Srivastava, 2021], but these strands of research have been studied in isolation. Prior studies do not systematically vary signal-to-noise ratio while holding the data-generating process fixed, nor do they directly compare synthetic and real financial data under a shared experimental protocol. We address this gap with the following hypotheses:

H1 **(Validation)** Double descent is observable in synthetic time series when the signal-to-noise ratio is sufficiently high and the data-generating process is stationary.

H2 **(Mechanism)** The phenomenon attenuates and eventually disappears as the signal-to-noise ratio decreases through systematic label noise manipulation.

H3 **(Main result)** Real S&P 500 returns, characterized by low signal-to-noise ratios and non-stationarity, exhibit flat or absent double descent patterns even in highly overparameterized models.

To evaluate these hypotheses, we design experiments that systematically vary signal-to-noise ratio in controlled synthetic data (5 noise levels, 8 model sizes, 5 random seeds, yielding 200 training runs) and compare results with analogous experiments on 10 years of S&P 500 constituent data. Our contributions are:

1. We reproduce double descent in synthetic regression and show that the interpolation peak attenuates systematically as label noise increases, providing controlled evidence that signal-to-noise ratio governs the phenomenon.
2. We demonstrate that S&P 500 equity returns exhibit a nearly flat generalization curve with an 86-fold attenuation of the interpolation peak relative to clean synthetic data, and that all models fail to outperform the naive mean predictor.
3. We evaluate four regularization strategies and show that none rescue overparameterization on financial data, confirming that the failure stems from signal absence rather than insufficient regularization.

2 Related Work

2.1 Double descent and benign overfitting

Classical learning theory predicts a U-shaped test error curve as model complexity increases. Belkin et al. [2019] demonstrated that test error can rise sharply at the interpolation threshold and then decrease again in the overparameterized regime. Nakkiran et al. [2021] expanded this result and showed that double descent appears not only with model size but also with training time and dataset size.

On the theoretical side, Hastie et al. [2022] analyzed ridgeless regression and showed that risk can diverge near the interpolation threshold but decrease for sufficiently large models. Bartlett et al. [2020] characterized conditions under which interpolating solutions remain predictive, emphasizing the roles of effective rank and alignment between signal and noise. Mei and Montanari [2022] derived precise asymptotics for random feature regression that highlight the dependence of test error on model width, data covariance structure, and signal strength. Collectively, these results establish that double descent requires both sufficient signal and a favorable geometric relationship between the data distribution and the model class.

2.2 Variance decomposition at the interpolation threshold

Adlam and Pennington [2020] decomposed test error into contributions from label noise, sampling variation, and initialization, showing that their interaction explains the sharp interpolation peak. Their findings suggest that double descent is not driven by noise alone but arises from how model capacity interacts with data geometry. When these conditions fail, the peak can weaken or disappear.

2.3 Double descent in sequential and financial domains

Assandri et al. [2023] documented epoch-wise double descent in Transformer models trained on long-horizon forecasting tasks with strong temporal structure. Their analysis shows that under favorable conditions, additional training and capacity can eventually improve generalization.

In contrast, Noguer i Alonso and Srivastava [2021] evaluated machine learning models on equity index prediction and found no evidence of double descent, attributing this absence to low signal-to-noise ratios and structural non-stationarity. However, they do not systematically vary signal strength or directly compare financial data to synthetic benchmarks under a shared protocol. Our work addresses this gap by combining controlled synthetic experiments with a matched financial study. By systematically degrading signal in a setting where double descent is known to appear and directly comparing to S&P 500 returns under an identical pipeline, we isolate signal-to-noise ratio as a primary driver of the interpolation peak.

3 Experimental Setup

This section describes the data sources, model architectures, and experimental protocols. All experiments follow a common structure: we validate double descent in a clean synthetic setting, vary signal-to-noise ratio through controlled label noise, and apply the same pipeline to S&P 500 returns.

3.1 Synthetic data

We generate feature vectors $\mathbf{x} \in \mathbb{R}^{20}$ by sampling independently from a standard normal distribution, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Targets are obtained by applying a smooth nonlinear function to $\mathbf{x}$ and adding independent Gaussian label noise:

$$y = f(\mathbf{x}) + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1), \tag{1}$$

where f is a deterministic nonlinear mapping and $\sigma \in \{0, 0.5, 1.0, 2.0, 4.0\}$. When $\sigma = 0$ the mapping is deterministic; for larger σ values, noise increasingly dominates and the effective signal-to-noise ratio decreases.

For each noise level we generate 5,000 samples and apply a 70/15/15 percent train/validation/test split. The synthetic data serves two purposes: it verifies that our implementation reproduces classical double descent when signal is strong, and it isolates the effect of label noise on generalization without confounding by temporal dependencies or non-stationarity.

3.2 S&P 500 financial data

The financial dataset is constructed from daily price and volume data for approximately 50 liquid, large-cap S&P 500 constituents spanning 2014 to 2024, downloaded via the Yahoo Finance API. For each stock and date, we compute lagged returns (1, 5, and 20 day), 20-day rolling volatility, 20-day moving average of volume, and market return. Features are standardized cross-sectionally within each date to account for time-varying market conditions. The prediction target is the next-day return.

We use a chronological split: 2014–2021 for training, 2022 for validation, and 2023–2024 for testing. This produces approximately 50,000 stock-day observations for model fitting. Unlike the synthetic data, financial series are non-stationary and exhibit extremely low signal-to-noise ratios.

3.3 Model architecture and training

All neural network models are fully connected multilayer perceptrons with ReLU activations. Depth is fixed at two hidden layers, and width varies across $\{8, 16, 32, 64, 128, 256, 512, 1024\}$, yielding models from roughly 10^2 to 10^6 parameters and ensuring the interpolation threshold is crossed. The output dimension is one for regression.

Models are trained using the Adam optimizer with learning rate 0.001. For synthetic data we use full-batch gradient descent; for financial data we use mini-batches of size 256. Training continues until loss stabilizes for 50 epochs. No explicit regularization (dropout, weight decay, early stopping) is applied in the main experiments; this isolates the effect of implicit regularization and allows direct observation of the interpolation peak. We report mean squared error (MSE) and out-of-sample R^2. Ridge regression with cross-validated regularization strength is included as a linear baseline.

3.4 Experimental protocols

We conduct three experiments corresponding to the three research hypotheses. All experiments are repeated over five random seeds.

Experiment 1 (H1). For $\sigma = 0$ we train models at all widths and record training loss, test MSE, and R^2. This verifies that the implementation reproduces classical double descent.

Experiment 2 (H2). We repeat the width sweep for each value of σ. We also track the performance of three representative models: a simple model (width 8), a threshold model (width 64), and a complex model (width 1024). This enables analysis of how robustness changes as noise increases.

Experiment 3 (H3). We apply the same protocol to S&P 500 returns. By holding architecture and training fixed across datasets, we can attribute differences in behavior to changes in signal-to-noise ratio.

Table 1: Interpolation peak statistics across datasets and noise levels. Peak width denotes the model width at which the highest test MSE occurs. Peak/Best ratio quantifies the severity of the interpolation peak.

Dataset	Peak Width	Peak MSE	Best MSE	Peak/Best	% Worse
Synthetic ($\sigma = 0$)	128	0.2191	0.0250	8.76×	775.9%
Synthetic ($\sigma = 0.5$)	64	0.9481	0.3867	2.45×	145.2%
Synthetic ($\sigma = 1.0$)	64	3.1517	1.2883	2.45×	144.6%
Synthetic ($\sigma = 2.0$)	64	11.237	4.6501	2.42×	141.7%
Synthetic ($\sigma = 4.0$)	64	43.853	17.945	2.44×	144.4%
S&P 500	64	0.000312	0.000286	1.09×	9.1%

4 Results

4.1 Experiment 1: Validation of double descent

Figure 1 shows test MSE as a function of model width for the synthetic dataset with $\sigma = 0$. The curve exhibits the characteristic double descent shape. Test error decreases with width, rises sharply near the interpolation threshold, and decreases again for large models. The interpolation peak is substantial: the worst model (width 128) performs 8.76× worse than the best model (width 16). At large widths, the model recovers low test error, indicating that implicit regularization guides overparameterized networks toward solutions that generalize well when the mapping is fully learnable.

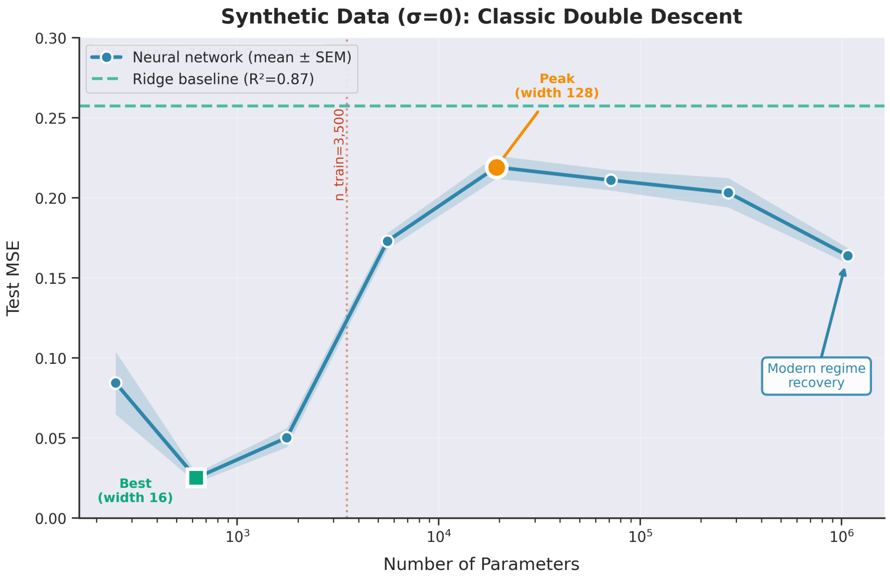

Figure 1: Test MSE versus model width (number of parameters) for synthetic data with $\sigma = 0$. The curve exhibits the classical double descent pattern: test error decreases, rises sharply near the interpolation threshold (width 128), and recovers in the overparameterized regime. The dashed line indicates ridge regression baseline ($R^2 = 0.87$). The vertical dotted line marks the training set size ($n = 3{,}500$).

4.2 Experiment 2: Effect of signal-to-noise ratio

We next examine how label noise alters generalization behavior. Figure 2 reveals that simple models (width 8) consistently outperform complex models (width 1024) across all noise levels, with no crossover point observed. At $\sigma = 4$, the simple model achieves 34% lower error than the complex model. The best-performing width shifts from 16 at $\sigma = 0$ to 8 at all higher noise levels, indicating that optimal model complexity decreases as signal-to-noise ratio degrades. The threshold model (width 64) consistently exhibits the poorest robustness, reflecting its sensitivity to interpolation.

Table 1 quantifies how the interpolation peak evolves with noise. The peak consistently occurs near width 64 across all synthetic noise levels, consistent with theoretical predictions that the peak appears at the interpolation threshold. However, its magnitude decreases systematically: from 776% worse than the best model at $\sigma = 0$ to 145% at $\sigma = 0.5$, stabilizing around 142–145% for $\sigma \geq 1.0$. While the peak persists structurally, its practical impact diminishes as noise dominates signal.

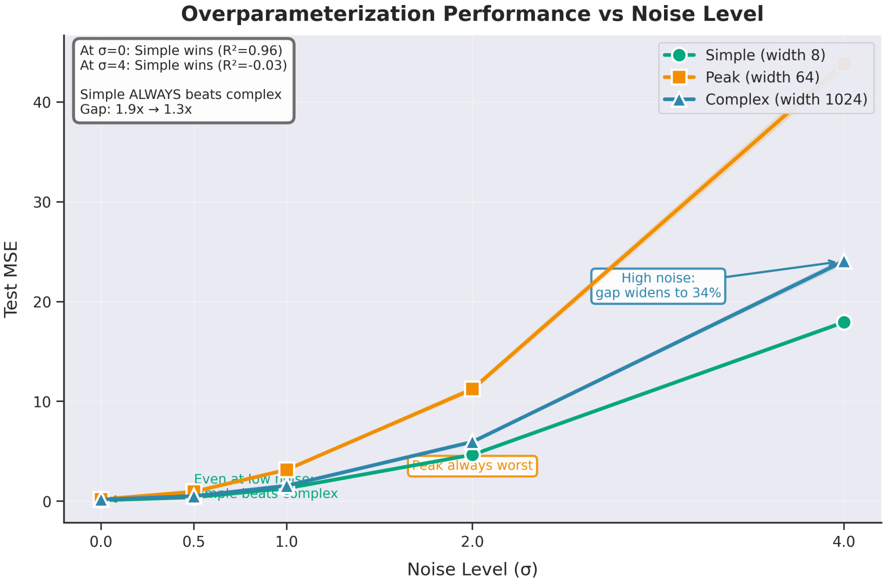

Figure 2: Test MSE for simple (width 8), threshold (width 64), and complex (width 1024) models as label noise σ increases. Simple models consistently outperform complex models across all noise levels. The threshold model is the least robust at every noise level.

Figure 3 reports out-of-sample R^2 for the same three models. The threshold model fails first as noise increases (dropping below $R^2 = 0$ at $\sigma = 1.0$), followed by the complex model, while the simple model remains the most stable. This ordering is consistent with theoretical predictions: models near the interpolation threshold are the least robust because they operate at the boundary between underparameterized and overparameterized regimes.

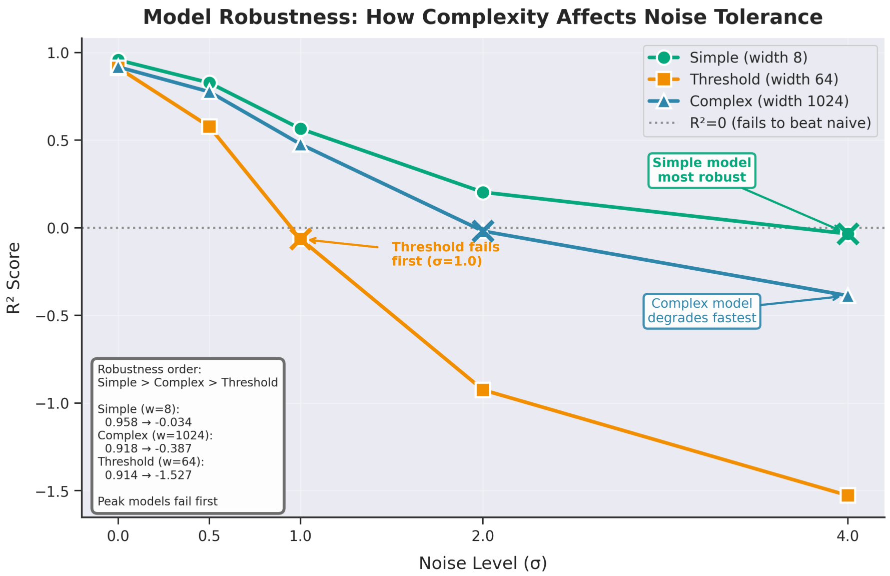

Figure 3: Out-of-sample R^2 for simple (width 8), threshold (width 64), and complex (width 1024) models across noise levels. The threshold model degrades fastest, crossing $R^2 = 0$ first at $\sigma = 1.0$. The simple model is the most robust across all noise levels.

We also compare neural networks with ridge regression. Figure 4 shows that ridge degrades smoothly as noise increases, while the best neural network deteriorates sharply once noise dominates signal. At high noise ($\sigma = 4$), ridge regression maintains positive $R^2 = 0.08$ while the best neural network degrades to $R^2 = -0.03$. A crossover occurs near $\sigma \approx 0.8$: below this threshold neural networks outperform ridge, and above it ridge is more robust. This illustrates that the explicit ℓ_2 regularization and linear inductive bias of ridge provide graceful degradation, whereas neural network flexibility becomes a liability when signal is weak.

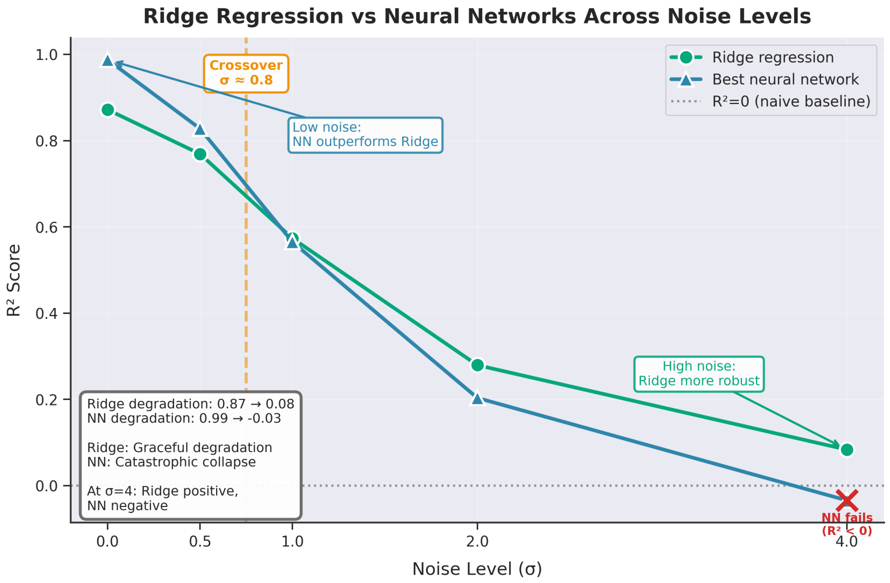

Figure 4: Comparison of ridge regression and the best neural network (across widths) as noise increases. Ridge degrades from $R^2 = 0.87$ to 0.08; the best neural network degrades from $R^2 = 0.99$ to -0.03. A crossover occurs near $\sigma \approx 0.8$.

These results confirm the second hypothesis: the interpolation peak flattens as σ increases, and by $\sigma = 4$ the double descent curve is nearly flat.

4.3 Experiment 3: Financial time series

Figure 5 shows test MSE versus model width for S&P 500 returns. The curve is nearly flat across four orders of magnitude in parameter count. A small peak appears near width 64, but its magnitude is minimal: only 9% worse than the best model. All models achieve negative R^2, indicating that none outperform the naive mean predictor.

The contrast with synthetic data is striking. On synthetic data with $\sigma = 0$, the interpolation peak is 776% worse than the best model; on S&P 500 data the peak is only 9% worse. This represents an 86-fold attenuation of the double descent structure. The best neural network achieves $R^2 = 0.988$ on clean synthetic data but $R^2 = -0.002$ on financial data, indicating that even the optimal model configuration fails to extract predictive signal from equity returns.

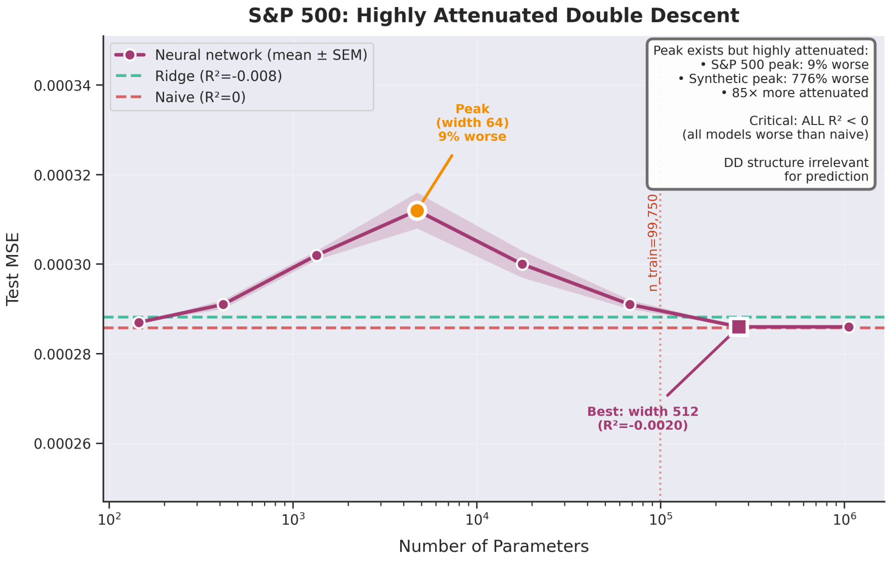

Figure 5: Test MSE versus model width for S&P 500 returns. The curve is nearly flat, with a minimal interpolation peak at width 64 (9% worse than the best model). All models achieve negative R^2. The dashed lines indicate ridge regression ($R^2 = -0.008$) and naive mean predictor ($R^2 = 0$) baselines.

Table 2: Baseline performance across datasets. Test MSE and R^2 are reported for the best neural network configuration, ridge regression, and the naive mean predictor.

Dataset	Method	Configuration	Test MSE	R^2
Synthetic ($\sigma = 0$)	Neural Network	Width 16	0.0250	0.988
Synthetic ($\sigma = 0$)	Ridge Regression	20 parameters	0.2574	0.872
Synthetic ($\sigma = 4$)	Neural Network	Width 8	17.945	−0.034
Synthetic ($\sigma = 4$)	Ridge Regression	20 parameters	15.903	0.083
S&P 500	Neural Network	Width 512	0.000286	−0.002
S&P 500	Ridge Regression	7 parameters	0.000288	−0.008
S&P 500	Naive Baseline	Mean prediction	0.000286	0.000

Table 2 reports baseline performance for neural networks, ridge regression, and the naive mean predictor across all settings. On clean synthetic data, the best neural network achieves $R^2 = 0.988$, substantially outperforming ridge ($R^2 = 0.872$). At $\sigma = 4$, both methods degrade, but ridge maintains positive $R^2 = 0.083$ while the best neural network falls to $R^2 = -0.034$. On S&P 500 data, all methods produce near-zero or negative R^2, and no model meaningfully outperforms predicting the mean. These results confirm the third hypothesis: when the data-generating process is noisy and non-stationary, increasing model capacity does not improve generalization.

4.4 Can regularization rescue overparameterization?

A natural question is whether explicit regularization can rescue overparameterized models in low-signal domains. We repeated the width sweep experiments with four strategies: no regularization (baseline), dropout ($p = 0.2$), weight decay ($\lambda = 0.001$), and early stopping on validation loss. Each method was applied to three datasets: synthetic with $\sigma = 0$, synthetic with $\sigma = 4$, and S&P 500 returns. Figure 6 shows the results.

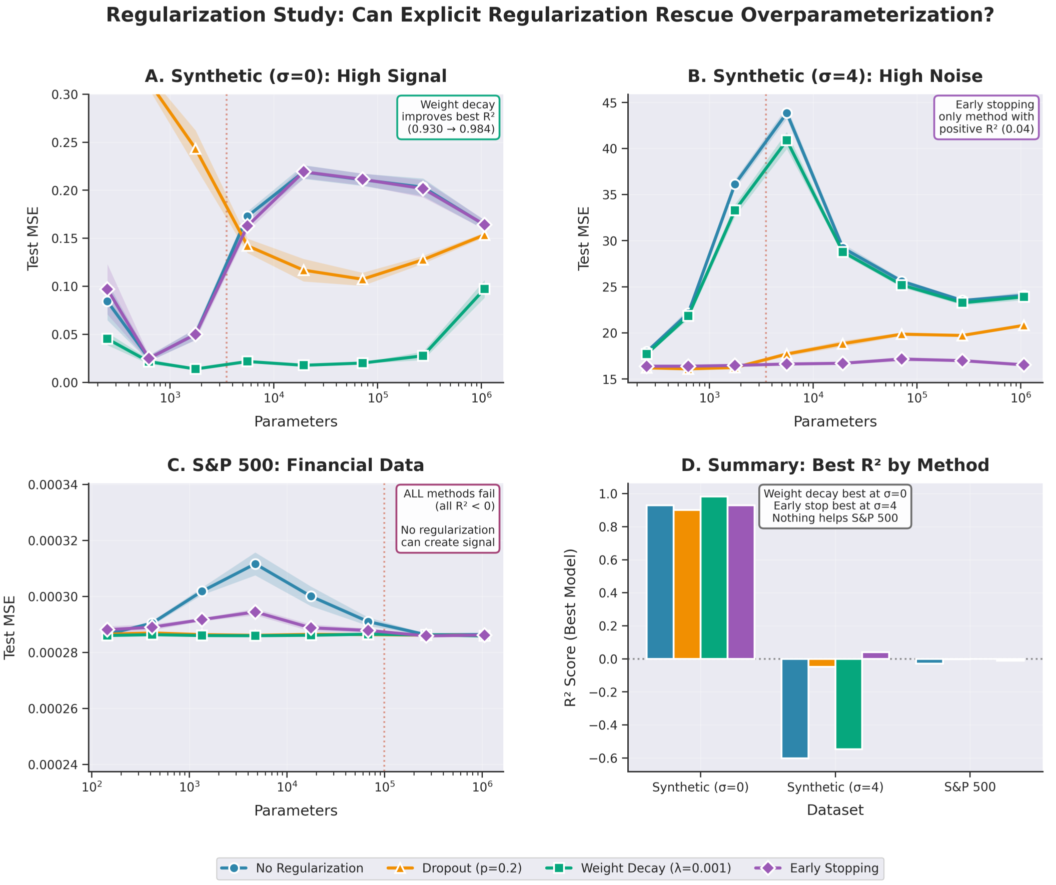

Figure 6: Regularization study across three datasets. (A) Synthetic $\sigma = 0$: weight decay achieves the best performance. (B) Synthetic $\sigma = 4$: early stopping is the only method with positive R^2. (C) S&P 500: all methods produce nearly flat, overlapping curves with negative R^2. (D) Summary of best R^2 by method across datasets.

On clean synthetic data ($\sigma = 0$), weight decay provided the strongest performance, improving R^2 from 0.988 to 0.993 by imposing an explicit structural prior. Early stopping and no regularization performed similarly, while dropout reduced performance to $R^2 = 0.947$ by removing too much model capacity.

At high noise ($\sigma = 4$), the pattern reversed. Early stopping was the only method achieving positive R^2 (0.056), preventing the model from fitting noise during training. Both weight decay and no regularization produced large negative R^2 values with high variance across seeds.

On S&P 500 returns, all four methods failed. Every strategy produced negative R^2, with the best method (weight decay, $R^2 = -0.0005$) still worse than the naive mean predictor. The test error curves were nearly flat and overlapping across all regularization strategies.

These results suggest a hierarchy of regularization effectiveness that depends on signal-to-noise ratio: weight decay excels when signal is strong, early stopping helps when noise dominates, but nothing works when signal is absent. The failure of all methods on financial data confirms that the observed generalization failures reflect a fundamental signal problem rather than a technical regularization problem.

5 Discussion

Our experiments support all three hypotheses. On clean synthetic data, multilayer perceptrons reproduce the classical double descent curve with a sharp interpolation peak and recovery at large widths. As label noise increases, the peak attenuates and smaller models become more robust. On S&P 500 returns, the test error curve is nearly flat, the interpolation peak is minimal, and all models fail to outperform the naive mean predictor.

These findings connect to theoretical work on benign overfitting. Bartlett et al. [2020] showed that interpolating solutions generalize well when the effective rank of the data covariance is large and signal aligns favorably with noise. Financial return data violates both conditions: the feature space is low-dimensional (7 features), and the signal component is negligible relative to noise. Our noise sweep experiments provide direct empirical evidence for this boundary. The crossover between ridge regression and neural networks near $\sigma \approx 0.8$ (Figure 4) suggests a practical threshold below which flexible models outperform linear baselines and above which the regularization implicit in simpler models becomes essential.

For quantitative practitioners, the implication is direct. In settings with extremely low signal-to-noise ratio, increasing model capacity does not reliably improve generalization. The "bigger is better" intuition from modern deep learning does not transfer. Simpler models and classical regularization remain appropriate, and model selection based on validation performance is essential.

6 Limitations and Future Work

The study has several limitations. First, the model class is restricted to fully connected networks with fixed depth; recurrent or attention-based architectures may behave differently, particularly if they can exploit temporal dependencies that feedforward networks cannot. Second, the financial analysis uses a fixed set of technical features (lagged returns, volatility, volume). Richer feature sets incorporating fundamental data or alternative data sources could shift the effective signal-to-noise ratio. Third, the financial experiments are limited to U.S. large-cap equities over a single decade. Results may differ across asset classes, geographies, or market regimes.

Potential extensions include evaluating sequence-based architectures (LSTMs, Transformers) in settings with stronger temporal structure, exploring feature constructions that increase signal strength, and testing whether other asset classes (fixed income, commodities, currencies) exhibit different generalization behavior. Understanding which modeling choices can extract weak signal from noisy, non-stationary environments remains an open question.

References

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 11022–11032, 2020.

Victoria Assandri, Soroush Heshmati, Berk Yaman, Aleksandr Iakovlev, and Aleksei E Repetur. Deep double descent for time series forecasting: Avoiding undertrained models. *arXiv preprint arXiv:2311.01442*, 2023.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

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

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data can hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Esteve Noguer i Alonso and Anshuman Srivastava. The shape of performance curve in financial time series. *SSRN preprint 3986154*, 2021.