Benign Overfitting and Double Descent in Overparameterized Linear Regression *

Emma Moncia
ENSAE

Lucía Criado del Rey Gutiérrez
ENSAE

Abstract

We study benign overfitting and double descent in a linear return-prediction setting using an equities panel. We keep a fixed set of real predictors and synthetically increase model dimension by appending irrelevant random features, which lets us sweep across the interpolation threshold while holding the underlying signal fixed. We document a test-error peak near $d \approx n$ alongside norm inflation and ill-conditioning, and we compare ridgeless minimum-norm interpolation against ridge baselines and early-stopped gradient descent.

1 Introduction

Classical learning theory suggests that highly overparameterized models are prone to overfitting, leading to poor generalization performance and motivating the use of explicit regularization to control model complexity. However, recent theoretical and empirical studies have shown that overparameterized models can achieve exact interpolation of the training data while still generalizing well. This phenomenon, known as benign overfitting, often appears together with the double descent curve, where test error increases as the model approaches the interpolation threshold and then decreases again as the model becomes even more overparameterized [1, 2, 3].

Financial data provides a particularly challenging and relevant setting for studying this phenomenon. Asset returns are notoriously noisy, exhibit low signal-to-noise ratios, heavy tails, heteroskedasticity and are subject to temporal dependence and cross-sectional correlation induced by common market factors. Moreover, they're often modeled using high-dimensional feature representations, which make financial prediction tasks especially susceptible to overfitting.

This project empirically investigates benign overfitting using U.S. equity data in a controlled linear regression framework. Starting from a low-dimensional set of economically meaningful predictors, we increase model complexity by adding irrelevant features in order to study the effect of increasing model size. We analyze test error behavior around the interpolation threshold, the role of parameter norms and conditioning, the importance of interpolator selection, implicit regularization through optimization, and robustness to label noise, with the goal of assessing whether theoretical mechanisms appear in real financial data.

To study these phenomena in a controlled way, we focus on linear regression with highly overparameterized feature spaces rather than more complex nonlinear models. The behavior of nonlinear models typically reflects a combination of feature learning, architectural choices, and optimization heuristics, which makes it difficult to isolate the effect of overparameterization. By contrast, linear models provide a clean empirical setting to test whether the mechanisms highlighted by benign overfitting theory appear in real financial data.

*Code available at `https://github.com/emmamaria123mon/Advanced-ML-benign-overfitting-finance`.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2 Theoretical Background

2.1 Linear regression and interpolation regimes

We consider a linear regression model with design matrix $X \in \mathbb{R}^{n \times d}$ and response vector $y \in \mathbb{R}^n$. The classical model is

$$y = Xw^\star + \varepsilon, \tag{1}$$

where $w^\star \in \mathbb{R}^d$ is an unknown parameter vector and $\varepsilon \in \mathbb{R}^n$ is mean-zero noise.

The (unregularized) least squares objective is

$$\hat{w} \in \arg\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2. \tag{2}$$

When $d < n$ and X has full column rank, the minimizer is unique and corresponds to the ordinary least squares (OLS) solution. When $d \geq n$ and X has full row rank, the training constraints can be satisfied exactly (or nearly exactly), and there are infinitely many solutions with zero training error. This corresponds to the interpolation regime. Intuitively, the training data imposes at most n independent constraints, leaving additional degrees of freedom when the parameter dimension d exceeds n.

2.2 Geometry of interpolating solutions

Suppose interpolation is possible, i.e. there exists at least one vector w_0 such that $Xw_0 = y$. Then the set of all interpolating solutions is

$$\{w \in \mathbb{R}^d : Xw = y\} = \{w_0 + v : v \in \mathrm{Ker}(X)\}, \tag{3}$$

where the nullspace (kernel) of X is

$$\mathrm{Ker}(X) = \{v \in \mathbb{R}^d : \ Xv = 0\}. \tag{4}$$

This decomposition shows that interpolation alone does not determine generalization. Directions in $\mathrm{Ker}(X)$ do not affect the training error but can substantially change the parameter norm and out-of-sample behavior.

2.3 The minimum-ℓ_2-norm interpolator

Among infinitely many interpolators, a canonical choice is the minimum-ℓ_2-norm interpolator

$$\hat{w}_{\min} = \arg\min_{w} \|w\|_2 \quad \text{s.t.} \quad Xw = y. \tag{5}$$

When X has full row rank (so that $XX^\top$ is invertible), it admits the closed form

$$\hat{w}_{\min} = X^\top (XX^\top)^{-1} y. \tag{6}$$

This solution plays a central role in benign overfitting theory [1, 3]. It is also the solution selected by gradient descent on the least squares objective in the overparameterized regime when initialized at $w_0 = 0$, providing an algorithmic mechanism that selects the minimum-norm interpolator among all solutions satisfying $Xw = y$.

2.4 Ridge regression

Ridge regression is a classical explicitly regularized estimator defined by

$$\hat{w}_\lambda \in \arg\min_{w \in \mathbb{R}^d} \ \|Xw - y\|_2^2 + \lambda \|w\|_2^2, \qquad \lambda \geq 0.$$

The ℓ_2 penalty shrinks the parameter norm and stabilizes estimation in ill-conditioned settings. Unlike ridgeless interpolation, ridge regression trades bias for reduced variance and typically avoids large coefficient norms near the interpolation threshold.

In this work, ridge regression serves as an explicitly regularized baseline against which we compare the behavior of the minimum-ℓ_2-norm interpolator in overparameterized regimes.

2.5 Conditioning, norm inflation, and double descent

To understand why ridgeless and ridge estimators behave differently near the interpolation threshold, it is useful to examine the conditioning of the training design. Equation (6) shows that the inverse $(XX^\top)^{-1}$ amplifies directions corresponding to small eigenvalues of $XX^\top$. Near the interpolation threshold $d \approx n$, the Gram matrix can become poorly conditioned, leading to large parameter norms and high variance, and thus a peak in test error. As d increases further, the distribution of eigenvalues can stabilize, and test error may decrease again, producing the double descent phenomenon documented in [2, 1, 3].

3 Data and Experimental Setup

We work with a processed U.S. equities panel. Each row corresponds to a (date, ticker) observation with a next-period forward return target. Stacking all observations yields a response vector $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n \times d_0}$.

- **Baseline feature space (d_0):** Each row of X contains a vector of standardized firm-level predictors, including past returns, volume, a market return proxy, momentum at multiple horizons, and realized volatility at multiple horizons. We denote the number of base features by d_0. (Set explicitly listed as `BASE_FEATURES` in the notebook.)
- **Target:** The target variable is the next-period forward return. Collecting targets across all observations defines the response vector $y \in \mathbb{R}^n$. In the notebook, this corresponds to `TARGET_COL = ret_fwd`. We evaluate performance using mean squared error.

All features are standardized using training set statistics only, and the same transformation is applied to validation and test data to avoid leakage. The target is centered using the training set mean. These steps stabilize optimization and ensure that comparisons across model sizes are meaningful.

We split the data chronologically into training, validation, and test sets. This is needed in financial applications to avoid look-ahead bias, as financial time series exhibit temporal dependence and distribution shift; random splitting would mix past and future information and lead to overly optimistic performance estimates.

Finally, to study interpolation phenomena, it is useful to operate in regimes where the model dimension d is comparable to the number of training observations n and to examine behavior as d exceeds n. To make this feasible and to enable controlled comparisons across model sizes, we subsample each data split to a fixed size of $n = 2000$. This fixes the effective sample size and allows us to vary the model dimension d up to several multiples of n in the experiments.

Controlled Overparameterization

We vary the model dimension while keeping the underlying signal fixed by appending irrelevant random features to the standardized base design. For each target dimension $d \geq d_0$, we construct

$$X^{(d)} = [X_{\text{base}} \;\; Z_{1:(d-d_0)}], \tag{7}$$

where Z consists of i.i.d. $\mathcal{N}(0, 1)$ noise features independent of the labels.

Using irrelevant Gaussian features provides a clean way to isolate the effect of overparameterization:

- The base features remain unchanged, so the signal content is fixed.
- Any change in performance as d increases is driven by geometry, conditioning, and estimator selection rather than additional predictive information.
- Gaussian noise features are analytically convenient and match the isotropic feature assumptions commonly used in theoretical studies of benign overfitting [1].

When sweeping d, we use nested random features. For each random seed, a single maximum noise matrix $Z_{\max}$ is generated, and for each d we select the prefix $Z_{1:(d-d_0)}$. This ensures that increasing d corresponds to adding features rather than replacing them, reducing unnecessary randomness and making the resulting curves easier to interpret.

We sweep d from d_0 up to $3n$, using a grid around the interpolation threshold $d \approx n$. The upper bound $3n$ reaches a clearly overparameterized regime where benign overfitting, if present, should manifest. It also remains computationally feasible for repeated seeds and multiple experiments.

Because the random features depend on the seed, results can vary across realizations. We therefore repeat the sweep across multiple seeds and report mean $\pm$ standard deviation bands to avoid over-interpreting a single random draw.

4 Double Descent and Benign Overfitting

4.1 Experimental Setup and Models

We study the behavior of overparameterized linear regression by expanding the base feature set using the mechanism described in Section 3. Starting from a fixed maximum noise matrix $Z_{\max}$, we construct for each target dimension $d \geq d_0$

$$X^{(d)} = [X_{\text{base}} \;\; Z_{1:(d-d_0)}]$$

For each value of d, we evaluate the following estimators:

- **Ridgeless regression**: the minimum-ℓ_2-norm interpolator when $d \geq n$, and the ordinary least squares solution when $d < n$ (implemented as ridge with $\lambda = 0$).
- **Ridge regression (augmented features)**: ridge regression fit on the expanded design $X^{(d)}$, used as an explicitly regularized comparator. We consider either a fixed penalty λ across all d, or a single λ tuned once per random-feature seed on a small subset of dimensions.
- **Ridge regression (base features)**: ridge regression fit only on the original predictors X_{base}, with λ selected by validation, serving as a regularized baseline without feature augmentation.

Computation: All ridge and ridgeless solutions are computed using closed-form solvers. When $d \leq n$, we use the primal formulation $(X^\top X + \lambda I_d)^{-1} X^\top y$; when $d > n$, we use the dual formulation $(XX^\top + \lambda I_n)^{-1}$. A small diagonal jitter is added for numerical stability near $d \approx n$. These choices affect numerical robustness but not the underlying estimator.

4.2 Double Descent in Test Error

Figure 1 plots test mean squared error as a function of d.

As the model dimension d increases toward the interpolation threshold $d \approx n$, the test error spikes and reaches its maximum. Beyond this threshold, test error decreases again as d grows, even though the additional features are pure noise. This behavior motivates a closer examination of interpolation, parameter norms, and conditioning in the following section.

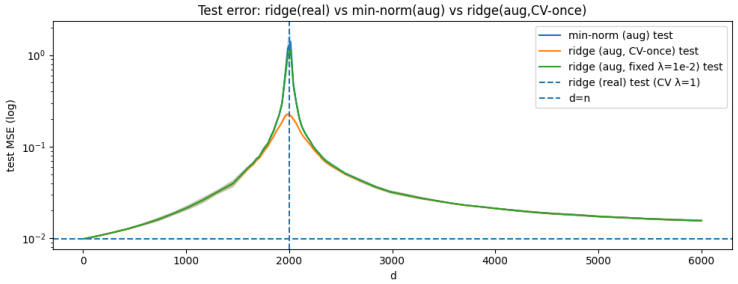
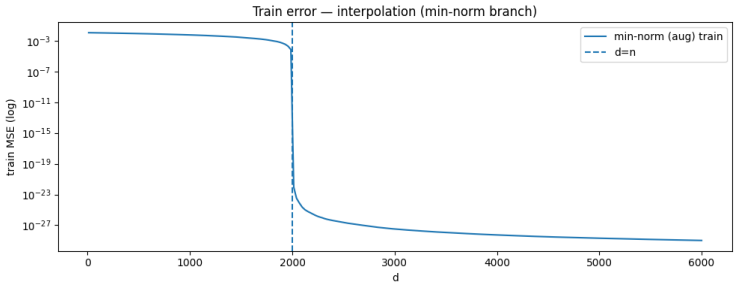


Figure 1: **Test error vs model dimension(Gaussian augmentation)** Test MSE as a function of d for the min-norm interpolator, fixed ridge, and ridge on real features. The vertical line marks the interpolation threshold $d = n$ and shaded bands represent variability across the random feature spaces.

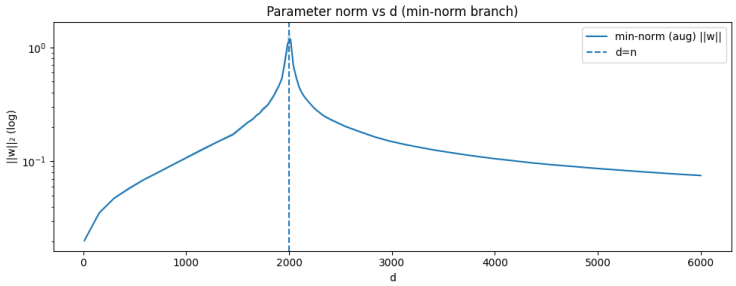
4.3 Interpolation, Norm Growth, and Conditioning

To understand the double-descent behavior observed in test error, we examine training error, parameter norms, and the conditioning of the training design.

Figure 2 shows that the training MSE drops sharply and becomes (near) zero once $d \geq n$, confirming exact interpolation in the overparameterized regime. At the same time, the ℓ_2 norm of the fitted parameters exhibits a pronounced spike near the interpolation threshold $d \approx n$.



(a) Training MSE (log scale) vs. d.



(b) $\|\hat{w}\|_2$ (log scale) vs. d.

Figure 2: **Interpolation and parameter norm.** Training error vanishes once $d \geq n$, while the parameter norm exhibits a sharp spike near the interpolation threshold.

This norm inflation is closely tied to the conditioning of the training Gram matrix $G = \frac{1}{n} XX^\top$. Figure 3 shows the smallest eigenvalue $\lambda_{\min}(G)$ and the condition number $\kappa(G)$ as functions of d.

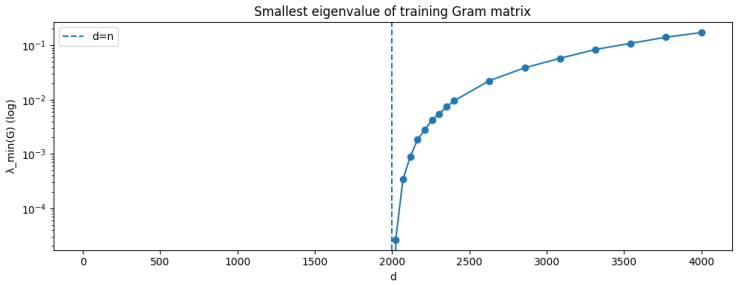
Around the interpolation threshold, the min-norm solution relies on directions associated with small singular values of X, which amplifies noise and leads to large coefficients. As d increases beyond n, $\lambda_{\min}(G)$ becomes positive and the condition number drops by several orders of magnitude, indicating a much better-conditioned problem. This coincides with reduced parameter norms and improved test performance.

Figure 4 visualizes this effect through the eigenvalue spectrum of G for representative values of d.

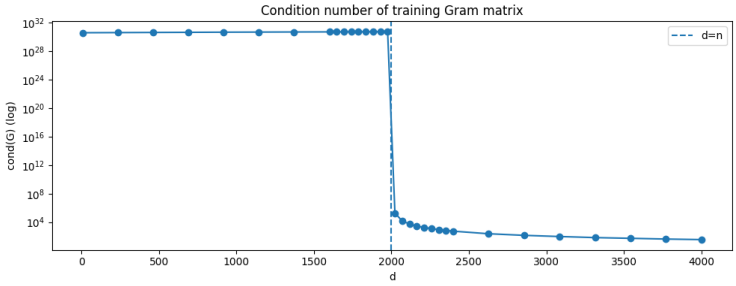
Overall, these diagnostics show that the interpolation threshold is a delicate point where the ridgeless estimator is highly sensitive to noise. The test-error peak around $d \approx n$ is thus linked to norm inflation and ill-conditioning, rather than to signal in the added features. This highlights that overparameterization alone does not determine generalization; the geometry of the design matrix and the choice of interpolating solution play a central role.

4.4 Robustness to Feature Representation

In the main experiments, overparameterization is achieved by appending i.i.d. Gaussian noise features. As a robustness check, we repeat the same dimension sweep using random ReLU features. The goal is not to improve prediction, but to test whether these effects persist when the additional features are generated nonlinearly.



(c) $\lambda_{\min}(G)$ vs. d.



(d) $\kappa(G)$ vs. d/n.

Figure 3: **Conditioning of the training Gram matrix.** Near $d \approx n$, the smallest eigenvalue becomes very small and the condition number grows large, indicating severe ill-conditioning.

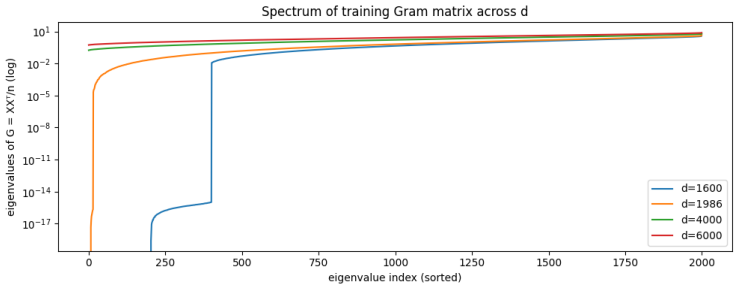


Figure 4: **Eigenvalue spectra of the training Gram matrix.** Very small eigenvalues appear near the interpolation threshold and disappear in the overparameterized regime.

For each seed, we generate a random ReLU map

$$\phi(x) = \mathrm{ReLU}(x^\top W + b),$$

with Gaussian weights and bias, and construct augmented designs $X^{(d)} = [X_{\text{base}}, \Phi_{1:(d-d_0)}]$. ReLU features are standardized using training-set statistics only.

Figure 5 compares test MSE as a function of d for Gaussian and ReLU augmentation. In both cases, test error increases as d approaches the interpolation threshold $d \approx n$, peaks near $d = n$, and then decreases again in the overparameterized regime. However, the peak is substantially larger for ReLU features, and test error remains higher across most values of d.

This difference is reflected in the parameter norm. Figure 6 shows that while norm inflation near $d \approx n$ occurs in both settings, it is much stronger for ReLU features.

Overall, this extension shows that the double-descent pattern and the associated spike in parameter norms near the interpolation threshold are not specific to Gaussian noise features, and that the severity of the instability depends on the geometry defined by the feature representation.

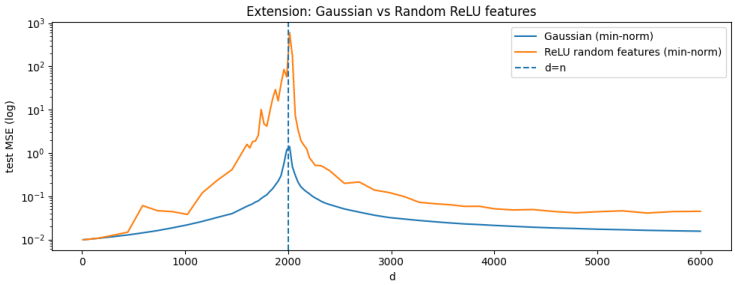


Figure 5: **Gaussian vs. random ReLU features: test error.** Test MSE as a function of model dimension d for the min-norm estimator. The vertical line indicates $d = n$.

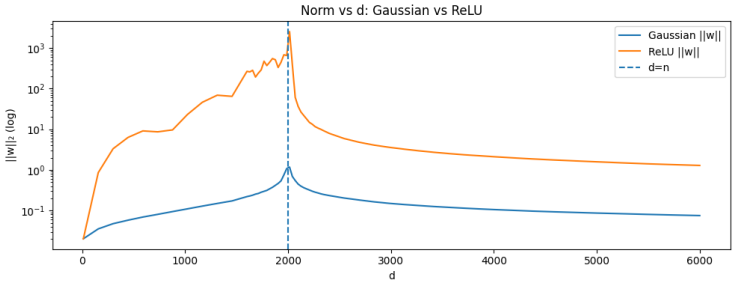


Figure 6: **Gaussian vs. random ReLU features: parameter norm.** $\|\hat{w}\|_2$ (log scale) as a function of d for the min-norm estimator.

4.5 Not All Interpolators Generalize

Finally, we demonstrate that interpolation alone does not guarantee good generalization. In the overparameterized regime, infinitely many parameter vectors satisfy $Xw = y$. If $\hat{w}_{\min}$ denotes the minimum-norm interpolator and $v \in \ker(X)$, then

$$w = \hat{w}_{\min} + \alpha v$$

also interpolates the training data for any α.

Because v lies in the nullspace of X, training predictions are unchanged, and training error remains (near) zero. However, increasing α inflates the parameter norm and can substantially degrade test performance, as shown in Figure 7.

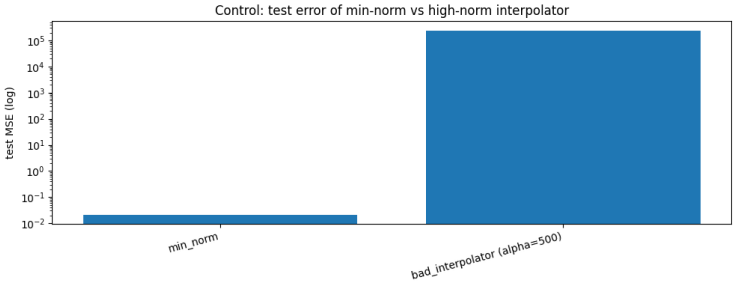


Figure 7: **Interpolation does not imply good generalization.** Test MSE (log scale) of the minimum-norm interpolator compared to a high-norm interpolator obtained by adding a large nullspace component ($\alpha = 500$). Both interpolate the training set, but the high-norm solution generalizes extremely poorly.

5 Implicit Regularization via Optimization

In this section, we connect the behavior of optimization algorithms to the selection of interpolating solutions. This complements the double-descent analysis by showing how, even at fixed overparameterization, different interpolating solutions can arise through optimization.

5.1 Gradient descent as an implicit regularizer

We run gradient descent on the squared loss starting from $w_0 = 0$ at a fixed overparameterized dimension, chosen as $d = 3n$. In overparameterized linear regression, gradient descent initialized at zero converges to the minimum-ℓ_2-norm interpolator under standard conditions. This explains why, in our experiments, the fully converged gradient descent solution coincides with the minimum-norm interpolator studied in the previous sections.

Early stopping and the norm viewpoint

Along the gradient descent path, the training error decreases toward zero while the parameter norm $\|w_t\|_2$ increases monotonically. Validation and test performance, however, can vary substantially along this path. Early stopping selects the iteration $t^\star$ that minimizes validation error, resulting in a solution with smaller norm than the fully converged interpolator.

This behavior can be interpreted as a form of implicit regularization: rather than adding an explicit penalty to the objective, the effective complexity of the model is controlled by stopping optimization early. Plotting mean squared error as a function of $\|w_t\|_2$ provides a complementary view of this effect, showing that generalization performance can deteriorate as the parameter norm grows along the optimization trajectory.

5.2 Interpretation in a Financial Prediction Regime

Equity return prediction is characterized by a very low signal-to-noise ratio. In such settings, validation performance can peak extremely early along the optimization path, while continued training increasingly adapts to noise rather than signal.

Figure 8 illustrates this behavior along the gradient descent trajectory. As shown in the top panel, training mean squared error decreases rapidly toward zero, while validation and test errors remain nearly flat. In this run, the best validation performance occurs at the beginning of training ($t^\star = 0$), indicating that further optimization does not improve out-of-sample performance.

The middle panel shows the evolution of the parameter norm $\|w_t\|_2$ as a function of iteration. Starting from zero, the norm increases quickly and then plateaus at the value corresponding to the fully converged interpolating solution. This highlights that continued optimization mainly increases model complexity, even when predictive performance does not improve.

The bottom panel plots validation and test MSE directly against the parameter norm along the optimization path. This view makes the implicit regularization effect explicit: as gradient descent moves from low-norm to high-norm solutions, generalization performance deteriorates. The selected early-stopping iterate corresponds to a much smaller norm than the fully converged interpolator, which achieves lower training error but worse out-of-sample performance.

Overall, this experiment shows that in low signal-to-noise financial prediction tasks, optimization dynamics play a critical role in determining generalization. Even when interpolation is possible, continued training can mainly fit noise rather than signal, and early stopping acts as an implicit regularizer by limiting the complexity of the fitted model.

5.3 Interpretation in a Financial Prediction Regime

Equity return prediction is characterized by a very low signal-to-noise ratio. In such settings, validation performance can peak extremely early along the optimization path, while continued training increasingly adapts to noise rather than signal. Figures 9, 10, 11 illustrate this behavior along the gradient descent trajectory.

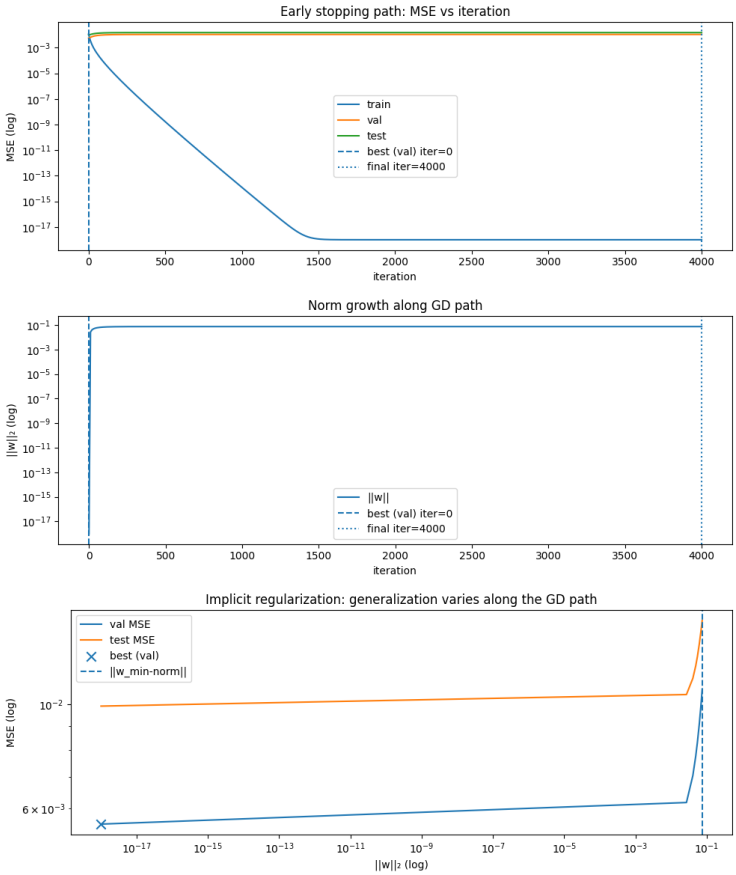


Figure 8: **Early stopping and implicit regularization along the gradient descent path.** *Top:* Training, validation, and test MSE versus iteration. *Middle:* Parameter norm $\|w_t\|_2$ versus iteration. *Bottom:* Validation and test MSE versus $\|w_t\|_2$ along the optimization trajectory.

As shown in Figure 9, training mean squared error decreases rapidly toward zero, while validation and test errors remain nearly flat. In this run, the best validation performance occurs at the beginning of training ($t^\star = 0$), indicating that further optimization does not improve out-of-sample performance.

Figure 10 shows the evolution of the parameter norm $\|w_t\|_2$ as a function of iteration. Starting from zero, the norm increases quickly and then plateaus at the value corresponding to the fully converged interpolating solution. This highlights that continued optimization mainly increases model complexity, even when predictive performance does not improve.

Figure 11 plots validation and test MSE directly against the parameter norm along the optimization path. This view makes the implicit regularization effect explicit: as gradient descent moves from low-norm to high-norm solutions, generalization performance deteriorates. The selected early-stopping iterate corresponds to a much smaller norm than the fully converged interpolator, which achieves lower training error but worse out-of-sample performance.

Overall, this experiment shows that in low signal-to-noise financial prediction tasks, optimization dynamics play a critical role in determining generalization. Even when interpolation is possible, continued training can mainly fit noise rather than signal, and early stopping acts as an implicit regularizer by limiting the complexity of the fitted model.

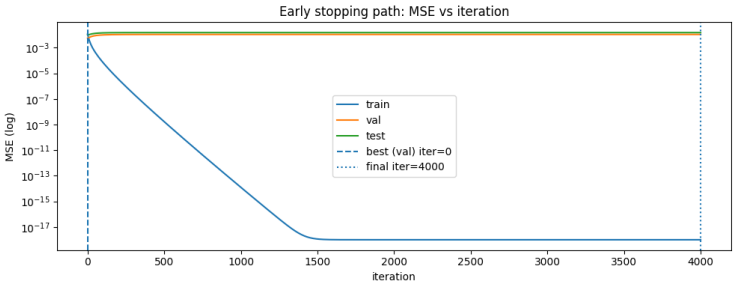


Figure 9: **Training, validation, and test MSE vs iteration.** Training error decreases rapidly toward zero, while validation and test errors remain nearly flat. The dashed vertical line marks the best validation iteration $t^\star$.

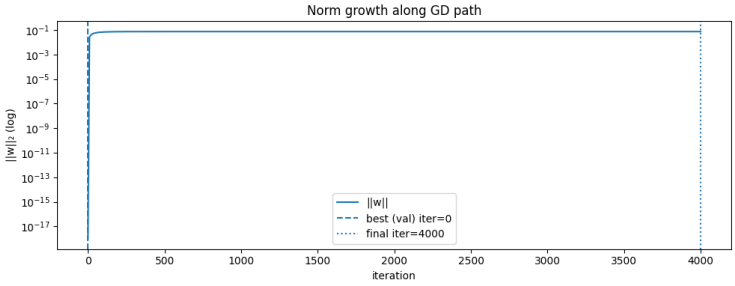


Figure 10: **Parameter norm vs iteration.** The ℓ_2 norm $\|w_t\|_2$ increases quickly from zero and plateaus at the value corresponding to the fully converged interpolating solution.

6 Effect of Label Noise on Interpolating Solutions

Benign overfitting theory emphasizes that the structure and level of noise in the labels can strongly affect generalization. To conclude, we test robustness to label noise by injecting additional Gaussian noise into the training labels:

$$\tilde{y} = y + \sigma\xi, \qquad \xi \sim \mathcal{N}(0, I_n), \tag{8}$$

and fit the min-norm interpolator in a fixed overparameterized regime (with d large, e.g. $d = 3n$). Evaluation is performed on clean validation and test labels, allowing us to link changes in performance to corruption in the training labels.

Why is the interpolating solution affected by label noise? From (6), the minimum-norm solution depends linearly on the labels through $(XX^\top)^{-1}y$. As a result, additional noise in y is propagated into the fitted parameters and can be amplified when the design is poorly conditioned. In

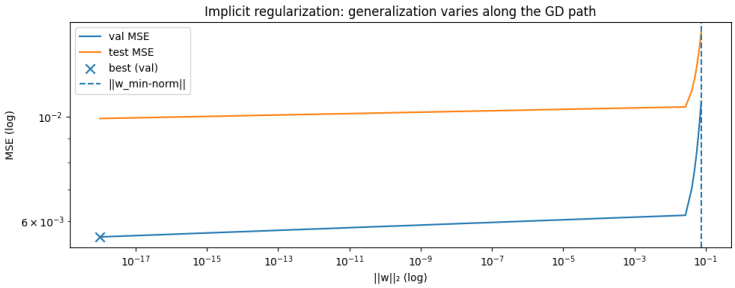


Figure 11: **Validation and test MSE vs parameter norm.** Plotting error against $\|w_t\|_2$ along the gradient descent path shows that generalization degrades as the norm increases. The early-stopped iterate corresponds to a substantially smaller norm than the fully converged interpolator.

benign regimes, generalization may remain relatively stable despite interpolation; in less favorable regimes, test performance is expected to degrade rapidly as the noise level σ increases.

Figures 12, 13 shows the empirical effect of increasing label noise. As σ grows, test MSE on clean labels increases, even though the model continues to interpolate the noisy training data. At the same time, the ℓ_2 norm of the fitted parameters increases and becomes more variable, indicating that the interpolating solution is increasingly dominated by noise rather than signal.

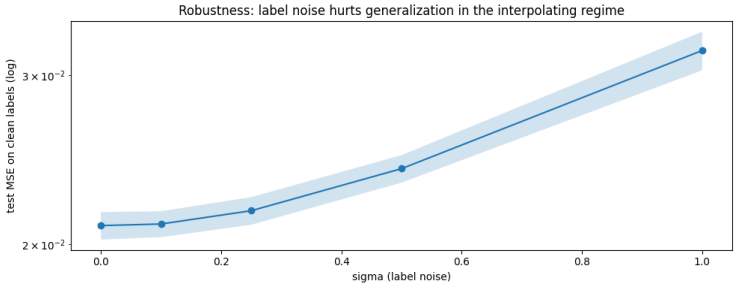


Figure 12: **Effect of label noise on test performance.** Test MSE on clean labels as a function of the training-label noise level σ. Shaded bands indicate variability across random seeds.

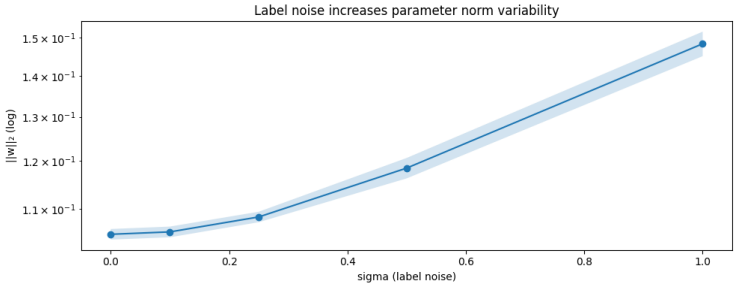


Figure 13: **Effect of label noise on parameter norm.** $\|\hat{w}\|_2$ as a function of the training-label noise level σ. Shaded bands indicate variability across random seeds.

In short, interpolation can remain benign only when noise amplification is limited. As label noise increases, the min-norm interpolator develops large coefficients and generalizes poorly, despite achieving zero training error.

7 Conclusion

We studied benign overfitting and double descent in a controlled linear regression setting using a U.S. equities panel. By appending irrelevant Gaussian features to a low-dimensional set of real predictors, we swept the model dimension across the interpolation threshold and evaluated how generalization changes with d. We observed the characteristic test-error peak near $d \approx n$ and linked it to poor conditioning and norm inflation of the fitted interpolator. This provides empirical support for the mechanism emphasized in theoretical analyses of ridgeless regression [3, 1].

We also demonstrated that interpolation alone is insufficient: by adding nullspace components to the minimum-norm interpolator, we constructed alternative interpolators with the same (near-zero) training error but significantly worse test performance. Finally, we examined implicit regularization via gradient descent and early stopping. Using a learning-rate bound based on the spectral norm of the training design, we traced the optimization path and showed how generalization varies as the parameter norm increases. In this low signal-to-noise financial setting, the best validation performance can occur extremely early, consistent with the idea that continued fitting rapidly adapts to noise. Adding label noise further degraded performance and increased variability, emphasizing that benign overfitting depends closely on noise structure and spectral properties.

Limitations and interpretation. Financial panel data do not satisfy many of the simplifying assumptions used in theory, such as independent and identically distributed samples, independent noise, or stationarity. We therefore interpret our results as showing that the studied phenomena (e.g., conditioning effects, norm inflation, and sensitivity to noise) can also be observed in practice, rather than as a direct validation of theoretical results.

References

[1] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[2] 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*, 2019.

[3] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 2022.