Limits To (Machine) Learning

Zhimin Chen, Bryan Kelly, and Semyon Malamud*

This version: January 26, 2026

Abstract

Machine learning (ML) methods are highly flexible, but their ability to approximate the true data-generating process is fundamentally constrained by finite samples. We characterize a universal lower bound, the Limits-to-Learning Gap (LLG), quantifying the unavoidable discrepancy between a model's empirical fit and the population benchmark. Recovering the true population R^2, therefore, requires correcting observed predictive performance by this bound. Using a broad set of variables, including excess returns, yields, credit spreads, and valuation ratios, we find that the implied LLGs are large. This indicates that standard ML approaches can substantially understate true predictability in financial data. We also derive LLG-based refinements to the classic Hansen and Jagannathan (1991) bounds, analyze implications for parameter learning in general-equilibrium settings, and show that the LLG provides a natural mechanism for generating excess volatility.

Keywords: machine learning, asset pricing, predictability, big data, limits to learning, excess volatility, stochastic discount factor, kernel methods
JEL: C13, C32, C55, C58, G12, G17

*Zhimin Chen is at Nanyang Technological University. Bryan Kelly is at AQR Capital Management, Yale School of Management, and NBER. Semyon Malamud is at the Swiss Finance Institute, EPFL, and CEPR, and is a consultant to AQR. Semyon Malamud gratefully acknowledges the financial support of the Swiss Finance Institute and the Swiss National Science Foundation, Grant 100018_192692. Zhimin Chen gratefully acknowledges the financial support of the Singapore MOE AcRF Tier 1 Grant #024891-00001. We are grateful to Mikhail Chernov, Andreas Fuster, Andreas Neuhierl, and Junwei Yang for excellent comments and suggestions. The implementation code is available at https://github.com/czm319319/CKM_LLG. AQR Capital Management is a global investment management firm that may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR.

Conflict-of-interest disclosure statement

Zhimin Chen

I have nothing to disclose.

Bryan Kelly

Bryan Kelly has received income from AQR Capital Management exceeding $10,000 over the past three years. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR.

Semyon Malamud

Semyon Malamud has received income from AQR Capital Management exceeding $10,000 over the past three years. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR.

1 Introduction

The modern approach to assessing whether an economic variable y is predictable from a set of variables X is to train a Machine Learning (ML) model and evaluate its out-of-sample performance. In practice, this approach often concludes that predictability is low or nonexistent. We show that such conclusions can be misleading. In high-dimensional, richly parameterized ML settings, estimators do not converge to the ground truth even in large samples, leading standard out-of-sample metrics to dramatically understate the true (population) degree of predictability. Our objective in this paper is to provide a simple and practical correction, thereby opening a new avenue for testing predictive relationships.

Classical econometric theory is designed for data-rich environments in which the number of parameters is small relative to the sample size. In such settings, the law of large numbers and central limit theorems ensure that estimators rapidly approach the true model.[1] Modern ML applications, however, operate in a fundamentally different regime. When the number of parameters is comparable to or exceeds the number of observations, classical asymptotics no longer guarantees consistency.

Under the data-generating process

$$y_{t+1} \;=\; f_t \;+\; \varepsilon_{t+1}, \tag{1}$$

modern ML estimators $\hat{f}_t$ typically exhibit substantial estimation error, which leads to large out-of-sample (OOS) Mean Squared Error (MSE):

$$MSE_{OOS} \;=\; \underbrace{E[(y_{t+1} - \hat{f}_t)^2]}_{out-of-sample\ error} \;=\; \underbrace{E[(f_t - \hat{f}_t)^2]}_{estimation\ error} + \underbrace{\sigma_\varepsilon^2}_{irreducible\ noise}, \tag{2}$$

where $\sigma_\varepsilon^2 = E[\varepsilon_{t+1}^2]$. In this paper, we derive a lower bound on the estimation error of any

[1] In the presence of model misspecification, estimators converge to a pseudo-true parameter; see White (1996).

linear estimator:

$$E[(f_t - \hat{f}_t)^2] \geq \sigma_\varepsilon^2 \widehat{\mathcal{L}}, \tag{3}$$

where $\widehat{\mathcal{L}}$ is the *Limits-to-Learning Gap (LLG)*, a fully data-driven lower bound that depends only on the observed sample and can be computed without estimating any predictive model.

LLG has direct implications for measured predictability. It yields a sharp lower bound for the true (population) R^2 (henceforth, R^2_*) in terms of the realized out-of-sample R^2 (henceforth, R^2_{OOS}). Namely, we show that

$$R^2_* \geq \frac{R^2_{OOS} + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}}. \tag{4}$$

We also derive a Central Limit Theorem (CLT), allowing us to construct a one-sided confidence interval for R^2_*.

Our bound provides a practical diagnostic tool. A researcher observing a large OOS MSE might conclude that there is no predictability. But if $\widehat{\mathcal{L}}$ is large, the true R^2_* may be far higher than the raw R^2_{OOS} suggests—meaning that substantial predictability exists in principle, even if standard ML models fail to uncover it. LLG, therefore, helps distinguish between cases where predictability is genuinely absent and those where existing models lack the capacity or architecture to detect it. It can thus guide variable selection and model design.

We provide extensive simulation evidence showing that our lower bound (CLT-adjusted for finite-sample error) closely approximates the true R^2_* from below. We then apply our framework to the classical dataset of Welch and Goyal (2008) to quantify the degree of predictability in US market returns and a broad set of financial indicators, including valuation ratios, bond yields, and spreads. We find that LLG is often large, even in settings traditionally viewed as weakly predictable. For example, although the R^2_{OOS} for market returns is

negative in our exercises, the LLG implies that the true population predictability is at least 20%. This result stands in sharp contrast to the 1–2% R^2_{OOS} achieved by state-of-the-art ML models at the monthly horizon (see Kelly et al. (2024); Kelly and Malamud (2025)), and poses a significant challenge for standard asset-pricing models. Likewise, LLG corrections indicate that the AR(1) residuals of many valuation ratios and spreads are more than 30% predictable, even though conventional linear ML methods fail to detect this structure.

These findings suggest that widely used ML methods may substantially understate predictability in macroeconomic and asset-return data. When the lower bound for R^2_* is large, more flexible econometric or structural approaches may be capable of recovering the underlying signal. As an illustration, we apply nonlinear models with variable selection to forecast each of the Welch and Goyal (2008) variables. For several targets with a high lower bound (4), more complex models uncover meaningful out-of-sample predictability. For instance, the lower bound implies that AR(1) residuals of the Treasury Bill rate are 10% predictable, and our best nonlinear specification indeed attains an R^2_{OOS} of 10%.

Our results also speak to theories of learning in financial markets. A growing literature, including Collin-Dufresne et al. (2016) and Farmer et al. (2024), shows that parameter learning can be extremely slow, challenging full-information rational expectations. We derive an LLG-adjusted version of the Hansen and Jagannathan (1991) bound that quantifies the role of parameter uncertainty in shaping asset prices.

In a simple equilibrium model with high-dimensional learning, we show that the LLG associated with predicting asset fundamentals naturally generates excess price volatility. This mechanism implies that rational overreaction to noisy signals can replicate the classic excess-volatility puzzle of Shiller (1981) without invoking behavioral departures from rationality.

Finally, our framework has implications for the economics of large-scale ML systems more broadly. Substantial computational and environmental resources are being devoted to scaling up ML models, yet empirical scaling laws (Kaplan et al., 2020) show sharply diminishing

returns, with convergence rates often following T^{α} for α near 0.1, far below the classical $\sqrt{T}$ rate. Our results show that LLG explains this slowdown: the limits-to-learning gap declines only slowly with sample size. By quantifying this gap, LLG provides guidance on when additional data or computation is likely to generate meaningful improvements and clarifies the fundamental limits of what ML can extract from finite data.

2 Literature Review

Machine learning methods have delivered remarkable performance in estimation, prediction, and portfolio optimization tasks within high-dimensional environments. See, for example, Jurado et al. (2015), Kleinberg et al. (2018), Gu et al. (2020), Bianchi et al. (2021), Cong et al. (2021), Bianchi et al. (2022), Fan et al. (2022), Kaniel et al. (2023), Fuster et al. (2022), Liao et al. (2023), Chen et al. (2024b), Kelly et al. (2024), Didisheim et al. (2024), and Chernov et al. (2025). However, the high complexity of these models renders them prone to overfitting and bias. A substantial literature, therefore, has focused on debiasing techniques (Chernozhukov et al., 2018, 2019) and on establishing technical conditions under which ML estimators achieve root-T consistency (Chen and White, 1999; Schmidt-Hieber, 2020; Kohler and Langer, 2021; Liao et al., 2025).

A more recent line of research highlights fundamental statistical limits to the ability of economic agents to learn high-dimensional models. For example, Da et al. (2022) show that investors may be unable to efficiently exploit the full cross-section of alphas, while Martin and Nagel (2021) demonstrate that strong in-sample predictability may coexist with little or no out-of-sample performance. Didisheim et al. (2024) and Kelly and Malamud (2025) refer to this phenomenon as "limits-to-learning." In this paper, we formalize and quantify these insights by introducing a tractable, easy-to-estimate lower bound on model performance—the Limits-to-Learning Gap (LLG). We show that our estimator is super-consistent and apply it to a range of widely studied ML models. Using both simulated and real data, we find that

LLG can be substantial even in models that incorporate sparsity-inducing structures, such as sparse neural networks (Chen and White, 1999; Schmidt-Hieber, 2020). In particular, we show how to compute LLG explicitly for any neural network trained via gradient descent, leveraging the recent empirical findings on the equivalence between neural networks and NTK-based ridge regressions. See Fort et al. (2020); Atanasov et al. (2021); Lauditi et al. (2025); Schwab et al. (2025). Our results also contribute to the ML literature examining the subtle relationship between overfitting and out-of-sample performance (Muthukumar et al., 2020; Holzmüller, 2020; Belkin et al., 2020; Ghosh and Belkin, 2023; Mohsin et al., 2025; Malach et al., 2025).

The LLG emerges due to the curse of dimensionality. The common way of dealing with this curse in modern econometrics is to assume sparsity of the underlying true model. However, recent theoretical and empirical evidence (see, e.g., Giannone et al. (2021); Avramov et al. (2025)) suggests that sparsity can be difficult to detect in economic datasets. Moreover, even when sparsity is present, dense models tend to outperform sparse ones in low signal-to-noise ratio environments (Shen and Xiu, 2024). Inference in dense, high-dimensional models, however, poses formidable challenges, as classical tools tend to fail in such settings. As Belloni et al. (2018) emphasize: *"Most of the work in the recent literature on high-dimensional estimation and inference relies on approximate sparsity to provide dimension reduction and the corresponding use of sparsity-based estimators. Dense models are appealing in many settings and may be usefully employed in more moderate-dimensional settings."* Our paper advances this literature by constructing pivotal confidence bands for key nuance parameters, σ_ε^2 and R_*^2, in dense, high-dimensional environments.

Our findings also relate to the notion of "dark matter" in asset pricing introduced by Chen et al. (2024a), who quantify the extent to which asset pricing models rely on latent, unobservable structure inferred from internal model restrictions. A large LLG implies that much of the underlying structure of the data is effectively unlearnable from finite samples—even if it

is theoretically exploitable—forcing ML models to rely on noisy approximations or implicitly inferred components, much like dark matter. This connection highlights how statistical limits to learning may underlie model fragility and overfitting, echoing the concerns raised in Chen et al. (2024a).

3 Limits to Learning

3.1 Environment

We consider the problem of predicting an economic variable y_{t+1} based on a (potentially) high-dimensional vector of signals $S_t \in \mathbb{R}^P$. The following assumption summarizes the basic properties of the data-generating process. While we focus our analysis on linear estimators, in Section 3.4 we demonstrate how our results naturally extend to nonlinear models, including kernel regressions and deep neural networks.

Assumption 1 *We have*

$$y_{t+1} \;=\; f_t \;+\; \varepsilon_{t+1}\;,\; t=0,\cdots,T+1\,, \tag{5}$$

for some f_t, where $E[\varepsilon_{t+1}]=0,\ E[\varepsilon_{t+1}^2]=\sigma_\varepsilon^2,\ E[\varepsilon_{t+1}^4]<\infty$ are independent and identically distributed, and are also independent of $(f_t,S_t)_{t\geq 0}$. Below, we frequently use the convenient matrix notation $y=(y_\tau)_{\tau=1}^T\in\mathbb{R}^T$, and $S=(S_\tau)_{\tau=0}^{T-1}\in\mathbb{R}^{T\times P}$ for the in-sample (training) data, and we use time $T+1$-data as the out-of-sample realization.

Importantly, no assumptions are made whatsoever about the nature of the joint dynamics of signals S and the predictable component f. In particular, this assumption allows for any form of non-stationarity, autocorrelation, heteroskedasticity, or model misspecification.

By (5), the second moment of y_{t+1} admits the standard decomposition

$$E[y_{t+1}^2] \;=\; \underbrace{E[f_t^2]}_{explained\ variance} \;+\; \underbrace{\sigma_\varepsilon^2}_{irreducible\ noise}\,. \tag{6}$$

Given an estimator $\hat{f}_t$, we are interested in the behavior of the expected out-of-sample error

$$MSE(\hat{f}) \;=\; E[(y_{T+1} - \hat{f}_T)^2]\,, \tag{7}$$

the corresponding R^2,

$$R^2(\hat{f}) \;=\; 1 - \frac{MSE(\hat{f})}{E[f_t^2] + \sigma_\varepsilon^2}\,, \tag{8}$$

and its relationship with the *infeasible* R^2 given by

$$R_*^2 \;=\; 1 \;-\; \frac{\sigma_\varepsilon^2}{E[f_t^2] \;+\; \sigma_\varepsilon^2} \tag{9}$$

and achievable by an econometrician who knows the true f_t. Our goal is to study limits-to-learning, defined as the gap between R_*^2 and $R^2(\hat{f})$.

3.2 Linear Estimators

Our focus in this paper is on linear estimators, admitting a form

$$\hat{f}_T \;=\; \mathcal{K}(S_T, S)\, y, \tag{10}$$

where $\mathcal{K}(S_T, S) \in \mathbb{R}^{1\times T}$ is a vector function, with $(\mathcal{K}(S_T, S))_t$ describing how much attention our estimator is paying to a particular observation at time t. A canonical example is the

ridge-regularized least-squares estimator,

$$\hat{f}_T \;=\; \underbrace{\left(S_T'(zI \;+\; \hat{\Psi})^{-1} \frac{1}{T} S' \right)}_{\mathcal{K}(S_T,S)} y\,, \tag{11}$$

where

$$\hat{\Psi} \;=\; \frac{1}{T}\sum_{\tau=0}^{T-1} S_\tau S_\tau' \;=\; \frac{1}{T} S'S \tag{12}$$

is the sample covariance matrix of signals. This simple class of estimators has become a workhorse theoretical framework for understanding the behavior of high-dimensional machine learning models[2] and has been shown to perform well even in environments with low signal-to-noise ratios (Shen and Xiu, 2024). Another, closely related linear estimator is a kernel ridge regression: given a positive definite kernel $k(\cdot,\cdot)$,[3] we define

$$\hat{f}_T \;=\; \underbrace{k(S_T,S)(zI + k(S,S))^{-1}}_{\mathcal{K}(S_T,S)} y\,. \tag{13}$$

In Appendix F, we show that when $y_{t+1} = \beta' S_t + \varepsilon_{t+1}$ and β is sampled at time zero so that β_i are i.i.d. and $E[\beta_i] = 0,\ E[\beta_i^2] = \sigma_\beta^2$, then the linear ridge regression estimator with $z = \dfrac{\sigma_\varepsilon^2}{\sigma_\beta^2}\dfrac{P}{T}$ is Bayes-optimal; that is, *no other ML model (linear or nonlinear) can achieve better performance than ridge.* As a result, our LLG bound applies to all ML models in this setting. We also show how our results extend to a larger class of distributions with arbitrary covariance structures.

[2] Hastie et al. (2019) explain how it approximates gradient descent in the "lazy training" regime, while Jacot et al. (2018) show that very wide neural nets can be closely approximated by high-dimensional linear ridge regressions with signals given by the gradients of the Neural Network (the so-called Neural Tangent Kernel (NTK)). Recently, a sequence of papers (Fort et al., 2020; Atanasov et al., 2021; Lauditi et al., 2025; Schwab et al., 2025) has made a surprising discovery that, in fact, the prediction of any neural net can be closely approximated by a kernel ridge regression. We use this discovery in Section 3.4 to compute LLG for any neural network.

[3] One popular choice is the Gaussian kernel, $k(x_1, x_2) = e^{-\|x_1 - x_2\|^2}$. See, e.g., Kelly et al. (2024); Kelly and Malamud (2025).

The following result presents a useful decomposition of the MSE (7). Given a partition of the data into T in-sample observations and T_{OOS} out-of-sample (OOS) observations, we define $E_{OOS}[X] = \frac{1}{T_{OOS}} \sum_{t=T}^{T+T_{OOS}-1} X_{t+1}$, and let MSE_{OOS} be the realized out-of-sample MSE,

$$MSE_{OOS}(\hat{f}) \;=\; \frac{1}{T_{OOS}} \sum_{t=T}^{T+T_{OOS}-1} (y_{t+1} - \hat{f}_t)^2 \,, \tag{14}$$

and let

$$R^2_{OOS}(\hat{f}) \;=\; 1 \;-\; \frac{MSE_{OOS}(\hat{f})}{E_{OOS}[y^2]} \tag{15}$$

be the realized out-of-sample R^2. The linearity of the estimator $\hat{f}_t$ leads to the natural decomposition,

$$\hat{f}_t \;=\; \hat{f}^s_t \;+\; \hat{f}^\varepsilon_t \;=\; \underbrace{\mathcal{K}(S_t, S) f}_{signal} \;+\; \underbrace{\mathcal{K}(S_t, S)\varepsilon}_{noise} \tag{16}$$

for each OOS time period t. Substituting this decomposition into (14) leads to the following result.

Proposition 1 *For the linear estimator, we have*

$$MSE_{OOS}(\hat{f}) \;=\; E_{OOS}[\varepsilon^2] \;+\; \widehat{\mathcal{B}} \;+\; \widehat{\mathcal{I}} \;+\; \widehat{\mathcal{V}} \,, \tag{17}$$

where

$$\begin{aligned} \widehat{\mathcal{B}} \;&=\; \underbrace{E_{OOS}[(f - \hat{f}^s)^2]}_{bias} \;\geq\; 0 \\ \widehat{\mathcal{V}} \;&=\; \underbrace{E_{OOS}[(\hat{f}^\varepsilon)^2]}_{variance} \;\geq\; 0, \end{aligned} \tag{18}$$

while

$$\widehat{\mathcal{I}} \;=\; \underbrace{-2E_{OOS}[(f-\hat{f}^s)(\hat{f}^\varepsilon-\varepsilon)+\varepsilon\hat{f}^\varepsilon]}_{interaction} \tag{19}$$

is the interaction term.

Standard law-of-large numbers arguments imply that the interaction term, $\widehat{\mathcal{I}}$, is asymptotically negligible, vanishing at the rate $O(T^{-1/2}+T_{OOS}^{-1/2})$. However, the terms $\widehat{\mathcal{B}}$, $\widehat{\mathcal{V}}$ are always nonnegative and do not vanish, creating limits-to-learning. Estimating the true bias term $\widehat{\mathcal{B}}$ in high-dimensional settings is highly non-trivial and requires novel techniques and additional assumptions. Ignoring this term, we get the lower bound for the MSE:

$$MSE_{OOS}(\hat{f}) \;\geq\; \sigma_\varepsilon^2 \;+\; \widehat{\mathcal{V}} \;+\; O(T^{-1/2})\,. \tag{20}$$

Let $\widehat{\mathcal{K}}=(\mathcal{K}(S_\tau,S))_{\tau=T}^{T+T_{OOS}-1}\in\mathbb{R}^{T_{OOS}\times T}$. Our key observation is that $\widehat{\mathcal{V}}$ can be written down as a *quadratic form,*

$$\widehat{\mathcal{V}} \;=\; \frac{1}{T}\,\varepsilon'\left(\frac{T}{T_{OOS}}\widehat{\mathcal{K}}'\widehat{\mathcal{K}}\right)\varepsilon \underbrace{\approx}_{Law\ of\ Large\ Numbers} \frac{1}{T}\,\sigma_\varepsilon^2\ \mathrm{tr}\left(\frac{T}{T_{OOS}}\widehat{\mathcal{K}}'\widehat{\mathcal{K}}\right). \tag{21}$$

Making this argument rigorous, we arrive at the main result of this section.

Theorem 2 *Suppose that, in the limit as* $T, T_{OOS}\to\infty$, [To Do: update the assumptions once Semyon confirms.]

$$\lim\frac{1}{T_{OOS}^2}\,\mathrm{tr}\,E\Big[(\widehat{\mathcal{K}}'\widehat{\mathcal{K}})^2\Big] \;=\; \lim\frac{1}{T_{OOS}^2}E\Big[E_{OOS}[\|(f-\hat{f}^s)\|^2]\Big] \;=\; \lim\frac{1}{T_{OOS}^2}E\Big[E_{OOS}[\|(f-\hat{f}^s)\widehat{\mathcal{K}}\|^2]\Big] \;=\; 0\,. \tag{22}$$

Then, the out-of-sample MSE from (14) satisfies

$$\liminf \frac{MSE_{OOS}(\hat{f})}{1+\widehat{\mathcal{L}}} \;\geq\; \underbrace{\sigma_\varepsilon^2}_{infeasible} \tag{23}$$

in probability,[4] *where*

$$\widehat{\mathcal{L}} \;=\; \frac{1}{T_{OOS}} \operatorname{tr}\left(\widehat{\mathcal{K}}'\widehat{\mathcal{K}}\right) \tag{24}$$

is the Limits-to-Learning Gap (LLG). This bound turns into an identity if $f_t = 0$.

Suppose now that $\lim_{T_{OOS}\to\infty} E_{OOS}[f^2] = E[f_t^2]$.[5] Then, $\lim E_{OOS}[y^2] = E[f_t^2] + \sigma_\varepsilon^2$ and we can rewrite (23) as

$$R_*^2 \;\geq\; \limsup \frac{R_{OOS}^2(\hat{f}) + \widehat{\mathcal{L}}}{1+\widehat{\mathcal{L}}}\,, \tag{25}$$

or, equivalently, (25) as

$$\underbrace{R_*^2 - R_{OOS}^2(\hat{f})}_{the\ gap} \geq \widehat{\mathcal{L}}(1 - R_*^2) \;+\; o(1)\,. \tag{26}$$

Thus, one can see explicitly how $\widehat{\mathcal{L}}$ also controls the gap between the infeasible and feasible R^2. This gap emerges because, in high-dimensional settings, regression overfits noise. Perhaps the most surprising implication of Theorem 2 is that LLG *only depends on the signals* S, and is completely independent of the structure of f_t. Thus, the nature of the dependent variable y_{t+1} is irrelevant: The same LLG will emerge no matter what stands on the left-hand side of the regression.

What defines the size of $\widehat{\mathcal{L}}$? The underlying mechanisms are particularly explicit for the

[4] We say that $X_T \geq Y_T$ holds in probability if $\lim_{T,T_{OOS}\to\infty} \text{Prob}(X_T < Y_T) = 0$.
[5] This holds, for example, if f_t^2 is stationary ergodic.

ridge estimator (11). In this case, by direct calculation, we have

$$\frac{T}{T_{OOS}}\widehat{\mathcal{K}}'\widehat{\mathcal{K}} \;=\; \frac{1}{T}S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S' \tag{27}$$

where $\hat{\Psi}_{OOS} = E_{OOS}[SS']$ is the out-of-sample signal covariance matrix and the technical condition of Theorem 2 holds when $E[\|E_T[\hat{\Psi}^2_{OOS}]\|] = o(\min\{T_{OOS}, T\})$ as $T_{OOS} \to \infty$. Then,

$$\widehat{\mathcal{L}}(z) \;=\; \frac{1}{T}\operatorname{tr}\Big(\hat{\Psi}_{OOS}\hat{\Psi}(zI+\hat{\Psi})^{-2}\Big)\,. \tag{28}$$

Here, $\hat{\Psi}_{OOS}\hat{\Psi}(zI+\hat{\Psi})^{-2}$ is a $P\times P$ matrix, and, hence, its trace grows approximately like P, so that $\widehat{\mathcal{L}} = O(c)$, where we have defined statistical complexity $c = \dfrac{P}{T}$ as in Kelly et al. (2024). When P is negligible relative to T, the LLG vanishes. By contrast, in the over-parametrized regime when $P > T$, $\widehat{\mathcal{L}}$ can get very large, creating significant limits to learning.

3.3 Asymptotic Normality

In this subsection, we establish asymptotic normality of our estimator, allowing us to compute confidence bands for the infeasible R^2_*. Without imposing additional technical conditions on f_t, we cannot determine the estimation error of its second moment, $E[f_t^2] - E_{OOS}[f^2]$. For this reason, instead of R^2_*, we work with a slightly modified object,

$$\tilde{R}^2_* \;=\; 1 \;-\; \frac{\sigma^2_\varepsilon}{E_{OOS}[f^2]+\sigma^2_\varepsilon} \tag{29}$$

The following is true.

Theorem 3 (Confidence Interval for The Infeasible R^2) *Suppose that* $E[\varepsilon_t^3] = 0$, $E[\varepsilon_t^4] =$

3. *Then,*

$$\frac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}}\,, \tag{30}$$

is a $T^{1/2}$*-consistent lower bound for* $\tilde{R}^2_*$ *in the following sense: The event* $\dfrac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}(z)}{1 + \widehat{\mathcal{L}}(z)} > R^2_*$ *occurs with vanishing probability:*

$$\limsup_{T,T_{OOS}\to\infty} \text{Prob}\left(\frac{T^{1/2}\left(\tilde{R}^2_* - \dfrac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}}\right)}{\sigma_{R^2}} < \alpha\right) \;\leq\; \Phi(\alpha) \tag{31}$$

for any $\alpha \leq 0$*, where* $\Phi(\cdot)$ *is the c.d.f. of the standard normal distribution. There exists an asymptotically consistent, pivotal estimator* $\hat{\sigma}_{R^2} = \hat{\sigma}_{R^2}(y, S)$ *of* σ_{R^2} *presented in the Appendix.*[6]

The proof of Theorem 3 is non-trivial. A key challenge is that σ_{R^2} is expressed in terms of the unobservable and impossible to estimate f_t and σ^2_ε. Overcoming this challenge requires some techniques that we develop in the Appendix D.

3.4 Limits to Learning for (Deep) Neural Networks

Deep Neural Networks (DNN) are complex, nonlinear families of functions, $f(x;\theta)$, where the parameter vector $\theta \in \mathbb{R}^P$ is typically very high-dimensional and $x \in \mathbb{R}^d$. Given some input data $X_t \in \mathbb{R}^d$ we define the in-sample MSE

$$\mathcal{L}(\theta) \;=\; \frac{1}{T}\sum_{t=0}^{T-1}(y_{t+1} - f(X_t;\theta))^2\,. \tag{32}$$

[6]See Theorem D.2. The expression for $\hat{\sigma}_{R^2}$ is too long for the main text.

Then, we pick a learning rate η, and the parameter vector θ is optimized recursively to achieve a low $\mathcal{L}(\theta)$ via gradient descent:

$$\theta_{e+1} \;=\; \theta_e \;-\; \eta\,\mathcal{L}(\theta_e),\;\; e=0,\cdots,\mathcal{E}\,. \tag{33}$$

Here, the number of gradient descent steps $\mathcal{E}$ is commonly referred to as the *number of epochs.* The prediction of the DNN is then given by $f(x;\theta_{\mathcal{E}})$. A surprising discovery made recently in a sequence of papers is that, in fact, the prediction of the DNN is closely approximated by the linear regression with signals

$$S_t \;=\; \nabla_\theta f(X_t;\theta_{\mathcal{E}}) \in \mathbb{R}^P\,. \tag{34}$$

See, e.g., Jacot et al. (2018); Fort et al. (2020); Geiger et al. (2020); Liu et al. (2020); Atanasov et al. (2021); Vyas et al. (2022); Lauditi et al. (2025); Schwab et al. (2025). Here, we appeal to the powerful result from Yang and Littwin (2021) showing that wide DNNs trained by gradient descent are approximately linear in y.

Proposition 4 (LLG for Wide Neural Networks) *Under the hypothesis of Yang and Littwin (2021), in the limit of large width, the prediction of a DNN trained by gradient descent is approximately linear in y with the $\mathcal{K}(S_T,S)$ in 10 that admits an analytical expression in terms of (34); hence, the LLG can be computed for the DNN using (24).*

4 Implications of Limits-to-Learning for Models of Dynamic Economies

4.1 Hansen-Jagannathan Bounds

Suppose that $y_{t+1} = R_{t+1}$ is the excess return on a security (e.g., a market index such as the SP500). In this case, we can define the timing strategy

$$R^{\pi}_{t+1} \;=\; \pi_t R_{t+1}\,, \tag{35}$$

where π_t is the conditionally efficient portfolio,

$$\pi_t \;=\; \frac{E_t[R_{t+1}]}{\mathrm{Var}_t[R_{t+1}]} \;=\; \frac{f_t}{\sigma^2_\varepsilon} \tag{36}$$

with the conditional squared Sharpe Ratio

$$\frac{E_t[R^{\pi}_{t+1}]^2}{\mathrm{Var}_t[R^{\pi}_{t+1}]} \;=\; f_t^2\sigma_\varepsilon^{-2} \tag{37}$$

satisfying (see (9))

$$E[f_t^2\sigma_\varepsilon^{-2}] \;=\; \frac{R^2_*}{1-R^2_*}\,. \tag{38}$$

This identity creates a direct link between the Sharpe ratio (a measure of economic significance) and R^2_* (a measure of statistical significance).[7] We now combine this intuition with Theorem 2 to derive implications of LLG for asset pricing.

Most equilibrium asset pricing models imply that a stochastic discount factor (SDF) can be constructed from the Intertemporal Marginal Rate of Substitution (IMRS) of economic

[7] It is instructive to relate our analysis to Barillas and Shanken (2018). While Barillas and Shanken (2018) focus on comparing models, our approach instead identifies the potential of the best possible model and the gap relative to what is empirically achievable. In this sense, we provide a cardinal measure of an asset pricing model's performance.

agents. Such an SDF $\widetilde{\mathcal{M}}_t$ satisfies the pricing equation

$$E_t[\widetilde{\mathcal{M}}_{t+1}R^{\pi}_{t+1}] \;=\; 0\,, \tag{39}$$

and the Cauchy-Schwarz inequality implies that

$$\frac{\mathrm{Var}_t[\widetilde{\mathcal{M}}_{t+1}]}{E_t[\widetilde{\mathcal{M}}_{t+1}]^2} \;\geq\; f_t^2\sigma_\varepsilon^{-2}\,. \tag{40}$$

This is commonly known as the Hansen and Jagannathan (1991) bound. Taking the expectation of this inequality and combining it with Theorem 2, we arrive at the following result, assuming that $\lim_{T_{OOS}\to\infty} E_{OOS}[f^2] = E[f_t^2]$.

Proposition 5 (LLG for Hansen-Jagannathan bounds) *Consider an equilibrium asset pricing model with d_{t+1} satisfying Assumption 1 and where an economic agent know f_t in (5). Then, this agent's IMRS is a non-tradable SDF $\widetilde{\mathcal{M}}_t$ and it satisfies*

$$E\left[\frac{\mathrm{Var}_t[\widetilde{\mathcal{M}}_{t+1}]}{E_t[\widetilde{\mathcal{M}}_{t+1}]^2}\right] \;\geq\; \lim \sup_{T,T_{OOS}\to\infty} \frac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}}{1 - R^2_{OOS}(\hat{f})}\,. \tag{41}$$

Proposition 5 shows how limits-to-learning directly translate into a lower bound on the variance of the SDF in a standard complete-information equilibrium. As an illustration, suppose that y_{t+1} is the return on the CRSP value-weighted index. Although realized out-of-sample R^2 from predicting market returns is typically small—around 1–2% at the one-month horizon—we find that $\widehat{\mathcal{L}}$ can be large. Consequently, Proposition 5 poses a significant challenge for models driven by macroeconomic shocks, since such models often struggle to generate a sufficiently high $\mathrm{Var}_t[\widetilde{\mathcal{M}}_{t+1}]$. As we argue, this puzzle can be partially resolved by accounting for high-dimensional parameter learning.

The key caveat is the assumption that agents know f_t. The classic rational-expectations approach in modern dynamic stochastic general equilibrium models typically presumes that

agents know all parameters of the data-generating process. This assumption has recently faced growing criticism (Han et al., 2021; Moll, 2024; Moll and Ryzhik, 2025; Dou et al., 2025) because these models often contain a very large number of parameters, many of which are difficult to estimate, leading to extremely slow learning. See, for example, Collin-Dufresne et al. (2016) and Farmer et al. (2024). Hence, it is natural to assume that, just like econometricians, economic agents face the same limits-to-learning in richly parametrized environments.

Consider now an econometrician who has access to the same number T of observations as the economic agent in our equilibrium model. Alternatively, we may assume that the econometrician evaluates the SDF on a purely out-of-sample basis: Even with access to $\bar{T} > T$ return realizations ex-post, only $t \leq T$ are used for time-T estimation. Let $\nu_*(R)dR$ be the true (unobservable and unknown, neither to the econometrician nor to economic agents) distribution of returns, and $\nu_T(R)dR$ the agent's posterior distribution given the observations $t \leq T$. We also denote by I_{T+1} the agent's IMRS (e.g., the ratio of marginal utilities in a Lucas Jr (1978) model). Then, the agent's inter-temporal optimality condition gives the *subjective* pricing equation

$$\int \underbrace{\nu_T(R_{T+1})}_{subjective\ probabilities} \underbrace{E[I_{T+1}|R_{T+1}]}_{IMRS} R_{T+1} dR_{T+1} \;=\; 0\,. \tag{42}$$

To get the *objective* pricing equation, we need to define the SDF under the *objective probability distribution*:

$$\widetilde{\mathcal{M}}_{T+1} \;=\; \underbrace{\ell_T(R_{T+1})}_{objective\ likelihood} \underbrace{E[I_{T+1}|R_{T+1}]}_{IMRS}, \qquad \ell_T(R_{T+1}) \;=\; \underbrace{\frac{\nu_T(R_{T+1})}{\nu_*(R_{T+1})}}_{unobservable}\,. \tag{43}$$

Mechanically, $\widetilde{\mathcal{M}}_{T+1}$ satisfies (39). However, when P is large enough, $\nu_*(R_{T+1})$ can be extremely difficult to estimate. As a result, even with a fully specified and calibrated equilib-

rium model (for example, Cogley and Sargent (2008); Johannes et al. (2016); Collin-Dufresne et al. (2016)), the true SDF $\widetilde{\mathcal{M}}_{T+1}$ exists but remains inaccessible to the econometrician. Substituting (43) into (41), we arrive at the following result.

Proposition 6 *Consider an equilibrium asset pricing model with an agent who needs to learn the true parameters of the model based on past T observations and ends up with a posterior $\nu_T(R_{T+1})$. Then, this agent's SDF $\widetilde{\mathcal{M}}_{T+1}$ is given by (43) and, hence,*

$$\frac{\left(\text{Var}[\ell_T(R_{T+1})E[I_{T+1}|R_{T+1}]]\right)^{1/2}}{E[\ell_T(R_{T+1})E[I_{T+1}|R_{T+1}]]} \geq \limsup \frac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}}{1 - R^2_{OOS}(\hat{f})} \tag{44}$$

Proposition 6 implies that the bound (41) is not just about "macro shocks being not volatile enough," it is also about "the objective likelihood being too volatile." In the language of Chen et al. (2024a), this unobservable objective likelihood constitutes a form of "dark matter", explaining the gap between the volatility of the true SDF and the volatility of IMRS. LLG emphasizes that economic agents face exactly the same limits-to-learning as academic econometricians and, as such, allows us to establish a lower bound for the size of this dark matter. For example, it can be used to measure the presence (and size) of fat tails in $\nu_T(R_{T+1})$, which naturally lead to large $\text{Var}[\ell_T(R_{T+1})E[I_{T+1}|R_{T+1}]]$, as well as test (rational or behavioral) belief formation models such as those in Farmer et al. (2024).

4.2 Parameter Learning And Equilibrium Excess Volatility

One of the most important empirical regularities in financial markets is excess volatility (Shiller, 1981): The fact that prices move too much to be justifiable by the volatility of fundamentals. The common explanation proposed in the literature relies either on irrationality (over-reaction, De Bondt and Thaler (1985)) or on complex preferences (Bansal and Yaron, 2004). In this section, we argue that excess volatility can emerge in a simple equilibrium model with *risk-neutral* agents due to limits-to-learning.

Consider a simple model where stocks live for one period and pay dividends

$$y_{t+1} \;=\; \beta' S_t + \varepsilon_{t+1}\,. \tag{45}$$

We assume that agents do not know the true value of β and have a rational, Gaussian prior about it, $\beta \sim N(0, \frac{\sigma_\beta^2}{P} I_{P\times P})$. In this case, standard calculations (see Appendix F) imply that the agents' posterior mean of β after observing (S, y) is given by $\hat{\beta}_T(cz)$, where $z = \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}$ and

$$\hat{\beta}_T(cz) \;=\; (czI + \hat{\Psi})^{-1} S' y\,. \tag{46}$$

If all agents are risk neutral and share this common prior, prices are given by

$$Q_T \;=\; E_T[y_{T+1}] \;=\; \hat{\beta}_T(cz)' S_T\,. \tag{47}$$

We can now characterize price variance and its relationship to fundamental variance.

Proposition 7 *Suppose that $E[S] = 0$, $E[SS'] = \Psi$ and that S_t are i.i.d. across t. We have*

$$\begin{aligned} \mathrm{Var}[y_{T+1}] \;&=\; \beta'\Psi\beta \;+\; \sigma_\varepsilon^2 \\ \mathrm{Var}_T[Q_{T+1}] \;&=\; \hat{\beta}_T(cz)'\Psi\hat{\beta}_T(cz) \end{aligned} \tag{48}$$

In the limit as $T, P \to \infty$, $\frac{P}{T} \to c$, we have

$$\mathrm{Var}_T[Q_{T+1}] \;\geq\; \sigma_\varepsilon^2 \widehat{\mathcal{L}}, \tag{49}$$

where $\widehat{\mathcal{L}} = \frac{1}{T}\,\mathrm{tr}\Big(\Psi\hat{\Psi}(zcI + \hat{\Psi})^{-2}\Big)$ is the LLG from (28), computed with the true Ψ.

Proposition 7 offers an elegant economic interpretation of LLG: It is the excess volatility generated due to the rational over-reaction of economic agents to noise while learning from high-dimensional data. When $P > T$, $\hat{\Psi}$ has at least $P - T$ zero eigenvalues and, as a result, $\widehat{\mathcal{L}}$ typically explodes when z is small enough. By Proposition 7, this always leads to excess volatility. This offers a novel potential solution to the Shiller (1981) puzzle: Prices move too much because economic agents face complexity and limits-to-learning.

Computing $\widehat{\mathcal{L}}$ in Proposition 7 formally requires the knowledge of the unobservable, true covariance matrix Ψ. However, Random Matrix Theory (RMT) can be used to compute it under more stringent conditions on S_t.

Proposition 8 *Suppose that $S_t = \Psi^{1/2} X_t$ where $X_t = (X_{i,t})$ has i.i.d. coordinates $X_{i,t}$ with $E[X_{i,t}] = 0$, $E[X_{i,t}^2] = 1$. Then, $\widehat{\mathcal{L}} - \widetilde{\mathcal{L}}$ converges to zero almost surely, where*

$$\widetilde{\mathcal{L}} \;=\; \frac{\frac{1}{T}\operatorname{tr}\big((zcI + SS'/T)^{-2}\big)}{\Big(\frac{1}{T}\operatorname{tr}\big((zcI + SS'/T)^{-1}\big)\Big)^2} - 1\,. \tag{50}$$

Thus, $\widetilde{\mathcal{L}}$ is the Herfindahl index of the eigenvalues of the matrix $(zcI + \frac{SS'}{T})^{-1} \in \mathbb{R}^{T\times T}$. How can this Herfindahl index matter if the law of large numbers (LLN) implies that $\frac{SS'}{T} \to 0$? The reason behind this is statistical complexity. While the off-diagonal elements of $\frac{SS'}{T}$ (i.e., cross-time signal covariances) are indeed small and deviate only slightly from zero, they aggregate a large number P of signals: For $t_1 \neq t_2$, we have $\frac{1}{T} S_{t_1}' S_{t_2} = \frac{1}{T}\sum_{i=1}^{P} S_{i,t_1} S_{i,t_2} = O(\frac{1}{T} P^{1/2})$ by the central limit theorem. Thus, while in the classical regime these elements decay like $\frac{1}{T}$, in the complex regime, when $c = \frac{P}{T} > 0$, they decay like $T^{-1/2}$. As a result, an econometrician observes systematic deviations from LLN and the eigenvalues of $\frac{SS'}{T}$ converge to non-zero limits when the statistical complexity $c = \frac{P}{T} > 0$. It is this form of spurious signal correlations across time periods that generates the LLG.

5 Empirics

In this section, we provide extensive empirical and simulated evidence illustrating the power of Theorems 2 and 3 to recover (a lower bound for) the true amount of predictability. As such, LLG can be used to extract important dynamic information about the economy.

While our analysis applies to any linear-in-y estimator, including nonlinear ones (Proposition 4), in this section we focus on ridge regression with random nonlinear features as the prototypical application of our theory. We proceed as follows:

- Given the data, X_t, we follow Kelly et al. (2024); Kelly and Malamud (2025) and generate $P = 20000$ random features

$$S_{k,t} \;=\; g(W_k' X_t),\; g \in \{\tanh, \mathrm{ReLu}\} \tag{51}$$

  from original data $X_t \in \mathbb{R}^d$, where W_k are sampled i.i.d. from $\mathcal{N}(0, I_{d\times d})$. The nonlinearity $g(x)$ is commonly referred to as the *activation function.* We study both $\tanh(x)$ and $\mathrm{ReLu}(x) = \max(x, 0)$ because of their distinct features: $\tanh(x)$ flattens out for large x and, hence, is less sensitive to outliers and tends to focus on the bulk of the distribution. By contrast, $\mathrm{ReLu}(x)$ grows indefinitely at ∞ and is therefore able to capture tail dependencies better.

  In our analysis, we always report the effect of statistical complexity on model performance by varying $P_1 = 100, \cdots, 20000$ and running the ridge regression using the first P_1 of the random features (51). Following Kelly et al. (2024), we refer to the curve showing model performance as a function of $c = P_1/T$ as a virtue of complexity (VoC) curve.

- Given the signals S and labels, y, we compute $\hat{\beta}(z)$ from (11). Since $\widehat{\mathcal{L}}(z)$ in (28) is monotone decreasing in z, we pick a relatively small z that "scales" with the size of

the signals. We set

$$z \;=\; z_{ref}\,\frac{1}{P}\,\mathrm{tr}(S'S),\;\; with\; z_{ref} \;=\; 0.01\,. \tag{52}$$

- We use Theorem 3 to compute R^2_{OOS}, its lower bound (25), as well as the one-sided 95% asymptotic confidence band for R^2_*,

$$\left[\frac{R^2_{OOS}(\hat{f})+\widehat{\mathcal{L}}(z)}{1+\widehat{\mathcal{L}}(z)} \;-\; 1.65T^{-1/2}\hat{\sigma}_{R^2}\,,\;1\right]. \tag{53}$$

5.1 Data

We consider the classic Welch and Goyal (2008) dataset with 14 different time series at monthly frequency covering the period from 1930-01-01 to 2020-11-01:

- **Group one: Equity Valuation and Market.** Excess returns (`retx`) are the excess returns on the CRSP value-weighted index. The Dividend Price Ratio (`dp`) is the difference between the log of dividends and the log of prices. The Dividend Yield (`dy`) is the difference between the log of dividends and the log of lagged prices.[8] The Earnings Price Ratio (`ep`) is the difference between the log of earnings and the log of prices.[9] The Dividend Payout Ratio (`de`) is the difference between the log of dividends and the log of earnings. The Book-to-Market Ratio (`bm`) is the ratio of book value to market value for the Dow Jones Industrial Average. The Net Equity Expansion (`ntis`) is the ratio of 12-month moving sums of net issues by NYSE-listed stocks divided by the total end-of-year market capitalization of NYSE stocks.

- **Group 2: Term structure, Credit, Inflation.** Treasury Bills (`tbl`) is the Treasury-bill rate. Long Term Yield (`lty`) is the U.S. Yield On Long-Term United States Bonds

[8] We exclude `dy` from our set of targets but use it in the set of signals because $dy_{t+1} = \log(d_{t+1}/p_t)$ mechanically co-moves with other time$-t$ variables involving p_t and `retx`.

[9] Earnings are 12-month moving sums of earnings on the S&P 500 index.

from the NBER's Macrohistory database. Long Term Rate of Returns (`ltr`) is from the same source. The Term Spread (`tms`) is the difference between the long-term yield on government bonds and the Treasury bill. The Default Yield Spread (`dfy`) is the difference between BAA and AAA-rated corporate bond yields. The Default Return Spread (`dfr`) is the difference between long-term corporate bond and long-term government bond returns. Inflation (`infl`) is the Consumer Price Index (All Urban Consumers) from the Bureau of Labor Statistics.

Our data sample spans almost 90 years, and non-stationarity considerations become essential. Furthermore, the variables from Welch and Goyal (2008) may not satisfy the technical conditions of Theorem 2, requiring that the residuals ε are uncorrelated and homoskedastic. To deal with these considerations, we pre-process the Welch and Goyal (2008) data using the following procedure.

Procedure 1. Construction of the Processed Series

- Demean and standardize each $X_{i,t}$ using its rolling 36-month mean and standard deviation, lagged by one month;[10]
- Clip it at $[-3, 3]$;
- Compute the AR(1) residuals of the resulting normalized variable using an expanding window to estimate the autocorrelation coefficient.

We use X_t to denote the Welch and Goyal (2008) data (including the excess returns) processed according to Procedure 1. By construction, these data have zero autocorrelation and are approximately homoskedastic.[11]

[10] In Python syntax, we perform the transformation

$$X \to (X - X.rolling(36).mean().shift(1))/X.rolling(36).std().shift(1).$$

[11] We have performed extensive simulations with y_t exhibiting diverse forms of autocorrelation and

5.2 Semi-Synthetic Simulations

We start with a semi-synthetic simulation where we use the 13 Welch and Goyal (2008) variables and excess returns as X_t, and simulate

$$y_{t+1}^{synthetic} \;=\; f_t \;+\; \varepsilon_{t+1},\; f_t \;=\; \gamma\, g(X_t' W)\,, \tag{54}$$

where $\varepsilon_{t+1} \sim \mathcal{N}(0,1)$ and $W \in \mathbb{R}^d$ is drawn from $\mathcal{N}(0, I)$, and $g \in \{\tanh,\ \text{ReLu}\}$. We also study a pure-noise benchmark, corresponding to the limiting semi-synthetic case with $\gamma = 0$, in which y_{t+1} follows a GARCH(1,1) process. Specifically, we generate a zero-mean GARCH(1,1) sequence $\{y_t^{GARCH(1,1)}\}_{t=1}^T$ following the procedure outlined in Appendix G. Since, formally, our theory does not apply to GARCH residuals, we test the efficiency of Procedure 1 and also study $y_{t,standardized}^{GARCH(1,1)}$ obtained from $y_t^{GARCH(1,1)}$ using the first two steps of Procedure 1.

The convenience of a semi-synthetic simulation is that we can estimate the infeasible R_*^2,

$$R_*^2 \;\approx\; \frac{E[f_t^2]}{E[y_{t+1}^2]} \tag{55}$$

and, hence, we can test our theory without sacrificing the highly complex nature of the real data. Varying γ in (54) allows us to control R_*^2 and, thus, test the efficiency of our lower bound for various degrees of predictability.

Our in-sample/out-of-sample split is set at January 1990. This period corresponds to the early phase of large-scale market electrification—characterized by the adoption of electronic trading platforms, faster information dissemination, and increased automation—which represents a structural break in how financial markets process information.

stochastic volatility. These simulations indicate that Theorems 2 and 3 continue to hold as long as the above-listed transformations are applied to the underlying predicted variable.

As is explained above, we construct $P = 20000$ random features S_t using (51), run regression with the first P_1 features, and report R^2_{OOS} as well as the lower bounds (25) and (53) as a function of complexity $c = P_1/T$, where T is the number of in-sample observations.

Figures 1-2 clearly demonstrate that our theory holds very well: First, although the lower bound (25) does cross the R^2_* sometimes (e.g., for $R^2_* = 0$), the lower confidence bound (53) is always below R^2_*, emphasizing the importance of the Central Limit Theorem correction to LLG. Second, the realized R^2_{OOS} decays very quickly with complexity c. An econometrician might interpret this as evidence against statistical complexity. However, once we correct for LLG, (25) increases with c (or stays flat) for a majority of plots, emphasizing a novel form of Virtue of Complexity: The decay in R^2_{OOS} is caused by accumulating estimation errors with growing P (models with more parameters are more difficult to estimate).

Figures 1-2 clearly show that approximating the ground truth requires complexity, and running a model with a larger P makes the lower bound (25) converge to the infeasible R^2_*. Third, although GARCH residuals violate the hypothesis of Theorems 2 and 3, Figures 1-2 show the theory holds both before and after the data is standardized using Procedure 1.

5.3 Real Data

We now use LLG to investigate the predictability of the Welch and Goyal (2008) variables transformed according to Procedure 1. For each $i = 1, \cdots, 13$, we set $y_{i,t+1} = X_{i,t+1}$ and use X_t as the signals. We then feed X_t into the $P = 20000$ random features in (51). Figures 3–6 summarize the results. Several patterns stand out clearly. First, R^2_* rises sharply for many target variables, displaying a strong virtue of complexity. Second, the LLG implied by ReLu features differs substantially from that implied by tanh features, illustrating how slightly different linear models can yield significantly different lower bounds.

For group 1 (Figures 3 and 4), the largest gains arise for the dividend-to-price `dp`, earnings-to-price `ep`, and dividends-to-earnings `de` ratios, whose lower bounds reach roughly

50% (and up to 70% for `de` under the *tanh* activation). We also find substantial gains for stock volatility `svar` and the book-to-market ratio `bm`, where the bounds reach 20–25%. Because we are predicting AR(1) residuals (Procedure 1), identifying a model capable of achieving such a high population R^2 would yield meaningful economic gains for investors and provide deeper insights into the underlying macroeconomic dynamics.

Surprisingly, the ReLu-based LLG implies a lower bound of about 20% for predicting U.S. market excess returns—around 10–20 times higher than a typical realized R^2_{OOS} at a monthly horizon documented, e.g., in Welch and Goyal (2008); Kelly et al. (2024); Kelly and Malamud (2025). This finding casts the common belief that market returns are essentially unpredictable in a new light: the LLG-based bound suggests that returns may be highly predictable, but that the underlying nonlinear function is extremely hard to learn due to limits-to-learning. This interpretation aligns with Kelly et al. (2024); Kelly and Malamud (2025), who argue that the true predictive relationship between returns and the Welch and Goyal (2008) variables is highly complex.

For group 2 (Figures 5 and 6), the effects are smaller but still economically meaningful: our theory implies an R^2_* of at least 10% for T-bills `tbl` and the long-term rate `ltr`; and at least 25% for the default yield spread `dfy` and the long-term yield `lty`.

In every target except `de`,[12] the strong predictability implied by the LLG-based lower bound coexists with a negative realized R^2_{OOS}. These findings match the semi-synthetic simulations in Figures 1–2, where a negative R^2_{OOS} often coexists with significant underlying true (population) predictability. Thus, although the simple random-feature ridge model fails to uncover predictability in Figures 3–6, some other model could achieve a realized R^2_{OOS} close to (or even above) the LLG-implied lower bound.

Identifying such a model directly may require nontrivial effort. Here, we consider two classes of related models: Ridge regression with a different ridge penalty and a model we refer to as "recursive ridge," which uses a simple feature selection and construction algorithm

[12] `de` is special because it is a "purely fundamental" variable that does not involve market prices or returns.

similar to that of Yan and Zheng (2017); Chen and Dim (2023); Li et al. (2025) before running the ridge regression. We describe this algorithm in Appendix H.

We begin by noting that, under Theorem 2, a ridge regression (11) estimated with a different penalty or a different feature set (e.g., activation tanh versus ReLu) constitutes a distinct ML model. There is no theoretical result tying the out-of-sample performance of such models to the performance of the reference model used to construct the LLG. For example, Figures 3–6 rely on $z_{ref} = 0.01$ in (52), and there is no ex-ante basis for expecting the corresponding LLG-based lower bound to be informative about the R^2_{OOS} of ridge regressions estimated with other penalties, even when the set of features is held fixed. This disconnect is even sharper for the recursive ridge model in Appendix H, which uses a different feature set and is nonlinear due to its variable-selection step. The distinction matters: the construction of the LLG relies on linearity, but the bound itself applies to any forecasting model.

For each variable $y_{i,t+1}$, we estimate ridge regressions with random features (51) using a grid of $z_{ref} \in 0.01, 0.1, 1., 10.$, and estimate recursive ridge models using the same grid. All specifications are trained on the 1933-01 to 1989-12 sample and evaluated out of sample from 1990-01 to 2024-12.

Our analysis focuses on two questions. First, which models deliver economically meaningful out-of-sample performance? Second, is the realized R^2_{OOS} systematically related to the LLG-based lower bound? To assess this, we report, for each model class—ridge with two types of random features (51) on $P = 100, \ldots, 20000$, and recursive ridge—the highest R^2_{OOS} attained within that class. Although these best-in-class values inevitably reflect model selection, they serve as a benchmark for the predictive content extractable by flexible (nonlinear) ML methods.

Table 1 reports the lower bounds (53), maximized over $P = 100, \ldots, 20000$, for tanh and ReLu activations,[13] along with the corresponding best R^2_{OOS} for each model class.

[13]The two LLGs should be interpreted as distinct signals. A combined lower bound incorporating their covariance is feasible but outside the scope of this analysis.

The results indicate substantial out-of-sample predictability. Both ridge and recursive ridge models forecasting `dfy`, `de`, `tbl`, and `ep` achieve sizable R^2_{OOS}, broadly consistent with the LLG-based bounds. For instance, for `dfy` and `ep`, the recursive ridge yields $R^2_{OOS} \approx 15\%$, compared with an LLG bound of roughly 25% from the tanh features. For `tbl`, recursive ridge achieves $R^2_{OOS} \approx 10\%$, matching the 10% bound. For `de`, the recursive ridge attains $R^2_{OOS} \approx 42\%$ versus a 67% bound.

The cases of `ep`, `tbl`, `tms`, and `dfr` are especially informative: the nonlinear recursive ridge consistently outperforms the linear ridge by a large margin and, in some instances, approaches the LLG-based benchmark. This pattern highlights a central implication of our theoretical results: *although LLG is built from a linear reference model, it captures information relevant for the performance frontier of nonlinear models.*

In contrast, for `retx`, `dp`, `bm`, `svar`, and `lty`, even the recursive ridge falls far short of the lower bound. This may indicate that other functional forms are required to approximate the true predictive structure, or that sampling variation imposes a hard limit on feasible predictive accuracy.

Table 2 documents a strong positive association between the LLG-based lower bounds and the realized out-of-sample performance of both classes of models. The tanh-based bound performs particularly well, exhibiting an 80% correlation with the best nonlinear R^2_{OOS}. This is noteworthy given that the bound is computed from a ridge model with a small penalty $z_{ref} = 0.01$ (52) that itself performs poorly out of sample. Yet, Theorem 2 successfully extracts information about the underlying degree of predictability.

Overall, the evidence suggests that LLG provides a powerful, model-agnostic diagnostic for detecting genuine structure in predictive regressions, even in long samples marked by structural breaks. It offers researchers a principled tool for identifying promising forecasting relationships and guiding subsequent theoretical and empirical inquiry.

6 Conclusion

This paper demonstrates that the apparent weakness of predictability in financial data is often a statistical illusion. In high-dimensional environments, even state-of-the-art ML estimators face intrinsic limits: finite samples prevent them from recovering the true signal, causing conventional out-of-sample metrics to systematically understate population predictability. We formalize this observation by deriving the Limits-to-Learning Gap (LLG)—a universal, data-driven lower bound on the discrepancy between empirical and population fit. The LLG provides a sharp, model-agnostic correction to measured R^2, yields confidence bounds for population predictability, and quantitatively identifies when weak OOS performance reflects a lack of signal versus the inability of an estimator to learn it.

Empirically, we find that LLGs are large across a broad set of classical financial predictors, implying that true R^2 values are often many times higher than their raw estimates. These results overturn the conventional view that return predictability is negligible. They further show that nonlinear or more expressive models can, in several cases, recover economically meaningful predictive structure precisely when the LLG indicates that such structure must exist.

Beyond empirical applications, the LLG has implications for asset-pricing theory. By refining the Hansen and Jagannathan (1991) bounds, our framework quantifies the role of parameter uncertainty in shaping equilibrium outcomes. In a simple general-equilibrium model, we show that high LLGs naturally generate excess volatility and slow learning, offering a rational benchmark for phenomena often attributed to behavioral biases or misperceptions. More broadly, the LLG sheds new light on why ML scaling laws exhibit sharply diminishing returns: when the limits-to-learning gap closes only slowly, additional data or computation cannot eliminate residual estimation error.

Taken together, these findings suggest a reorientation of empirical practice. Instead of evaluating predictability solely through realized OOS performance, researchers should assess

whether the observed performance is informative about population fit at all. The LLG offers precisely this diagnostic. It clarifies when predictability is truly absent, when existing ML architectures are inadequate, and when economic signals are fundamentally obscured by finite-sample noise. As economic data continue to grow in dimensionality and complexity, incorporating limits-to-learning into empirical design, model evaluation, and asset-pricing theory will be essential. The framework introduced here provides a tractable and broadly applicable foundation for doing so.

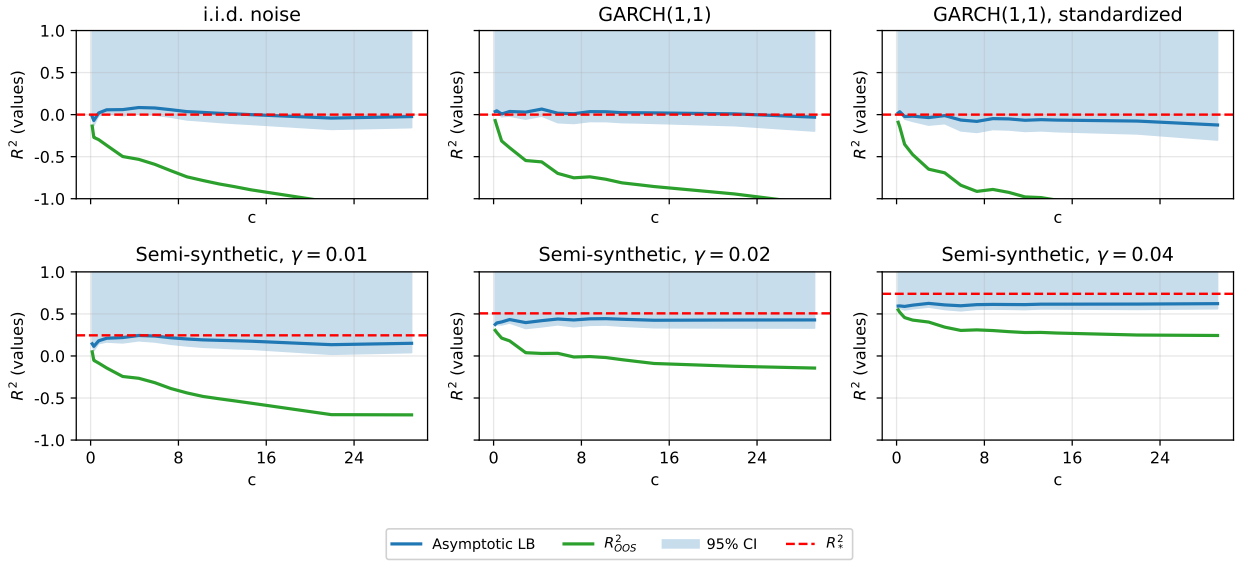


FIGURE 1 – Semi-synthetic simulation (54), with activation=ReLu. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. i.i.d. noise has $y_{t+1} = \varepsilon_{t+1} \sim \mathcal{N}(0, 1)$. GARCH(1,1) has $y_{t+1} = \varepsilon_{t+1}$ being a GARCH(1,1) noise defined in Appendix G. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51), increasing from $P_1 = 100$ to $P_1 = 20000$. R^2_* is computed in (55). Values of $R^2_{OOS} < -1$ are not shown. γ values are selected to achieve R^2_* of 0, 0.25, 0.5, 0.75, respectively.

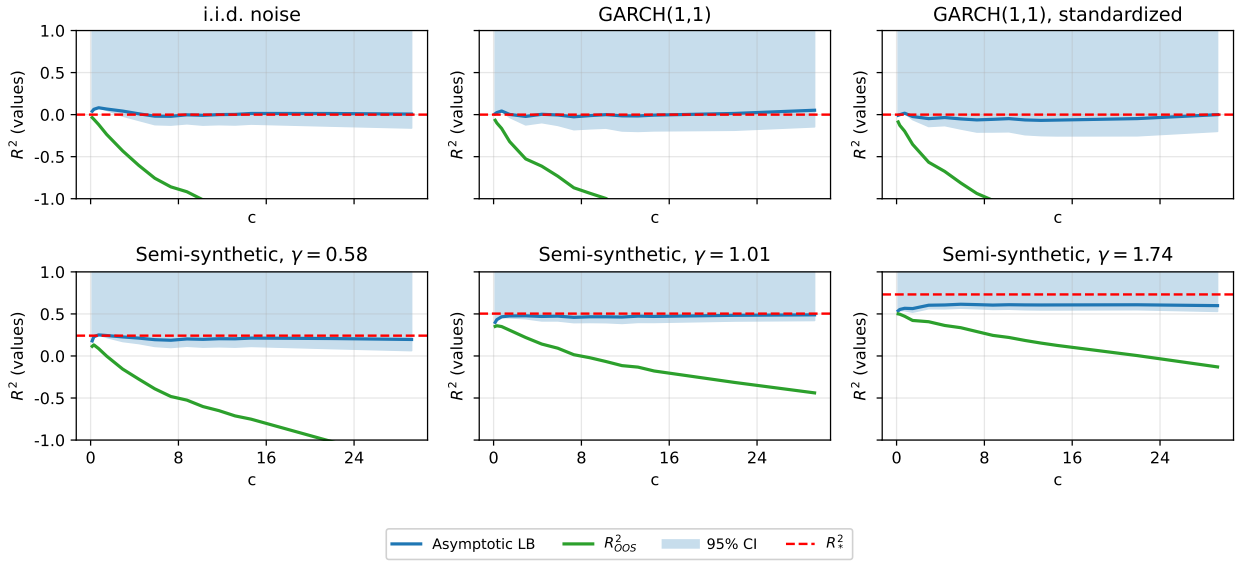


FIGURE 2 – Semi-synthetic simulation (54), with activation=tanh. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. i.i.d. noise has $y_{t+1} = \varepsilon_{t+1} \sim \mathcal{N}(0,1)$. GARCH(1,1) has $y_{t+1} = \varepsilon_{t+1}$ being a GARCH(1,1) noise defined in Appendix G. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51), increasing from $P_1 = 100$ to $P_1 = 20000$. R^2_* is computed in (55). Values of $R^2_{OOS} < -1$ are not shown. γ values are selected to achieve R^2_* of 0, 0.25, 0.5, 0.75, respectively.

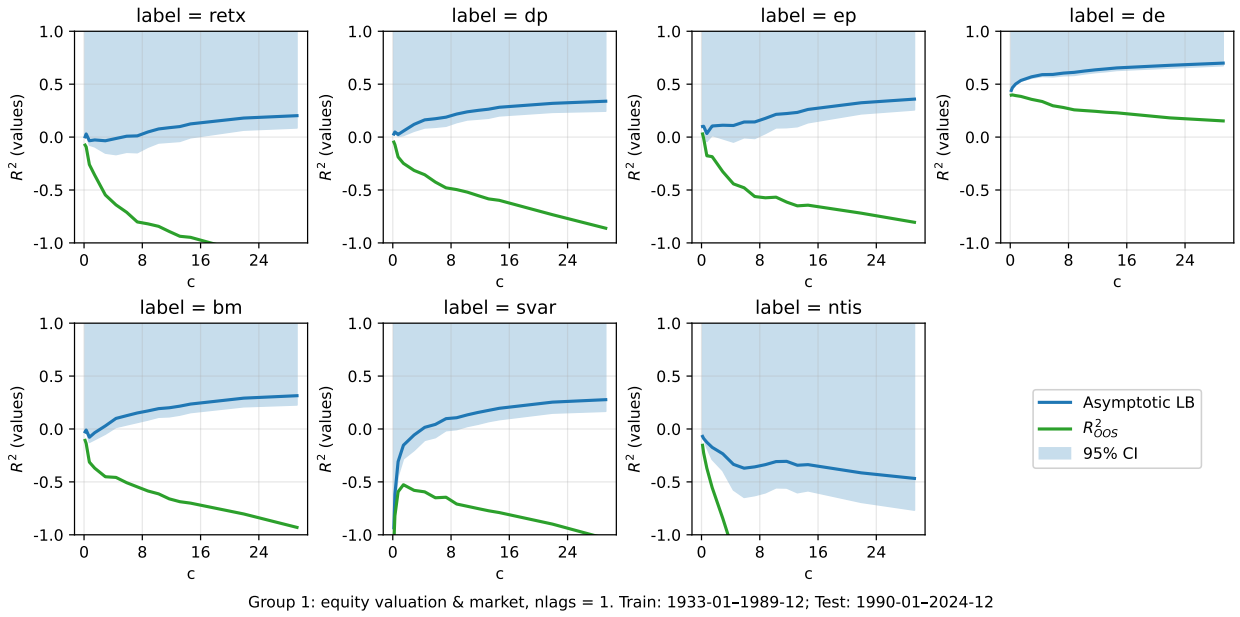


FIGURE 3 – Predicting Welch and Goyal (2008) variables from **Group one** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation**=tanh, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.

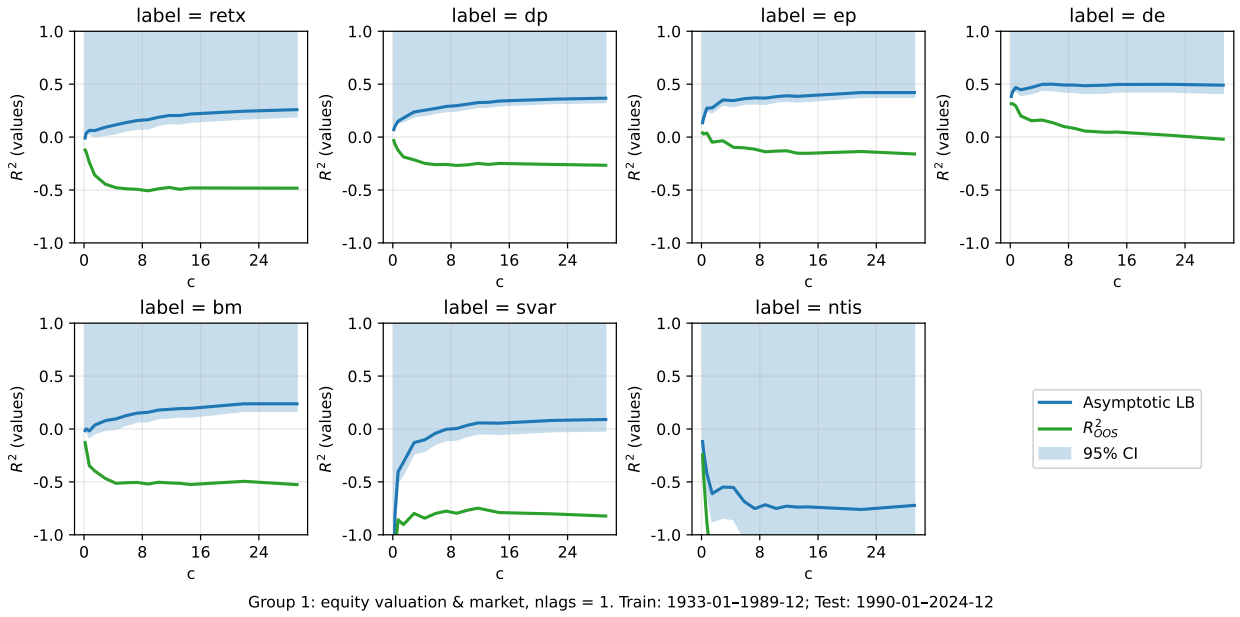


FIGURE 4 – Predicting Welch and Goyal (2008) variables from **Group one** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation**=ReLu, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.

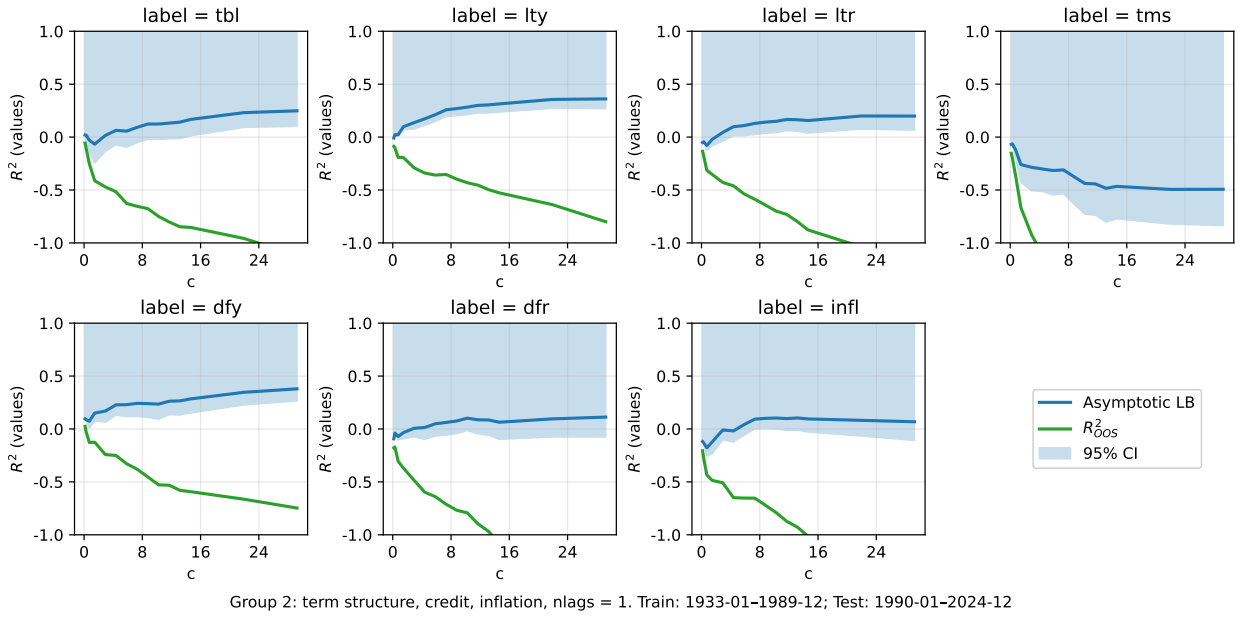


FIGURE 5 – Predicting Welch and Goyal (2008) variables from **Group two** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation**=tanh, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.

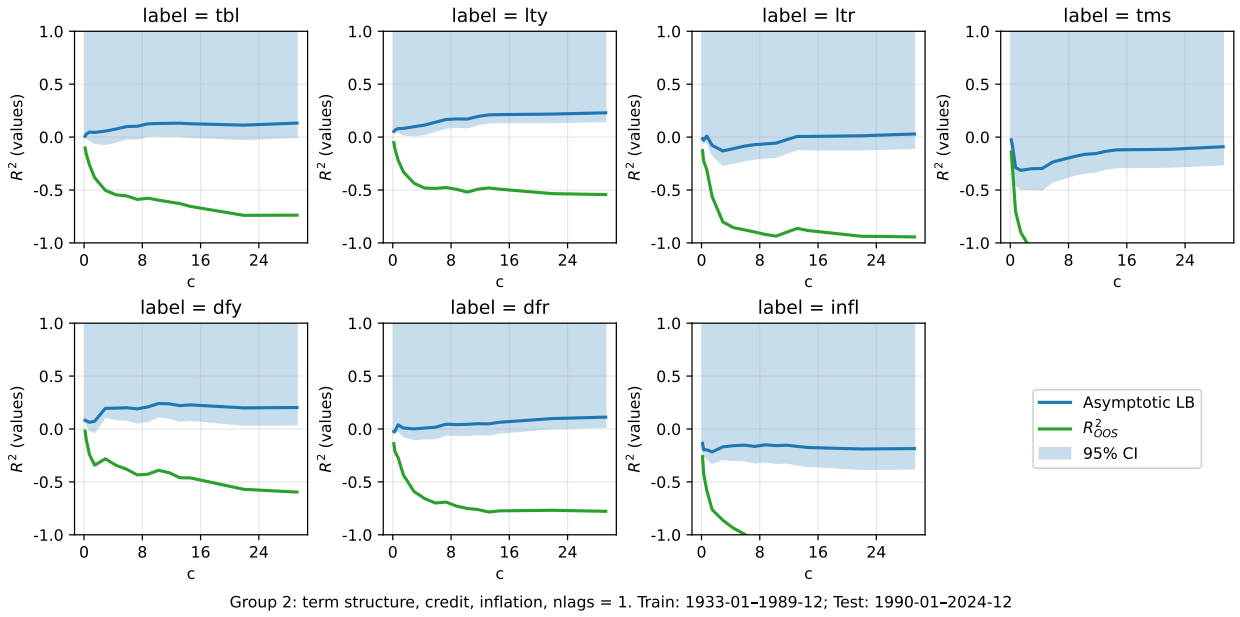


FIGURE 6 – Predicting Welch and Goyal (2008) variables from **Group two** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation**=ReLu, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.

TABLE 1 – 95% probability lower bound (53) vs R^2_{OOS} from best benchmarks

	tanh	ReLu	Best	
Label	95% prob. bound	95% prob. bound	Ridge	Recursive Ridge
retx	9%	19%	1%	2%
dp	24%	33%	2%	2%
ep	26%	38%	10%	15%
de	67%	41%	42%	42%
bm	23%	17%	0%	0%
svar	17%	0%	0%	0%
ntis	0%	0%	0%	1%
tbl	10%	0%	5%	10%
lty	27%	15%	1%	1%
ltr	7%	0%	1%	1%
tms	0%	0%	5%	8%
dfy	27%	4%	13%	14%
dfr	0%	2%	0%	3%
infl	0%	0%	0%	0%

TABLE 2 – Correlation matrix of 95% probability lower bound (53) and R^2_{OOS} from best benchmarks

		tanh	ReLu	Best	
		95% prob. bound	95% prob. bound	Ridge	Recursive Ridge
tanh	95% prob. bound	1.00			
ReLu	95% prob. bound	0.76	1.00		
Best	Ridge	0.67	0.59	1.00	
	Recursive Ridge	0.80	0.57	0.86	1.00

References

Atanasov, Alexander, Blake Bordelon, and Cengiz Pehlevan, "Neural networks as kernel learners: The silent alignment effect," *arXiv preprint arXiv:2111.00034*, 2021.

Avramov, Doron, Guanhao Feng, Jingyu He, and Shuhua Xiao, "Schrödinger's Sparsity in the Cross Section of Stock Returns," *Available at SSRN 5370960*, 2025.

Bai, Zhidong and Hewa Saranadasa, "Effect of high dimension: by an example of a two sample problem," *Statistica Sinica*, 1996, pp. 311–329.

Bansal, Ravi and Amir Yaron, "Risks for the long run: A potential resolution of asset pricing puzzles," *Journal of Finance*, 2004, *59* (4), 1481–1509.

Barillas, Francisco and Jay Shanken, "Comparing asset pricing models," *The Journal of Finance*, 2018, *73* (2), 715–754.

Belkin, Mikhail, Daniel Hsu, and Ji Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, 2020, *2* (4), 1167–1180.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato, "High-dimensional econometrics and regularized GMM," *arXiv preprint arXiv:1806.01888*, 2018.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni, "Bond risk premiums with machine learning," *The Review of Financial Studies*, 2021, *34* (2), 1046–1089.

Bianchi, Francesco, Sydney C Ludvigson, and Sai Ma, "Belief distortions and macroeconomic fluctuations," *American Economic Review*, 2022, *112* (7), 2269–2315.

Billingsley, Patrick, *Probability and Measure* Wiley Series in Probability and Mathematical Statistics, 3 ed., New York: John Wiley & Sons, 1995.

Bondt, Werner F. M. De and Richard Thaler, "Does the stock market overreact?," *Journal of Finance*, 1985, *40* (3), 793–805.

Chen, Andrew Y and Chukwuma Dim, "High-throughput asset pricing," *arXiv preprint arXiv:2311.10685*, 2023.

Chen, Hui, Winston Wei Dou, and Leonid Kogan, "Measuring "Dark Matter" in Asset Pricing Models," *The Journal of Finance*, 2024, *79* (2), 843–902.

Chen, Luyang, Markus Pelger, and Jason Zhu, "Deep learning in asset pricing," *Management Science*, 2024, *70* (2), 714–750.

Chen, Xiaohong and Halbert White, "Improved rates and asymptotic normality for nonparametric neural network estimators," *IEEE Transactions on Information Theory*, 1999, *45* (2), 682–691.

Chernov, Mikhail, Bryan T Kelly, Semyon Malamud, and Johannes Schwab, "A Test of the Efficiency of a Given Portfolio in High Dimensions," *Swiss Finance Institute Research Paper*, 2025, (25-26).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.

___, Whitney K Newey, James Robins, and Rahul Singh, "Double/de-biased machine learning of global and local parameters using regularized Riesz representers," *stat*, 2019, *1050* (9).

Cogley, Timothy and Thomas J Sargent, "The market price of risk and the equity premium: A legacy of the Great Depression?," *Journal of Monetary Economics*, 2008, *55* (3), 454–476.

Collin-Dufresne, Pierre, Michael Johannes, and Lars A Lochstoer, "Parameter learning in general equilibrium: The asset pricing implications," *American Economic Review*, 2016, *106* (3), 664–698.

Cong, Lin William, Ke Tang, Jingyuan Wang, and Yang Zhang, "AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI," *SSRN Electronic Journal. https://doi. org/10.2139/ssrn*, 2021, *3554486.*

Da, Rui, Stefan Nagel, and Dacheng Xiu, "The Statistical Limit of Arbitrage," Technical Report, Technical Report, Chicago Booth 2022.

Didisheim, Antoine, Shikun Barry Ke, Bryan T Kelly, and Semyon Malamud, "APT or "AIPT"? the surprising dominance of large factor models," Technical Report, National Bureau of Economic Research 2024.

Dou, Winston Wei, Itay Goldstein, and Yan Ji, "Ai-powered trading, algorithmic collusion, and price efficiency," *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper, The Wharton School Research Paper*, 2025.

Fan, Jianqing, Zheng Tracy Ke, Yuan Liao, and Andreas Neuhierl, "Structural Deep Learning in Conditional Asset Pricing," *Available at SSRN 4117882*, 2022.

Farmer, Leland E, Emi Nakamura, and Jón Steinsson, "Learning about the long run," *Journal of Political Economy*, 2024, *132* (10), 3334–3377.

Fort, Stanislav, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli, "Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel," *Advances in Neural Information Processing Systems*, 2020, *33*, 5850–5861.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably unequal? The effects of machine learning on credit markets," *The Journal of Finance*, 2022, *77* (1), 5–47.

Geiger, Mario, Stefano Spigler, Arthur Jacot, and Matthieu Wyart, "Disentangling feature and lazy training in deep neural networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2020, *2020* (11), 113301.

Ghosh, Nikhil and Mikhail Belkin, "A universal trade-off between the model size, test loss, and training loss of linear predictors," *SIAM Journal on Mathematics of Data Science*, 2023, *5* (4), 977–1004.

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri, "Economic predictions with big data: The illusion of sparsity," *Econometrica*, 2021, *89* (5), 2409–2437.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Han, Jiequn, Yucheng Yang et al., "Deepham: A global solution method for heterogeneous agent models with aggregate shocks," *arXiv preprint arXiv:2112.14377*, 2021.

Hansen, Lars Peter and Ravi Jagannathan, "Implications of security market data for models of dynamic economies," *Journal of political economy*, 1991, *99* (2), 225–262.

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *arXiv preprint arXiv:1903.08560*, 2019.

Holzmüller, David, "On the universality of the double descent peak in ridgeless regression," *arXiv preprint arXiv:2010.01851*, 2020.

Jacot, Arthur, Franck Gabriel, and Clément Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *arXiv preprint arXiv:1806.07572*, 2018.

Johannes, Michael, Lars A Lochstoer, and Yiqun Mou, "Learning about consumption dynamics," *The Journal of finance*, 2016, *71* (2), 551–600.

Jong, Peter De, "A central limit theorem for generalized quadratic forms," *Probability Theory and Related Fields*, 1987, *75* (2), 261–277.

Jr, Robert E Lucas, "Asset prices in an exchange economy," *Econometrica: journal of the Econometric Society*, 1978, pp. 1429–1445.

Jurado, Kyle, Sydney C Ludvigson, and Serena Ng, "Measuring uncertainty," *American Economic Review*, 2015, *105* (3), 1177–1216.

Kaniel, Ron, Zihan Lin, Markus Pelger, and Stijn Van Nieuwerburgh, "Machine-learning the skill of mutual fund managers," *Journal of Financial Economics*, 2023, *150* (1), 94–138.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, "The virtue of complexity in return prediction," *The Journal of Finance*, 2024, *79* (1), 459–503.

Kelly, Bryan T and Semyon Malamud, "Understanding The Virtue of Complexity," *Available at SSRN 5346842*, 2025.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, 2018, *133* (1), 237–293.

Knowles, Antti and Jun Yin, "Anisotropic local laws for random matrices," *Probability Theory and Related Fields*, 2017, *169*, 257–352.

Kohler, Michael and Sophie Langer, "On the rate of convergence of fully connected deep neural network regression estimates," *The Annals of Statistics*, 2021, *49* (4), 2231–2249.

Lauditi, Clarissa, Blake Bordelon, and Cengiz Pehlevan, "Adaptive kernel predictors from feature-learning infinite limits of neural networks," *arXiv preprint arXiv:2502.07998*, 2025.

Li, Bin, Alberto G Rossi, Xuemin Sterling Yan, and Lingling Zheng, "Machine learning from a "Universe" of signals: The role of feature engineering," *Journal of Financial Economics*, 2025, *172*, 104138.

Li, Haoran, Alexander Aue, Debashis Paul, Jie Peng, and Pei Wang, "An Adaptable Generalization of Hotelling's T^2 Test in High Dimension," *The Annals of Statistics*, 2020, *48* (3), 1815–1847.

Liao, Yuan, Xinjie Ma, Andreas Neuhierl, and Linda Schilling, "The Uncertainty of Machine Learning Predictions in Asset Pricing," *arXiv preprint arXiv:2503.00549*, 2025.

___ , ___ , ___ , **and Zhentao Shi**, "Does Noise Hurt Economic Forecasts?," *Available at SSRN 4659309*, 2023.

Liu, Chaoyue, Libin Zhu, and Misha Belkin, "On the linearity of large non-linear models: when and why the tangent kernel is constant," *Advances in Neural Information Processing Systems*, 2020, *33*, 15954–15964.

Liu, Haoyang, Alexander Aue, and Debashis Paul, "On the Marčenko–Pastur law for linear time series," 2015.

Malach, Eran, Omid Saremi, Sinead Williamson, Arwen Bradley, Aryo Lotfi, Emmanuel Abbe, Josh Susskind, and Etai Littwin, "To Infinity and Beyond: Tool-Use Unlocks Length Generalization in State Space Models," *arXiv preprint arXiv:2510.14826*, 2025.

Martin, Ian WR and Stefan Nagel, "Market efficiency in the age of big data," *Journal of Financial Economics*, 2021.

Mohsin, Muhammad Ahmed, Muhammad Umer, Ahsan Bilal, Zeeshan Memon, Muhammad Ibtsaam Qadir, Sagnik Bhattacharya, Hassan Rizwan, Abhiram R Gorle, Maahe Zehra Kazmi, Ayesha Mohsin et al., "On the Fundamental Limits of LLMs at Scale," *arXiv preprint arXiv:2511.12869*, 2025.

Moll, Benjamin, "The Trouble with Rational Expectations in Heterogeneous Agent Models: A Challenge for Macroeconomics," *London School of Economics, mimeo, available at https://benjaminmoll. com*, 2024.

___ **and Lenya Ryzhik**, "Mean Field Games without Rational Expectations," *arXiv preprint arXiv:2506.11838*, 2025.

Muthukumar, Vidya, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai, "Harmless interpolation of noisy data in regression," *IEEE Journal on Selected Areas in Information Theory*, 2020, *1* (1), 67–83.

Schmidt-Hieber, Johannes, "Nonparametric regression using deep neural networks with ReLU activation function," *The Annals of Statistics*, 2020, (4), 1875–1897.

Schwab, Johannes, Bryan Kelly, and Semyon Malamud, "Training NTK to Generalize with KARE," *working paper*, 2025.

Shen, Zhouyu and Dacheng Xiu, "Can Machines Learn Weak Signals?," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2024, (2024-29).

Shiller, Robert J., "Do stock prices move too much to be justified by subsequent changes in dividends?," *American Economic Review*, 1981, *71* (3), 421–436.

Srivastava, Muni S, "A test for the mean vector with fewer observations than the dimension under non-normality," *Journal of Multivariate Analysis*, 2009, *100* (3), 518–532.

Vershynin, Roman, *High-dimensional probability: An introduction with applications in data science*, Vol. 47, Cambridge university press, 2018.

Vyas, Nikhil, Yamini Bansal, and Preetum Nakkiran, "Limitations of the NTK for understanding generalization in deep learning," *arXiv preprint arXiv:2206.10012*, 2022.

Welch, Ivo and Amit Goyal, "A comprehensive look at the empirical performance of equity premium prediction," *The Review of Financial Studies*, 2008, *21* (4), 1455–1508.

White, Halbert, *Estimation, inference and specification analysis* number 22, Cambridge university press, 1996.

Yan, Xuemin and Lingling Zheng, "Fundamental analysis and the cross-section of stock returns: A data-mining approach," *The Review of Financial Studies*, 2017, *30* (4), 1382–1423.

Yang, Greg and Etai Littwin, "Tensor programs iib: Architectural universality of neural tangent kernel training dynamics," in "International conference on machine learning" PMLR 2021, pp. 11762–11772.

A Preliminaries on Random Matrix Theory

A.1 Concentration of Quadratic Forms

Our key assumption in the main text is that the residuals, ε_t, are i.i.d. Hence, for any matrix A independent of ε, we have

$$\begin{aligned}
\frac{1}{T}\varepsilon' A\varepsilon &= \frac{1}{T}\sum_{t_1,t_2}\varepsilon_{t_1}\varepsilon_{t_2}A_{t_1,t_2} \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 A_{t,t}}_{\approx\frac{1}{T}\sigma_\varepsilon^2\sum_{t=1}^{T}A_{t,t}\ because\ E[\varepsilon_t^2]=\sigma_\varepsilon^2} + \underbrace{\frac{1}{T}\sum_{t_1\neq t_2}\varepsilon_{t_1}\varepsilon_{t_2}A_{t_1,t_2}}_{\approx 0\ because\ E[\varepsilon_{t_1}\varepsilon_{t_2}]=0} \\
&\approx \frac{1}{T}\sigma_\varepsilon^2\sum_{i=1}^{N}A_{t,t} \\
&= \frac{1}{T}\sigma_\varepsilon^2\,\mathrm{tr}(A)\,.
\end{aligned} \tag{A.1}$$

Many proofs in this paper are based on the classic "concentration of quadratic forms" lemma that makes the above argument rigorous.

Lemma A.1 *Suppose that $u=(u_i)_{i=1}^P$ where u_i are i.i.d., with $E[u_i]=0$, $E[u_i^2]=\sigma^2$, $E[u_i^4]<\infty$. Suppose also A_P is a sequence of symmetric random matrices that is independent of u. Then,*

$$\begin{aligned}
E[(\frac{1}{P}u'A_Pu)^2-(P^{-1}\sigma^2\,\mathrm{tr}(A_P))^2] &= E[(\frac{1}{P}u'A_Pu-(P^{-1}\sigma^2\,\mathrm{tr}(A_P)))^2] \\
&\leq (E[u_i^4]-\sigma^4)P^{-2}E[\mathrm{tr}(A_P^2)]
\end{aligned} \tag{A.2}$$

In $E[u_i^4]=3\sigma^4$, then this identity is exact:

$$\begin{aligned}
E[(\frac{1}{P}u'A_Pu)^2-(P^{-1}\sigma^2\,\mathrm{tr}(A_P))^2] &= E[(\frac{1}{P}u'A_Pu-(P^{-1}\sigma^2\,\mathrm{tr}(A_P)))^2] \\
&= 2\sigma^4P^{-2}E[\mathrm{tr}(A_P^2)]\,.
\end{aligned} \tag{A.3}$$

If $\lim P^{-2}E[\mathrm{tr}(A_P^2)] = 0$, *then*

$$\frac{1}{P}u'A_Pu \;-\; P^{-1}\sigma^2\,\mathrm{tr}(A_P) \;\to\; 0 \tag{A.4}$$

in L_2 *and, hence, in probability.*

A.2 Spectral Concentration for Sub-Gaussian Designs

In this section we collect a concentration result for empirical feature covariance matrices generated by high-dimensional sub-Gaussian designs. This lemma is used repeatedly in later proofs to control higher-order spectral terms.

A real-valued random variable Z is called sub-Gaussian if there exists a finite constant $K > 0$ such that $\mathbb{E}\big[e^{Z^2/K^2}\big] \le 2$. We then write $\|Z\|_{\psi_2} \le K$. A random vector $x \in \mathbb{R}^P$ is called sub-Gaussian with parameter K if every one-dimensional projection is sub-Gaussian, that is, $\sup_{u\in\mathbb{S}^{P-1}} \|\langle x, u\rangle\|_{\psi_2} \le K$. This implies in particular that all coordinates of x are sub-Gaussian with parameter bounded by a constant multiple of K.

Let $T, P \in \mathbb{N}$ and let $S \in \mathbb{R}^{T\times P}$ denote a random matrix whose rows $s_t' \in \mathbb{R}^P$ (the feature vectors at time t) satisfy the following assumptions. We define the empirical feature covariance as $\hat{\Psi} = \dfrac{1}{T}S'S \in \mathbb{R}^{P\times P}$.

(A1) The rows $(s_t)_{t=1}^T$ are independent.

(A2) Each row is isotropic: $\mathbb{E}[s_t] = 0$ and $\mathbb{E}[s_t s_t'] = I_P$.

(A3) Each s_t is sub-Gaussian with parameter K in the above sense.

Lemma A.2 (Spectral concentration for sub-Gaussian designs) *Under* (A1)–(A3)*, there exist constants* $a > 0$ *and* $C_\Psi > 0$*, depending only on* K*, such that for all* $t \ge 0$*,*

$$\mathbb{P}\Big(\big\|\hat{\Psi} - I_P\big\| > C_\Psi\Big(\sqrt{P/T} + t/\sqrt{T}\Big)\Big) \;\le\; 2e^{-at^2}. \tag{A.5}$$

Consequently, for any fixed integer $k \geq 1$, there exists $C_k < \infty$ depending only on k and K such that

$$\mathbb{E}\Big[\big\|S'S\big\|^k\Big] \;\leq\; C_k T^k, \tag{A.6}$$

and hence

$$\mathbb{E}\big[\|\hat{\Psi}\|^k\big] = \mathbb{E}\Big[\Big\|\frac{1}{T}S'S\Big\|^k\Big] \;\leq\; C_k. \tag{A.7}$$

Proof of Lemma A.2. By (A1)–(A3), the rows $s_t \in \mathbb{R}^P$ are independent, mean-zero, isotropic, sub-Gaussian with $\|s_t\|_{\psi_2} \leq K$. Since S is a $T \times P$ matrix with independent isotropic sub-Gaussian rows, by Theorem 4.6.1 of Vershynin (2018), there exist constant $C_\Psi > 0$, depending only on K, such that for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\frac{1}{T}S'S - I_P\right\| > C_\Psi \max\big(\delta, \delta^2\big)\right) \;\leq\; 2e^{-t^2}, \qquad \delta = \sqrt{\frac{P}{T}} + \frac{t}{\sqrt{T}}. \tag{A.8}$$

Our first goal is to convert (A.8) into a deviation inequality stated directly in the level s. Observe that if $\left\|\frac{1}{T}S'S - I_P\right\| > s$, then necessarily $s > C_\Psi \max(\delta, \delta^2)$. For $s > C_\Psi$, this forces $\delta \geq 1$, hence $\delta = \sqrt{\dfrac{s}{C_\Psi}}$ and $t = \sqrt{T}\Big(\sqrt{\dfrac{u}{C_\Psi}} - \sqrt{\dfrac{P}{T}}\Big)$. Since $\|SS'\| \geq 0$, we may use the standard representation

$$\mathbb{E}[Z^k] = \int_0^\infty k s^{k-1}\, \mathbb{P}(Z > s)\, ds, \qquad Z \geq 0. \tag{A.9}$$

Let $Z = \|XX'\|$. For $0 \leq s \leq C_\Psi T$ (with C_Ψ depending only on K), the trivial bound $\mathbb{P}(Z > s) \leq 1$ implies that

$$\int_0^{C_\Psi T} k s^{k-1}\, \mathbb{P}(Z > s)\, ds \;\leq\; \int_0^{C_\Psi T} k s^{k-1}\, ds \;=\; C_\Psi^k T^k, \tag{A.10}$$

which is of order T^k. For $s > C_\Psi T$, the above concentration bound gives

$$\mathbb{P}(Z > s) \;=\; \mathbb{P}\Big(\Big\|\frac{1}{T}SS'\Big\| > s/T\Big) \;\le\; 2\exp\big[-T\,(s/T - C_\Psi)^2\big] \;=\; 2\exp\big[-T^{-1}(s - C_\Psi T)^2\big]. \tag{A.11}$$

Hence the tail contribution satisfies

$$\int_{C_\Psi T}^{\infty} k s^{k-1}\,\mathbb{P}(Z > s)\,ds \;\le\; 2k\int_{C_\Psi T}^{\infty} s^{k-1}\exp\big[-T^{-1}(s - C_\Psi T)^2\big]\,ds. \tag{A.12}$$

With the change of variables $u = (s - C_\Psi T)/\sqrt{T}$, so that $s = C_\Psi T + \sqrt{T}\,u$ and $ds = \sqrt{T}\,du$, we obtain

$$\begin{aligned}\int_{C_\Psi T}^{\infty} s^{k-1}\exp\big[-T^{-1}(s - C_\Psi P)^2\big]\,ds &= \sqrt{T}\int_0^{\infty}(C_\Psi T + \sqrt{T}\,u)^{k-1}e^{-u^2}\,du \\ &= T^{k-\frac{1}{2}}\int_0^{\infty}(C_\Psi + \frac{u}{\sqrt{T}})^{k-1}e^{-u^2}\,du \\ &\le C_\Psi^* T^{k-\frac{1}{2}}\end{aligned} \tag{A.13}$$

with the constant C_Ψ^* depending only on k and K. The inequality comes from the fact that Gaussian decay term shrinks much faster than any polynomial grows. Combining the two parts of the integral (A.10) and (A.13), we arrive at

$$\mathbb{E}\big[\|SS'\|^k\big] \;\le\; C_k T^k, \qquad \mathbb{E}\big[\|\hat{\Psi}\|^k\big] = \mathbb{E}\Big[\Big\|\frac{1}{T}S'S\Big\|^k\Big] \;\le\; C_k. \tag{A.14}$$

for some finite constant C_k depending only on k and K. The proof of Lemma A.2 is complete. □

A.3 Main RMT Asymptotic Results

We will need the Stieltjes transform and its derivative:

$$\begin{aligned} m(-z) &= \lim_{P\to\infty} P^{-1}\operatorname{tr}((zI+\Psi_P)^{-1}) \\ m'_\Psi(-z) &= \lim_{P\to\infty} P^{-1}\operatorname{tr}((zI+\Psi_P)^{-2}) \end{aligned} \tag{A.15}$$

and

$$\begin{aligned} \hat{m}(-z) &= P^{-1}\operatorname{tr}((zI+\hat{\Psi})^{-1}) \\ \hat{m}'(-z) &= P^{-1}\operatorname{tr}((zI+\hat{\Psi})^{-2}) \end{aligned} \tag{A.16}$$

Let also $m(-z;c)$ be the unique, positive solution to the fixed point equation[14]

$$m(-z;c) = \frac{1}{1-c+czm(-z;c)} m_\Psi\left(\frac{-z}{1-c+czm(-z;c)}\right). \tag{A.17}$$

We will now need the following result from Kelly et al. (2024).

Proposition A.1 *We have*

$$\lim_{T\to\infty}\frac{1}{T}\operatorname{tr}((zI+\hat{\Psi})^{-1}\Psi) \;\to\; \xi(z;c) \tag{A.18}$$

almost surely and

$$\lim_{T\to\infty}\frac{1}{T}S'_{T+1}(zI+\hat{\Psi})^{-1}S_{T+1} \;\to\; \xi(z;c) \tag{A.19}$$

in probability, where

$$\xi(z;c) = \frac{1-zm(-z;c)}{c^{-1}-1+zm(-z;c)}. \tag{A.20}$$

[14] See, for example, Chernov et al. (2025) for a proof that this equation indeed has a unique solution.

Similarly,

$$\lim_{T\to\infty} \frac{1}{T} \operatorname{tr}((zI + \hat{\Psi})^{-2}\Psi) \;\to\; -\xi'(z;c) \tag{A.21}$$

almost surely, where

$$\begin{aligned} \xi'(z;c) &= \frac{d}{dz}\left(\frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}\right) \\ &= \frac{d}{dz}\left(-1 + \frac{1}{1 - c + czm(-z;c)}\right) \\ &= -\frac{c\Big(m(-z;c) - zm'(-z;c)\Big)}{\Big(1 - c + czm(-z;c)\Big)^2} \end{aligned} \tag{A.22}$$

We will also define

$$Z_*(z;c) \;=\; \frac{z}{1 - c + czm(-z;c)} \;=\; z(1 + \xi(z;c)),\; Z'_* = 1 + \xi + z\xi'\,. \tag{A.23}$$

to be the implicit shrinkage and

$$\begin{aligned} \underline{m}(z;c) \;&=\; 1/Z_*(z;c) \;=\; z^{-1}(1 - c + czm(-z;c)) \\ \widetilde{m}(z;c) \;&=\; 1/Z_*(z;c) \;=\; z^{-1}(1 - c + cz\hat{m}(-z;c))\,. \end{aligned} \tag{A.24}$$

Then, the results of Bai and Saranadasa (1996); Li et al. (2020); Liu et al. (2015); Knowles and Yin (2017); Hastie et al. (2019) imply that the following is true:

Theorem A.2 *Suppose that* $S_t = \Psi^{1/2}X_t$, *where* $X_t = (X_{i,t})$ *where* $E[X_{i,t}] = 0,\; E[X_{i,t}^2] = 1,\; E[X_{i,t}^4] < \infty$ *are independent and identically distributed, and* $\Psi = E[S_tS_t']$ *is a positive semi-definite, uniformly bounded signal covariance matrix. Then, in the limit as* $P, T \to$

$\infty, P/T \to c$, *for any uniformly bounded sequence of non-random matrices* A_P, *we have*

$$P^{-1}z\,\mathrm{tr}(A_P(zI + \underbrace{\hat{\Psi}}_{random})^{-1}) \; - P^{-1}Z_*\,\mathrm{tr}(A_P(Z_*I + \underbrace{\Psi}_{deterministic})^{-1}) \; \to \; 0 \tag{A.25}$$

almost surely. Similarly, for any sequence of uniformly bounded vectors β, *we have*

$$z\beta'(zI + \underbrace{\hat{\Psi}}_{random})^{-1}\beta \; - Z_*\beta'(Z_*I + \underbrace{\Psi}_{deterministic})^{-1}\beta \; \to \; 0 \tag{A.26}$$

almost surely. Furthermore, the same result holds for $\tilde{\Psi} = \hat{\Psi} - \bar{S}\bar{S}'$, *where* $\bar{S} = T^{-1}\sum_t S_t$:

$$\begin{aligned} &P^{-1}z\,\mathrm{tr}(A_P(zI + \tilde{\Psi})^{-1}) \; - P^{-1}Z_*\,\mathrm{tr}(A_P(Z_*I + \Psi)^{-1}) \; \to \; 0 \\ &z\beta'(zI + \tilde{\Psi})^{-1}\beta \; - \; Z_*\beta'(Z_*I + \Psi)^{-1}\beta \; \to \; 0 \end{aligned} \tag{A.27}$$

almost surely.

Informally, we can rewrite this theorem as

$$\begin{aligned} z(zI + \hat{\Psi})^{-1} \; &\approx \; Z_*(Z_*I + \Psi)^{-1} \\ z(zI + \tilde{\Psi})^{-1} \; &\approx \; Z_*(Z_*I + \Psi)^{-1}\,. \end{aligned} \tag{A.28}$$

A.4 Auxiliary CLT results

First, we will need the following

Lemma A.3 *Suppose that* $E[\varepsilon_t^3] = 0$ *and* $E[\varepsilon_t^4] = 3$. *Let* a_T *be a sequence of bounded vectors vectors and* A_T *a sequence of bounded random matrices, all of them being independent of* ε_t. *Then,* $(a_T'\varepsilon\|a_T\|^{-1}, \; T^{-1/2}(\varepsilon' A_T\varepsilon - \sigma_\varepsilon^2\,\mathrm{tr}(A_T))/\sqrt{2\tfrac{1}{T}\sigma_\varepsilon^4\,\mathrm{tr}(A_T^2)})$ *converges to a* $\mathcal{N}(0, I)$ *distribution. That is, these two random variables are asymptotically independent standard Normals.*

Proof of Lemma A.3. The proof of Lemma A.3 is completely analogous to that of the

main result in Srivastava (2009) and follows closely the classical argument from De Jong (1987), plus the Cramer-Wold device (Billingsley (1995)). Without loss of generality, we normalize $\sigma_\varepsilon^2 = 1$. Let $\tilde{a}_T = a_T/\|a_T\|$, $\tilde{A}_T = A_T/\sqrt{2\frac{1}{T}\operatorname{tr}(A_T^2)}$.

With Cramer-Wold, we build for any vector (q_1, q_2)

$$S(T) = q_1 \tilde{a}_T' \varepsilon + q_2 T^{-1/2}(\varepsilon' \tilde{A}_T \varepsilon - \operatorname{tr}(\tilde{A}_T)) \tag{A.29}$$

As in De Jong (1987), we define

$$S(t) = \sum_{\tau \leq t} \Big(q_1 \tilde{a}(\tau)\varepsilon_\tau + q_2 T^{-1/2}(\varepsilon_\tau^2 - 1)\tilde{A}_{\tau,\tau} + q_2 T^{-1/2} 2 \sum_{\tau_1 < \tau} \varepsilon_{\tau_1}\varepsilon_\tau \tilde{A}_{\tau_1,\tau} \Big) \tag{A.30}$$

By direct calculation, $S(t)$ is a Martingale and then, standard arguments based on the Lindeberg CLT implies asymptotic normality. □

We will also use the following multi-variate extension.

Lemma A.4 *Suppose that $E[\varepsilon_t^3] = 0$ and $E[\varepsilon_t^4] = 3$. Let $a_{k,T}$ be a sequence of bounded vectors vectors and $A_{k,T}$ a sequence of bounded, symmetric random matrices, all of them being independent of ε_t; $k = 1, \cdots, K$. Then, $((a_{k,T}'\varepsilon \|a_{k,T}\|^{-1},\ T^{-1/2}(\varepsilon' A_{k,T}\varepsilon - \sigma_\varepsilon^2 \operatorname{tr}(A_{k,T}))/\sqrt{2\frac{1}{T}\sigma_\varepsilon^4 \operatorname{tr}(A_{k,T}^2)}))_{k=1}^K$ converges to a $\mathcal{N}(0, \Sigma(\sigma_\varepsilon^2))$ distribution. The structure of the covariance matrix $\Sigma(\sigma_\varepsilon)$ is as follows: any of the linear components, $a_{k,T}'\varepsilon\|a_{k,T}\|^{-1}$, has zero covariance with any quadratic component; by contrast, quadratic terms are correlated with*

$$\operatorname{Cov}(\varepsilon' A\varepsilon,\ \varepsilon' B\varepsilon) = 2\operatorname{tr}(AB) \tag{A.31}$$

while linear terms give

$$\operatorname{Cov}(\varepsilon' a,\ \varepsilon' b) = a'b\,. \tag{A.32}$$

Proof of Lemma A.4. The joint normality follows by the same Cramer-Wold argument.

The only claim to prove is the covariance formula. We have

$$\mathrm{Var}[\varepsilon'(A+B)\varepsilon] \;=\; 2\,\mathrm{tr}((A+B)^2) \;=\; 2\,\mathrm{tr}(A^2+B^2+2AB) \tag{A.33}$$

but, by the variance decomposition

$$\mathrm{Var}[\varepsilon'(A+B)\varepsilon] \;=\; \mathrm{Var}[\varepsilon'A\varepsilon]+\mathrm{Var}[\varepsilon'B\varepsilon]+2\mathrm{Cov}(\varepsilon'A\varepsilon,\varepsilon'B\varepsilon) \tag{A.34}$$

Comparing, we get the required identity. □

B Proof of Proposition 1 and Theorem 2

The goal of this section is to prove Theorem 2. We also proof the special case of the ridge regression as a separate corollary.

Proof of Proposition 1. Recall that for each OOS period t,

$$\hat{f}_t \;=\; \hat{f}_t^s \;+\; \hat{f}_t^\varepsilon \;=\; \mathcal{K}(S_t,S)f \;+\; \mathcal{K}(S_t,S)\varepsilon\,. \tag{B.1}$$

Hence, period by period we have

$$\begin{aligned}(y_{t+1}-\hat{f}_t)^2 \;&=\; \Big[f_t+\varepsilon_{t+1}-(\hat{f}_t^s+\hat{f}_t^\varepsilon)\Big]^2\\ &=\; \Big(f_t-\hat{f}_t^s-\hat{f}_t^\varepsilon+\varepsilon_{t+1}\Big)^2\\ &=\; (f_t-\hat{f}_t^s)^2+\varepsilon_{t+1}^2+(\hat{f}_t^\varepsilon)^2-2(f_t-\hat{f}_t^s)(\hat{f}_t^\varepsilon-\varepsilon_{t+1})-2\varepsilon_{t+1}'\hat{f}_t^\varepsilon\end{aligned} \tag{B.2}$$

Averaging over $t = T, \ldots, T + T_{\text{OOS}} - 1$ yields

$$\begin{aligned} & MSE_{OOS}(\hat{f}) \\ &= \frac{1}{T_{OOS}} \sum_{t=T}^{T+T_{OOS}-1} (y_{t+1} - \hat{f}_t)^2 \\ &= E_{OOS}[\varepsilon^2] + \underbrace{E_{\text{OOS}}\left[(f - \hat{f}^s)^2\right]}_{\widehat{\mathcal{B}}} + \underbrace{E_{\text{OOS}}\left[(\hat{f}^\varepsilon)^2\right]}_{\widehat{\mathcal{V}}} - \underbrace{2E_{OOS}\left[(f - \hat{f}^s)(\hat{f}^\varepsilon - \varepsilon) + \varepsilon \hat{f}^\varepsilon\right]}_{\widehat{\mathcal{I}}} \end{aligned} \tag{B.3}$$

The proof of Proposition 1 is complete.

□

Proof of Theorem 2. We have

$$\begin{aligned} \frac{1}{2}\widehat{\mathcal{I}} \;&=\; -E_{OOS}[(f - \hat{f}^s)\hat{f}_\varepsilon] \;+\; E_{OOS}[(f - \hat{f}^s)\varepsilon] \;-\; E_{OOS}[\varepsilon \hat{f}_\varepsilon] \\ &= \widehat{\mathcal{I}}_1 + \widehat{\mathcal{I}}_2 + \widehat{\mathcal{I}}_3 \,. \end{aligned} \tag{B.4}$$

We will use the inequality

$$E[(\varepsilon' A \varepsilon)^2] \;- \sigma_\varepsilon^2 E[(\text{tr}(A))^2] \;\leq\; CE[\text{tr}(A^2)] \tag{B.5}$$

where C is a constant (see Lemma A.1). Let $\varepsilon = \binom{\varepsilon_{OOS}}{\varepsilon_{IS}}$ where we use the obvious notations $\varepsilon_{OOS}, \varepsilon_{IS}$ to denote subsets of indices. Then,

$$\widehat{\mathcal{I}}_3 \;=\; \frac{1}{T_{OOS}} \varepsilon'_{OOS} \widehat{\mathcal{K}} \varepsilon_{IS} \tag{B.6}$$

Then, by the independence of ε, we get

$$\begin{aligned} E[\widehat{\mathcal{I}}_3^2] &= \frac{1}{T_{OOS}^2} E[(\varepsilon'_{OOS}\widehat{\mathcal{K}}\varepsilon_{IS})^2] \\ &= \frac{1}{T_{OOS}^2} E[\varepsilon'_{OOS}\widehat{\mathcal{K}}\varepsilon_{IS}\varepsilon'_{IS}\widehat{\mathcal{K}}'\varepsilon_{OOS}] \\ &= \frac{1}{T_{OOS}^2}\sigma_\varepsilon^2 E[\mathrm{tr}(\widehat{\mathcal{K}}\varepsilon_{IS}\varepsilon'_{IS}\widehat{\mathcal{K}}')] \\ &= \frac{1}{T_{OOS}^2}\sigma_\varepsilon^4 E[\mathrm{tr}(\widehat{\mathcal{K}}\widehat{\mathcal{K}}')] \end{aligned} \tag{B.7}$$

by assumption. Note that

$$\begin{aligned} E[\widehat{\mathcal{I}}_1^2] = &\ E[(E_{OOS}[(f-\hat{f}^s)f_\varepsilon])^2] \\ = &\ E\Big[\Big(\frac{1}{T_{OOS}}(f_{OOS}-\hat{f}^s_{OOS})'\widehat{\mathcal{K}}\varepsilon_{IS}\Big)^2\Big] \\ = &\ \frac{1}{T_{OOS}^2}E\Big[\Big((f_{OOS}-\hat{f}^s_{OOS})'\widehat{\mathcal{K}}\varepsilon_{IS}\Big)^2\Big] \\ = &\ \frac{1}{T_{OOS}^2}\sigma_\varepsilon^2 E\Big[(f_{OOS}-\hat{f}^s_{OOS})'\widehat{\mathcal{K}}\widehat{\mathcal{K}}'(f_{OOS}-\hat{f}^s_{OOS})\Big] \\ = &\ \frac{1}{T_{OOS}^2}\sigma_\varepsilon^2 E\Big[\sum_{j=1}^{T}\Big(\sum_{t=T}^{T+T_{OOS}-1}\big(f_t-\hat{f}^s_t\big)\left(\mathcal{K}(S_t,S)\right)_j\Big)^2\Big] \end{aligned} \tag{B.8}$$

$$\begin{aligned} E[\widehat{\mathcal{I}}_1^2] = &\ E[(E_{OOS}[(f-\hat{f}^s)\varepsilon])^2] \\ = &\ E\Big[\Big(\frac{1}{T_{OOS}}(f_{OOS}-\hat{f}^s_{OOS})'\varepsilon_{OOS}\Big)^2\Big] \\ = &\ \frac{1}{T_{OOS}^2}E\Big[\Big((f_{OOS}-\hat{f}^s_{OOS})'\varepsilon_{OOS}\Big)^2\Big] \\ = &\ \frac{1}{T_{OOS}^2}\sigma_\varepsilon^2 E\Big[(f_{OOS}-\hat{f}^s_{OOS})'(f_{OOS}-\hat{f}^s_{OOS})\Big] \\ = &\ \frac{1}{T_{OOS}^2}\sigma_\varepsilon^2 E\Big[\sum_{t=T}^{T+T_{OOS}-1}(f_t-\hat{f}^s_t)^2\Big] \end{aligned} \tag{B.9}$$

and the claim follows from the hypotheses of the theorem. Since $\widehat{\mathcal{B}}(z), \widehat{\mathcal{V}}(z) \geq 0$, this implies

$$MSE_{OOS}(\hat{f}) \;\geq\; E_{OOS}[\varepsilon^2] \;+\; \widehat{\mathcal{V}} \;+\; o(1) \;=\; \sigma_\varepsilon^2 \;+\; \widehat{\mathcal{V}} \;+\; o(1), \tag{B.10}$$

where the last equality follows from the definition of the OOS expectation. Furthermore,

$$\widehat{\mathcal{V}} \;=\; \varepsilon' \left(\frac{1}{T_{OOS}} \widehat{\mathcal{K}}' \widehat{\mathcal{K}} \right) \varepsilon \,. \tag{B.11}$$

and, hence,

$$\widehat{\mathcal{V}} \;-\; \sigma_\varepsilon^2 \, \widehat{\mathcal{L}} \;\to\; 0 \tag{B.12}$$

in L_2 and in probability by Lemma A.1. Substituting this expression into (B.10) yields

$$MSE_{OOS}(\hat{f}) \;\geq\; \sigma_\varepsilon^2 \Big(1 + \widehat{\mathcal{L}} \Big) \;+\; o(1), \tag{B.13}$$

and dividing both sides by $1 + \widehat{\mathcal{L}}$ and taking $\liminf_{T, T_{OOS} \to \infty}$ (appealing to the continuous mapping theorem) gives the desired inequality

$$\liminf \; \frac{MSE(\hat{f})}{1 + \widehat{\mathcal{L}}} \;\geq\; \sigma_\varepsilon^2. \tag{B.14}$$

When $f_t = 0$, the bias term $\widehat{\mathcal{B}}$ vanishes, and the bound becomes tight. In that case,

$$MSE_{OOS}(\hat{f}) \;=\; \sigma_\varepsilon^2 + \widehat{\mathcal{V}} + o(1) \;=\; \sigma_\varepsilon^2 \Big(1 + \widehat{\mathcal{L}} \Big) + o(1),$$

so that the inequality holds with equality asymptotically. The proof of Theorem 2 is complete. □

Corollary B.1 *For the ridge estimator, we have*

$$MSE_{OOS}(\hat{\beta}) \;=\; E_{OOS}[\varepsilon^2] \;+\; \widehat{\mathcal{B}}(z) \;-\; \widehat{\mathcal{I}}(z) \;+\; \widehat{\mathcal{V}}(z)\,, \tag{B.15}$$

where

$$\begin{aligned} \widehat{\mathcal{B}}(z) \;&=\; \underbrace{z^2\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta}_{bias} \;\geq\; 0 \\ \widehat{\mathcal{V}}(z) \;&=\; \underbrace{\frac{1}{T^2}\,\varepsilon' S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon}_{variance} \;\geq\; 0, \end{aligned} \tag{B.16}$$

while

$$\widehat{\mathcal{I}}(z) \;=\; \underbrace{\frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon + 2E_{OOS}[\varepsilon' S(\beta-\hat{\beta})]}_{interaction} \tag{B.17}$$

Proof of Corollary B.1. Recall that ridge estimator $\hat{\beta}(z)$

$$\hat{\beta}(z) \;=\; (zI \;+\; \hat{\Psi})^{-1}T^{-1}S'd\,, \tag{B.18}$$

implies the decomposition

$$\hat{\beta}(z) \;=\; \beta \;-\; \underbrace{z\,(zI+\hat{\Psi})^{-1}\beta}_{bias} \;+\; \underbrace{T^{-1}(zI+\hat{\Psi})^{-1}S'\varepsilon}_{noise}\,. \tag{B.19}$$

We have the realized out-of-sample MSE

$$\begin{aligned} MSE_{OOS}(\hat{\beta}) \;&= E_{OOS}[(d-\hat{\beta}'S)^2] \\ &= E_{OOS}[(\beta'S-\hat{\beta}'S+\varepsilon)^2] \\ &= (\beta-\hat{\beta})'\Psi_{OOS}(\beta-\hat{\beta}) + 2E_{OOS}[\varepsilon' S(\beta-\hat{\beta})] +\; E_{OOS}[\varepsilon^2] \end{aligned} \tag{B.20}$$

Plugging the decomposition into the first term, we obtain

$$\begin{aligned}
&(\beta-\hat{\beta})'\Psi_{OOS}(\beta-\hat{\beta}) \\
&= \Big(z\,(zI+\hat{\Psi})^{-1}\beta - T^{-1}(zI+\hat{\Psi})^{-1}S'\varepsilon\Big)'\hat{\Psi}_{OOS}\Big(z\,(zI+\hat{\Psi})^{-1}\beta - T^{-1}(zI+\hat{\Psi})^{-1}S'\varepsilon\Big) \\
&= z^2\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta \;+\; \frac{1}{T^2}\,\varepsilon'S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon \\
&\quad - \frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon
\end{aligned} \tag{B.21}$$

The proof of the Corollary B.1 is complete. □

Corollary B.2 *Suppose that*

$$E[\big\|E_T[\hat{\Psi}^2_{OOS}]\big\|] = o(\min(T_{OOS},T)) \quad \text{as } T_{OOS}\to\infty. \tag{B.22}$$

In the limit as $T, T_{OOS}\to\infty$, the MSE from (B.20) satisfies

$$\liminf \frac{MSE_{OOS}(\hat{\beta})}{1+\widehat{\mathcal{L}}(z)} \;\geq\; \underbrace{\sigma^2_\varepsilon}_{infeasible}, \tag{B.23}$$

in probability, where

$$\widehat{\mathcal{L}}(z) \;=\; \frac{1}{T}\,\mathrm{tr}\Big(\hat{\Psi}_{OOS}\hat{\Psi}(zI+\hat{\Psi})^{-2}\Big) \tag{B.24}$$

is the Limits-To-Learning Gap (LLG). This bound turns into an identity for $\beta=0$.

Proof of Corollary B.2. We have

$$\frac{1}{2}\widehat{\mathcal{I}}(z) \;=\; \frac{z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon + E_{OOS}[\varepsilon'S(\beta-\hat{\beta})] \tag{B.25}$$

Note that

$$\begin{aligned} & E_T\Big[\big(\frac{z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\big)^2\Big] \\ &= \frac{\sigma_\varepsilon^2 z^2}{T^2}\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta \\ &= \frac{\sigma_\varepsilon^2 z^2}{T}\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta \end{aligned} \tag{B.26}$$

Using the inequality $x'ABAx \leq \|B\|\, x'A^2x$ for any symmetric B and any vector x, we obtain

$$\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS} \;\preceq\; \|(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\|\,\hat{\Psi}_{OOS}^2. \tag{B.27}$$

Since $\|(zI+\hat{\Psi})^{-1}\hat{\Psi}\| \leq 1$ and $\|(zI+\hat{\Psi})^{-1}\| \leq 1/z$, it follows

$$\begin{aligned} & E_T\Big[\big(\frac{z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\big)^2\Big] \\ &\leq \frac{\sigma_\varepsilon^2 z^2}{T}\frac{1}{z}\beta'(zI+\hat{\Psi})^{-1}\,\hat{\Psi}_{OOS}^2\,(zI+\hat{\Psi})^{-1}\beta \\ &\leq \frac{\sigma_\varepsilon^2\|(zI+\hat{\Psi})^{-1}\beta\|^2}{zT}\|E_T[\hat{\Psi}_{OOS}^2]\| \end{aligned} \tag{B.28}$$

By the law of iterated expectations and the assumption $\|E_T[\hat{\Psi}_{OOS}^2]\| = o(\min(T_{OOS}, T))$ as $T_{OOS} \to \infty$, we have

$$E\Big[\big(\frac{z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\big)^2\Big] \;\leq\; \frac{\sigma_\varepsilon^2\|\beta\|^2}{zT}E[\|E_T[\hat{\Psi}_{OOS}^2]\|] = o(1) \tag{B.29}$$

Thus, the first term in (B.25) is negligible in L^2. We next bound the square of the out-of-sample term in (B.25). Recall that

$$E_{OOS}[\varepsilon'S(\beta-\hat{\beta})] = \frac{1}{T_{OOS}}\varepsilon'_{OOS}S_{OOS}\Big(z\,(zI+\hat{\Psi})^{-1}\beta - T^{-1}(zI+\hat{\Psi})^{-1}S'\varepsilon\Big) \tag{B.30}$$

We have

$$
\begin{aligned}
&\Big(E_{OOS}[\varepsilon' S(\beta-\hat{\beta})]\Big)^2 \\
&= \frac{1}{T_{OOS}^2}\varepsilon'_{OOS}S_{OOS}\Big(z\,(zI+\hat{\Psi})^{-1}\beta - T^{-1}(zI+\hat{\Psi})^{-1}S'\varepsilon\Big)\Big(z\,(zI+\hat{\Psi})^{-1}\beta - T^{-1}(zI+\hat{\Psi})^{-1}S'\varepsilon\Big)' S'_{OOS}\varepsilon_{OOS} \\
&= \frac{z^2}{T_{OOS}^2}\varepsilon'_{OOS}S_{OOS}(zI+\hat{\Psi})^{-1}\beta\beta'(zI+\hat{\Psi})^{-1}S'_{OOS}\varepsilon_{OOS} \\
&\quad - \frac{2z}{TT_{OOS}^2}\varepsilon'_{OOS}S_{OOS}(zI+\hat{\Psi})^{-1}\beta\varepsilon' S(zI+\hat{\Psi})^{-1}S'_{OOS}\varepsilon_{OOS} \\
&\quad + \frac{1}{T^2T_{OOS}^2}\varepsilon'_{OOS}S_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\varepsilon' S(zI+\hat{\Psi})^{-1}S'_{OOS}\varepsilon_{OOS}
\end{aligned}
\tag{B.31}
$$

Note that, for any unit vector h, we have

$$
\begin{aligned}
\|E_T[\hat{\Psi}_{OOS}]h\|^2 &\le E_T[\|\hat{\Psi}_{OOS}h\|^2] \\
&= E_T[h'\hat{\Psi}_{OOS}^2 h] \\
&= h' E_T[\hat{\Psi}_{OOS}^2]h \\
&\le \|E_T[\hat{\Psi}_{OOS}^2]\|,
\end{aligned}
\tag{B.32}
$$

so that, taking supremum over all unit h,

$$
\|E_T[\hat{\Psi}_{OOS}]\| \le \|E_T[\hat{\Psi}_{OOS}^2]\|^{\frac{1}{2}} \tag{B.33}
$$

Taking conditional expectations of the first term in (B.31) on the training sample, we obtain

$$
\begin{aligned}
& E_T\Big[\frac{z^2}{T_{OOS}^2}\varepsilon_{OOS}' S_{OOS}(zI+\hat{\Psi})^{-1}\beta\beta'(zI+\hat{\Psi})^{-1}S_{OOS}'\varepsilon_{OOS}\Big] \\
&= \frac{z^2\sigma_\varepsilon^2}{T_{OOS}^2}\,\beta'(zI+\hat{\Psi})^{-1}\,E_T[S_{OOS}'S_{OOS}]\,(zI+\hat{\Psi})^{-1}\beta \\
&= \frac{z^2\sigma_\varepsilon^2}{T_{OOS}}\,\beta'(zI+\hat{\Psi})^{-1}\,E_T[\hat{\Psi}_{OOS}]\,(zI+\hat{\Psi})^{-1}\beta \\
&\le \frac{z^2\sigma_\varepsilon^2}{T_{OOS}}\,\|(zI+\hat{\Psi})^{-1}\|^2\|\beta\|^2\,\|E_T[\hat{\Psi}_{OOS}]\| \\
&\le \frac{\sigma_\varepsilon^2\|\beta\|^2}{T_{OOS}}\|E_T[\hat{\Psi}_{OOS}^2]\|^{\frac{1}{2}} \\
&= o(\frac{1}{\sqrt{T_{OOS}}})
\end{aligned}
\tag{B.34}
$$

Then taking unconditional expectation, we have

$$
E\Big[\frac{z^2}{T_{OOS}^2}\varepsilon_{OOS}' S_{OOS}(zI+\hat{\Psi})^{-1}\beta\beta'(zI+\hat{\Psi})^{-1}S_{OOS}'\varepsilon_{OOS}\Big] = o(1) \tag{B.35}
$$

Taking conditional expectations of the second term in (B.31) on the training sample, by the independence of ε, we obtain

$$
\begin{aligned}
& E_T\Big[\frac{2z}{TT_{OOS}^2}\varepsilon_{OOS}' S_{OOS}(zI+\hat{\Psi})^{-1}\beta\varepsilon' S(zI+\hat{\Psi})^{-1}S_{OOS}'\varepsilon_{OOS}\Big] \\
&= \frac{2z\sigma_\varepsilon^2}{TT_{OOS}^2}E_T\Big[\,\mathrm{tr}(S_{OOS}(zI+\hat{\Psi})^{-1}\beta\varepsilon' S(zI+\hat{\Psi})^{-1}S_{OOS}')\Big] \\
&= \frac{2z\sigma_\varepsilon^2}{TT_{OOS}}E_T\Big[\varepsilon' S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta\Big] \\
&= 0
\end{aligned}
\tag{B.36}
$$

Thus, we have $E\Big[\frac{2z}{TT_{OOS}^2}\varepsilon_{OOS}' S_{OOS}(zI+\hat{\Psi})^{-1}\beta\varepsilon' S(zI+\hat{\Psi})^{-1}S_{OOS}'\varepsilon_{OOS}\Big] = 0$. Finally, for

the third term in (B.31), we obtain the conditional expectations

$$
\begin{aligned}
& E_T\Big[\frac{1}{T^2T_{OOS}^2}\varepsilon_{OOS}'S_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\varepsilon'S(zI+\hat{\Psi})^{-1}S_{OOS}'\varepsilon_{OOS}\Big] \\
&= \frac{\sigma_\varepsilon^2}{T^2T_{OOS}^2}E_T\Big[\operatorname{tr}(S_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\varepsilon'S(zI+\hat{\Psi})^{-1}S_{OOS}')\Big] \\
&= \frac{\sigma_\varepsilon^2}{T^2T_{OOS}}E_T\Big[\varepsilon'S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\Big] \\
&= \frac{\sigma_\varepsilon^4}{T^2T_{OOS}}E_T\Big[\operatorname{tr}(S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S')\Big] \\
&= \frac{\sigma_\varepsilon^4}{TT_{OOS}}E_T\Big[\operatorname{tr}((zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi})\Big] \\
&\le \frac{\sigma_\varepsilon^4}{zTT_{OOS}}E_T[\operatorname{tr}(\hat{\Psi}_{OOS})] \\
&\le \frac{\sigma_\varepsilon^4}{zTT_{OOS}}(E_T[\operatorname{tr}(\hat{\Psi}_{OOS})^2])^{\frac{1}{2}} \\
&\le \frac{\sigma_\varepsilon^4}{zTT_{OOS}}P^{\frac{1}{2}}(E_T[\operatorname{tr}(\hat{\Psi}_{OOS}^2)])^{\frac{1}{2}} \\
&\le \frac{\sigma_\varepsilon^4}{zTT_{OOS}}P^{\frac{1}{2}}(P\|E_T[\hat{\Psi}_{OOS}^2]\|)^{\frac{1}{2}} \\
&= \frac{\sigma_\varepsilon^4P}{zTT_{OOS}}\|E_T[\hat{\Psi}_{OOS}^2]\|^{\frac{1}{2}} \\
&= o(\frac{P}{T}\frac{1}{\sqrt{T_{OOS}}})
\end{aligned}
\tag{B.37}
$$

Then taking unconditional expectation, we have

$$
E\Big[\frac{1}{T^2T_{OOS}^2}\varepsilon_{OOS}'S_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\varepsilon'S(zI+\hat{\Psi})^{-1}S_{OOS}'\varepsilon_{OOS}\Big] = o(1) \tag{B.38}
$$

Thus the third term in (B.25) is negligible in L^2. Since $\widehat{\mathcal{B}}(z), \widehat{\mathcal{V}}(z) \ge 0$, this implies

$$
MSE_{OOS}(\hat{\beta}) \;\ge\; E_{OOS}[\varepsilon^2] \;+\; \widehat{\mathcal{V}}(z) + \;o(1) \;=\; \sigma_\varepsilon^2 \;+\; \widehat{\mathcal{V}}(z) + \;o(1), \tag{B.39}
$$

Recall that

$$\widehat{\mathcal{V}}(z) \;=\; \frac{1}{T^2}\,\varepsilon' S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon. \tag{B.40}$$

Note that we have

$$\begin{aligned}
&\frac{1}{T}\sigma_\varepsilon^2\,\mathrm{tr}\Big(\frac{1}{T}S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\Big)\\
&=\frac{1}{T^2}\sigma_\varepsilon^2\,\mathrm{tr}\Big(S'S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\Big)\\
&=\frac{1}{T}\sigma_\varepsilon^2\,\mathrm{tr}\Big(\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\Big)\\
&=\frac{1}{T}\sigma_\varepsilon^2\,\mathrm{tr}\Big(\hat{\Psi}_{OOS}\hat{\Psi}(zI+\hat{\Psi})^{-2}\Big)\\
&=\sigma_\varepsilon^2\widehat{\mathcal{L}}(z).
\end{aligned} \tag{B.41}$$

Lemma A.1 implies that

$$\widehat{\mathcal{V}}(z) - \;\sigma_\varepsilon^2\,\widehat{\mathcal{L}}(z) \to\; 0 \tag{B.42}$$

in L_2 and in probability. Substituting this expression into (B.39) yields

$$MSE_{OOS}(\hat{\beta}) \;\geq\; \sigma_\varepsilon^2\Big(1+\widehat{\mathcal{L}}(z)\Big) \;+\; o(1), \tag{B.43}$$

and dividing both sides by $1+\widehat{\mathcal{L}}(z)$ and taking $\liminf_{T,T_{OOS}\to\infty}$ (appealing to the continuous mapping theorem) gives

$$\liminf \;\frac{MSE(\hat{\beta})}{1+\widehat{\mathcal{L}}(z)} \;\geq\; \sigma_\varepsilon^2. \tag{B.44}$$

When $\beta = 0$, the bias term $\widehat{\mathcal{B}}(z)$ vanishes, and the bound becomes tight. In that case,

$$MSE_{OOS}(\hat{\beta}) \;=\; \sigma_\varepsilon^2 + \widehat{\mathcal{V}}(z) + o(1) \;=\; \sigma_\varepsilon^2\Big(1+\widehat{\mathcal{L}}(z)\Big) + o(1),$$

so that the inequality holds with equality asymptotically. The proof of Corollary B.2 is complete. □

C CLT

C.1 CLT For MSE

The following Lemma is a direct consequence of Proposition 1 and Lemma A.4.

Lemma C.1 *We have*

$$T^{1/2}\frac{\frac{MSE_{OOS}(\hat{f})-\widehat{\mathcal{B}}}{1+\widehat{\mathcal{L}}}-\sigma_\varepsilon^2}{\sigma^2_{MSE}} \;\to\; \mathcal{N}(0,1) \tag{C.1}$$

in distribution, where

$$\sigma^2_{MSE} \;=\; \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4+\sigma_\varepsilon^4\sigma_V^2+\sigma_\varepsilon^2\sigma_I^2+\sigma_\varepsilon^2\sigma^2_{I,OOS}}{(1+\widehat{\mathcal{L}})^2}, \tag{C.2}$$

where

$$\begin{aligned}\sigma_V^2 &= 2TT_{OOS}^{-2}\,\mathrm{tr}((\widehat{\mathcal{K}}'\widehat{\mathcal{K}})^2)\\ \sigma_I^2 &= 4TT_{OOS}^{-2}\|(f-\hat{f}^s)\widehat{\mathcal{K}}\|^2\\ \sigma^2_{I,OOS} &= 4\frac{T}{T_{OOS}}\left(\widehat{\mathcal{B}}+\;\sigma_\varepsilon^2\widehat{\mathcal{L}}\right)\end{aligned} \tag{C.3}$$

Proof of Lemma C.1. By Proposition 1,

$$MSE_{OOS}(\hat{f})-\widehat{\mathcal{B}} \;=\; E_{OOS}[\varepsilon^2] \;+\; \widehat{\mathcal{I}} \;+\; \widehat{\mathcal{V}}\,, \tag{C.4}$$

where

$$\widehat{\mathcal{V}} \;=\; T_{OOS}^{-1}\varepsilon'\widehat{\mathcal{K}}'\widehat{\mathcal{K}}\varepsilon\,. \tag{C.5}$$

By Lemma A.4, these three terms satisfy

$$\begin{aligned} T_{OOS}^{1/2}(E_{OOS}[\varepsilon^2]-\sigma_\varepsilon^2) \;&\sim \mathcal{N}(0,2\sigma_\varepsilon^4) \\ \frac{T^{1/2}(\widehat{\mathcal{V}}-\sigma_\varepsilon^2\widehat{\mathcal{L}})}{\sigma_V\sigma_\varepsilon^2} \;&\sim \mathcal{N}(0,1) \end{aligned} \tag{C.6}$$

where, under the made assumptions of $E[\varepsilon_t^4]=3$,

$$\sigma_V^2 \;=\; 2TT_{OOS}^{-2}\,\mathrm{tr}((\widehat{\mathcal{K}}'\widehat{\mathcal{K}})^2)\,. \tag{C.7}$$

Similarly, under the made assumptions of $E[\varepsilon_t^3]=0,\ E[\varepsilon_t^4]=3$, we have that the three terms (B.4) are jointly Gaussian,

$$\widehat{\mathcal{I}} \;=\; 2\hat{\mathcal{I}}_1 \;+\; (2\hat{\mathcal{I}}_2+2\hat{\mathcal{I}}_3) \tag{C.8}$$

and asymptotically uncorrelated, so that, by (**??**)-(B.7),

$$\begin{aligned} E_\varepsilon[\widehat{\mathcal{I}}^2] \;&= 4E_\varepsilon[\widehat{\mathcal{I}}_1^2] \;+\; 4E_\varepsilon[\widehat{\mathcal{I}}_2^2] \;+\; 4E_\varepsilon[\widehat{\mathcal{I}}_3^2] \\ &= 4\sigma_\varepsilon^2\frac{1}{T_{OOS}^2}\|(f-\hat{f}^s)\widehat{\mathcal{K}}\|^2 + \;4\sigma_\varepsilon^2\frac{1}{T_{OOS}^2}\|(f-\hat{f}^s)\|^2 + \;4\sigma_\varepsilon^4\frac{1}{T_{OOS}^2}\,\mathrm{tr}(\widehat{\mathcal{K}}\widehat{\mathcal{K}}') \\ &= 4\sigma_\varepsilon^2\frac{1}{T_{OOS}^2}\|(f-\hat{f}^s)\widehat{\mathcal{K}}\|^2 + \;4\sigma_\varepsilon^2\frac{1}{T_{OOS}}\widehat{\mathcal{B}} + \;4\sigma_\varepsilon^4\frac{1}{T_{OOS}}\widehat{\mathcal{L}}\,. \end{aligned} \tag{C.9}$$

Thus, Lemma A.4 implies

$$\frac{T^{1/2}\widehat{\mathcal{I}}(z)}{(\sigma_\varepsilon^2\sigma_I^2+\sigma_\varepsilon^2\sigma_{I,OOS}^2)^{1/2}} \;\to\; \mathcal{N}(0,1), \tag{C.10}$$

The proof of Lemma C.1 is complete. □

Corollary C.1 *We have*

$$T^{1/2}\frac{\frac{MSE_{OOS}(\hat{f})-\widehat{\mathcal{B}}(z)}{1+\widehat{\mathcal{L}}}-\sigma_\varepsilon^2}{\sigma_{MSE}^2} \;\to\; \mathcal{N}(0,1) \tag{C.11}$$

in distribution, where

$$\sigma_{MSE}^2 \;=\; \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4+\sigma_\varepsilon^4\sigma_V^2+\sigma_\varepsilon^2\sigma_I^2+\sigma_\varepsilon^2\sigma_{I,OOS}^2}{(1+\widehat{\mathcal{L}})^2}, \tag{C.12}$$

where

$$\begin{aligned}
\sigma_V^2 \;&=2\frac{1}{T}\operatorname{tr}\Big((zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}\Big)\\
\sigma_I^2 \;&=4z^2\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta\\
\sigma_{I,OOS}^2 \;&=4\frac{T}{T_{OOS}}\Big(\widehat{\mathcal{B}}(z)+\;\sigma_\varepsilon^2\widehat{\mathcal{L}}\Big)
\end{aligned} \tag{C.13}$$

Proof of Corollary C.1. By Proposition 1,

$$MSE_{OOS}(\hat{f})-\widehat{\mathcal{B}}(z) \;=\; E_{OOS}[\varepsilon^2] \;-\; \widehat{\mathcal{I}}(z) \;+\; \widehat{\mathcal{V}}(z)\,, \tag{C.14}$$

where

$$\begin{aligned}
\widehat{\mathcal{B}}(z) \;&=\; z^2\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta \;\geq\; 0\\
\widehat{\mathcal{V}}(z) \;&=\; \frac{1}{T^2}\,\varepsilon' S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon \;\geq\; 0,
\end{aligned} \tag{C.15}$$

while

$$\widehat{\mathcal{I}}(z) \;=\; \frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon+2E_{OOS}[\varepsilon' S(\beta-\hat{\beta})] \tag{C.16}$$

By Lemma A.4, these three terms satisfy

$$\begin{aligned} T_{OOS}^{1/2}(E_{OOS}[\varepsilon^2] - \sigma_\varepsilon^2) &\sim \mathcal{N}(0, 2\sigma_\varepsilon^4) \\ \frac{T^{1/2}(\widehat{\mathcal{V}}(z) - \sigma_\varepsilon^2 \widehat{\mathcal{L}})}{\sigma_V \sigma_\varepsilon^2} &\sim \mathcal{N}(0,1) \end{aligned} \tag{C.17}$$

where

$$\begin{aligned} \sigma_V^2 &= 2\frac{1}{T^3}\operatorname{tr}\Big((S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S')^2\Big) \\ &= 2\frac{1}{T^3}\operatorname{tr}\Big(S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\Big) \\ &= 2\frac{1}{T^3}\operatorname{tr}\Big((zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'S\Big) \\ &= 2\frac{1}{T}\operatorname{tr}\Big((zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}\Big), \end{aligned} \tag{C.18}$$

and

$$\frac{T^{1/2}\widehat{\mathcal{I}}(z)}{(\sigma_\varepsilon^2\sigma_I^2 + \sigma_\varepsilon^2\sigma_{I,OOS}^2)^{1/2}} \;\to\; \mathcal{N}(0,1), \tag{C.19}$$

where

$$\sigma_I^2 \;=\; 4z^2\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta \tag{C.20}$$

and

$$\begin{aligned} &\sigma_{I,OOS}^2 \\ &= 4\frac{T}{T_{OOS}}(\beta-\hat{\beta})'\hat{\Psi}_{OOS}(\beta-\hat{\beta}) \\ &= 4\frac{T}{T_{OOS}}\Big(z^2\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta + \;\sigma_\varepsilon^2\frac{1}{T}\operatorname{tr}((zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi})\Big) \;+\; O(T^{-1/2}) \\ &= 4\frac{T}{T_{OOS}}\Big(\widehat{\mathcal{B}}(z) + \;\sigma_\varepsilon^2\widehat{\mathcal{L}}\Big) \;+\; O(T^{-1/2}) \end{aligned} \tag{C.21}$$

where we have used the decomposition

$$\beta \; - \hat{\beta} \;=\; \underbrace{z\,(zI + \hat{\Psi})^{-1}\beta}_{bias} \;-\; \underbrace{(zI + \hat{\Psi})^{-1}\frac{1}{T}S'\varepsilon}_{noise}\,. \tag{C.22}$$

Furthermore, the three Normals are asymptotically independent, conditional on S. We have

$$T^{1/2}\Big(\frac{MSE_{OOS}(\hat{f}) - \hat{B}(z)}{1+\widehat{\mathcal{L}}} - \sigma_\varepsilon^2\Big) = \; T^{1/2}\frac{MSE_{OOS}(\hat{f}) - \hat{B}(z) - \sigma_\varepsilon^2 - \widehat{\mathcal{L}}\sigma_\varepsilon^2}{1+\widehat{\mathcal{L}}}\,. \tag{C.23}$$

The claim follows from the continuous mapping theorem. □

D Proof of Theorem 3

Theorem D.1 (Probabilistic Lower Bound for R^2_*) *Suppose that $E[\varepsilon_t^3] = 0,\ E[\varepsilon_t^4] = 3$. Let*

$$R^2_{OOS}(\hat{f}) \;=\; 1 \;-\; \frac{MSE_{OOS}(\hat{f})}{E_{OOS}[y^2]} \tag{D.1}$$

be the realized OOS R^2. Then,

$$\frac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}(z)}{1+\widehat{\mathcal{L}}(z)}\,, \tag{D.2}$$

is a $T^{1/2}$-consistent upper bound for $\tilde{R}^2_*$ in the following sense: The event $\frac{R^2_{OOS}(\hat{f})+\widehat{\mathcal{L}}(z)}{1+\widehat{\mathcal{L}}(z)} > \tilde{R}^2_*$ occurs with vanishing probability:

$$\limsup_{T,T_{OOS}\to\infty} \text{Prob}\left(\frac{T_{OOS}^{1/2}\left(\tilde{R}^2_* - \dfrac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}(z)}{1+\widehat{\mathcal{L}}(z)}\right)}{\hat{\sigma}_{R^2}} < \alpha\right) \;\leq\; \Phi(\alpha) \tag{D.3}$$

for any $\alpha \leq 0$, where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. Here,

$$\begin{aligned}
\hat{\sigma}_{R^2} &= \frac{\frac{1}{MSE(0)^2}\Sigma_{1,1}+\Big(\frac{\widehat{MSE}}{MSE(0)^2}\Big)^2\Sigma_{2,2} - 2\frac{\widehat{MSE}}{MSE(0)^3}\Sigma_{1,2}}{(1+\widehat{\mathcal{L}})^2} \\
\Sigma_{1,1} &= \frac{T_{OOS}}{T}\sigma^2_{MSE}(1+\widehat{\mathcal{L}})^2 \\
\Sigma_{1,2} &= 2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon E_{OOS}[f(f-\hat{f})] \\
\Sigma_{2,2} &= 2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon E_{OOS}[f^2] \\
\widehat{MSE} &= \widehat{\mathcal{B}} \;+\; \sigma^2_\varepsilon(1+\widehat{\mathcal{L}}) \\
\sigma^2_{MSE} &= \frac{2\frac{T}{T_{OOS}}\sigma^4_\varepsilon + \sigma^4_\varepsilon\sigma^2_V + \sigma^2_\varepsilon\sigma^2_I + \sigma^2_\varepsilon\sigma^2_{I,OOS}}{(1+\widehat{\mathcal{L}})^2} \\
\sigma^2_V &= 2TT^{-2}_{OOS}\,\mathrm{tr}((\widehat{\mathcal{K}}'\widehat{\mathcal{K}})^2) \\
\sigma^2_I &= 4TT^{-2}_{OOS}\|(f-\hat{f}^s)\widehat{\mathcal{K}}\|^2 \\
\sigma^2_{I,OOS} &= 4\frac{T}{T_{OOS}}\Big(\widehat{\mathcal{B}} + \;\sigma^2_\varepsilon\widehat{\mathcal{L}}\Big)
\end{aligned} \tag{D.4}$$

Proof of Theorem D.2. We have

$$R^2_{OOS} \;=\; 1 \;-\; \frac{MSE_{OOS}(\hat{f})}{MSE_{OOS}(0)}\,. \tag{D.5}$$

As above, we suppose that $E[\varepsilon^3_t] = 0,\ E[\varepsilon^4_t] = 3$. We have

$$E_{OOS}[y^2] \;= E_{OOS}[\varepsilon^2] \;+\; 2E_{OOS}[\varepsilon f] \;+\; E_{OOS}[f^2] \tag{D.6}$$

Now, asymptotic normality and asymptotic independence of all the terms follow by the same argument as in the proof of Lemma A.3. All we need to do is compute Σ^2_{OOS}. We have

$$T^{1/2}_{OOS}(E_{OOS}[\varepsilon^2] - \sigma^2_\varepsilon) \;\to\; \mathcal{N}(0, 2\sigma^4_\varepsilon) \tag{D.7}$$

and

$$\frac{T_{OOS}^{1/2} 2E_{OOS}[\varepsilon f]}{(4T_{OOS}^{-1}\|f\|^2)^{1/2}} \;\to\; \mathcal{N}(0,1)\,, \tag{D.8}$$

Thus,

$$\begin{pmatrix} MSE_{OOS}(\hat{f}) \\ E_{OOS}[y^2] \end{pmatrix} \;=\; \begin{pmatrix} \widehat{MSE} \\ MSE(0) \end{pmatrix} \;+\; \frac{1}{T_{OOS}}\mathcal{N}(0, \Sigma_{Joint}) \tag{D.9}$$

where

$$\Sigma_{Joint} \;=\; \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2} & \Sigma_{2,2} \end{pmatrix} \tag{D.10}$$

where

$$\begin{aligned} \Sigma_{1,1} &= \frac{T_{OOS}}{T}\sigma_{MSE}^2(1+\widehat{\mathcal{L}})^2 \\ \Sigma_{1,2} &= 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2 E_{OOS}[f(f-\hat{f})] \\ \Sigma_{2,2} &= 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2 E_{OOS}[f^2]\,. \end{aligned} \tag{D.11}$$

Here,

$$E_{OOS}[f(f-\hat{f})] \;=\; E_{OOS}[f^2] - E_{OOS}[f\hat{f}] \tag{D.12}$$

and $E_{OOS}[f\hat{f}]$ admits a pivotal estimator

$$E_{OOS}[d\hat{f}] \;=\; \frac{1}{T_{OOS}}(f_{OOS}+\varepsilon_{OOS})'\hat{f}_{OOS} \;\approx\; E_{OOS}[f\hat{f}]\,. \tag{D.13}$$

Thus,

$$\begin{aligned}\frac{MSE_{OOS}(\hat{f})}{E_{OOS}[y^2]} &\approx \frac{\frac{\widehat{MSE}}{MSE(0)} + T_{OOS}^{-1/2}\frac{Error_1}{MSE(0)}}{1 + T_{OOS}^{-1/2}\frac{Error_2}{MSE(0)}} \\ &\approx \Big(\frac{\widehat{MSE}}{MSE(0)} + T_{OOS}^{-1/2}\frac{Error_1}{MSE(0)}\Big)\Big(1 - T_{OOS}^{-1/2}\frac{Error_2}{MSE(0)}\Big) \\ &\approx \frac{\widehat{MSE}}{MSE(0)} + T_{OOS}^{-1/2}\Big(\frac{Error_1}{MSE(0)} - \frac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)\end{aligned} \tag{D.14}$$

Thus,

$$T_{OOS}\text{Var}\Big[\frac{MSE_{OOS}}{E_{OOS}[y^2]}\Big] = \frac{1}{MSE(0)^2}\Sigma_{1,1} + \Big(\frac{\widehat{MSE}}{MSE(0)^2}\Big)^2\Sigma_{2,2} - 2\frac{\widehat{MSE}}{MSE(0)^3}\Sigma_{1,2}\,. \tag{D.15}$$

Thus, we get

$$\begin{aligned}\frac{R^2_{OOS} + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}} &= \frac{1 - \frac{MSE_{OOS}}{E_{OOS}[y^2]} + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}} \\ &= \frac{1 - \frac{\widehat{MSE}}{MSE(0)} - T_{OOS}^{-1/2}\Big(\frac{Error_1}{MSE(0)} - \frac{\widehat{MSE}}{MSE(0)^2}Error_2\Big) + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}} \\ &\leq \frac{1 - \frac{\widehat{MSE} - \widehat{\mathcal{B}}}{MSE(0)} - T_{OOS}^{-1/2}\Big(\frac{Error_1}{MSE(0)} - \frac{\widehat{MSE}}{MSE(0)^2}Error_2\Big) + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}} \\ &= \frac{\frac{E_{OOS}[f_t^2] - \sigma_\varepsilon^2\widehat{\mathcal{L}}}{E_{OOS}[f_t^2] + \sigma_\varepsilon^2} + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}} - \frac{T_{OOS}^{-1/2}\Big(\frac{Error_1}{MSE(0)} - \frac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)}{1 + \widehat{\mathcal{L}}} \\ &= R^2_* - \frac{T_{OOS}^{-1/2}\Big(\frac{Error_1}{MSE(0)} - \frac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)}{1 + \widehat{\mathcal{L}}}\end{aligned} \tag{D.16}$$

The proof of Theorem 3 is complete. □

For the reader's convenience, we state the special case of the ridge regression as a separate proposition.

Corollary D.2 (Probabilistic Lower Bound for R_*^2 for a Ridge Regression) *Suppose that* $E[\varepsilon_t^3] = 0,\ E[\varepsilon_t^4] = 3$. *Then,*

$$\frac{R_{OOS}^2(\hat{f}) + \widehat{\mathcal{L}}(z)}{1 + \widehat{\mathcal{L}}(z)}\,, \tag{D.17}$$

is a $T^{1/2}$*-consistent upper bound for* $\tilde{R}_*^2$ *in the following sense: The event* $\frac{R_{OOS}^2(\hat{f}) + \widehat{\mathcal{L}}(z)}{1 + \widehat{\mathcal{L}}(z)} > \tilde{R}_*^2$ *occurs with vanishing probability:*

$$\limsup_{T, T_{OOS} \to \infty} \text{Prob}\left(\frac{T_{OOS}^{1/2}\left(\tilde{R}_*^2 - \dfrac{R_{OOS}^2(\hat{f}) + \widehat{\mathcal{L}}(z)}{1 + \widehat{\mathcal{L}}(z)} \right)}{\hat{\sigma}_{R^2}} < \alpha \right) \;\leq\; \Phi(\alpha) \tag{D.18}$$

for any $\alpha \leq 0$, where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. Here,

$$
\begin{aligned}
\hat{\sigma}_{R^2} &= \frac{\frac{1}{MSE(0)^2}\Sigma_{1,1}+\Big(\frac{\widehat{MSE}}{MSE(0)^2}\Big)^2\Sigma_{2,2}-2\frac{\widehat{MSE}}{MSE(0)^3}\Sigma_{1,2}}{(1+\widehat{\mathcal{L}})^2} \\
\Sigma_{1,1} &= \frac{T}{T_{OOS}}\sigma^2_{MSE}(1+\widehat{\mathcal{L}})^2 \\
\Sigma_{1,2} &= 2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon \beta'\hat{\Psi}_{OOS}(\beta-\hat{\beta}) \\
\Sigma_{2,2} &= 2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon \beta'\Psi_{OOS}\beta \\
\widehat{MSE} &= \widehat{\mathcal{B}} \;+\; \sigma^2_\varepsilon(1+\widehat{\mathcal{L}}) \\
\sigma^2_{MSE} &= \frac{2\frac{T}{T_{OOS}}\sigma^4_\varepsilon + \sigma^4_\varepsilon\sigma^2_V + \sigma^2_\varepsilon\sigma^2_I + \sigma^2_\varepsilon\sigma^2_{I,OOS}}{(1+\widehat{\mathcal{L}})^2} \\
\sigma^2_V &= 2\frac{1}{T}\operatorname{tr}\Big((zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}\Big) \\
\sigma^2_I &= 4z^2\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta \\
\sigma^2_{I,OOS} &= 4\frac{T}{T_{OOS}}\Big(\widehat{\mathcal{B}}(z) + \;\sigma^2_\varepsilon\widehat{\mathcal{L}}\Big)
\end{aligned}
\tag{D.19}
$$

Proof of Corollary D.2. We have

$$
R^2_{OOS} \;=\; 1 \;-\; \frac{MSE_{OOS}(\hat{f})}{MSE_{OOS}(0)}\,. \tag{D.20}
$$

As above, we suppose that $E[\varepsilon_t^3]=0,\ E[\varepsilon_t^4]=3$. We have

$$
\begin{aligned}
E_{OOS}[y^2] &= E_{OOS}[\varepsilon^2] \;+\; 2E_{OOS}[\varepsilon' S\beta] \;+\; \beta'\hat{\Psi}_{OOS}\beta \\
MSE_{OOS}(\hat{f}) &= E_{OOS}[\varepsilon^2] \;+\; 2E_{OOS}[\varepsilon' S(\beta-\hat{\beta})] \\
&\quad +\; \widehat{\mathcal{V}}(z) \;+\; \widehat{\mathcal{B}}(z) \;+\; \frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\varepsilon\,.
\end{aligned}
\tag{D.21}
$$

Now, asymptotic normality and asymptotic independence of all the terms follow by the same

argument as in the proof of Lemma A.3. All we need to do is compute Σ^2_{OOS}. We have

$$T_{OOS}^{1/2}(E_{OOS}[\varepsilon^2] - \sigma_\varepsilon^2) \;\to\; \mathcal{N}(0, 2\sigma_\varepsilon^4) \tag{D.22}$$

and

$$\frac{T_{OOS}^{1/2} 2E_{OOS}[\varepsilon' S\beta]}{(4\beta'\hat{\Psi}_{OOS}\beta)^{1/2}} \;\to\; \mathcal{N}(0,1)\,, \tag{D.23}$$

Thus,

$$\begin{pmatrix} MSE_{OOS}(\hat{f}) \\ E_{OOS}[y^2] \end{pmatrix} \;=\; \begin{pmatrix} \widehat{MSE} \\ MSE(0) \end{pmatrix} \;+\; \frac{1}{T_{OOS}}\mathcal{N}(0, \Sigma_{Joint}) \tag{D.24}$$

where

$$\Sigma_{Joint} \;=\; \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2} & \Sigma_{2,2} \end{pmatrix} \tag{D.25}$$

where

$$\begin{aligned} \Sigma_{1,1} &= \frac{T_{OOS}}{T}\sigma^2_{MSE}(1+\widehat{\mathcal{L}})^2 \\ \Sigma_{1,2} &= 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\beta'\hat{\Psi}_{OOS}(\beta - \hat{\beta}) \\ \Sigma_{2,2} &= 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\beta'\Psi_{OOS}\beta\,. \end{aligned} \tag{D.26}$$

Here,

$$\beta'\hat{\Psi}_{OOS}(\beta - \hat{\beta}) \;\approx\; \beta'\hat{\Psi}_{OOS}\beta - \beta'\hat{\Psi}_{OOS}\hat{\beta} \tag{D.27}$$

and $\beta'\hat{\Psi}_{OOS}\hat{\beta}$ admits a pivotal estimator

$$E_{OOS}[y'S]\hat{\beta} \;=\; \frac{1}{T_{OOS}}(S_{OOS}\beta+\varepsilon_{OOS})'S_{OOS}\hat{\beta} \;\approx\; \beta'\Psi\hat{\beta}\,. \tag{D.28}$$

Thus,

$$\begin{aligned}\frac{MSE_{OOS}}{E_{OOS}[y^2]} &\approx \frac{\dfrac{\widehat{MSE}}{MSE(0)}+T_{OOS}^{-1/2}\dfrac{Error_1}{MSE(0)}}{1+T_{OOS}^{-1/2}\dfrac{Error_2}{MSE(0)}}\\ &\approx \Big(\frac{\widehat{MSE}}{MSE(0)}+T_{OOS}^{-1/2}\frac{Error_1}{MSE(0)}\Big)\Big(1-T_{OOS}^{-1/2}\frac{Error_2}{MSE(0)}\Big)\\ &\approx \frac{\widehat{MSE}}{MSE(0)}+\;T_{OOS}^{-1/2}\Big(\frac{Error_1}{MSE(0)}-\frac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)\end{aligned} \tag{D.29}$$

Thus,

$$T_{OOS}\mathrm{Var}\Big[\frac{MSE_{OOS}}{E_{OOS}[y^2]}\Big] = \frac{1}{MSE(0)^2}\Sigma_{1,1}+\Big(\frac{\widehat{MSE}}{MSE(0)^2}\Big)^2\Sigma_{2,2}-2\frac{\widehat{MSE}}{MSE(0)^3}\Sigma_{1,2}\,. \tag{D.30}$$

Thus, we get

$$\begin{aligned}
\frac{R^2_{OOS}+\widehat{\mathcal{L}}}{1+\widehat{\mathcal{L}}} &= \frac{1-\dfrac{MSE_{OOS}}{E_{OOS}[y^2]}+\widehat{\mathcal{L}}}{1+\widehat{\mathcal{L}}} \\
&= \frac{1-\dfrac{\widehat{MSE}}{MSE(0)}-\ T_{OOS}^{-1/2}\Big(\dfrac{Error_1}{MSE(0)}-\dfrac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)+\widehat{\mathcal{L}}}{1+\widehat{\mathcal{L}}} \\
&\leq \frac{1-\dfrac{\widehat{MSE}-\widehat{\mathcal{B}}}{MSE(0)}-\ T_{OOS}^{-1/2}\Big(\dfrac{Error_1}{MSE(0)}-\dfrac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)+\widehat{\mathcal{L}}}{1+\widehat{\mathcal{L}}} \qquad \text{(D.31)} \\
&= \frac{\dfrac{\beta'\Psi_{OOS}\beta-\sigma_\varepsilon^2\widehat{\mathcal{L}}}{\beta'\Psi_{OOS}\beta+\sigma_\varepsilon^2}+\widehat{\mathcal{L}}}{1+\widehat{\mathcal{L}}}-\frac{T_{OOS}^{-1/2}\Big(\dfrac{Error_1}{MSE(0)}-\dfrac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)}{1+\widehat{\mathcal{L}}} \\
&= R^2_*\ -\ \frac{T_{OOS}^{-1/2}\Big(\dfrac{Error_1}{MSE(0)}-\dfrac{\widehat{MSE}}{MSE(0)^2}Error_2\Big)}{1+\widehat{\mathcal{L}}}
\end{aligned}$$

The proof of Theorem 3 is complete. □

D.1 Pivotal Bounds for the Variance

Proposition D.3 *Let*

$$\hat{q}_{OOS}\ =\ \widehat{\mathcal{K}}'(\widehat{\mathcal{K}}y-y_{OOS})\,. \qquad \text{(D.32)}$$

Suppose that

$$TT_{OOS}^{-2}E[(f_{OOS}-\hat{f}^s)'\widehat{\mathcal{K}}\widehat{\mathcal{K}}'\big(\widehat{\mathcal{K}}\widehat{\mathcal{K}}'+I\big)\widehat{\mathcal{K}}\widehat{\mathcal{K}}'(f_{OOS}-\hat{f}^s)]\ \to\ 0 \qquad \text{(D.33)}$$

as $T, T_{OOS}\to\infty$. *Then,*

$$\frac{1}{4}\sigma_I^2 =\ TT_{OOS}^{-2}\|(f-\hat{f}^s)\widehat{\mathcal{K}}\|^2\approx\ TT_{OOS}^{-2}\|\hat{q}_{OOS}\|^2\ -\ \sigma_\varepsilon^2\Big(\frac{1}{2}\sigma_V^2\ +\ \frac{T}{T_{OOS}}\widehat{\mathcal{L}}\Big)\,. \qquad \text{(D.34)}$$

Thus,

$$\sigma^2_{MSE} = \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 - \sigma_\varepsilon^4\sigma_V^2 + 4\sigma_\varepsilon^2 T T_{OOS}^{-2}\|\hat{q}_{OOS}\|^2 + 4\sigma_\varepsilon^2\frac{T}{T_{OOS}}\widehat{\mathcal{B}}}{(1+\widehat{\mathcal{L}})^2} \tag{D.35}$$

Proof of Proposition D.3. Recall that

$$y = f + \varepsilon, \qquad y_{OOS} = f_{OOS} + \varepsilon_{OOS}. \tag{D.36}$$

Then,

$$\begin{aligned} \widehat{\mathcal{K}}y - y_{OOS} &= \widehat{\mathcal{K}}(f+\varepsilon) - (f_{OOS} + \varepsilon_{OOS}) \\ &= (\widehat{\mathcal{K}}f - f_{OOS}) + (\widehat{\mathcal{K}}\varepsilon - \varepsilon_{OOS}) \\ &= (\widehat{\mathcal{K}}f - f_{OOS}) + \hat{\varepsilon}_{OOS}, \end{aligned} \tag{D.37}$$

and we have defined

$$\hat{\varepsilon}_{OOS} \equiv \widehat{\mathcal{K}}\varepsilon - \varepsilon_{OOS}. \tag{D.38}$$

By definition of $\hat{q}_{OOS}$,

$$\begin{aligned} \hat{q}_{OOS} &= \widehat{\mathcal{K}}'(\widehat{\mathcal{K}}y - y_{OOS}) \\ &= \widehat{\mathcal{K}}'\left[(\widehat{\mathcal{K}}f - f_{OOS}) + \hat{\varepsilon}_{OOS}\right] \\ &= \widehat{\mathcal{K}}'(\widehat{\mathcal{K}}f - f_{OOS}) + \widehat{\mathcal{K}}'\hat{\varepsilon}_{OOS} \\ &= -\widehat{\mathcal{K}}'(f_{OOS} - \widehat{\mathcal{K}}f) + \tilde{\varepsilon}_{OOS} \\ &= -\widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s) + \tilde{\varepsilon}_{OOS}, \end{aligned} \tag{D.39}$$

where we have defined

$$\tilde{\varepsilon}_{OOS} \equiv \widehat{\mathcal{K}}' \hat{\varepsilon}_{OOS}. \tag{D.40}$$

Taking quadratic forms in (D.39) yields

$$\begin{aligned} TT_{OOS}^{-2} \|\hat{q}_{OOS}\|^2 &= TT_{OOS}^{-2} \left\| - \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s) + \tilde{\varepsilon}_{OOS} \right\|^2 \\ &= TT_{OOS}^{-2} \left\| \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s) \right\|^2 + TT_{OOS}^{-2} \|\tilde{\varepsilon}_{OOS}\|^2 - 2TT_{OOS}^{-2} \langle \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s),\, \tilde{\varepsilon}_{OOS} \rangle. \end{aligned} \tag{D.41}$$

We first show that the cross term in (D.41) is asymptotically negligible. Using $\tilde{\varepsilon}_{OOS} = \widehat{\mathcal{K}}' \hat{\varepsilon}_{OOS}$ and the fact that $E[\hat{\varepsilon}_{OOS} \mid S] = 0$, we have

$$E\Big[\big\langle \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s),\, \tilde{\varepsilon}_{OOS} \big\rangle \;\Big|\; S \Big] = \big\langle \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s),\, E[\tilde{\varepsilon}_{OOS} \mid S] \big\rangle = 0. \tag{D.42}$$

Therefore, by conditional variance,

$$E\Big[\big(2TT_{OOS}^{-2} \, \langle \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s),\, \tilde{\varepsilon}_{OOS} \rangle \big)^2 \;\Big|\; S \Big] = 4T^2 T_{OOS}^{-4} \, \hat{v}(S), \tag{D.43}$$

where

$$\hat{v}(S) := \big\langle \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s),\, E[\tilde{\varepsilon}_{OOS} \tilde{\varepsilon}_{OOS}' \mid S] \, \widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s) \big\rangle. \tag{D.44}$$

Since

$$E[\hat{\varepsilon}_{OOS} \hat{\varepsilon}_{OOS}' \mid S] = \sigma_\varepsilon^2 \big(\widehat{\mathcal{K}} \widehat{\mathcal{K}}' + I \big), \qquad \tilde{\varepsilon}_{OOS} = \widehat{\mathcal{K}}' \hat{\varepsilon}_{OOS}, \tag{D.45}$$

we obtain

$$E[\tilde{\varepsilon}_{OOS}\tilde{\varepsilon}'_{OOS} \mid S] = \sigma_\varepsilon^2\, \widehat{\mathcal{K}}'\big(\widehat{\mathcal{K}}\widehat{\mathcal{K}}' + I\big)\widehat{\mathcal{K}}. \tag{D.46}$$

Hence

$$\hat{v}(S) \;=\; \sigma_\varepsilon^2 (f_{OOS} - \hat{f}^s)'\widehat{\mathcal{K}}\widehat{\mathcal{K}}'\big(\widehat{\mathcal{K}}\widehat{\mathcal{K}}' + I\big)\widehat{\mathcal{K}}\widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s) \tag{D.47}$$

By (D.33), the cross term in (D.41) converges to zero in L_2. Now, we have

$$TT_{OOS}^{-2}\|\hat{q}_{OOS}\|^2 \approx TT_{OOS}^{-2}\Big\|\widehat{\mathcal{K}}'(f_{OOS} - \hat{f}^s)\Big\|^2 + TT_{OOS}^{-2}\|\tilde{\varepsilon}_{OOS}\|^2. \tag{D.48}$$

By the concentration of quadratic forms (Lemma A.1),

$$\begin{aligned}
TT_{OOS}^{-2}\,\|\tilde{\varepsilon}_{OOS}\|^2 &= TT_{OOS}^{-2}\,\hat{\varepsilon}'_{OOS}\,\widehat{\mathcal{K}}\widehat{\mathcal{K}}'\,\hat{\varepsilon}_{OOS} \\
&= TT_{OOS}^{-2}\,\operatorname{tr}\Big[\widehat{\mathcal{K}}\widehat{\mathcal{K}}'\,\hat{\varepsilon}_{OOS}\hat{\varepsilon}'_{OOS}\Big] \\
&\approx TT_{OOS}^{-2}\,\operatorname{tr}\Big[\widehat{\mathcal{K}}\widehat{\mathcal{K}}'\,E\big[\hat{\varepsilon}_{OOS}\hat{\varepsilon}'_{OOS} \mid S\big]\Big] \\
&= TT_{OOS}^{-2}\,\operatorname{tr}\Big[\widehat{\mathcal{K}}\widehat{\mathcal{K}}'\,\sigma_\varepsilon^2\big(\widehat{\mathcal{K}}\widehat{\mathcal{K}}' + I\big)\Big] \\
&= \sigma_\varepsilon^2\,\underbrace{TT_{OOS}^{-2}\,\operatorname{tr}\Big[(\widehat{\mathcal{K}}\widehat{\mathcal{K}}')^2\Big]}_{=\frac{1}{2}\sigma_V^2 \;\text{by (C.3)}} + \sigma_\varepsilon^2\,TT_{OOS}^{-2}\,\operatorname{tr}\Big[\widehat{\mathcal{K}}\widehat{\mathcal{K}}'\Big] \\
&= \sigma_\varepsilon^2\,\frac{1}{2}\,\sigma_V^2 \;+\; \sigma_\varepsilon^2\,\frac{T}{T_{OOS}}\,\widehat{\mathcal{L}} \\
&= \sigma_\varepsilon^2\Big(\frac{1}{2}\,\sigma_V^2 \;+\; \frac{T}{T_{OOS}}\,\widehat{\mathcal{L}}\Big).
\end{aligned} \tag{D.49}$$

Thus, by (C.3) and (D.48),

$$\frac{1}{4}\sigma_I^2 = \; TT_{OOS}^{-2}\|(f_{OOS} - \hat{f}^s)\widehat{\mathcal{K}}\|^2 \approx \; TT_{OOS}^{-2}\|\hat{q}_{OOS}\|^2 \;-\; \sigma_\varepsilon^2\Big(\frac{1}{2}\sigma_V^2 \;+\; \frac{T}{T_{OOS}}\widehat{\mathcal{L}}\Big). \tag{D.50}$$

Plugging this expression into (C.2), we have

$$\begin{aligned}\sigma^2_{MSE} &= \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 + \sigma_\varepsilon^4\sigma_V^2 + \sigma_\varepsilon^2\sigma_I^2 + \sigma_\varepsilon^2\sigma_{I,OOS}^2}{(1+\widehat{\mathcal{L}})^2} \\ &= \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 + \sigma_\varepsilon^4\sigma_V^2 + 4\sigma_\varepsilon^2\Big[TT_{OOS}^{-2}\|\hat{q}_{OOS}\|^2 \;-\; \sigma_\varepsilon^2\Big(\frac{1}{2}\sigma_V^2 \;+\; \frac{T}{T_{OOS}}\widehat{\mathcal{L}}\Big)\Big] + 4\sigma_\varepsilon^2\frac{T}{T_{OOS}}\Big(\widehat{\mathcal{B}} + \;\sigma_\varepsilon^2\widehat{\mathcal{L}}\Big)}{(1+\widehat{\mathcal{L}})^2} \\ &= \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 - \sigma_\varepsilon^4\sigma_V^2 + 4\sigma_\varepsilon^2 TT_{OOS}^{-2}\|\hat{q}_{OOS}\|^2 \;+ 4\sigma_\varepsilon^2\frac{T}{T_{OOS}}\widehat{\mathcal{B}}}{(1+\widehat{\mathcal{L}})^2}\end{aligned} \tag{D.51}$$

The proof of Proposition D.3 is complete. □

For the reader's convenience, we state the special case of the ridge regression as a separate proposition.

Proposition D.4 *Let*

$$\hat{q}_{OOS} \;=\; S(zI+\hat{\Psi})^{-1}\frac{1}{T_{OOS}}S'_{OOS}\Big(S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'y - y_{OOS}\Big)\,. \tag{D.52}$$

Then,

$$\begin{aligned}\frac{1}{4}\sigma_I^2 &= z^2\beta'(zI+\hat{\Psi})^{-1}\Psi_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi_{OOS}(zI+\hat{\Psi})^{-1}\beta \\ &\approx \frac{1}{T}\|\hat{q}_{OOS}\|^2 \;-\; \sigma_\varepsilon^2\Big(\frac{1}{2}\sigma_V^2 \;+\; \frac{T}{T_{OOS}}\widehat{\mathcal{L}}\Big)\,.\end{aligned} \tag{D.53}$$

Thus,

$$\sigma^2_{MSE} \;=\; \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 - \sigma_\varepsilon^4\sigma_V^2 + 4\sigma_\varepsilon^2\frac{1}{T}\|\hat{q}_{OOS}\|^2 \;+ 4\sigma_\varepsilon^2\frac{T}{T_{OOS}}\widehat{\mathcal{B}}(z)}{(1+\widehat{\mathcal{L}})^2} \tag{D.54}$$

Proof of Proposition D.4. Now,

$$\begin{aligned} & S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'y - y_{OOS} \\ &= S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'(S\beta+\varepsilon) - (S_{OOS}\beta+\varepsilon_{OOS}) \\ &= S_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}\beta - S_{OOS}\beta + \ S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'\varepsilon - \varepsilon_{OOS} \\ &= \ -zS_{OOS}(zI+\hat{\Psi})^{-1}\beta \ + \ \hat{\varepsilon}_{OOS} \end{aligned} \tag{D.55}$$

where we have defined

$$\hat{\varepsilon}_{OOS} \ = \ S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'\varepsilon - \varepsilon_{OOS} \tag{D.56}$$

and, hence,

$$\begin{aligned} \hat{q}_{OOS} \ &= S(zI+\hat{\Psi})^{-1}\frac{1}{T_{OOS}}S'_{OOS}\Big(S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'y - y_{OOS}\Big) \\ &= S(zI+\hat{\Psi})^{-1}\frac{1}{T_{OOS}}S'_{OOS}\Big(-zS_{OOS}(zI+\hat{\Psi})^{-1}\beta \ + \ \hat{\varepsilon}_{OOS}\Big) \\ &= \ -zS(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta + \tilde{\varepsilon}_{OOS}, \end{aligned} \tag{D.57}$$

where

$$\tilde{\varepsilon}_{OOS} \ = \ S(zI+\hat{\Psi})^{-1}\frac{1}{T_{OOS}}S'_{OOS}\hat{\varepsilon}_{OOS}, \tag{D.58}$$

so that

$$\begin{aligned} & \frac{1}{T}\hat{q}'_{OOS}\hat{q}_{OOS} \\ &= \frac{1}{T}\Big(-zS(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta + \tilde{\varepsilon}_{OOS}\Big)'\Big(-zS(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta + \tilde{\varepsilon}_{OOS}\Big) \\ &= z^2\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta + \ \frac{1}{T}\|\tilde{\varepsilon}_{OOS}\|^2 \\ & \ - \ \frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\tilde{\varepsilon}_{OOS}\,. \end{aligned}$$

(D.59)

We now show that the cross term in (D.59)

$$-\frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\tilde{\varepsilon}_{OOS} \tag{D.60}$$

is asymptotically negligible. Recall that

$$\hat{\Psi}=\frac{1}{T}S'S, \qquad \hat{\Psi}_{OOS}=\frac{1}{T_{OOS}}S'_{OOS}S_{OOS}. \tag{D.61}$$

Using the fact that $\tilde{\varepsilon}_{OOS} = S(zI+\hat{\Psi})^{-1}\frac{1}{T_{OOS}}S'_{OOS}\hat{\varepsilon}_{OOS}$ and $E[\hat{\varepsilon}_{OOS} \mid S, S_{OOS}] = 0$, we have

$$E\left[-\frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\tilde{\varepsilon}_{OOS}\,\middle|\, S, S_{OOS}\right]=0. \tag{D.62}$$

Its conditional second moment is

$$\begin{aligned} &E\Big[\Big(-\frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\tilde{\varepsilon}_{OOS}\Big)^2\,\Big|\, S, S_{OOS}\Big] \\ &=\frac{4z^2}{T^2}\,E\Big[\big(\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\tilde{\varepsilon}_{OOS}\big)^2\,\Big|\, S, S_{OOS}\Big]. \end{aligned} \tag{D.63}$$

Since

$$\begin{aligned} E[\hat{\varepsilon}_{OOS}\hat{\varepsilon}'_{OOS} \mid S, S_{OOS}] &= \sigma_\varepsilon^2\Big(S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T^2}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}+I\Big) \\ &= \sigma_\varepsilon^2\Big(\frac{1}{T}S_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}S'_{OOS}+I\Big) \end{aligned} \tag{D.64}$$

and by (D.58), we obtain

$$
\begin{aligned}
&E[\tilde{\varepsilon}_{OOS}\tilde{\varepsilon}'_{OOS} \mid S, S_{OOS}] \\
&= \frac{\sigma_\varepsilon^2}{T_{OOS}^2} S(zI+\hat{\Psi})^{-1} S'_{OOS}\Big(\frac{1}{T} S_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}S'_{OOS} + I\Big) S_{OOS}(zI+\hat{\Psi})^{-1}S' \\
&= \frac{\sigma_\varepsilon^2}{TT_{OOS}^2} S(zI+\hat{\Psi})^{-1} S'_{OOS}S_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}S'_{OOS}S_{OOS}(zI+\hat{\Psi})^{-1}S' \\
&\quad + \frac{\sigma_\varepsilon^2}{T_{OOS}^2} S(zI+\hat{\Psi})^{-1} S'_{OOS}S_{OOS}(zI+\hat{\Psi})^{-1}S' \\
&= \frac{\sigma_\varepsilon^2}{T} S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S' \\
&\quad + \frac{\sigma_\varepsilon^2}{T_{OOS}} S(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'
\end{aligned}
\tag{D.65}
$$

For simplicity, let $R := (zI+\hat{\Psi})^{-1}$. Thus,

$$
\begin{aligned}
&E\Big[\big(\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\tilde{\varepsilon}_{OOS}\big)^2 \;\Big|\; S, S_{OOS}\Big] \\
&= \beta' R\hat{\Psi}_{OOS}RS'\, E[\tilde{\varepsilon}_{OOS}\tilde{\varepsilon}'_{OOS} \mid S, S_{OOS}]\, SR\hat{\Psi}_{OOS}R\beta \\
&= \beta' R\hat{\Psi}_{OOS}RS'\,\frac{\sigma_\varepsilon^2}{T} SR\hat{\Psi}_{OOS}R\hat{\Psi}R\hat{\Psi}_{OOS}RS'SR\hat{\Psi}_{OOS}R\beta + \beta' R\hat{\Psi}_{OOS}RS'\,\frac{\sigma_\varepsilon^2}{T_{OOS}} R\hat{\Psi}_{OOS}RS'SR\hat{\Psi}_{OOS}R\beta \\
&= \sigma_\varepsilon^2 T\,\beta' M_1\beta \;+\; \sigma_\varepsilon^2\frac{T^2}{T_{OOS}}\,\beta' M_2\beta,
\end{aligned}
\tag{D.66}
$$

where

$$
\begin{aligned}
M_1 &= R\hat{\Psi}_{OOS}R\hat{\Psi}R\hat{\Psi}_{OOS}R\hat{\Psi}R\hat{\Psi}_{OOS}R\hat{\Psi}R\hat{\Psi}_{OOS}R, \\
M_2 &= R\hat{\Psi}_{OOS}R\hat{\Psi}R\hat{\Psi}_{OOS}R\hat{\Psi}R\hat{\Psi}_{OOS}R.
\end{aligned}
\tag{D.67}
$$

We now bound $\beta' M_1\beta$ and $\beta' M_2\beta$ explicitly. Since M_1 and M_2 are symmetric, for any vector

β, we have

$$E[\beta' M_i \beta] \;\le\; \|\beta\|^2 \, E[\|M_i\|], \qquad i = 1, 2. \tag{D.68}$$

Note that, if λ_i are the eigenvalues of $\hat{\Psi}$, then the eigenvalues of $\hat{\Psi}R$ are $\dfrac{\lambda_i}{z+\lambda_i} \in [0,1)$, so $\|R\hat{\Psi}\| \le 1$. From the structure of M_1 and M_2 we then obtain

$$\begin{aligned} \|M_1\| &\;\le\; \|R\hat{\Psi}_{OOS}\|\|R\hat{\Psi}\|\|R\hat{\Psi}_{OOS}\|\|R\hat{\Psi}\|\|R\hat{\Psi}_{OOS}\|\|R\hat{\Psi}\|\|R\hat{\Psi}_{OOS}R\| \le \frac{\|\hat{\Psi}_{OOS}\|^4}{z^5}, \\ \|M_2\| &\;\le\; \|R\hat{\Psi}_{OOS}\|\|R\hat{\Psi}\|\|R\hat{\Psi}_{OOS}\|\|R\hat{\Psi}\|\|R\hat{\Psi}_{OOS}R\| \le \frac{\|\hat{\Psi}_{OOS}\|^3}{z^4}. \end{aligned} \tag{D.69}$$

Under our sub-Gaussian assumptions on the rows of S_{OOS}, Lemma A.2 applied to S_{OOS} with $k = 3, 4$, implies that

$$E[\|\hat{\Psi}_{OOS}\|^3] \le C_3, \qquad E[\|\hat{\Psi}_{OOS}\|^4] \le C_4 \tag{D.70}$$

where C_3 and C_4 depend only on the sub-Gaussian parameter K but not on T_{OOS} or P. Hence,

$$E[\beta' M_1 \beta] \;\le\; \frac{C_4}{z^5}\|\beta\|^2, \qquad E[\beta' M_2 \beta] \;\le\; \frac{C_3}{z^4}\|\beta\|^2. \tag{D.71}$$

Plugging these bounds into the expression for the conditional second moment and then taking expectations, we obtain

$$E\Big[\Big(-\frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}S'\tilde{\varepsilon}_{OOS}\Big)^2\Big] \;\le\; 4\sigma_\varepsilon^2\|\beta\|^2\left[\frac{1}{T}\,\frac{C_4}{z^3} + \frac{1}{T_{OOS}}\,\frac{C_3}{z^2}\right]. \tag{D.72}$$

In particular, if $T \to \infty$ and $T_{OOS} \to \infty$, the right-hand side converges to zero, so the cross

term in (D.59) converges to zero in L_2. Now, we have

$$\frac{1}{T}\hat{q}'_{OOS}\hat{q}_{OOS} \approx \; z^2\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta + \frac{1}{T}\|\tilde{\varepsilon}_{OOS}\|^2. \tag{D.73}$$

Here, by the concentration of quadratic forms (Lemma A.1),

$$\begin{aligned}
&\frac{1}{T}\|\tilde{\varepsilon}_{OOS}\|^2 \\
&= \frac{1}{T_{OOS}^2}\hat{\varepsilon}'_{OOS}S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}\hat{\varepsilon}_{OOS} \\
&= \frac{1}{T_{OOS}^2}\operatorname{tr}[S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}\hat{\varepsilon}_{OOS}\hat{\varepsilon}'_{OOS}] \\
&\approx \frac{1}{T_{OOS}^2}\operatorname{tr}\Big[S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}\sigma_\varepsilon^2\Big(S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T^2}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}+I\Big)\Big] \\
&= \sigma_\varepsilon^2\frac{1}{T_{OOS}^2}\operatorname{tr}\Big[S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}\Big(S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T^2}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}\Big)\Big] \\
&\quad + \sigma_\varepsilon^2\frac{1}{T_{OOS}^2}\operatorname{tr}[S_{OOS}(zI+\hat{\Psi})^{-1}\frac{1}{T}S'S(zI+\hat{\Psi})^{-1}S'_{OOS}] \\
&= \sigma_\varepsilon^2\underbrace{\frac{1}{T}\operatorname{tr}[(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}]}_{=0.5\sigma_V^2 \;\; by \;\; (C.13)} \\
&\quad + \sigma_\varepsilon^2\frac{1}{T_{OOS}}\operatorname{tr}[(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}] \\
&= \sigma_\varepsilon^2\Big(\frac{1}{2}\sigma_V^2 \; + \; \frac{T}{T_{OOS}}\widehat{\mathcal{L}}\Big).
\end{aligned} \tag{D.74}$$

Thus, by (C.13) and (D.59),

$$\begin{aligned}
\frac{1}{4}\sigma_I^2 &= z^2\beta'(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\hat{\Psi}_{OOS}(zI+\hat{\Psi})^{-1}\beta \\
&\approx \frac{1}{T}\|\hat{q}_{OOS}\|^2 - \frac{1}{T}\|\tilde{\varepsilon}_{OOS}\|^2.
\end{aligned} \tag{D.75}$$

Plugging this expression into (C.12), we have

$$\begin{aligned}\sigma^2_{MSE} &= \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 + \sigma_\varepsilon^4\sigma_V^2 + \sigma_\varepsilon^2\sigma_I^2 + \sigma_\varepsilon^2\sigma^2_{I,OOS}}{(1+\widehat{\mathcal{L}})^2} \\ &= \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 + \sigma_\varepsilon^4\sigma_V^2 + 4\sigma_\varepsilon^2\Big[\frac{1}{T}\|\hat{q}_{OOS}\|^2 \;-\; \sigma_\varepsilon^2\Big(\frac{1}{2}\sigma_V^2 \;+\; \frac{T}{T_{OOS}}\widehat{\mathcal{L}}\Big)\Big] + 4\sigma_\varepsilon^2\frac{T}{T_{OOS}}\Big(\widehat{\mathcal{B}}(z) + \;\sigma_\varepsilon^2\widehat{\mathcal{L}}\Big)}{(1+\widehat{\mathcal{L}})^2} \\ &= \frac{2\frac{T}{T_{OOS}}\sigma_\varepsilon^4 - \sigma_\varepsilon^4\sigma_V^2 + 4\sigma_\varepsilon^2\frac{1}{T}\|\hat{q}_{OOS}\|^2 \;+ 4\sigma_\varepsilon^2\frac{T}{T_{OOS}}\widehat{\mathcal{B}}(z)}{(1+\widehat{\mathcal{L}})^2}\end{aligned} \tag{D.76}$$

The proof of Proposition D.4 is complete. □

D.2 The Final Pivotal Estimator

By the continuous mapping theorem, in estimating σ_{R^2}, we can replace any quantity with its consistent estimator, and we can replace σ_{R^2} with any consistent upper bound. We use this observation repeatedly below. We have

$$\begin{aligned}\Sigma_{1,2} &= 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\beta'\hat{\Psi}_{OOS}(\beta-\hat{\beta}) \\ &\underbrace{\approx}_{\text{(D.13)}} 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\Big(\beta'\hat{\Psi}_{OOS}\beta - E_{OOS}[y'S]\hat{\beta}\Big) \\ &\approx 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\Big(MSE_{OOS}(0) - \sigma_\varepsilon^2 - E_{OOS}[y'S]\hat{\beta}\Big) \\ &= -2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\Big(MSE_{OOS}(0) - E_{OOS}[y'S]\hat{\beta}\Big) \\ \Sigma_{2,2} &= 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\beta'\Psi_{OOS}\beta \\ &\approx 2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2\Big(MSE_{OOS}(0) - \sigma_\varepsilon^2\Big) \\ &\approx -2\sigma_\varepsilon^4 + 4\sigma_\varepsilon^2 MSE(0) \\ \widehat{\mathcal{B}}(z) &\approx \widehat{MSE} - (1+\widehat{\mathcal{L}})\sigma_\varepsilon^2\end{aligned} \tag{D.77}$$

Thus,

$$
\begin{aligned}
&(1+\widehat{\mathcal{L}})^2\hat{\sigma}_{R^2}MSE(0)^4 \\
&= MSE(0)^2\Sigma_{1,1} + (\widehat{MSE})^2\Sigma_{2,2} - 2MSE(0)\widehat{MSE}\Sigma_{1,2} \\
&= MSE(0)^2\frac{T_{OOS}}{T}\sigma^2_{MSE}(1+\widehat{\mathcal{L}})^2 + (\widehat{MSE})^2\Big(-2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon MSE(0)\Big) \\
&\quad - 2MSE(0)\widehat{MSE}\Big[-2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon\Big(MSE_{OOS}(0) - E_{OOS}[y'S]\hat{\beta}\Big)\Big] \\
&= MSE(0)^2\frac{T_{OOS}}{T}\Big[2\frac{T}{T_{OOS}}\sigma^4_\varepsilon - \sigma^4_\varepsilon\sigma^2_V + 4\sigma^2_\varepsilon\frac{1}{T}\|\hat{q}_{OOS}\|^2 \; + 4\sigma^2_\varepsilon\frac{T}{T_{OOS}}\Big(\widehat{MSE} - (1+\widehat{\mathcal{L}})\sigma^2_\varepsilon\Big)\Big] \\
&\quad + (\widehat{MSE})^2\Big(-2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon MSE(0)\Big) - 2MSE(0)\widehat{MSE}\Big[-2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon\Big(MSE_{OOS}(0) - E_{OOS}[y'S]\hat{\beta}\Big)\Big] \\
&= A_2\sigma^4_\varepsilon \; + \; A_1\sigma^2_\varepsilon
\end{aligned}
\tag{D.78}
$$

where

$$
\begin{aligned}
A_2 \; &= MSE(0)^2\frac{T}{T_{OOS}}\Big(2\frac{T}{T_{OOS}} - \sigma^2_V - 4\frac{T}{T_{OOS}}(1+\widehat{\mathcal{L}})\Big) \\
&\quad - \; 2(\widehat{MSE})^2 \; + \; 4MSE(0)\widehat{MSE}; \\
A_1 \; &= MSE(0)^2\frac{T}{T_{OOS}}\Big(4\frac{1}{T}\|\hat{q}_{OOS}\|^2 + 4\frac{T}{T_{OOS}}\widehat{MSE}\Big) \\
&\quad + 4(\widehat{MSE})^2MSE(0) \; - \; 8MSE(0)\widehat{MSE}\Big(MSE_{OOS}(0) - E_{OOS}[y'S]\hat{\beta}\Big)\,.
\end{aligned}
\tag{D.79}
$$

We can now build an asymptotically consistent, pivotal upper bound on $\sigma^2_{R^2}$ as

$$
\sigma^2_{R^2} \; \leq \min_{\sigma^2_\varepsilon \in [0, \min_z MSE_{OOS}(\hat{\beta}(z))/(1+\hat{\mathcal{L}}(z))]} (A_2\sigma^4_\varepsilon \; + \; A_1\sigma^2_\varepsilon) \; + \; O(T^{-1/2})\,. \tag{D.80}
$$

E Proof of Proposition 8

Our goal is to show that

$$\widehat{\mathcal{L}} = \frac{1}{T}\operatorname{tr}\Big(\Psi\hat{\Psi}(zcI+\hat{\Psi})^{-2}\Big) \tag{E.1}$$

and

$$\widetilde{\mathcal{L}} \;=\; \frac{\frac{1}{T}\operatorname{tr}\big((zcI+SS'/T)^{-2}\big)}{\Big(\frac{1}{T}\operatorname{tr}\big((zcI+SS'/T)^{-1}\big)\Big)^2} - 1\,. \tag{E.2}$$

satisfy $\widehat{\mathcal{L}} \approx \widetilde{\mathcal{L}}$. From Proposition A.1, we know that

$$\frac{1}{T}\operatorname{tr}\Big(\Psi(zcI+\hat{\Psi})^{-1}\Big) \;\approx\; \hat{\xi}(z;c) \;=\; -1 \;+\; \frac{1}{1-c+cz\hat{m}(-z)}\,. \tag{E.3}$$

and

$$-\frac{1}{T}\operatorname{tr}\Big(\Psi(zcI+\hat{\Psi})^{-2}\Big) \;=\; \hat{\xi}'(z;c) \;=\; -\,\frac{c(\hat{m}(-z)-z\hat{m}'(-z))}{(1-c+cz\hat{m}(-z))^2}\,, \tag{E.4}$$

so that

$$\begin{aligned} 1+\widehat{\mathcal{L}} =&\; 1+\hat{\xi}(z;c)+z\hat{\xi}'(z;c) \\ \approx&\; \frac{1}{1-c+cz\hat{m}(-z)} - \frac{cz(\hat{m}(-z)-z\hat{m}'(-z))}{(1-c+cz\hat{m}(-z))^2} \\ =&\; \frac{1-c+cz^2\hat{m}'(-z)}{(1-c+cz\hat{m}(-z))^2}\,. \end{aligned} \tag{E.5}$$

Let

$$\tilde{m}(z) \;=\; (1-c)z^{-1}+c\hat{m}(-z)\,. \tag{E.6}$$

Then,

$$\frac{d}{dz}\tilde{m}(-z) \;=\; -(1-c)z^{-2} - c\hat{m}'(-z) \tag{E.7}$$

and, hence, we get

$$1+\widehat{\mathcal{L}} \approx \; \frac{-\frac{d}{dz}\tilde{m}(-z)}{\tilde{m}(-z)^2}\,. \tag{E.8}$$

Now, by direct calculation, the matrices SS' and $S'S$ have the same eigenvalues, up to $P-T$ zero eigenvalues. As a result,

$$\tilde{m}(-z) \;=\; (1-c)z^{-1} + cP^{-1}\operatorname{tr}((zI+S'S/T)^{-1}) \;=\; T^{-1}\operatorname{tr}((zI+SS'/T)^{-1})\,, \tag{E.9}$$

and the claim follows.

F Bayesian Risk and Optimal Ridge

In this section, we provide Bayesian foundation for ridge regression and discuss when ridge and its modifications are Bayes-optimal and, hence, cannot be dominated by any other machine learning model (linear or nonlinear). What is special about our setting is that, in high dimensions, due to concentration phenomena (where random objects in high dimensions become non-random), our bounds hold almost surely and not just in expectation.

F.1 Bayesian Optimality

Suppose that nature samples a parameter vector $\theta \in \mathbb{R}^D$ from a distribution $p(\theta)d\theta$. The agent observes i.i.d. samples $X_t \sim p(X|\theta),\ \mathbf{X} = (X_t)_{t=1}^T$. His objective is to solve a utility

optimization problem

$$\max_{\pi(\mathbf{X})} E[U(\pi(\mathbf{X}),\theta)] \;=\; \max_{\pi(\mathbf{X})} \int E[U(\pi(\mathbf{X}),\theta)|\theta]p(\theta)d\theta\,. \tag{F.1}$$

In the real world, the the dimension D is large enough, no agent can know (or have any reasonable algorithm to finy) the true distribution $p(\theta)$ from which nature samples θ. A rational agent is thus forced to choose a prior $p^{subjective}(\theta)$, and then learn from the observations of $\mathbf{X}$. When D is large relative to T, this learning will not converge due to limits to learning. This convergence breaks down *even if the agent has the optimal prior* $p(\theta)$. In this section, we discuss the implications of these fundamental results for limits to learning.

Suppose that the agent does have the (infeasible) optimal prior $p(\theta)$. By the law of iterated expectations, we can rewrite his objective as

$$E[U(\pi(\mathbf{X}),\theta)] \;=\; E[E[U(\pi(\mathbf{X}),\theta)|\mathbf{X}]]\,. \tag{F.2}$$

Let

$$\pi_*(\mathbf{X}) \;=\; \arg\max_{\pi} E[U(\pi,\theta)|\mathbf{X}] \;=\; \arg\max_{\pi} \int U(\pi,\theta)p(\theta|\mathbf{X})d\theta \tag{F.3}$$

be the optimal Bayesian policy. Then, we get the simple, classical result.

Theorem F.1 (Bayesian Policies are Optimal) *Suppose that*

$$E[|E[U(\pi_*(\mathbf{X}),\theta)|\mathbf{X}]|] \;<\; \infty\,. \tag{F.4}$$

Then,

$$\pi_*(\mathbf{X}) \;\in\; \arg\max_{\pi(\mathbf{X})} E[U(\pi(\mathbf{X}),\theta)] \tag{F.5}$$

F.2 Bayesian Optimality and The Best Feasible Policy

In this subsection, we apply Theorem F.1 to the linear predictive setting studied in the main body of the paper.

Lemma F.1 (Bayesian updating) *Suppose that*

$$y_t \;=\; \beta' S_{t-1} \;+\; \varepsilon_t\,. \tag{F.6}$$

Suppose also that the agent's prior about β *is* $\beta \sim N(0,\Sigma_\beta)$. *Let* $y = (y_\tau)_{\tau=1}^t$, *and* $S = (S_\tau)_{\tau=0}^{t-1} \in \mathbb{R}^{t\times P}$. *The agent's posterior distribution is Gaussian,* $\beta|\mathcal{F}_t \;\sim\; N(\hat\beta_t,\; \hat\Sigma_t)$, *with*

$$\begin{aligned}
\hat\beta_{1,t} \;&=\; (\sigma_\varepsilon^2\Sigma_\beta^{-1} + S'S)^{-1}S'y \\
\hat\Sigma_{1,t} \;&=\; (\sigma_\varepsilon^2\Sigma_\beta^{-1} + S'S)^{-1}\sigma_\varepsilon^2\,.
\end{aligned} \tag{F.7}$$

Proof of Lemma F.1. In the vector form, we have

$$d \;=\; S\beta \;+\; \varepsilon \tag{F.8}$$

The agent believes that the vector $\beta \in \mathbb{R}^P$ is sampled at time zero from $\mathcal{N}(0,\Sigma_\beta)$. Signals have a covariance matrix ψ_t. By the Gaussian projection theorem:

$$\begin{aligned}
E[\beta|d] \;&=\; \mathrm{Cov}[\beta,d|S]\mathrm{Cov}[d|S]^{-1}d \\
\mathrm{Var}[\beta|d] \;&=\; \mathrm{Var}[\beta] \;-\; \mathrm{Cov}[\beta,d|S]\mathrm{Cov}[d|S]^{-1}\mathrm{Cov}[\beta,d|S]'
\end{aligned} \tag{F.9}$$

We have

$$\begin{aligned} \text{Cov}[d|S] &= (\sigma_\varepsilon^2 I_{t\times t} + S\Sigma_\beta S') \\ \text{Cov}[\beta, d|S] &= E[\beta d'] = \Sigma_\beta S' \end{aligned} \tag{F.10}$$

and hence

$$\text{Var}[\beta|d] = \Sigma_\beta - \Sigma_\beta S'(\sigma_\varepsilon^2 I_{t\times t} + S\Sigma_\beta S')^{-1} S\Sigma_\beta. \tag{F.11}$$

Thus, his posterior after t observations is thus $\beta \approx N(\hat{\beta}_{1,t}, \hat{\Sigma}_{1,t})$, where

$$\begin{aligned} \hat{\beta}_{1,t} &= \Sigma_\beta S'(\sigma_\varepsilon^2 I_{t\times t} + S\Sigma_\beta S')^{-1} d \\ \hat{\Sigma}_{1,t} &= \Sigma_\beta - \Sigma_\beta S'(\sigma_\varepsilon^2 I_{t\times t} + S\Sigma_\beta S')^{-1} S\Sigma_\beta \end{aligned} \tag{F.12}$$

Now, define $\tilde{S} = S\Sigma_\beta^{1/2}$. Then,

$$\begin{aligned} &\Sigma_\beta S'(\sigma_\varepsilon^2 I_{t\times t} + S\Sigma_\beta S')^{-1} \\ &= \Sigma_\beta^{1/2} \tilde{S}'(\sigma_\varepsilon^2 I_{t\times t} + \tilde{S}\tilde{S}')^{-1} S' \\ &= \Sigma_\beta^{1/2} (\sigma_\varepsilon^2 I_{P\times P} + \tilde{S}'\tilde{S})^{-1} \tilde{S}' \\ &= (\sigma_\varepsilon^2 \Sigma_\beta^{-1} + S'S)^{-1} S' \end{aligned} \tag{F.13}$$

so that

$$\begin{aligned} \hat{\beta}_{1,t} &= (\sigma_\varepsilon^2 \Sigma_\beta^{-1} + S'S)^{-1} S'y \\ \hat{\Sigma}_{1,t} &= \Sigma_\beta - \Sigma_\beta S'(\sigma_\varepsilon^2 I_{t\times t} + S\Sigma_\beta S')^{-1} S\Sigma_\beta = (\sigma_\varepsilon^2 \Sigma_\beta^{-1} + S'S)^{-1} \sigma_\varepsilon^2 . \end{aligned} \tag{F.14}$$

□

We can now prove the following result.

Proposition F.2 *Suppose that β is sampled at time zero from $N(0, \Sigma_\beta)$. Then,*

$$\pi_t^* \;=\; S_t'(\sigma_\varepsilon^2 \Sigma_\beta^{-1} + S'S)^{-1} S'y \tag{F.15}$$

is Bayes optimal in the sence that

$$E_\beta[(y_{t+1} - \pi_t^*)^2] \;\leq\; E_\beta[(y_t - \Pi(S, y))^2] \tag{F.16}$$

for any map $\Pi : \mathbb{R}^{t(P+1)} \to \mathbb{R}$.

The proposition F.2 states a classic result: If an economic agent knows the true optimal prior from which the β vector is sampled, then the Bayes rule is optimal on average: No other machine learning algorithm $\Pi(\cdot, \cdot)$ can beat Bayesian learning with an optimally chosen prior.

By direct calculation, (F.13) implies the following result.

Lemma F.2 *Prediction π_t^* coincides with that in a linear ridge regression with $z = \sigma_\varepsilon^2$ and S_t replaced by $\tilde{S}_t = \Sigma_\beta^{1/2} S_t$:*

$$\pi_t^* \;=\; \tilde{S}_t'(\sigma_\varepsilon^2 I + \tilde{S}'\tilde{S})^{-1} \tilde{S}'y \tag{F.17}$$

Lemma F.2 combined with Theorem 2 implies the following result.

Theorem F.3 *For any Machine Learning model $\Pi(S, y)$, we have*

$$E_\beta[(y_{T+1} - \Pi(S, y))^2] \;\geq\; \underbrace{E_\beta[\liminf (y_{T+1} - \pi_T^*)^2]}_{best\ \ feasible\ \ MSE} \;\geq\; (1 + \mathcal{L}_\beta(z; c))\sigma_\varepsilon^2 \,, \tag{F.18}$$

where

$$\mathcal{L}_\beta(z;c) \;=\; \lim \frac{\frac{1}{T}\operatorname{tr}((\sigma_\varepsilon^2 I + \tilde{S}\tilde{S}')^{-2}))}{(\frac{1}{T}\operatorname{tr}((\sigma_\varepsilon^2 I + \tilde{S}\tilde{S}')^{-1}))^2} - 1\,. \tag{F.19}$$

Furthermore, the best feasible MSE asymptotics is given by

$$E_T[(y_{T+1} - \pi_T^*)^2] \;\underbrace{-\,(1+\mathcal{L}_\beta(z;c))((Z_*^\beta)^2 \operatorname{tr}\Big(\tilde{\Psi}(\tilde{\Psi}+Z_*^\beta I)^{-2}\Sigma_\beta\Big) + \sigma_\varepsilon^2)}_{best\ \ feasible\ \ asymptotic\ \ MSE} \;\to\; 0 \tag{F.20}$$

in probability, where Z_*^β is defined with $\tilde{S}$ instead of S, and where $\tilde{\Psi} \;=\; \Sigma_\beta^{1/2}\Psi\Sigma_\beta^{1/2} = E[\tilde{S}'\tilde{S}]$.

Proof of Theorem F.3. From the proof of Theorem 2, we know that the asymptotic MSE is given by

$$(1+\mathcal{L}_\beta(z;c))(Z_*^2\beta'\Sigma_\beta^{1/2}\Psi\Sigma_\beta^{1/2}(\Sigma_\beta^{1/2}\Psi\Sigma_\beta^{1/2} + Z_*^\beta I)^{-2}\beta + \sigma_\varepsilon^2)\,. \tag{F.21}$$

By the concentration of quadratic forms, the first term is converging to

$$\beta'\tilde{\Psi}(\tilde{\Psi}+Z_*^\beta I)^{-2}\beta \approx \;\operatorname{tr}\Big(\tilde{\Psi}(\tilde{\Psi}+Z_*^\beta I)^{-2}\Sigma_\beta\Big) \tag{F.22}$$

where we have defined

$$\tilde{\Psi} \;=\; \Sigma_\beta^{1/2}\Psi\Sigma_\beta^{1/2}\,. \tag{F.23}$$

□

It would be great to develop techniques for computing the best feasible MSE in (F.20). However, this would require estimating Σ_β and this is a highly complex task that we leave for future research.

G GARCH Simulation

We draw innovations

$$z_t \sim \mathcal{N}(0,1), \qquad t = 1, \ldots, T, \tag{G.1}$$

and compute the conditional variance according to

$$\sigma_t^2 \;=\; \omega \;+\; \alpha\, y_{t-1}^2 \;+\; \beta\, \sigma_{t-1}^2, \qquad t \geq 2, \tag{G.2}$$

with an initialization such as

$$\sigma_1^2 \;=\; \frac{\omega}{1-\alpha-\beta}. \tag{G.3}$$

The GARCH(1,1) observations are then generated as

$$y_t^{GARCH(1,1)} \;=\; \sigma_t\, z_t, \qquad t = 1, \ldots, T. \tag{G.4}$$

We use $\omega = 0.5, \alpha = 0.05, \beta = 0.9$.

H Recursive Ridge

We proceed following the approach of Yan and Zheng (2017); Chen and Dim (2023); Li et al. (2025): Given the basic Welch and Goyal (2008) signals transformed using the Procedure 1, we build pairwise sums and products of these variables and their non-linear transformations. Then, we pre-select the most powerful signals based on their in-sample importance using a modification of the empirical Bayes methodology of Chen and Dim (2023). And then, we

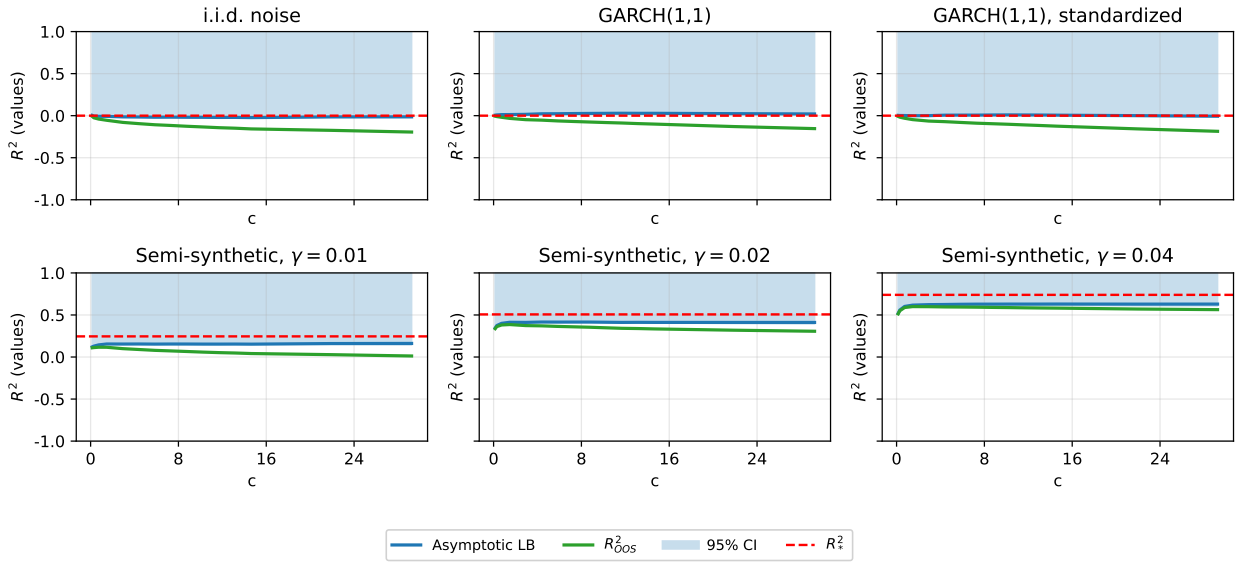


FIGURE 7 – Semi-synthetic simulation (54), with activation=ReLu and $z_{ref} = 1$ in (52). In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. i.i.d. noise has $y_{t+1} = \varepsilon_{t+1} \sim \mathcal{N}(0, 1)$. GARCH(1,1) has $y_{t+1} = \varepsilon_{t+1}$ being a GARCH(1,1) noise defined in Appendix G. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51), increasing from $P_1 = 100$ to $P_1 = 20000$. R^2_* is computed in (55). Values of $R^2_{OOS} < -1$ are not shown. γ values are selected to achieve R^2_* of 0, 0.25, 0.5, 0.75, respectively.

run a ridge regression on these pre-selected signals and follow the same procedure as with the random feature ridge regression from the main text.

I Plots with $z_{ref} = 1$

In the main text, we use (52) with $z_{ref} = 0.01$ because, to achieve a large $\widehat{\mathcal{L}}$, we need a small z. Here, we report the results with $z_{ref} = 1$ in (52) to show how R^2_{OOS} improves, while the LLG-correction becomes negligible.

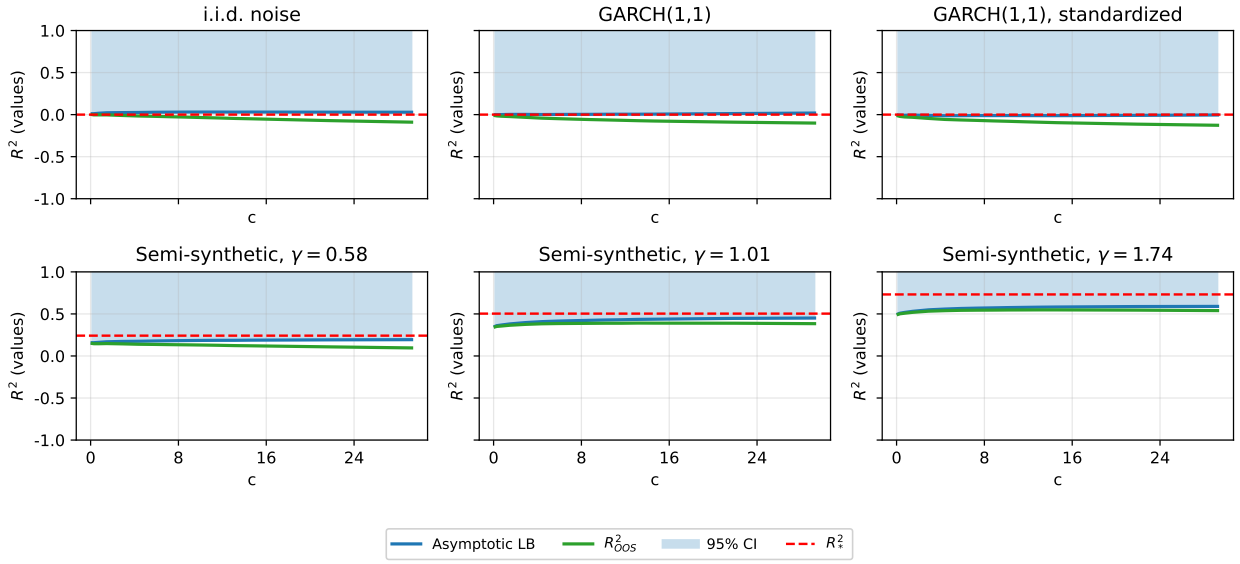


FIGURE 8 – Semi-synthetic simulation (54), with activation=tanh and $z_{ref} = 1$ in (52). In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. i.i.d. noise has $y_{t+1} = \varepsilon_{t+1} \sim \mathcal{N}(0, 1)$. GARCH(1,1) has $y_{t+1} = \varepsilon_{t+1}$ being a GARCH(1,1) noise defined in Appendix G. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51), increasing from $P_1 = 100$ to $P_1 = 20000$. R^2_* is computed in (55). Values of $R^2_{OOS} < -1$ are not shown. γ values are selected to achieve R^2_* of 0, 0.25, 0.5, 0.75, respectively.

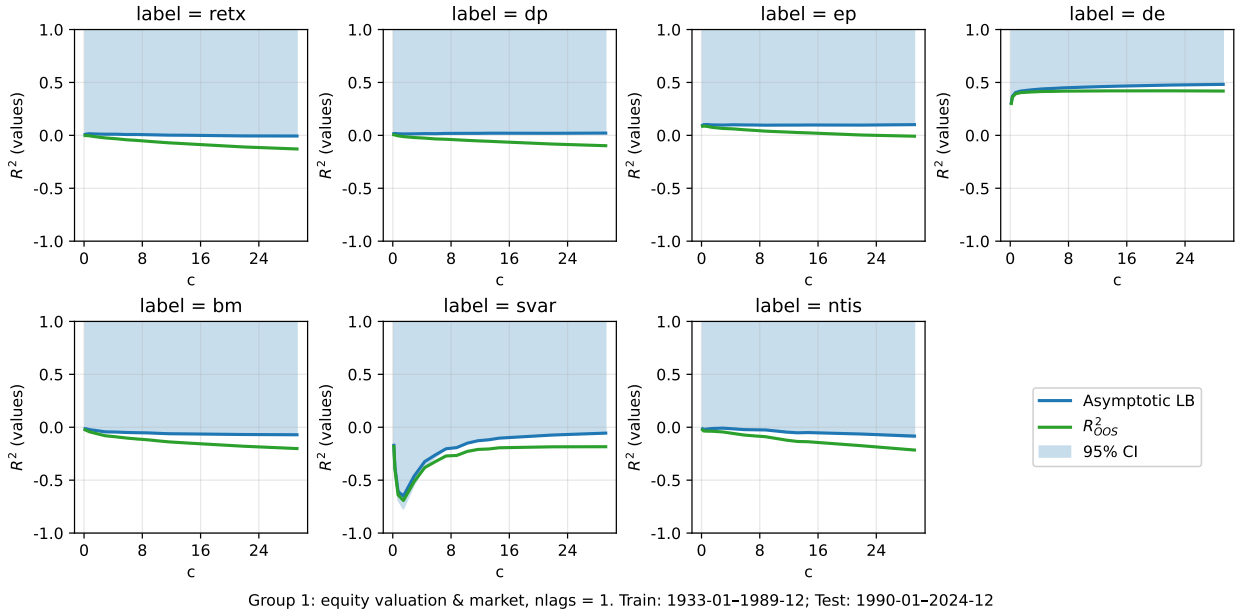


FIGURE 9 – Predicting Welch and Goyal (2008) variables from **Group one** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation**=tanh and $z_{ref} = 1$, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.

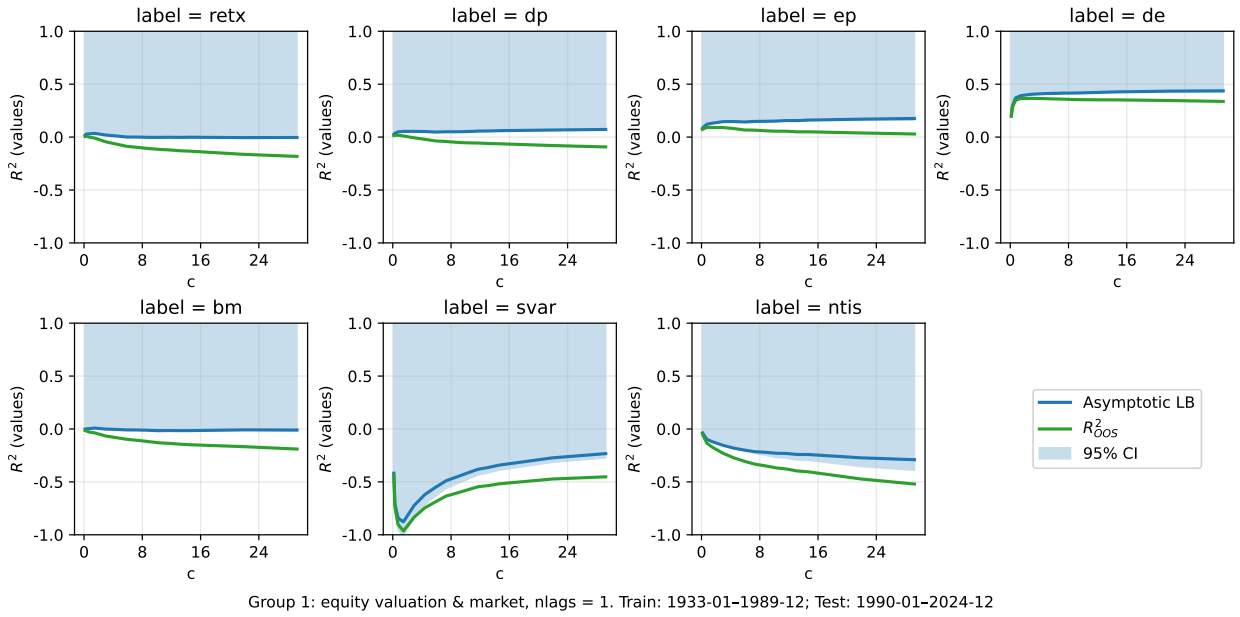


FIGURE 10 – Predicting Welch and Goyal (2008) variables from **Group one** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation**=ReLu and $z_{ref} = 1$, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.

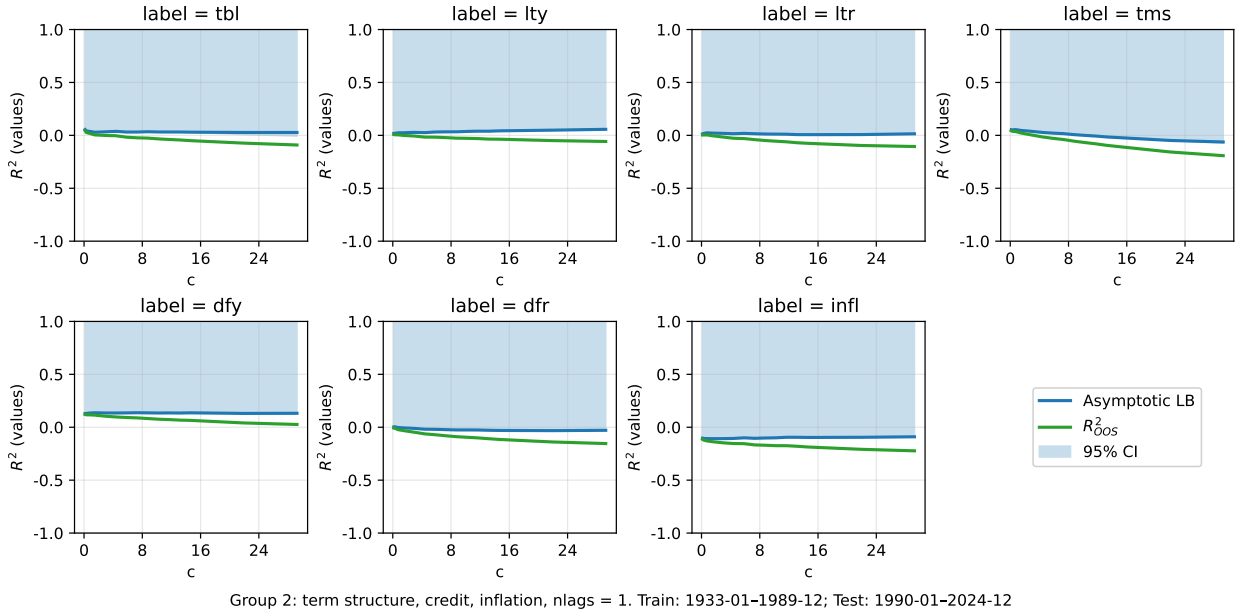


FIGURE 11 – Predicting Welch and Goyal (2008) variables from **Group two** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation=**tanh and $z_{ref} = 1$, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.

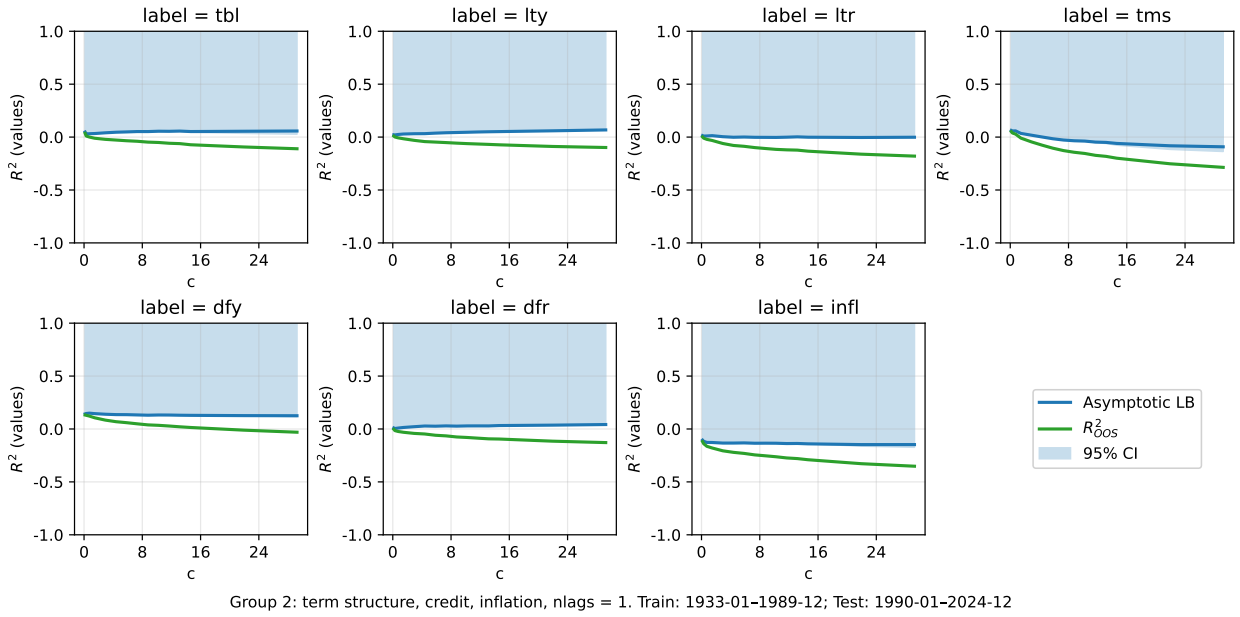


FIGURE 12 – Predicting Welch and Goyal (2008) variables from **Group two** processed according to Procedure 1. Signals are the Welch and Goyal (2008) 14 variables and excess returns. In-sample period is 1933-01 to 1989-12; OOS period is 1990-01 to 2024-12. Asymptotic Lower Bound is given by (25). The shaded region is the one-sided confidence band for R^2_*. The lower bound of the shaded region is (53). Horizontal axis is statistical complexity $c = P_1/T$, where P_1 is the number of random features (51) with **activation**=ReLu and $z_{ref} = 1$, increasing from $P_1 = 100$ to $P_1 = 20000$. Values of $R^2_{OOS} < -1$ are not shown.