Limits To (Machine) Learning

Zhimin Chen, Bryan Kelly, and Semyon Malamud

May 17, 2026

The Billion Dollar Question

- We have features X and labels y.
- Before committing dollars and years: *how do we know if X predicts y?*
  - Should we license this $2M dataset?
  - Should a team of PhDs spend a year hunting for alpha here?
  - Academics: Is there a link between X and y to study (theoretically/empirically)

How do We Attack This Question? ML!

- Run a few regressions, XGBoost, Deep Learning, ...
  - If it works, put more effort and build a better model
  - If it does not, abandon the project
- But what if we are wrong?

Why are We Likely to be Wrong? Complexity!

- Statistical Complexity: too many Xs and too small T (number of observations)
- Long periods where "nothing works" do *not* imply that no signal exists.
  - limited sample size T
  - limited model complexity
  - Poor model design (bad alignment)

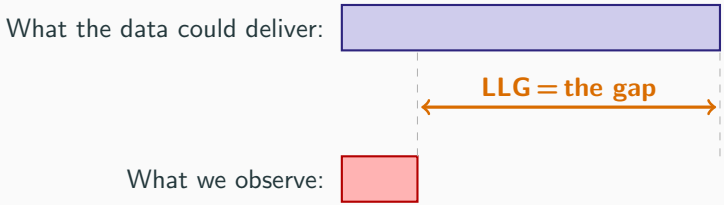
A Little History — Long Plateaus, Sudden Breakthroughs

Task	Before	After breakthrough
Vision (2012 break)	26.2% top-5 error	15.3% top-5 error
Translation (2016 break)	Phrase-based plateau	$\sim$ 60% fewer errors
Speech (2016–17 break)	$\sim$ 70% Error Rate	5.1% Error Rate

Progress can look flat for decades, then jump once the right model class, data scale, and compute threshold align.

What We Do in the Paper

- Complexity imposes a **Limits-to-Learning Gap (LLG)** between what we observe and what the data can potentially achieve:



- Our paper **quantifies** this gap
  - The New Paradigm: adjust to the gap to know the true potential !

Limits to Learning

The Predictability Equation You Already Know

$$y_{t+1} = \underbrace{f_t}_{\text{true predictable part}} + \underbrace{\varepsilon_{t+1}}_{\text{nobody can predict}} \tag{1}$$

- y: returns, dividends, spreads — whatever you predict
- f_t: the true conditional mean (what we wish we knew)
- ε_{t+1}: irreducible noise, $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$

Linear Estimators

- A linear estimator is of the form

$$\hat{f}_T \;=\; \mathcal{K}(S_T, S)\, y, \tag{2}$$

  where $\mathcal{K}(S_T, S) \in \mathbb{R}^{1\times T}$ is a vector function, with $(\mathcal{K}(S_T, S))_t =$ attention to time t

- **Example 1:** ridge

$$\hat{f}_T \;=\; \underbrace{\left(S_T'(zI \;+\; S'S/T)^{-1}\frac{1}{T}S' \right)}_{\mathcal{K}(S_T,S)} y\,, \tag{3}$$

- **Example 2:** kernel ridge. Given a positive definite kernel $k(\cdot,\cdot)$,

$$\hat{f}_T \;=\; \underbrace{k(S_T, S)(zI + k(S,S))^{-1}}_{\mathcal{K}(S_T,S)} y\,. \tag{4}$$

◂ DNN

$$MSE_{OOS}(\hat{f}) \;=\; \underbrace{E[(y_{t+1} - \hat{f}_t)^2]}_{out-of-sample\ \ error} \;=\; \underbrace{E[(f_t - \hat{f}_t)^2]}_{estimation\ \ error} + \underbrace{\sigma_\varepsilon^2}_{irreducible\ \ noise}, \qquad (5)$$

- We study

$$R^2(\hat{f}) \;=\; 1 - \frac{MSE_{OOS}(\hat{f})}{E[f_t^2] + \sigma_\varepsilon^2}, \qquad (6)$$

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

$$R^2_* \;=\; 1 \;-\; \frac{\sigma_\varepsilon^2}{E[f_t^2] \;+\; \sigma_\varepsilon^2} \qquad (7)$$

LLG

Let $\widehat{\mathcal{K}} = (\mathcal{K}(S_\tau, S))_{\tau=T}^{T+T_{OOS}-1} \in \mathbb{R}^{T_{OOS} \times T}$.

$$\widehat{\mathcal{L}} = \frac{1}{T_{OOS}} \operatorname{tr}\left(\widehat{\mathcal{K}}'\widehat{\mathcal{K}}\right) \tag{8}$$

is the **Limits-to-Learning Gap (LLG).**

- $\widehat{\mathcal{L}}$: depends only on signals S, *not on* y $\rightarrow$ compute once, applies to anything.
- Intuition:
  - LLG = **capacity to overfit noise**
  - asymptotically, a Herfindahl index of the eigenvalues

Assumptions

A Lower Bound

Theorem

The OOS MSE and R^2 satisfy

$$\liminf \frac{MSE_{OOS}(\hat{f})}{1+\widehat{\mathcal{L}}} \;\geq\; \sigma_\varepsilon^2, \qquad R_*^2 \;\geq\; \limsup \frac{R^2_{OOS}(\hat{f})+\widehat{\mathcal{L}}}{1+\widehat{\mathcal{L}}} \tag{9}$$

Numerical example:

- If $R^2_{OOS} = -5\% \rightarrow$ your model said **no** signal
- If $\widehat{\mathcal{L}} = 0.25 \rightarrow R_*^2 \geq \dfrac{-0.05+0.25}{1+0.25} = 16\% \rightarrow$ the data says at least **16%** predictability

Confidence Band for R^2_*

Theorem (Pivotal CLT)

$$R^2_* \geq \underbrace{\frac{R^2_{OOS} + \widehat{\mathcal{L}}}{1 + \widehat{\mathcal{L}}}}_{\textit{point estimate}} - 1.65 \cdot \frac{\hat{\sigma}_{R^2}}{\sqrt{T}} \quad \textit{at 95\%} \tag{10}$$

- $\hat{\sigma}_{R^2}$ computed from (y, S) alone
- Pivotal under standard moment conditions

Robustness

Hypothesis Testing: Do I Have The Final Model?

Corollary

Hypothesis $\hat{f} = f_t$ is rejected if

$$\underbrace{\frac{\widehat{\mathcal{L}}(1 - R^2_{OOS})}{1 + \widehat{\mathcal{L}}}}_{\text{point estimate of } \mathrm{R}^2 \text{ gap}} - 1.65 \cdot \frac{\tilde{\sigma}_{R^2}}{\sqrt{T}} > 0\,.$$

Empirics

Setup & Reading the Figures

Data: Goyal–Welch monthly predictors; train: 1933–1989; OOS: 1990–2024 process

- **G1** : `dp, ep, de, bm, ntis, svar`, excess returns
- **G2** : `tbl, lty, ltr, tms, dfy, dfr, infl`

Two exercises

(1) Semi-synthetic: known R^2_*

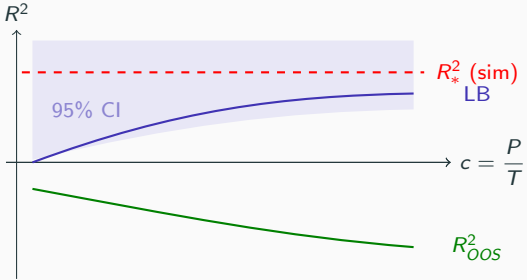
$$y_{t+1} = \gamma\, g(X_t' W) + \varepsilon_{t+1}$$

(2) Real: predict each variable from all 14

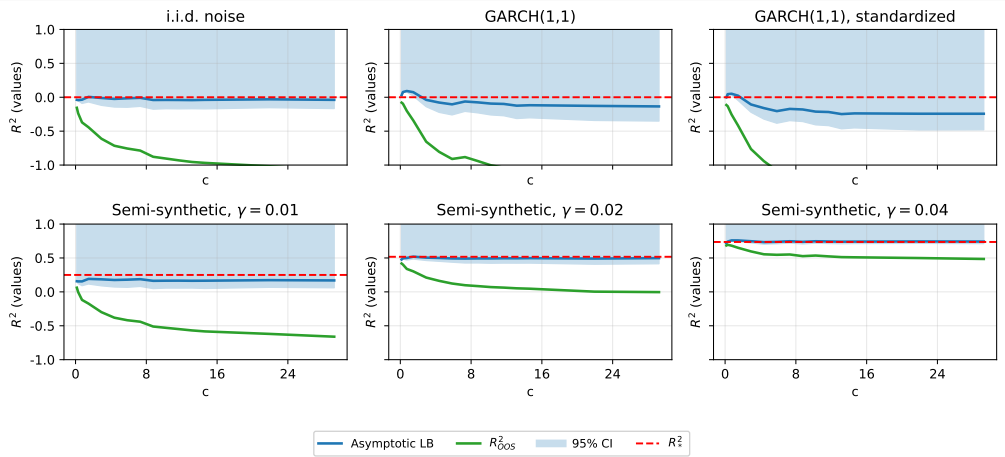
Ridge on $P = 100, \ldots, 20{,}000$ random features

x-axis: complexity $c = \frac{P}{T}$.

What each figure shows



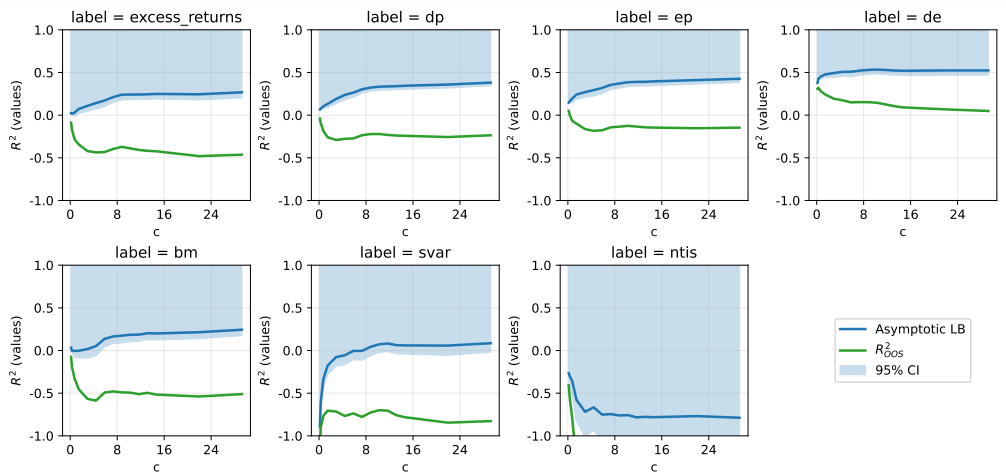
Semi-synthetic Simulation (ReLu)



Noise and semi-synthetic benchmarks: activation=relu, scale=1.0, seed=41, nlags=1. Train: 1933-01–1989-12; Test: 1990-01–2024-12

▸ Compare with tanh

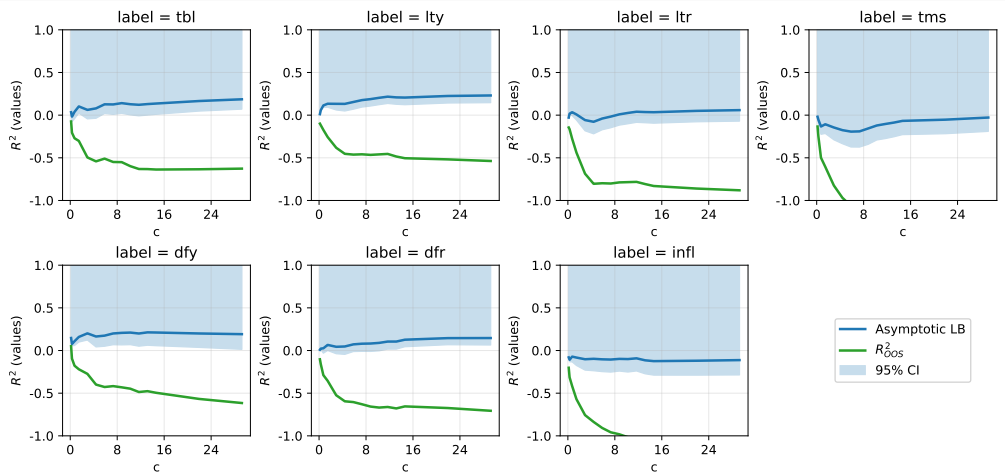
Goyal-Welch Group 1 — Equity Valuation & Market (ReLu)



Group 1: equity valuation & market, nlags = 1. Train: 1933-01–1989-12; Test: 1990-01–2024-12

▸ Compare with tanh

Goyal-Welch Group 2 — Term Structure, Credit, Inflation (ReLu)



Group 2: term structure, credit, inflation, nlags = 1. Train: 1933-01–1989-12; Test: 1990-01–2024-12

▸ Compare with tanh

Lower Bounds vs Best Out-of-Sample R-squared

Label	tanh 95%	ReLU 95%	Best Ridge	Best Rec. Ridge
Excess Returns	6%	20%	2%	3%
dp	24%	35%	3%	2%
ep	24%	38%	10%	16%
de	69%	47%	42%	42%
bm	16%	17%	0%	0%
svar	16%	0%	0%	0%
ntis	0%	0%	1%	1%
tbl	13%	7%	4%	7%
lty	24%	15%	1%	1%
ltr	9%	2%	1%	1%
tms	0%	0%	7%	7%
dfy	30%	15%	14%	16%
dfr	0%	7%	0%	2%
infl	0%	0%	0%	0%

Asset Pricing Implications

Asset Pricing Reminder

- SDF M_{t+1} prices any asset excess return R_{t+1} **conditionally**:

$$E_t[M_{t+1}R_{t+1}] = 0$$

- In equilibrium, the Intertemporal Marginal Rate of Substitution (IMRS) of any agent is an SDF

LLG Sharpens the Hansen-Jagannathan Bound ii

Theorem (LLG-corrected HJ bound)

Suppose

$$R_{t+1} \;=\; f_t + \varepsilon_{t+1}$$

comes from an equilibrium model with agents knowing $f_t = E_t[R_{t+1}]$. *The IMRS-based SDF* M_{t+1} *satisfies*

$$E\left[\frac{\mathrm{Var}_t[M_{t+1}]}{E_t[M_{t+1}]^2}\right] \;\geq\; \limsup\; \frac{R^2_{OOS}(\hat{f}) + \widehat{\mathcal{L}}}{1 - R^2_{OOS}(\hat{f})}. \tag{11}$$

- Classical HJ (1991) bound for Sharpe ratio
- Our refinement: incorporate the LLG $\widehat{\mathcal{L}}$

LLG Sharpens the Hansen-Jagannathan Bound iii

- Implication: SDF volatility lower bound is *much higher* than previously believed when the LLG is large
- This holds even for the **projection of the SDF** on the asset return

Excess Volatility from Rational Learning

Setup: Risk-neutral agents, Gaussian prior on β, posterior price $Q_T = \hat{\beta}_T' S_T$.

Theorem (LLG generates excess volatility)

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

- Excess volatility from rational over-reaction to noise in high-dimensional learning
- *No behavioral departures needed*
- Resolves Shiller (1981) puzzle without invoking irrationality

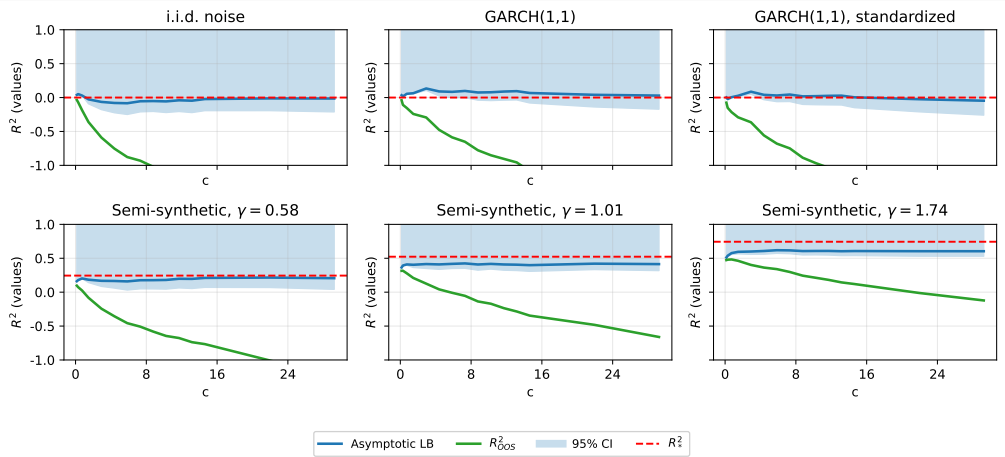
Conclusion

The New Paradigm

- Train a model and evaluate it Out-Of-Sample (OOS): Get the MSE_{OOS}
- Adjust to the Limits to Learning Gap (LLG): Get the $AdjustedMSE_{OOS}$
- If it is high, look for a better model

Thank You

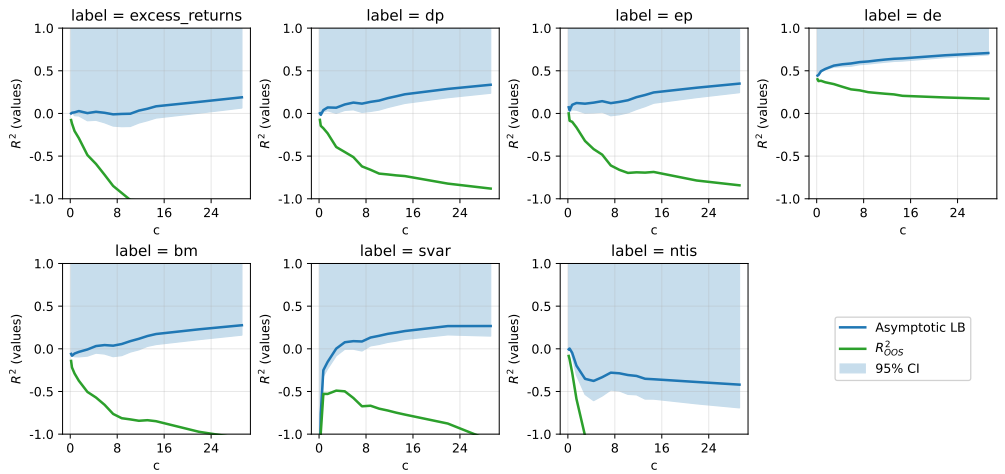
Semi-synthetic Simulation (tanh)



Noise and semi-synthetic benchmarks: activation=tanh, scale=1.0, seed=41, nlags=1. Train: 1933-01–1989-12; Test: 1990-01–2024-12

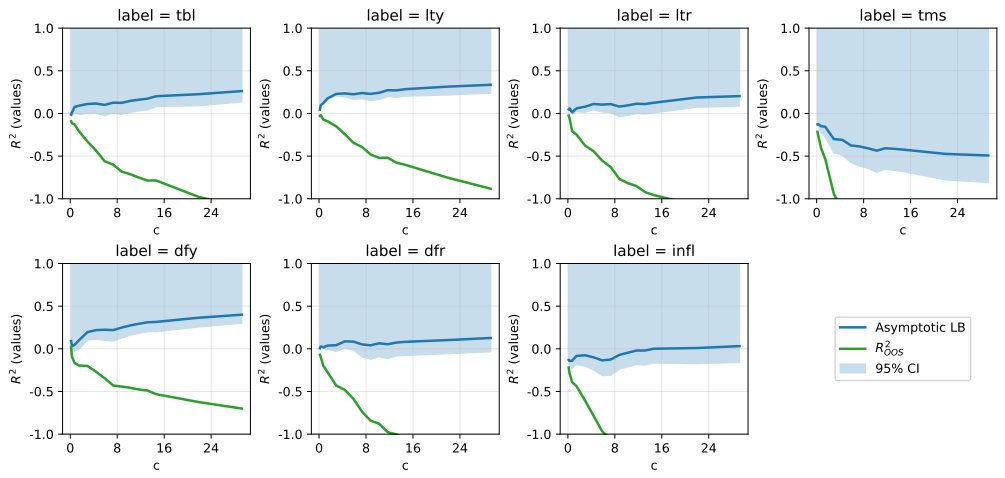
Back to ReLu

Goyal-Welch Group 1 (tanh)



Back to ReLu

Backup: Goyal-Welch Group 2 (tanh)



Group 2: term structure, credit, inflation, nlags = 1. Train: 1933-01–1989-12; Test: 1990-01–2024-12

Back to ReLu

NTK $\rightarrow$ DNN Extension

- Deep Neural Network (DNN) $f(x;\theta),\ \theta \in \mathbb{R}^P$.
- Neural Tangent Kernel (NTK): $K(x,\tilde{x};\theta) = \nabla_\theta f(x;\theta)' \nabla_\theta f(\tilde{x};\theta)$.
- Wide DNN trained by gradient descent $\approx$ kernel ridge with trained NTK:

$$f(x;\theta_{\mathcal{E}}) \approx T^{-1}K(x;X;\theta_{\mathcal{E}})(zI + T^{-1}K(X;X;\theta_{\mathcal{E}}))^{-1}y$$

- Equivalently, linear regression with signals $S_t = \nabla_\theta f(X_t;\theta_{\mathcal{E}}) \in \mathbb{R}^P$.
- $\Rightarrow$ LLG formula applies to wide neural nets.

Back

Why AR(1) Residuals?

- Theory needs uncorrelated, ∼homoskedastic residuals.
- For **excess returns**: AR(1) coefficient ≈ 0, so residual $\approx$ raw return
- For **persistent variables** (`dp`, `bm`, `lty`, `tbl`): residualization isolates cross-variable predictability beyond own-lag.

Back to setup

Technical Conditions

Assumption 1 (DGP):

- $y_{t+1} = f_t + \varepsilon_{t+1}$, with ε_{t+1} i.i.d., zero mean, finite fourth moment, independent of $(f_t, S_t)_{t \geq 0}$
- **No restriction on** (f_t, S_t): non-stationarity, autocorrelation, heteroskedasticity, and misspecification all allowed on the signal side

For Theorem 1 (the bound): three trace conditions on $\widehat{\mathcal{K}}'\widehat{\mathcal{K}}$ ensuring the LLN-type argument applies to the quadratic form in ε. Hold e.g. when $E[\|\widehat{\Psi}^2_{OOS}\|] = o(\min\{T_{OOS}, T\})$.

For Theorem 2 (CLT): $E[\varepsilon^3] = 0$ and $E[\varepsilon^4] = 3$ for the displayed standard error formula. Generalizes to arbitrary fourth moment with a modified $\hat{\sigma}_{R^2}$ (paper).

◂ Back to LLG

Robustness

Heavy tails / non-Gaussian residuals:

- GARCH(1,1) simulations (with and without standardization) shown earlier — bound remains valid
- *Procedure 1* (rolling standardize + clip at ± 3) handles empirical fat tails in ε

One-sided bound (bias dropped):

- Price of being fully data-driven, no specification of f_t
- Tight when bias is small; remains a *valid* lower bound when it isn't

◂ Back to CLT

Correlation: Lower Bounds vs Benchmarks

	tanh (95%)	ReLU (95%)	Ridge	Rec. Ridge
tanh (95%)	1.00			
ReLU (95%)	0.79	1.00		
Ridge	0.87	0.67	1.00	
Rec. Ridge	0.85	0.69	0.99	1.00