Self-Fulfilling Accuracy

Dmitry Livdan* Bernardo Maciel† Norman Schürhoff‡

July 2026§

Abstract

A model that allocates resources changes the outcomes on which it is evaluated and retrained, an increasingly important concern for deployed AI systems. We characterize repeated model retraining under feedback. The stable predictor is calibrated against deployment outcomes and is a strictly increasing transformation of expected fundamentals, preserving information while raising measured fit under monotone amplification. This self-fulfilling accuracy is not identified from deterministic deployment data, while randomized holdouts recover its informational and causal components. In a static linear-Gaussian economy, archived model vintages follow an autoregression that identifies the feedback gain and yields an exact stability frontier. Below it, retraining is ergodic, and coefficient dispersion diverges near the boundary, while above it coefficients diverge almost surely. Gaussian features imply that feedback rescales stable linear coefficients without rotation. In a dynamic economy, durable allocations and persistent fundamentals compound the loop gain, and sufficiently large panels are stable below the population frontier.

*Haas School of Business, University of California, Berkeley and CEPR. Email livdan@haas.berkeley.edu.
†Haas School of Business, University of California, Berkeley. Email bernardo_maciel@berkeley.edu.
‡University of Lausanne and Swiss Finance Institute and CEPR. Email norman.schuerhoff@unil.ch.
§New draft, comments welcome.

A credit scoring model assigns a high score to a borrower. The score improves the borrower's financing terms, the resulting investment raises repayment, and the model is subsequently retrained on an outcome that its own deployment helped produce. A recommender system raises a product's visibility and thereby contributes to the sales used to evaluate and update the ranking. Similar feedback arises when policy rules change macroeconomic outcomes or clinical scores trigger treatments that alter patient health. Artificial intelligence makes these feedback loops increasingly consequential because predictions can be deployed at scale, updated frequently, and connected directly to lending, hiring, pricing, ranking, and other allocation decisions. In each case, a prediction determines an allocation, the allocation changes the outcome, and the changed outcome enters the next round of estimation.

This feedback complicates the interpretation of predictive performance. Accuracy, calibration, and fairness are commonly assessed using outcomes observed after deployment, when the model is already influencing the relevant allocation. The issue is especially salient for AI systems whose predictions are continuously incorporated into operational decisions and whose performance is monitored using the data generated by those decisions. Measured fit therefore combines information about underlying fundamentals with the causal effect of the decision attached to the prediction. A model may fit deployment data increasingly well because it has become more consequential, even when the information contained in its score has not improved. We call this divergence between fit and information self-fulfilling accuracy.

We study a model with a latent fundamental, observed features, and a score that determines an allocation. The allocation affects the realized outcome, and the model is retrained by projecting that outcome on the same features. This structure applies to traditional statistical scores as well as AI systems whose predictions are converted into economically consequential decisions. It separates the component of predictive performance attributable to information about fundamentals from the component produced by deployment.

Our first result characterizes performatively stable predictors. When the slope of the feedback response is uniformly below one, the unique stable predictor is a strictly increasing transformation of the posterior mean of the fundamental. Feedback changes the scale on which information is expressed but preserves the ordering of units, the conditional distribution of fundamentals, mutual information, and rank dependence. A credit model may spread borrowers farther apart because favorable scores improve financing and repayment, yet it does not learn anything about repayment capacity that was absent from the original features.

The stable predictor is also calibrated against the outcomes generated by its deployment. Under a monotone amplification condition, stronger feedback increases its measured fit while leaving the conditional residual distribution unchanged. The condition holds for affine feedback with nonnegative slope and for nonnegative convex responses. The allocation expands the predictable component of the outcome without reducing uncertainty about the fundamental. Calibration remains exact, measured R^2 rises, and fundamental information remains fixed.

This result changes the interpretation of model comparisons. Two predictors can rank every unit identically and contain the same information, while the predictor attached to the stronger allocation displays higher deployment accuracy. Comparisons across institutions, policy regimes, AI systems, or time periods may therefore confound changes in predictive technology with changes in the economic influence of the score.

Our second result concerns identification. Under a single deterministic deployment, the conditional mean of the outcome admits observationally equivalent decompositions into a fundamental component and a feedback component. Any suitable function of the deployed score can be shifted from the latent fundamental into the feedback response without changing the joint distribution of observed features and outcomes. Deployment data identify the reduced form but do not reveal how much of it reflects information and how much reflects the effect of the allocation.

The source of the problem is an absence of overlap. Conditional on the features, the allocation is fixed by the score, so the data contain no otherwise comparable units exposed to different allocations. This is the performative counterpart of the identification failure that arises in treatment-effect models when treatment probabilities are zero or one at each covariate value (Rosenbaum and Rubin, 1983; Hahn, 1998; Crump et al., 2009; D'Amour et al., 2021).

Randomized holdouts restore the missing variation. All units are scored, while a randomly selected group receives a fixed baseline allocation. Outcomes in the holdout group identify the posterior mean up to an additive normalization, and differences between deployed and holdout outcomes identify the feedback response over the observed score range. The design randomizes the allocation while preserving the prediction for every unit. Known threshold rules provide a more local source of identification. When the latent score-conditional mean is continuous at the cutoff, a discrete change in the allocation identifies the corresponding treatment effect.

We next study retraining within a linear class. With Gaussian features, Stein's lemma implies that the covariance generated by feedback is aligned with the deployed coefficient vector. Every stable linear coefficient is therefore collinear with the coefficient obtained without feedback. Feedback changes the magnitude of the coefficient but cannot rotate it. The result extends to elliptical feature distributions through the linearity of conditional expectations.

The no-rotation result provides an empirical restriction. After coefficients have been placed on a common normalization and changes in the feature distribution have been accounted for, changes in feedback strength should move stable coefficients along a fixed ray. Persistent rotation points toward feature-specific treatment effects, feedback into future features, changes in the information structure, or departures from ellipticity. The restriction is particularly useful when a complex AI system feeds into a stable low-dimensional scoring or allocation layer whose coefficients are retained for auditing.

Our third set of results studies finite-sample retraining. In the static linear-Gaussian economy, retraining on fresh cross-sections produces an exact first-order autoregression in the deployed slope. The autoregressive coefficient equals the feedback gain, while the innovations are independent scaled Student t variables generated by estimation error. Archived model vintages therefore identify the

economic loop. Below the stability frontier, the stationary autocorrelation identifies the feedback gain, the stationary mean identifies feature reliability, and the stationary variance supplies an overidentifying restriction.

Sampling error becomes useful because each estimation sample produces a slightly different coefficient and hence a slightly different deployment. Under Gaussianity, these perturbations are exogenous to the previously deployed coefficient. A model archive consequently contains a sequence of small experiments generated by ordinary retraining noise. This implication is relevant for automated AI pipelines, which routinely retain model versions, validation records, and deployment histories that can be used to measure the feedback generated by repeated updating.

The autoregressive representation also yields an exact stochastic phase transition. Below the frontier, fresh-sample retraining is geometrically ergodic around the performatively stable coefficient. At the frontier, the coefficient follows a random walk with positive drift. Above it, the coefficient diverges almost surely at a geometric rate. The stationary distribution below the frontier inherits the regularly varying tails of the finite-sample innovations, while its dispersion rises without bound as the feedback gain approaches one.

The archive also provides an early-warning statistic. Near the frontier, coefficient levels become more persistent and more dispersed across vintages, even though adjacent revisions become smaller. A model sequence can therefore appear locally calm while moving over an increasingly wide range at longer horizons. Under a stable sampling environment, the joint behavior of persistence, level dispersion, and revision variance reveals strengthening feedback.

Pooling all historical observations changes the speed but not the location of the frontier. Below the frontier, pooled least squares converges almost surely to the stable coefficient. Above it, the coefficient diverges at a polynomial rather than geometric rate. Historical averaging slows the instability because the effective learning gain declines, but it cannot eliminate feedback embedded in the data already collected.

The dynamic economy introduces persistent fundamentals and durable allocations. Past scores affect current outcomes through accumulated capital, so the state must include both the deployed coefficient and the cross-sectional distribution of capital. Durability and persistence compound the contemporaneous loop gain into an effective gain. The population system is stable exactly when this effective gain is below one, and even weak contemporaneous feedback can become destabilizing when allocations are sufficiently durable and fundamentals sufficiently persistent. This mechanism is relevant for AI applications in which recommendations, credit decisions, treatments, or investments have effects that persist long after the model is updated.

Finite-panel dynamic retraining is governed by the top Lyapunov exponent of an exact random coefficient recurrence. A negative exponent yields a unique causal stationary and ergodic solution, and paths driven by the same primitive shocks converge to one another. A positive exponent makes homogeneous coefficient products expand exponentially. We also establish a stable-side bridge between the population and finite-panel systems. Every population economy strictly below the

frontier has a negative finite-panel Lyapunov exponent for all sufficiently large panels. The exact large-panel exponent and the positive side remain conjectural. Simulations show the estimated finite-panel zero crossing approaching the population frontier and associate positive exponents with divergence of the affine state and large clustered coefficient excursions.

Related literature. The paper contributes to the literature on performative prediction initiated by Perdomo et al. (2020), which studies settings in which deployment changes the distribution used for evaluation or retraining. Brown et al. (2022) introduce persistent states, Hardt and Mendler-Dünner (2023) survey the theory, Perdomo (2025) connects performative stability to multicalibration, and Kim and Perdomo (2023) studies decisions under outcome performativity. This literature has become increasingly relevant as AI systems move from passive prediction to repeated deployment in economic environments. Our framework introduces a latent fundamental and an allocation that affects realized outcomes. This structure delivers an explicit representation of the stable predictor, separates information from fit, and links population stability to the stochastic behavior of archived model vintages.

The identification results connect performative prediction to the econometrics of treatment effects. Deterministic deployment violates overlap by assigning a unique allocation at each covariate value. Randomized holdouts restore the missing counterfactual variation, while threshold rules identify discrete effects under continuity and functional-form restrictions. The resulting observational equivalence is closely related to the structural identification failure studied by Mendler-Dünner et al. (2022).

The analysis also applies the logic of the Lucas critique to repeatedly estimated predictive systems. Lucas (1976) emphasizes that policy changes the behavior and distribution underlying an econometric relationship. Here the feedback operates through the outcome attached to an individual score, which permits a complete characterization of the stable predictor and of the data generated along the retraining path. The same logic applies to AI systems whose predictions become inputs into the economic behavior that later appears in their training data.

Our results complement work on AI, data, and economic decision making. Acemoglu and Restrepo (2018) study the effects of AI on tasks and labor demand, Farboodi and Veldkamp (2023) analyze data as an endogenous input, and Chen et al. (2024) distinguish model fit from the information required for structural validity. We provide a theory of deployed AI and predictive systems in which a model's economic influence changes its measured success without changing its information about fundamentals.

The paper is also related to strategic scoring (Frankel and Kartik, 2022; Hennessy and Goodhart, 2023; Ball, 2025). Strategic responses may reduce informativeness by distorting the data supplied to the model. Our mechanism operates through productive responses to the allocation. Stronger feedback improves measured fit while preserving the information contained in the score.

Finally, the stochastic analysis connects to adaptive learning and random recurrence models.

Pooled retraining builds on the decreasing-gain methods of Marcet and Sargent (1989), Sargent (1999), and Evans and Honkapohja (2001). Fresh-sample retraining produces an exact constant-gain autoregression, while the dynamic economy generates a random coefficient system governed by a Lyapunov exponent. These results characterize when repeatedly deployed predictive and AI systems converge, fluctuate around a stable predictor, or become unstable.

1 Population Theory of Stable Predictors

This section characterizes the population theory of stable predictors, which establishes the fixed point of the retraining operator and its stability.

1.1 Setup

A unit may represent a borrower whose repayment capacity is imperfectly observed, a firm whose future productivity is inferred from accounting and market information, a product whose underlying demand is estimated from consumer characteristics, or a patient whose health risk is summarized by clinical data. The unit is described by a latent fundamental $\theta \in \mathbb{R}$, an observed feature x taking values in a Polish space $\mathcal{X}$, and an outcome shock $\varepsilon \in \mathbb{R}$. Their joint law μ satisfies $\mathbb{E}[\theta^2] < \infty$, $\mathbb{E}[\varepsilon^2] < \infty$, and $\mathbb{E}[\varepsilon \mid \theta, x] = 0$. The final condition allows realized outcomes to contain transitory variation while requiring that this variation have conditional mean zero once the fundamental and the observed feature are given.

We write

$$u(x) := \mathbb{E}[\theta \mid x] \tag{1}$$

for the posterior mean of the fundamental. This object summarizes the information about θ contained in the observed feature and provides the benchmark prediction in the absence of deployment feedback.

The modeler deploys a score $s = m(x)$, where $m\colon \mathcal{X} \to \mathbb{R}$ is measurable. The score determines an allocation, the allocation induces an action, and the action changes the realized outcome. A higher credit score may expand a borrower's credit line, which finances investment and improves repayment. A higher product ranking may increase visibility, which raises demand and subsequent sales. A clinical risk score may trigger treatment, which changes the patient's health outcome. In each case, the model affects the outcome against which its predictive performance is subsequently evaluated.

We represent this sequence through a feedback response $f\colon \mathbb{R} \to \mathbb{R}$ and study outcomes of the form

$$y = \theta + f(m(x)) + \varepsilon. \tag{2}$$

The function f is a reduced-form representation of the economic mechanism through which the score affects the outcome. Suppose, for example, that the score induces an action according to $a = g(s)$

and that the action affects the outcome through a production response $b(a)$. The feedback response is then given by

$$f(s) = b(g(s)) \,. \tag{3}$$

In a credit market, g may describe how the score changes the availability or cost of financing, while b describes how financing affects investment and repayment. On a platform, g may map a ranking into product visibility, while b maps visibility into demand. In a clinical application, g may map a risk score into treatment intensity, while b maps treatment into health outcomes.

A useful linear specification takes $g(s) = a_0 + \gamma s$ and $b(a) = \beta a$. The resulting feedback response is

$$f(s) = \beta\gamma s + \beta a_0. \tag{4}$$

The no-feedback benchmark corresponds to a constant function f, since changes in the deployed score then have no effect on the realized outcome.

Assumption 1. *The feedback response f is differentiable and satisfies $f'(t) \le \kappa^+ < 1$ for every $t \in \mathbb{R}$. Define the local loop gain by $k(t) \coloneqq f'(t)$ and the uniform absolute gain by $\overline{\kappa}_{\mathrm{abs}} \coloneqq \sup_{t\in\mathbb{R}} |f'(t)|$.*

The local loop gain measures the amplification generated by deployment at a given score level. Under the decomposition in equation (3), the chain rule gives $k(t) = b'(g(t))\, g'(t)$ whenever the relevant derivatives exist. The term $g'(t)$ measures how strongly the score changes the induced action, while $b'(g(t))$ measures how strongly the action changes the outcome. Their product measures how much of a marginal change in the score returns through the outcome used in the next round of estimation.

A loop gain close to one means that a marginal increase in the score produces an almost equally large increase in the outcome on which the model is subsequently retrained. In a lending application, this occurs when the credit allocation reacts strongly to the score, and investment responds strongly to the resulting financing. In a recommendation system, it occurs when rankings have a large effect on visibility and visibility has a large effect on demand. The one-sided bound in Assumption 1 keeps this amplification below the threshold at which the mapping from fundamentals into stable scores loses invertibility.

The stronger condition $\overline{\kappa}_{\mathrm{abs}} < 1$ will be imposed when a contraction argument is required. This condition limits feedback in both directions and ensures that repeated retraining converges globally to the stable predictor. Differentiability simplifies the exposition. The arguments extend to Lipschitz feedback responses whose one-sided slopes are uniformly bounded above by $\kappa^+ < 1$. Under that formulation, statements involving derivatives can be interpreted using almost-everywhere derivatives.

The modeler retrains nonparametrically using squared loss. Given a deployed score m, let $y^{(m)}$ denote the outcome generated under that deployment. The next predictor is the population regression function

$$(Tm)(x) \coloneqq \mathbb{E}\big[y^{(m)} \mid x\big] = u(x) + f(m(x)) \,. \tag{5}$$

The operator T combines two sources of predictive content. The first is the posterior mean of the latent fundamental, which reflects the information contained in the observed feature. The second is the expected effect of the allocation generated by the deployed score. A regression estimated on deployment data treats both components as predictive because both enter the conditional mean of the realized outcome.

This decomposition clarifies the central economic issue. The term $u(x)$ measures what the model knows about the underlying state. The term $f(m(x))$ measures what the deployed model causes through the allocation attached to its prediction. Retraining on realized outcomes does not distinguish between these two sources of fit. It incorporates both into the next prediction.

Definition 1. *A score m^* is performatively stable if $T\,m^* = m^*$ μ-almost surely.*

A performatively stable score reproduces itself under population retraining. Its deployment changes the outcome through the feedback response, and the regression of that outcome on the observed feature returns the same score. This is the fixed-point concept introduced by Perdomo et al. (2020), specialized to squared loss and an unrestricted class of measurable prediction functions. Section 3.1 studies the corresponding problem when retraining is restricted to linear scores.

Conditional expectations and population regression functions are defined only up to sets of measure zero. Stability and uniqueness are therefore understood up to μ-almost-sure equality. For the pointwise arguments below, we fix an everywhere-defined measurable version of the posterior mean in equation (1) and define the operator in equation (5) using that version. Statements concerning stability, uniqueness, and information remain almost-sure statements, while pointwise identities and convergence bounds are understood relative to the selected versions and hold for every $x \in \mathcal{X}$.

1.2 The structure theorem

The retraining operator in equation (5) acts pointwise in the feature. For a unit with feature x, a stable score must satisfy

$$m(x) = u(x) + f\big(m(x)\big). \tag{6}$$

The posterior mean $u(x)$ captures the information about the latent fundamental that is available before deployment. The second term records the effect of the allocation induced by the score itself. Stability therefore requires the score to equal the sum of what the model knows and what its deployment causes. This equation has a direct economic interpretation. Consider a borrower whose observable characteristics imply an expected repayment capacity $u(x)$. A favorable score lowers the cost of financing and raises investment, which improves realized repayment through the response f. Retraining attributes both the borrower's underlying repayment capacity and the induced improvement to the observed characteristics. A stable score is the value that reproduces itself after this feedback has been incorporated.

The fixed-point equation can be rewritten as

$$u(x) = m(x) - f\big(m(x)\big). \tag{7}$$

This representation separates the informational component from the feedback component. Define a map

$$h(t) \coloneqq t - f(t). \tag{8}$$

The function h measures the portion of a score that remains after subtracting the outcome effect generated by deploying that score. Assumption 1 implies that h is strictly increasing. A higher score always corresponds to a higher inferred fundamental because the feedback response grows less than one-for-one with the score. The next theorem formalizes this observation and characterizes the stable predictor.

Theorem 1 (Structure)**.** *Under Assumption 1, the following statements hold.*

(i) The map $h(t)$ is a strictly increasing bijection. Its inverse $\Phi \coloneqq h^{-1}$ is strictly increasing and differentiable, with

$$\Phi'(u) = \frac{1}{1 - k\big(\Phi(u)\big)} > 0. \tag{9}$$

(ii) There is a unique performatively stable score up to μ-almost-sure equality given by

$$m^*(x) = \Phi\big(u(x)\big). \tag{10}$$

(iii) Suppose in addition that $\overline{\kappa}_{\text{abs}} < 1$. For any initial score m_0, the sequence defined by $m_{n+1} = Tm_n$ converges pointwise to m^* at a geometric rate. In particular,

$$|m_n(x) - m^*(x)| \le \overline{\kappa}_{\text{abs}}^{\,n} \, |m_0(x) - m^*(x)| \ \text{ for every } x \in \mathcal{X}. \tag{11}$$

Theorem 1 provides a complete representation of the stable predictor. The fixed-point equation in (7) becomes

$$h\big(m^*(x)\big) = u(x), \tag{12}$$

and inversion gives equation (10). The stable score is therefore obtained by applying the same scalar transformation Φ to the posterior mean for every unit. The distribution of features may be high dimensional, and the posterior mean may be generated by a complicated prediction technology, yet feedback enters the stable score through a one-dimensional map.

The economic content of the result is easiest to see in a lending example. Two borrowers with posterior repayment capacities $u(x_1)$ and $u(x_2)$ receive stable scores

$$m^*(x_1) = \Phi\big(u(x_1)\big) \quad \text{and} \quad m^*(x_2) = \Phi\big(u(x_2)\big). \tag{13}$$

Since Φ is strictly increasing, the borrower with the higher expected repayment capacity always receives the higher stable score. Feedback changes the numerical distance between their scores because the induced financing affects their subsequent outcomes, but it does not reverse their ranking.

Equation (9) describes the local amplification created by this feedback. A marginal increase in the posterior mean raises the stable score by

$$\Phi'(u) = \frac{1}{1 - k\big(\Phi(u)\big)}. \tag{14}$$

When the local loop gain is positive, the multiplier exceeds one. A unit of information about the fundamental then produces more than a unit change in the stable score because the score generates an allocation that reinforces the predicted outcome. In a credit market, a stronger repayment signal raises the score, the higher score improves financing terms, and the resulting investment further improves repayment. In a recommendation system, higher predicted demand raises visibility, which generates additional demand and magnifies the final score.

The local gain may vary across score levels. Financing conditions may respond strongly to a score near an approval threshold and more weakly elsewhere, while a platform ranking may have a larger effect on demand near the top of a search page. The transformation Φ can therefore be nonlinear even when the information summarized by $u(x)$ is one-dimensional. Its slope remains positive throughout, so the transformation may stretch different parts of the score distribution by different amounts without folding the distribution or reversing the ordering of units.

Part (iii) separates existence and uniqueness from convergence of repeated retraining. The one-sided restriction in Assumption 1 guarantees that the stable score exists and is unique. Global convergence additionally requires the absolute feedback gain to be uniformly below one. Under this stronger condition, the map $t \mapsto u(x) + f(t)$ is a contraction for every fixed x. Each round of retraining reduces the distance to the stable score to at most $\overline{\kappa}_{\text{abs}}$ times its previous value.

The distinction matters when feedback is negative. A clinical intervention may reduce the outcome associated with a high risk score because treatment lowers the patient's subsequent risk. Retraining can then oscillate around the stable score. The absolute-gain condition ensures that these oscillations shrink rather than grow.

The uniform gap from one is also essential for global existence. If $\sup_t f'(t) = 1$, the derivative may remain strictly below one at every finite point while approaching one sufficiently quickly that h fails to cover the real line. For example, let $f(t) = t - \arctan(t)$. Then $h(t) = \arctan(t)$, whose range is bounded. Posterior means outside this range cannot be reconciled with any stable score. The uniform bound in Assumption 1 rules out this possibility by ensuring that $h(t)$ grows at least linearly in both tails.

These conclusions are the nonparametric counterparts of the stability condition in the affine

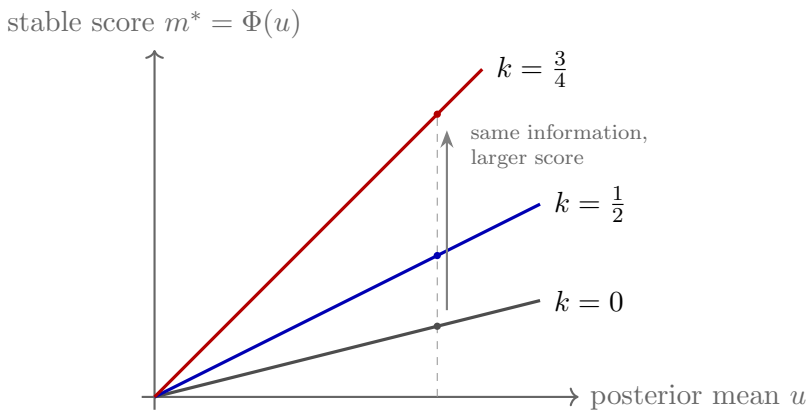


Figure 1: The structure theorem as a picture. Each stable score is the posterior mean u passed through the strictly increasing map Φ, drawn for the affine case in which Φ is a ray of slope $1/(1-k)$. Stronger feedback steepens the map and disperses the score, which manufactures fit, while the ordering of units and the information the score carries coincide with those of the posterior mean at every feedback strength.

model. When

$$f(t) = a + kt, \tag{15}$$

the inverse map is affine and has slope $1/(1-k)$. The condition $k < 1$ guarantees the existence of a stable score, while $|k| < 1$ guarantees convergence of repeated retraining. Related contraction conditions appear in Perdomo et al. (2020) and, in environments with a persistent state, in Brown et al. (2022). Theorem 1 adds an explicit representation of the stable predictor and makes its information content transparent.

Figure 1 illustrates the affine case. Each line maps the same posterior mean into a stable score. A stronger positive loop gain steepens the line, so a given difference in fundamentals produces a larger difference in scores. The ranking of units remains unchanged because all of the lines are increasing.

The monotonicity of Φ has a stronger implication than preservation of rankings. Since the transformation is invertible, observing the stable score reveals exactly the same information as observing the posterior mean. The score may place the information on a different numerical scale, but no information is created or destroyed by the feedback transformation. This observation yields the next result.

Corollary 1 (Information invariance)**.** *Under Assumption 1,*

$$\sigma\big(m^*(x)\big) = \sigma\big(u(x)\big). \tag{16}$$

Consequently, the following statements hold.

(i) The conditional law of θ given $m^*(x)$ equals the conditional law of θ given $u(x)$.

(ii) The mutual information satisfies

$$I\big(m^*(x);\theta\big) = I\big(u(x);\theta\big). \tag{17}$$

(iii) Every rank statistic of $\big(m^*(x),\theta\big)$ equals the corresponding rank statistic of $\big(u(x),\theta\big)$ including the Spearman and Kendall correlations.

These quantities do not depend on the feedback response f.

Corollary 1 distinguishes the information contained in a score from the scale on which that information is expressed. Since

$$m^*(x) = \Phi\big(u(x)\big) \tag{18}$$

and Φ is invertible, the posterior mean can be recovered from the stable score through

$$u(x) = \Phi^{-1}\big(m^*(x)\big). \tag{19}$$

An observer who knows the stable score therefore has exactly the same information about the fundamental as an observer who knows the posterior mean.

In a lending application, feedback may substantially widen the distribution of credit scores because high scores improve financing conditions and thereby improve repayment outcomes. Even so, the score does not reveal any additional information about borrowers' latent repayment capacity. Borrowers are ordered in the same way, their posterior distributions of repayment capacity are unchanged, and every information measure invariant to one-to-one transformations remains fixed.

The result applies to nonlinear feedback responses. A nonlinear transformation can change variances, covariances, and Pearson correlations because these objects depend on cardinal scale. It cannot change mutual information or rank dependence because these objects depend only on the information set or ordering generated by the score. The distinction explains why Pearson correlation is excluded from the general statement.

In the affine Gaussian model, the stable score is an affine transformation of the posterior mean. The squared Pearson correlation with the fundamental is then also preserved and satisfies $\mathrm{Corr}\big(m^*(x),\theta\big)^2 = \rho$. This equality relies on the affine structure. Corollary 1 gives the distribution-free statement that survives under nonlinear monotone transformations.

Information invariance does not imply that the stable score performs poorly when evaluated against realized outcomes. The score affects those outcomes through the allocation, and the regression used for retraining incorporates this causal effect. At a fixed point, the resulting score satisfies an exact calibration condition. The next corollary records this property.

Corollary 2 (Calibration at stability)**.** *Let m^* be a performatively stable score satisfying $\mathbb{E}\left[|m^*(x)|\right] < \infty$. Then*

$$\mathbb{E}\left[y \mid m^*(x)\right] = m^*(x). \tag{20}$$

The calibration result follows from the defining property of a population squared-loss fixed point. Stability gives $m^*(x) = \mathbb{E}[y \mid x]$. Applying iterated expectations conditional on the stable score yields equation (20). The argument does not require Assumption 1, differentiability of the feedback response, or the additive outcome equation (2). It applies to any integrable stable point of nonparametric squared-loss retraining.

This property has an immediate interpretation. Among borrowers receiving a stable score of s, average realized repayment equals s. Among products receiving the same predicted sales score, average realized sales equal that score. A calibration exercise conducted on deployment data therefore finds no systematic discrepancy between predictions and outcomes.

The result is known in the performative prediction literature. It is implicit in the discussion of stable predictors in Hardt and Mendler-Dünner (2023) and is developed formally in Perdomo (2025). Related equilibrium conditions appear in Kim and Perdomo (2023) for decision making under outcome performativity and in Ball (2025) for scoring under strategic manipulation.

Its role in the present analysis follows from combining it with Corollary 1. Feedback leaves the information about fundamentals unchanged, yet the stable score remains exactly calibrated against outcomes generated under its own deployment. Calibration therefore does not reveal whether predictive success comes from information about the fundamental or from the causal effect of the allocation attached to the prediction. The next section shows that measured fit can rise with feedback even though the information characterized in Corollary 1 remains fixed.

1.3 The feedback-invariant residual

The structure theorem and Corollaries 1 and 2 separate three properties of a stable predictor. Feedback preserves the information contained in the posterior mean, the deployed score remains calibrated against realized outcomes, and the numerical scale of the score may change. The remaining question is whether this change in scale improves conventional measures of predictive fit. The answer follows from a simple accounting property of the outcome equation.

For any deployed score m, equation (2) implies

$$y - \mathbb{E}[y \mid x] = \theta - u(x) + \varepsilon. \tag{21}$$

The feedback term disappears because it is measurable with respect to x. A lender may improve a borrower's repayment outcome by extending cheaper credit, and a platform may raise a product's sales by increasing its visibility, but these effects are predictable from the score and are therefore absorbed by the fitted conditional mean. They do not enter the regression residual.

Lemma 1 (Feedback-invariant residual)**.** *Under equation (2), for any deployed score m, the conditional law of $y - \mathbb{E}[y \mid x]$ given x equals the conditional law of $\theta - u(x) + \varepsilon$ given x. This law does not depend on the feedback response f or on the deployed score m. In particular, the irreducible*

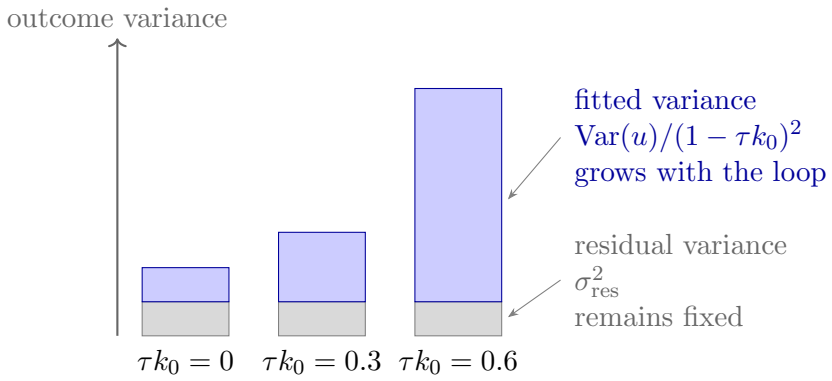


Figure 2: The variance decomposition under affine feedback. The figure sets $\mathrm{Var}(u) = \sigma^2_{\mathrm{res}} = 1$ and rescales bar heights for presentation. The fitted component equals $\mathrm{Var}(u)/(1-\tau k_0)^2$ and rises with feedback strength, while the residual component remains equal to σ^2_{res}. Deployment R^2 therefore rises even though the information about fundamentals remains unchanged.

risk satisfies $\sigma^2_{\mathrm{res}} := \mathbb{E}[\mathrm{Var}(y \mid x)] = \mathbb{E}[\mathrm{Var}(\theta \mid x)] + \mathbb{E}[\mathrm{Var}(\varepsilon \mid x)]$*, and is the same in every feedback regime.*

Lemma 1 provides the accounting identity that drives the fit comparison. Feedback changes the conditional mean of the outcome while leaving the conditional residual distribution fixed. In a credit application, repayment capacity that can be inferred from borrower characteristics belongs to $u(x)$, while the repayment improvement caused by financing belongs to $f(m(x))$. Both components are fitted by the regression. Unobserved heterogeneity in repayment capacity and transitory repayment shocks remain in the residual, and their distribution is unaffected by the strength of the allocation rule.

The result is stronger than a variance decomposition because the entire conditional residual law is invariant. This property later delivers the exact finite-sample results in Section 4. Sampling error in each retraining regression is generated by the same residual distribution, so changes in the deployed coefficient alter the conditional mean without changing the law of the estimation disturbance.

Figure 2 illustrates the variance accounting in the affine case. The residual component remains fixed while the fitted component expands with the loop gain. The bar heights are scaled for presentation, while their relative movement reflects the exact formulas reported in the caption.

1.4 Self-fulfilling accuracy

Lemma 1 fixes the component of outcome variation that cannot be predicted from the feature. To determine how measured fit changes with feedback, it remains to study the dispersion of the fitted conditional mean. Parameterize feedback strength by $\tau \geq 0$ and write

$$f_\tau(s) = \tau f_0(s). \tag{22}$$

Let Φ_τ denote the inverse transformation from Theorem 1, and let the corresponding stable score be

$$m^*_\tau(x) = \Phi_\tau\big(u(x)\big). \tag{23}$$

Deployment fit in regime τ is measured by

$$R^2(\tau) := \frac{\mathrm{Var}\big(m^*_\tau(x)\big)}{\mathrm{Var}\big(m^*_\tau(x)\big) + \sigma^2_{\mathrm{res}}}. \tag{24}$$

This expression is the population coefficient of determination from the regression of the deployed outcome on x. It follows because the stable score is the conditional mean of the outcome, while Lemma 1 fixes the irreducible residual variance. The definition is understood for regimes in which the denominator in (24) is positive.

A stronger feedback response does not automatically increase score dispersion when the response is nonlinear. The stable score itself moves with τ, so the relevant derivative is evaluated at a different point of the feedback function in each regime. The following condition requires the effective local gain along the stable path to rise with the overall strength of feedback.

Definition 2 (Monotone amplification)**.** *The family $\{f_\tau = \tau f_0\}_{\tau\in[0,\overline{\tau}]}$ satisfies monotone amplification if, for every u, the map $\tau \longmapsto \tau f_0'\big(\Phi_\tau(u)\big)$ is nondecreasing on $[0,\overline{\tau}]$.*

The condition has a direct economic interpretation. Consider a borrower with posterior repayment capacity u. As the lender places more weight on the score, the borrower receives more favorable financing and the stable score moves to a new point on the response schedule. Monotone amplification requires the total feedback operating at that new point to be at least as strong as before. The condition holds automatically under affine feedback because the local slope is constant. It also holds for nonnegative convex responses, which describe environments in which the marginal effect of an allocation becomes stronger as the allocation expands.

Theorem 2 (Self-fulfilling accuracy)**.** *Let f_0 be nondecreasing, and suppose that*

$$\overline{\tau}\sup_{t\in\mathbb{R}} f_0'(t) < 1, \tag{25}$$

and assume that $\mathrm{Var}(u(x)) < \infty$. If monotone amplification holds, then $\mathrm{Var}(m^*_\tau(x))$ and $R^2(\tau)$ are nondecreasing on $[0,\overline{\tau}]$. Monotone amplification holds in each of the following cases.

(a) The function f_0 is affine with nonnegative slope.

(b) The function f_0 is convex and nonnegative.

Suppose in case (a) that $f_0'(t) = k_0 > 0$ and $\mathrm{Var}(u(x)) > 0$. Then

$$\mathrm{Var}\big(m^*_\tau(x)\big) = \frac{\mathrm{Var}(u(x))}{(1-\tau k_0)^2} \tag{26}$$

is strictly increasing in τ. If $\sigma^2_{\text{res}} > 0$, then $R^2(\tau)$ is also strictly increasing. Throughout the comparison, every information measure in Corollary 1 remains constant in τ, and the calibration result in Corollary 2 holds in every regime.

Theorem 2 follows by comparing stable scores across two feedback regimes. Under monotone amplification, the transformation that maps the stable score in the weaker regime into the stable score in the stronger regime has derivative at least one. The stronger regime therefore stretches the distribution of scores without changing their ordering. This expansion in pairwise distances weakly increases the variance of the fitted conditional mean, while Lemma 1 leaves the residual variance unchanged.

The affine case makes the mechanism transparent. A borrower whose posterior repayment capacity rises by one unit receives a stable-score increase of $1/(1 - \tau k_0)$. As the lender's response to the score becomes stronger, the same information about repayment capacity produces a larger difference in financing and therefore a larger difference in realized repayment. The regression records the entire difference as predictive fit even though the information contained in borrower characteristics has not changed.

The result also clarifies the role of nonlinear feedback. Convexity allows amplification to be stronger for units with high posterior fundamentals, as when highly rated borrowers can convert additional financing into investment more effectively or when products near the top of a ranking receive disproportionately more attention. The stable transformation may then stretch the upper part of the score distribution more than the lower part. Its monotonicity preserves rankings and information, while its increasing steepness raises the dispersion of fitted outcomes.

Corollaries 1 and 2, Lemma 1, and Theorem 2 together characterize self-fulfilling accuracy. The stable predictor remains exactly calibrated against outcomes observed under deployment, its measured fit rises as the allocation loop strengthens, and its information about latent fundamentals remains unchanged. The additional fit is generated by the economic response to the score and is therefore manufactured by deployment rather than acquired through improved inference.

Related concerns appear in Perdomo (2025), who studies self-validation through performative stability, in Liley et al. (2021), who studies intervention-induced bias in model updating, and in Scarone and Baeza-Yates (2026), who studies selective observation in credit scoring. Theorem 2 provides a comparative-static characterization in an economically interpretable feedback parameter. The mechanism operates under correct specification, complete observation of outcomes, and population retraining.

2 Identification

2.1 Observational equivalence under a single deployment

Deployment data combine information about fundamentals with the consequences of the allocation generated by the model. In a lending application, borrower characteristics predict repayment capacity, while the credit terms assigned by the score affect investment and subsequent repayment. The conditional mean observed by the lender is therefore

$$\mathbb{E}[y \mid x] = u(x) + f\big(m(x)\big). \tag{27}$$

The first term reflects what the feature reveals about the borrower's underlying repayment capacity. The second reflects the improvement or deterioration in repayment caused by the financing decision attached to the score. Data generated under a single deployment reveal their sum, while the decomposition into these two components requires additional variation or restrictions.

Formally, the modeler observes the joint law P of (x, y) generated by equation (2) under a fixed deployed score m. The score need not be performatively stable. A structure is a pair $S = (\mu, f)$, where μ is a joint law of (θ, x, ε) satisfying the moment conditions in the setup, and $\mathbb{E}[\varepsilon \mid \theta, x] = 0$ while f is a measurable feedback response. A structure is consistent with P if equation (2), evaluated under that structure and the deployed score m, reproduces the observed law P.

The source of the identification problem can be seen by considering any square-integrable function of the deployed score. An observed repayment advantage among highly scored borrowers can be attributed partly to stronger latent borrowers or partly to the favorable financing induced by their scores. Moving a score-dependent component from the fundamental into the feedback response leaves every observed outcome unchanged. The following theorem characterizes this observational equivalence.

Theorem 3 (Observational equivalence under deployment)**.** *Consider the unrestricted class of structures in which f is any measurable function and μ is any joint law satisfying the setup moment conditions. Assumption 1 is not imposed, and the transformations below need not preserve differentiability, monotonicity, or the loop-gain bound. Let $S = (\mu, f)$ be consistent with P, and let $c\colon \mathbb{R} \to \mathbb{R}$ be any measurable function satisfying $\mathbb{E}\Big[c\big(m(x)\big)^2\Big] < \infty$. Define $S_c = (\mu_c, f + c)$, where μ_c is the law of $\big(\theta - c\big(m(x)\big), x, \varepsilon\big)$. Then S_c is also consistent with P. Consequently, the following statements hold.*

(i) The posterior mean is identified only up to transformations of the form

$$u(x) \longmapsto u(x) - c\big(m(x)\big). \tag{28}$$

If m is injective on the support of x, every measurable function of x can be represented as a measurable function of $m(x)$ on that support. The decomposition in equation (27) is then

unrestrictedly unidentified.

(ii) Every functional of the observed law P is constant across the identified set. This includes calibration statistics, R^2, proper scoring rules, and any other validation criterion computed solely from deployment data.

(iii) The transformations

$$(\theta, u, f) \longmapsto \big(\theta - c\big(m(x)\big), u - c \circ m, f + c\big) \tag{29}$$

generate a subset of the identified set. Any structural functional that changes under one admissible transformation is therefore not identified. In particular, the fundamental-information quantities in Corollary 1 are not identified in general. Invariance under the shift family is necessary for identification, although it is not sufficient because the identified set may also contain redecompositions of the latent fundamental and the outcome shock that are not generated by deterministic score-dependent shifts.

The equivalence follows from an exact cancellation. Under the transformed structure, the observed outcome is

$$\begin{aligned} y &= \theta - c\big(m(x)\big) + f\big(m(x)\big) + c\big(m(x)\big) + \varepsilon \\ &= \theta + f\big(m(x)\big) + \varepsilon. \end{aligned} \tag{30}$$

The transformation changes the economic interpretation of the outcome without changing its observed distribution. It also preserves the conditional mean restriction on ε because, conditional on x, the original fundamental can be recovered from the transformed fundamental by adding $c(m(x))$.

Theorem 3 is an absence-of-overlap result. The deployed allocation is a deterministic function of the feature because it is determined by $m(x)$. Conditional on x, the data contain no otherwise comparable units exposed to a different allocation. In a lending application, every borrower with the same observed characteristics receives the same score-dependent credit terms. The data therefore cannot reveal how the same borrower would have performed under a different credit allocation. The observed conditional mean combines borrower quality with the causal effect of financing.

This interpretation connects the theorem to Proposition 3 of Mendler-Dünner et al. (2022), which establishes the analogous identification failure in a structural causal model. The equivalence in Theorem 3 concerns an unrestricted structural class. Restrictions on the latent conditional mean or on the feedback technology can reduce the identified set by preventing the fundamental from absorbing arbitrary functions of the score.

Variation across deployments provides another route. Suppose outcomes are observed under two scoring rules, m_1 and m_2, while the fundamental environment and feedback response remain unchanged. Their conditional means satisfy

$$\mathbb{E}\big[y^{(m_1)} \mid x\big] - \mathbb{E}\big[y^{(m_2)} \mid x\big] = f\big(m_1(x)\big) - f\big(m_2(x)\big). \tag{31}$$

The posterior mean cancels from this comparison, so differences across deployments reveal differences in the feedback response. Archived model vintages can therefore act as quasi-experiments when their variation is exogenous to fundamentals. Section 4 shows that fresh-sample retraining noise generates such variation in the static linear-Gaussian economy.

Theorem 3 also clarifies why high deployment accuracy cannot resolve the problem. Calibration, fit, and proper scoring rules are all functionals of the observed distribution, which remains unchanged across observationally equivalent structures. A model that predicts strong fundamentals and has little effect on outcomes can generate the same deployment data as a model that predicts weaker fundamentals and creates a large outcome response through its allocation. The next subsection studies designs and institutional rules that supply the variation missing from a single deterministic deployment.

2.2 Restoring identification

The identification failure arises because the allocation does not vary among units with the same feature. Random assignment restores this variation directly. A lender can continue to score every applicant while assigning a randomly selected group to a fixed baseline credit policy. The outcomes of this group reveal the conditional mean of fundamentals under a common allocation, while the difference between the deployed and baseline groups reveals the effect of the score-dependent policy.

Proposition 1 (Holdout identification)**.** *Augment the environment with a randomization indicator $D \in \{0,1\}$ that is independent of (θ, x, ε) and satisfies $\mathbb{P}(D=1) = h$, $h \in (0,1)$. Units with $D=1$ form the holdout group. They are scored but receive a fixed baseline allocation associated with the score s_0, so their outcome is*

$$y = \theta + f(s_0) + \varepsilon. \tag{32}$$

Units with $D=0$ receive the score-dependent allocation in equation (2). The joint law of (x, y, D) then identifies the following objects.

(i) The conditional mean in the holdout group satisfies

$$\mathbb{E}[y \mid x, D=1] = u(x) + f(s_0). \tag{33}$$

The posterior mean $u(x)$ is therefore identified on the full support of x, up to the additive constant $f(s_0)$.

(ii) The difference between the deployed and holdout conditional means satisfies

$$\mathbb{E}[y \mid x, D=0] - \mathbb{E}[y \mid x, D=1] = f\big(m(x)\big) - f(s_0). \tag{34}$$

The feedback response is therefore identified on the range of m, up to the same additive constant.

(iii) Suppose that f is differentiable and that the support of $m(x)$ contains an interval I. The local loop gain $k(s) = f'(s)$ is point-identified on the interior of I. If I covers the support of $m(x)$ up to null sets, then the deployment-average gain $\overline{k} = \mathbb{E}\big[f'\big(m(x)\big)\big]$ is also point-identified. A sufficient condition for the support requirement is that $m(x)$ has a density that is positive almost everywhere on I.

The holdout design randomizes the allocation while preserving the score. Every unit can therefore be evaluated by the model, and the comparison between the deployed and baseline groups is available throughout the support of the feature. This differs from designs that randomize whether the model is evaluated at all, which would remove the score for some units and weaken the comparison across feature values.

The additive constant reflects the absence of an observed allocation with a known zero effect. If the baseline allocation satisfies $f(s_0) = 0$, then both the posterior mean and the feedback response are identified in levels. In a lending application, this normalization may correspond to a benchmark financing policy whose effect on repayment is normalized to zero. Without such a benchmark, the data identify how outcomes change with the allocation while leaving the common level of the fundamental and feedback components undetermined.

Randomized intervention-free holdouts are studied by Liley et al. (2021) and Haidar-Wehbe et al. (2025) as a method for preventing intervention-induced bias in predictive model updating. Munro (2025) studies related treatment allocation problems when agents respond strategically. Proposition 1 characterizes the structural objects recovered by such a design in the present environment. The applied companion paper studies permanent and rotating holdouts in a dynamic environment in which current outcomes depend on capital accumulated under earlier deployments, so the two designs recover different components of the feedback there.

Randomization may be costly or institutionally infeasible. Many economic allocation rules nevertheless contain discrete changes that create local variation. Credit approval, insurance eligibility, regulatory classification, and platform access often change discontinuously when a score crosses a known threshold. A discontinuity can identify the feedback effect when the latent fundamental varies smoothly through the cutoff, and the allocation technology is known to operate through the jump.

Proposition 2 (Threshold allocations)**.** *Suppose that the allocation is binary at a known cutoff c and that the feedback response belongs to the class*

$$f_\Delta(s) = a + \Delta \mathbf{1}\{s \geq c\}, \tag{35}$$

where a and Δ are unknown constants. Assume that m is continuous, that $m(x)$ has a density that is positive in a neighborhood of c, and that the score-conditional fundamental mean

$$r(s) := \mathbb{E}[\theta \mid m(x) = s] \tag{36}$$

is continuous at c. Then deployment data identify the jump through

$$\Delta = \lim_{s \downarrow c} \mathbb{E}[y \mid m(x) = s] - \lim_{s \uparrow c} \mathbb{E}[y \mid m(x) = s]\,. \tag{37}$$

Once Δ is known, the posterior mean $u(x)$ is identified on the full support of x up to the additive constant a. If the feedback class is unrestricted beyond the threshold jump, the observational equivalence in Theorem 3 continues to apply.

The two designs rely on different sources of identifying variation. The holdout creates variation in the allocation by random assignment and therefore requires little knowledge of the shape of the feedback response. The threshold design uses a known discontinuity in the allocation rule and requires continuity of the latent score-conditional mean. Its identifying content consequently depends on the maintained restriction that the feedback response changes only through the discrete jump. Appendix A illustrates the identifying configuration and develops two economic caveats, a smooth score-dependent component of the allocation that remains observationally equivalent to a smooth component of borrower quality, and a discontinuous change in the composition of units at the cutoff when the score compresses multidimensional information.

Taken together, Theorem 3 and Propositions 1 and 2 show that separating information from feedback requires variation that exposes comparable units to different allocations. A single deterministic deployment does not contain this variation. Randomized holdouts create it throughout the support of the feature, while threshold rules create it locally under continuity and functional-form restrictions. The next section studies how feedback affects the coefficients of a restricted predictive model once the economic structure is specified.

3 Linear Scores

The preceding analysis retrains over an unrestricted class of measurable predictors, with deployment affecting a unit only through its score. Institutions often restrict prediction to a low-dimensional model instead, because coefficients must be interpretable, auditable, and comparable across model vintages. A lender may use a linear index of income, leverage, and repayment history, while a regulator may examine changes in the corresponding coefficient vector to understand how the model has evolved. Restricting the model class raises a new question. Feedback may amplify the score, but it may also appear to change which characteristics the model uses. This section shows that, under Gaussian or elliptical features, feedback operating through a scalar score cannot rotate a stable linear coefficient. Feedback changes the strength of the model's response to its information while preserving the relative weights assigned to the features, so rotation across vintages provides a diagnostic for mechanisms outside the baseline allocation loop.

3.1 Linear retraining and coefficient direction

Linear scores are the leading restricted class. Let $\widetilde{x} \in \mathbb{R}^d$ be centered Gaussian with nonsingular covariance matrix Σ, and let θ be a square-integrable fundamental. Define $q \coloneqq \mathrm{Cov}(\widetilde{x}, \theta),\ q \neq 0$. The modeler deploys the linear score $s = \lambda'\widetilde{x}$ and the resulting outcome is

$$y = \theta + f\big(\lambda'\widetilde{x}\big) + \varepsilon, \quad \mathbb{E}[\varepsilon\widetilde{x}] = 0. \tag{38}$$

The feedback response f is Lipschitz and satisfies

$$\sup_{t\in\mathbb{R}} f'(t) \leq \kappa^+ < 1. \tag{39}$$

Retraining uses population least squares, so the coefficient update is

$$T\lambda = \Sigma^{-1}\,\mathrm{Cov}\big(\widetilde{x}, y^{(\lambda)}\big). \tag{40}$$

A stable coefficient satisfies $\lambda^* = T\lambda^*$. The coefficient in the absence of feedback is $v \coloneqq \Sigma^{-1}q$.

Under Gaussian features, the covariance between the feature vector and any differentiable function of the score is aligned with $\Sigma\lambda$. Feedback that operates only through the scalar score therefore has no direction in feature space other than the direction already chosen by the model. This observation yields the no-rotation result.

Theorem 4 (No rotation)**.** *Every stable coefficient is collinear with the no-feedback coefficient. Any stable coefficient satisfies*

$$\lambda^* = \frac{v}{1 - \overline{k}(\lambda^*)}, \quad \overline{k}(\lambda) \coloneqq \mathbb{E}\big[f'\big(\lambda'\widetilde{x}\big)\big], \tag{41}$$

with $\overline{k}(\lambda^*) < 1$. At least one stable coefficient exists. If f is affine with slope k, the stable coefficient is unique and satisfies

$$\lambda^* = \frac{v}{1-k}. \tag{42}$$

Theorem 4 follows from Stein's lemma. For any deployed coefficient λ,

$$\mathrm{Cov}\big(\widetilde{x}, f\big(\lambda'\widetilde{x}\big)\big) = \Sigma\lambda\,\mathbb{E}\big[f'\big(\lambda'\widetilde{x}\big)\big]. \tag{43}$$

The feedback term in the population regression therefore points in the same direction as the deployed coefficient. At a fixed point, the regression combines the no-feedback coefficient v with a scalar multiple of λ^*, which forces λ^* to lie on the ray spanned by v.

The economic interpretation is immediate. Suppose a credit score places positive weight on income and negative weight on leverage. When better scores reduce financing costs and thus improve repayment, retraining increases the strength with which the score responds to the same index of borrower characteristics. It cannot change the relative weights on income and leverage because

the outcome response is generated by the scalar score rather than by feature-specific interventions. Feedback changes the length of the coefficient vector while preserving its direction. The stable coefficient also admits an interpretation as a Berk–Nash equilibrium under a Gaussian subjective model, which Appendix A develops together with the reading of the fixed point as equilibrium learning under misspecification.

Gaussianity provides a convenient covariance identity, although the absence of rotation relies on a broader property. Elliptical distributions preserve linear conditional expectations along one-dimensional projections, which is sufficient to keep the feedback covariance aligned with the deployed coefficient.

Proposition 3 (No rotation under elliptical features)**.** *Let $\widetilde{x}$ be elliptically distributed with mean zero, nonsingular covariance matrix Σ, and finite second moments. Define, for $\lambda \neq 0$,*

$$\overline{k}(\lambda) := \frac{\mathrm{Cov}\big(\lambda'\widetilde{x}, f\big(\lambda'\widetilde{x}\big)\big)}{\mathrm{Var}\big(\lambda'\widetilde{x}\big)}. \tag{44}$$

Every stable coefficient satisfies

$$\lambda^* = \frac{v}{1-\overline{k}(\lambda^*)}, \quad \overline{k}(\lambda^*) \leq \kappa^+ < 1, \tag{45}$$

and at least one stable coefficient exists. Under Gaussian features,

$$\overline{k}(\lambda) = \mathbb{E}\big[f'\big(\lambda'\widetilde{x}\big)\big], \tag{46}$$

so Theorem 4 is recovered.

Proposition 3 shows that no rotation is a restriction on the geometry of the feature distribution rather than a peculiarity of Gaussian integration by parts. Under ellipticity, the regression of the feature vector on the scalar score is linear, so the covariance induced by feedback remains proportional to $\Sigma\lambda$. The covariance ratio in equation (44) replaces the average derivative in the Gaussian case, while the Lipschitz bound ensures that this effective gain remains below one.

The result gives coefficient rotation an empirical interpretation. If archived model vintages move primarily along a fixed ray, the pattern is consistent with changes in the strength of score-mediated feedback. Persistent changes in coefficient direction point toward mechanisms outside this class. A lender may alter specific contract terms in response to income or collateral rather than to the aggregate score, a platform may allow rankings to change future features through user engagement, or the feature distribution may depart substantially from ellipticity. In each case, feedback can load differently across characteristics and rotate the coefficient vector. The diagnostic concerns rotation beyond ordinary estimation error, after estimated coefficients have been placed on a common normalization and changes in the feature distribution have been accounted for, because sampling variation alone rotates estimated vectors even when the population mechanism satisfies the theorem.

The additive specification underlying these results assumes that units receiving the same score have the same expected response to the resulting allocation. Appendix A relaxes this restriction and allows the response to interact with the latent fundamental, so that two borrowers receiving the same credit line may invest differently because they differ in private productivity. When posterior distributions are ordered by a one-dimensional index, the stable score remains a strictly increasing transformation of that index, and information invariance and calibration continue to hold. Residual invariance and the corresponding fit comparison generally do not survive, because private responses add score-dependent residual variation, and a randomized holdout identifies the average private response rather than the underlying allocation and production technology.

3.2 The linear-Gaussian benchmark

The linear-Gaussian economy provides the case in which the general transformations become affine, and every object has a closed form. Let $\theta = \mu + \widetilde{\theta}$, $\widetilde{\theta} \sim N(0, \sigma_\theta^2)$ and suppose $x = \theta + \eta$, $\eta \sim N(0, \sigma_\eta^2)$ with $\widetilde{\theta}$ independent of η. Define $V := \sigma_\theta^2 + \sigma_\eta^2$, $\rho := \frac{\sigma_\theta^2}{V}$, and $\widetilde{x} := x - \mu$. Let the feedback response be $f(s) = ks + \beta a_0$, $k = \beta\gamma$, which is the reduced form of a linear allocation rule and a linear outcome response.

Example 1 (Linear-Gaussian benchmark)**.** *The posterior mean is $u(x) = \mu + \rho\widetilde{x}$. The structure theorem gives $\Phi(u) = \frac{u+\beta a_0}{1-k}$ and*

$$m^*(x) = \frac{\mu + \beta a_0 + \rho\widetilde{x}}{1-k}. \tag{47}$$

The stable slope is $\rho/(1-k)$, and the variance of the stable score is

$$\mathrm{Var}\big(m^*(x)\big) = \frac{\rho^2 V}{(1-k)^2}. \tag{48}$$

By Theorem 2, this variance and the associated deployment fit increase as k approaches one from below, while the information content of the score remains fixed. The stable score is affine in the posterior mean, so $\mathrm{Corr}\big(m^*(x), \theta\big)^2 = \rho$ for every $k < 1$. Theorem 3 reduces to observational equivalence among primitive pairs that generate the same deployed conditional mean, while Proposition 1 reduces to the corresponding holdout moment restrictions.

The benchmark makes the two forces in the model transparent. Measurement error shrinks the posterior mean relative to the feature by the reliability coefficient ρ. Feedback then expands the stable score by the multiplier $1/(1-k)$. The stable score becomes steeper than the feature when $k > 1 - \rho$. Figure 3 illustrates this comparison. The feature, posterior mean, and stable score are affine functions of the same scalar signal, so they rank units identically and share the same squared correlation with the fundamental. Their slopes differ because measurement error attenuates the posterior mean while deployment feedback amplifies the stable score. The linear-Gaussian economy is

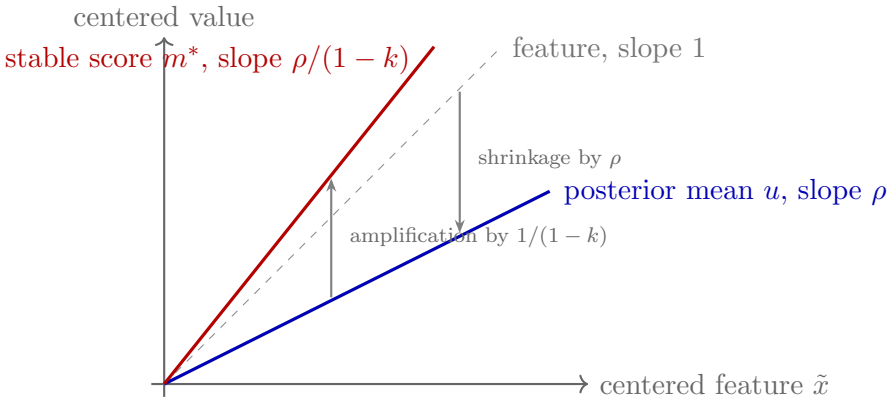


Figure 3: The linear-Gaussian boundary case of Example 1, drawn for $\rho = 0.5$ and $k = 0.6$. Measurement noise shrinks the posterior mean below the feature, the retraining loop re-amplifies the stable score by the multiplier $1/(1-k)$, and the score becomes steeper than the feature exactly when $k > 1-\rho$. All three lines are affine in the same feature, so they order units identically and share the squared correlation ρ with the fundamental, while their slopes, and with them measured fit, differ.

therefore the benchmark in which the general information-invariance results coexist with closed-form fit and stability calculations.

4 The Exact Finite-Sample Theory of Retraining

The population analysis in Section 1 characterizes the fixed point of the retraining operator and the conditions under which repeated population retraining converges to it. Finite-sample retraining introduces a new source of variation because each estimated model responds to sampling error in the cross-section on which it is trained. This section studies whether the population stability frontier continues to govern the resulting sequence of deployed coefficients, how those coefficients behave below the frontier, and whether sampling variation reveals features of the feedback mechanism that are hidden in deployment levels. In the linear-Gaussian economy of Section 3.2, each of these questions admits an exact answer.

4.1 Setting

Consider the static linear-Gaussian economy. In period t the modeler deploys the linear score $s_{i,t} = \bar{s}_t + \lambda_t \tilde{x}_{i,t}$. A fresh cross-section of $n \geq 2$ units is drawn, with $\tilde{x}_{i,t} = \tilde{\theta}_{i,t} + \eta_{i,t}$, where across units and periods the shocks $(\tilde{\theta}_{i,t}, \eta_{i,t}, \varepsilon_{i,t})$ are mutually independent Gaussian variables, $\tilde{\theta}_{i,t} \sim N(0, \sigma_\theta^2)$, $\eta_{i,t} \sim N(0, \sigma_\eta^2)$, $\varepsilon_{i,t} \sim N(0, \sigma_\varepsilon^2)$, with successive cross-sections independent over time, $V = \sigma_\theta^2 + \sigma_\eta^2$, and $\rho = \sigma_\theta^2 / V$.

Joint Gaussianity is the main distributional restriction in this section. The population theory requires only that the feature be orthogonal to the regression residual. Gaussianity strengthens this

orthogonality to independence, which makes the finite-sample regression disturbance independent of the sample feature variance. This property produces the exact innovation distribution derived below.

Under the deployed coefficient λ_t, the realized outcome is

$$y_{i,t} = \overline{y}_t + \big(1 + k\lambda_t\big)\widetilde{\theta}_{i,t} + k\lambda_t\eta_{i,t} + \varepsilon_{i,t}, \tag{49}$$

where $k = \beta\gamma$ is the loop gain. A higher credit-score coefficient, for example, makes financing more responsive to borrower characteristics. The induced investment then makes realized repayment more responsive to the same characteristics. Equation (49) captures this reinforcement through the term $k\lambda_t$.

We maintain $k \geq 0$ throughout this section and the dynamic extension. This is the sign implied by a productive allocation that raises the outcome the model is trained to predict. The subcritical regime is therefore $k \in [0, 1)$. Negative gains would generate oscillatory retraining dynamics and require a separate analysis.

The modeler estimates a regression with an intercept using the current cross-section and deploys the estimated slope in the following period. Let $\overline{x}_t = \frac{1}{n}\sum_{i=1}^n \widetilde{x}_{i,t}$ and $\overline{y}_t^{\text{obs}} = \frac{1}{n}\sum_{i=1}^n y_{i,t}$. The coefficient update is

$$\lambda_{t+1} = \widehat{\lambda}_t^{\text{OLS}} = \frac{\sum_{i=1}^n \big(\widetilde{x}_{i,t} - \overline{x}_t\big)\big(y_{i,t} - \overline{y}_t^{\text{obs}}\big)}{\sum_{i=1}^n \big(\widetilde{x}_{i,t} - \overline{x}_t\big)^2}. \tag{50}$$

Each model vintage is therefore estimated from data generated by the preceding vintage. A favorable sampling realization raises the estimated coefficient, which strengthens the next period's allocation response and changes the outcomes used in the following regression.

The initial coefficient λ_0 is deterministic, or more generally independent of all subsequent cross-sections. This nonanticipative initialization is needed for the divergence results. An initial condition constructed using future samples could be selected to offset the random component of a supercritical path and would therefore obscure the instability generated by retraining itself.

Two properties inherited from the population economy drive the finite-sample analysis. First, the population slope under a deployed coefficient λ is $\frac{\text{Cov}\left(\widetilde{x}_{i,t}, y_{i,t}\right)}{V} = \rho + k\lambda$. The regression therefore combines the reliability of the feature with the feedback generated by the currently deployed coefficient.

Second, the population regression residual does not depend on the deployment regime. For any coefficient λ, define

$$u_{i,t} \coloneqq y_{i,t} - \mathbb{E}[y_{i,t}] - (\rho + k\lambda)\widetilde{x}_{i,t}. \tag{51}$$

Substituting equation (49) gives

$$u_{i,t} = (1 - \rho)\widetilde{\theta}_{i,t} - \rho\eta_{i,t} + \varepsilon_{i,t}. \tag{52}$$

The residual is independent of λ and satisfies $\text{Cov}\big(\widetilde{x}_{i,t}, u_{i,t}\big) = 0$. Joint Gaussianity then implies

$u_{i,t} \perp \widetilde{x}_{i,t}$.

The variance of the residual is $\sigma_u^2 \coloneqq \mathrm{Var}(u_{i,t}) = \rho(1-\rho)V + \sigma_\varepsilon^2$. The deployment coefficient changes the conditional mean of the outcome but leaves the regression disturbance unchanged. In a lending application, stronger score-based financing increases the part of repayment explained by borrower characteristics, while unobserved repayment capacity and transitory repayment shocks continue to generate the same residual variation. This invariance makes the sampling error in each regression exogenous to the coefficient deployed in the preceding period.

Throughout the analysis, we assume $V > 0$, $\sigma_\theta^2 > 0$, and $\sigma_u^2 > 0$. These restrictions imply $\rho > 0$ and ensure that the finite-sample innovations are nondegenerate. If $\rho = 1$ and $\sigma_\varepsilon^2 = 0$, the regression residual vanishes and the coefficient update becomes deterministic. The stationary distribution below the frontier then collapses to a point mass, while an initialization exactly at a supercritical fixed point remains there indefinitely. If $\rho = 0$, the feature contains no information about the fundamental, and the positive drift that governs the critical case disappears. The maintained restrictions exclude these degenerate boundary cases and isolate the interaction between informative features, sampling variation, and deployment feedback.

4.2 Fresh-sample retraining as an exact autoregression

Finite-sample retraining turns the population fixed-point problem into a time-series problem. Each cross-section produces a new estimate of the slope, and that estimate determines the allocation applied to the next cross-section. The sequence of model vintages therefore records both the deterministic feedback in the population retraining operator and the sampling variation generated by estimation.

In the linear-Gaussian economy, this sequence has an exact autoregressive representation. The autoregressive coefficient is the economic loop gain, while the innovation is the ordinary least-squares estimation error. The result provides the foundation for the stability, diagnostic, and identification results that follow.

Theorem 5 (Exact autoregressive representation)**.** *For every t,*

$$\lambda_{t+1} = \rho + k\lambda_t + U_t, \tag{53}$$

where the innovations $\{U_t\}_{t\geq 0}$ are independent and identically distributed across periods and are independent of $(\lambda_0, U_0, \ldots, U_{t-1})$, with the exact distribution

$$U_t \overset{d}{=} \frac{\sigma_u}{\sqrt{V(n-1)}} T_{n-1}, \quad T_{n-1} \sim \text{Student-}t_{n-1}. \tag{54}$$

Consequently, $\mathbb{E}[U_t] = 0$ for $n \geq 3$, and $\mathrm{Var}(U_t) = \dfrac{\sigma_u^2}{V(n-3)}$ for $n \geq 4$.

To see the source of the representation, substitute the population regression decomposition into the least-squares estimator. The estimated slope can be written as

$$\widehat{\lambda}_t^{\mathrm{OLS}} = \rho + k\lambda_t + \frac{\sum_{i=1}^{n}\left(\widetilde{x}_{i,t} - \overline{x}_t\right)u_{i,t}}{\sum_{i=1}^{n}\left(\widetilde{x}_{i,t} - \overline{x}_t\right)^2}. \tag{55}$$

Conditional on the observed features, the numerator of the final term is Gaussian with mean zero and variance $\sigma_u^2 \sum_{i=1}^{n}\left(\widetilde{x}_{i,t} - \overline{x}_t\right)^2$. The denominator satisfies $\frac{1}{V}\sum_{i=1}^{n}\left(\widetilde{x}_{i,t} - \overline{x}_t\right)^2 \sim \chi^2_{n-1}$. Independence between the Gaussian regression residual and the observed feature then converts the ratio into the scaled Student distribution in equation (54). Fresh and independent cross-sections make these innovations independent across vintages.

Theorem 5 has a direct economic interpretation. A lending system that repeatedly estimates the sensitivity of repayment to borrower characteristics generates a first-order autoregression in its own slope coefficient. The persistence of the coefficient archive is the strength of the allocation loop. A higher value of k means that a sampling-induced increase in the current coefficient produces a stronger credit response, which raises the covariance between borrower characteristics and subsequent repayment and therefore persists into the next estimate.

The innovation scale reflects the statistical environment in which the model is retrained. It falls with the cross-sectional sample size and the variance of the observed characteristic and rises with the residual uncertainty in the outcome. The model archive therefore combines an economic parameter, represented by the autoregressive coefficient, with a statistical disturbance generated by the sampling design.

Two properties make the representation exact. Gaussianity turns the orthogonality condition $\mathrm{Cov}(\widetilde{x}_{i,t}, u_{i,t}) = 0$ into independence, which removes finite-sample bias from the regression slope. Residual invariance ensures that the law of the estimation error does not depend on the previously deployed coefficient. The second property is structural and follows from the additive feedback environment. The first is distributional and can be relaxed at the cost of introducing a finite-sample intercept bias.

Proposition 4 (Robustness beyond Gaussianity)**.** *Retain the environment of the setting and suppose that $\widetilde{\theta}_{i,t}$, $\eta_{i,t}$, and $\varepsilon_{i,t}$ are mutually independent within units, have mean zero and finite fourth moments, and are independent across units and periods. Let $S_t = \sum_{i=1}^{n}\left(\widetilde{x}_{i,t} - \overline{x}_t\right)^2$ and suppose that $\mathbb{E}[S_t^{-2}] < \infty$. Then the retraining path satisfies*

$$\lambda_{t+1} = (\rho + b_n) + k\lambda_t + \widetilde{U}_t, \tag{56}$$

where $\{\widetilde{U}_t\}_{t\geq 0}$ are independent and identically distributed, have mean zero, and are independent of the past. The constant $b_n = \mathbb{E}[U_t]$ captures the finite-sample bias generated by dependence between the regression residual and the design. The following statements hold.

(i) If $k \in [0, 1)$, the process has a unique stationary distribution with mean $\mathbb{E}[\lambda_\infty] = \frac{\rho + b_n}{1-k}$. If

$\mathrm{Var}(\widetilde{U}_t) < \infty$, *its variance is* $\mathrm{Var}(\lambda_\infty) = \frac{\mathrm{Var}(\widetilde{U}_t)}{1-k^2}$.

(ii) The estimator $\widehat{k}_T$ *in Proposition 5(i) remains consistent. If* $\mathbb{E}\left[|\widetilde{U}_t|^{2+\varsigma}\right] < \infty$ *for some* $\varsigma > 0$, *then* $\sqrt{T}\left(\widehat{k}_T - k\right) \Rightarrow N(0, 1-k^2)$. *The corresponding intercept estimator identifies* $\rho + b_n$ *rather than* ρ.

(iii) If $k > 1$ *and the distribution of* $\widetilde{U}_t$ *has an atomless component, then* $|\lambda_t| \longrightarrow \infty$ *almost surely.*

The symmetry, Student form, exact innovation variance, regularly varying tail index, and exact moment characterization obtained under Gaussianity need not hold under the weaker assumptions.

Proposition 4 separates the implications of linear feedback from those of Gaussian sampling. Linearity and fresh cross-sections preserve the autoregressive law, the exogeneity of vintage innovations, and the location of the stability frontier. Gaussianity delivers the stronger finite-sample statements. It eliminates the intercept bias, fixes the innovation distribution, and ties the stationary tail index directly to the cross-sectional sample size.

This distinction determines which diagnostics are robust. Persistence in the coefficient archive continues to identify the loop gain under the conditions of Proposition 4. Diagnostics based on symmetry, exact variances, or tail indices require the Gaussian specification. Slope-based auditing therefore rests on a broader set of assumptions than tail-based auditing.

The autoregressive representation also converts the population frontier into a sharp stochastic phase transition.

Theorem 6 (Stochastic phase transition)**.** *Let* λ_t *satisfy equation* (53) *from any nonanticipative initial condition* λ_0.

(i) Suppose that $|k| < 1$. *The process is geometrically ergodic and has a unique stationary distribution represented by*

$$\lambda_\infty = \lambda^* + \sum_{j=0}^{\infty} k^j U_{-j}, \quad \lambda^* = \frac{\rho}{1-k}. \tag{57}$$

The stationary distribution is symmetric about λ^*. *For* $n \geq 3$, $\mathbb{E}[\lambda_\infty] = \lambda^*$ *and for* $n \geq 4$, $\mathrm{Var}(\lambda_\infty) = \frac{\sigma_u^2}{V(n-3)} \frac{1}{1-k^2}$. *The stationary distribution has regularly varying tails with index* $n-1$. *Consequently,* $\mathbb{E}\left[|\lambda_\infty|^p\right] < \infty \Longleftrightarrow p < n-1$. *The stationary distribution exists for every* $n \geq 2$, *although its mean and variance require the stronger sample-size conditions stated above.*

(ii) Suppose that $k = 1$. *For* $n \geq 3$,

$$\lambda_t = \lambda_0 + \rho t + \sum_{j=0}^{t-1} U_j, \tag{58}$$

and $\frac{\lambda_t}{t} \to \rho$ *almost surely. Since* $\rho > 0$ *under the maintained assumptions, the coefficient diverges at the linear rate* ρt*. If* $\rho = 0$*, the coefficient instead follows a centered random walk with unbounded positive and negative excursions.*

(iii) Suppose that $k > 1$*. Define the repelling population fixed point* $\widetilde{\lambda} = \frac{\rho}{1-k} < 0$*. Then*

$$\lambda_t - \widetilde{\lambda} = k^t \left[\lambda_0 - \widetilde{\lambda} + M_t\right], \tag{59}$$

where $M_t = \sum_{j=0}^{t-1} k^{-j-1} U_j \to M_\infty$ *almost surely. Moreover,* $\Pr\left(\lambda_0 - \widetilde{\lambda} + M_\infty = 0\right) = 0$*. It follows that* $|\lambda_t| \longrightarrow \infty$ *almost surely, and the divergence occurs at the geometric rate* k^t*. This conclusion also holds when the process is initialized at the repelling fixed point.*

Regular variation in part (i) follows because geometric sums of independent regularly varying random variables preserve the innovation tail index (Embrechts et al., 1997, Appendix A3.3). The persistence parameter changes the scale of the stationary distribution but does not change the power-law exponent inherited from the Student innovation.

Theorem 6 shows that sampling noise does not move the stability frontier. Below the frontier, each sample perturbs the coefficient away from the population fixed point, while contraction eventually absorbs the effect of the perturbation. At the frontier, the positive informational component ρ accumulates without mean reversion. Above the frontier, even a small estimation error is repeatedly magnified, so the path leaves the repelling fixed point and diverges geometrically.

The supercritical result is stronger than its population counterpart. Population retraining initialized exactly at the repelling fixed point remains there forever. Finite-sample retraining has no such knife-edge refuge because the first nondegenerate estimation error moves the coefficient away from the fixed point with probability one.

The stationary regime also produces a measurable warning signal before instability occurs. Figure 4 illustrates the phase transition. Below the frontier, the slope fluctuates ergodically around the performative fixed point, at the frontier it drifts at the linear rate ρt, and above the frontier it grows geometrically even when initialized at the repelling fixed point.

Corollary 3 (Noise amplification and early warning)**.** *Suppose that* $n \geq 4$ *and* $k \in [0, 1)$*. The stationary standard deviation of the deployed slope is*

$$\mathrm{sd}(\lambda_\infty) = \frac{\sigma_u}{\sqrt{V(n-3)}} \frac{1}{\sqrt{1-k^2}}. \tag{60}$$

It is strictly increasing in k *and satisfies* $\mathrm{sd}(\lambda_\infty) \longrightarrow \infty$ *as* $k \uparrow 1$*. Holding the sample size, feature variance, reliability, and outcome-noise variance fixed, increasing dispersion in coefficient levels therefore indicates an increase in the feedback gain.*

The early-warning statistic is the dispersion of coefficient levels across model vintages. It differs

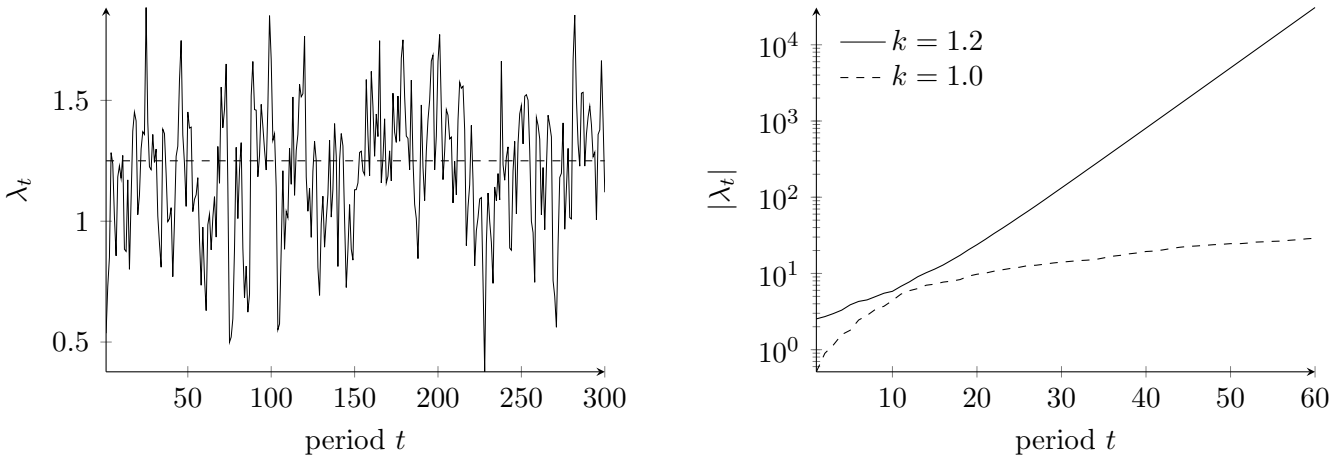


Figure 4: The phase transition of fresh-sample retraining ($n = 12$, $\rho = 0.5$). Left panel, the subcritical path at $k = 0.6$ fluctuates ergodically around $\lambda^* = 1.25$, the dashed line. Right panel, log scale, the critical path at $k = 1.0$ drifts linearly while the supercritical path at $k = 1.2$, initialized at the repelling fixed point, diverges geometrically.

from the volatility of adjacent coefficient revisions. Under stationarity,

$$\operatorname{Var}(\lambda_{t+1} - \lambda_t) = \frac{2\sigma_u^2}{V(n-3)} \frac{1}{1+k}. \tag{61}$$

This variance falls as the loop gain rises. Near the frontier, the coefficient changes little from one vintage to the next because it is highly persistent, even though its long-run level varies widely across the archive. A sequence of models can therefore appear locally stable while drifting over a much wider range at longer horizons.

An auditor who monitors only adjacent coefficient revisions may interpret this local smoothness as evidence of stability. The model instead predicts that proximity to the frontier appears jointly as rising persistence, rising level dispersion, and falling revision variance. A rolling implementation requires the underlying gain and sampling environment to remain approximately stable over the estimation window. Changes in sample size, feature variance, feature reliability, or residual outcome uncertainty must also be controlled before a dispersion trend is attributed to feedback.

The same archive provides identifying variation that is absent from a single deterministic deployment.

Corollary 4 (Identification from model vintages)**.** *Suppose that $n \geq 4$ and that the coefficient process is observed in its subcritical stationary distribution. Then* $\operatorname{Corr}(\lambda_{t+1}, \lambda_t) = k$ *and* $\mathbb{E}[\lambda_t] = \frac{\rho}{1-k}$*. The loop gain and feature reliability are therefore identified by* $k = \operatorname{Corr}(\lambda_{t+1}, \lambda_t)$ *and* $\rho = (1-k)\mathbb{E}[\lambda_t]$*. The stationary variance provides the additional restriction*

$$\operatorname{Var}(\lambda_t) = \frac{\rho(1-\rho)V + \sigma_\varepsilon^2}{V(n-3)(1-k^2)}. \tag{62}$$

Corollary 4 gives an operational counterpart to the identification results in Section 2. A single deterministic deployment does not separate fundamental information from the effect of the allocation. Fresh-sample retraining generates a sequence of slightly different deployments because every estimated slope contains sampling variation. Under Gaussianity, this variation is independent of the previously deployed model and therefore acts as a sequence of exogenous perturbations.

A lender, platform, or regulator that retains archived model coefficients can use the persistence of those coefficients to estimate the feedback gain. Their average level then identifies the reliability of the underlying feature, while their variance supplies an overidentifying restriction involving the residual outcome uncertainty. Sampling error, which is ordinarily treated as a nuisance, becomes the variation that reveals the economic loop.

This identification argument is parametric. It relies on the linear dependence of the population regression coefficient on the deployed slope. Exact recovery of ρ from the stationary mean additionally uses the Gaussian implication $\mathbb{E}[U_t] = 0$. Under the broader conditions of Proposition 4, the archive continues to identify k, while the intercept identifies $\rho + b_n$. The distinction reinforces the broader division between robust persistence-based diagnostics and distribution-specific level and tail restrictions.

4.3 Estimation and testing from model archives

The identification results above translate directly into estimators and specification tests. Suppose an auditor observes $T+1$ consecutive coefficient vintages $\lambda_1, \ldots, \lambda_{T+1}$ from the stationary subcritical regime. Let $(\widehat{\rho}_T, \widehat{k}_T)$ denote the intercept and slope from the time-series regression

$$\lambda_{t+1} = a + b\lambda_t + e_{t+1}, \quad t = 1, \ldots, T. \tag{63}$$

Under the exact autoregressive law in Theorem 5, the population coefficients in this regression are $a = \rho$ and $b = k$. The model archive can therefore be used to estimate the reliability of the feature and the strength of the allocation loop without observing the latent fundamental.

The same environment also permits an exact finite-sample test of whether feedback is present. The test uses the cross-sectional design sums stored with each model vintage and exploits the independence created by retaining alternate observations from the archive. A randomized holdout provides a complementary estimator based on contemporaneous rather than intertemporal variation.

Proposition 5 (Inference from model archives and holdouts)**.** *Suppose the coefficient process is in the stationary regime of Theorem 6(i), with $k \in [0, 1)$ and $\sigma_u^2 > 0$. The following statements hold.*

(i) Suppose that $n \geq 4$. Then $(\widehat{\rho}_T, \widehat{k}_T) \to (\rho, k)$ almost surely and $\sqrt{T}\left(\widehat{k}_T - k\right) \xrightarrow{d} N(0, 1 - k^2)$. The asymptotic variance of the slope estimator is free of nuisance parameters. A feasible asymptotic standard error is therefore $\mathrm{se}(\widehat{k}_T) = \sqrt{\frac{1-\widehat{k}_T^2}{T}}$.

(ii) Suppose that $n \geq 2$ and that the archive also records $S_t = \sum_{i=1}^{n} (\widetilde{x}_{i,t} - \overline{x}_t)^2$, the design sum of the cross-section used to estimate λ_{t+1}. Retain only the observations with odd indices $t \in \{1, 3, 5, \ldots\}$ and let $m = \lceil T/2 \rceil$ be their number. Estimate equation (63) by weighted least squares using weight S_t on observation t. Under $H_0 \colon k = 0$, the conventional t-statistic for the slope has the exact distribution

$$t_{\widehat{k}} \sim t_{m-2} \tag{64}$$

for every $m \geq 3$ and for all admissible values of the remaining parameters.

(iii) Consider a single cross-section of size n in which the deployed slope is λ and a fraction $h \in (0, 1)$ of the units is randomly assigned to a holdout group. The holdout group receives the baseline allocation, while the remaining units receive the score-dependent allocation. If the treated and holdout slopes are estimated separately by OLS with intercepts, then

$$\sqrt{n} \left(\widehat{\beta}_{\text{treated}} - \widehat{\beta}_{\text{holdout}} - k\lambda \right) \xrightarrow{d} N \left(0, \frac{\sigma_u^2}{V} \left[\frac{1}{1-h} + \frac{1}{h} \right] \right). \tag{65}$$

Part (i) shows that an auditor can estimate the loop gain and attach a standard error using the archived coefficient series alone. The nuisance-free variance follows from the stationary autoregressive structure. Since $\mathrm{Var}(\lambda_t) = \frac{\mathrm{Var}(U_t)}{1-k^2}$, the innovation variance that enters the numerator of the usual autoregressive standard error cancels against the stationary variance of the regressor. The precision of $\widehat{k}_T$ therefore depends only on the persistence of the archive and the number of observed vintages.

In a lending application, this estimator uses the history of deployed credit-score coefficients rather than individual borrower outcomes. A strong feedback loop causes a sampling-induced increase in one vintage to persist into later vintages because the larger coefficient changes financing and thereby changes the outcomes used in subsequent estimation. The autocorrelation of the archived coefficients records this persistence and identifies the loop gain.

Part (ii) provides an exact answer to the narrower question of whether feedback is present. Under the null $k = 0$, the vintage law becomes $\lambda_{t+1} = \rho + U_t$. Successive coefficient levels are then independent, although adjacent observations in the regression share a coefficient. The variable λ_{t+1} is the dependent variable in observation t and becomes the regressor in observation $t + 1$. Conditioning on all design sums in the full archive would consequently condition on variables that embed the errors of neighboring observations and would not yield a classical fixed-design regression.

Retaining alternate periods removes this overlap. For an odd observation t, the regressor λ_t depends on innovations dated no later than $t - 1$, while the regression error is U_t. The errors associated with the retained observations, $U_1, U_3, U_5, \ldots$, are independent of their regressors and independent of one another. Conditional on the corresponding design sums, they are Gaussian with variances proportional to S_t^{-1}. Weighting observation t by S_t therefore produces a homoskedastic normal linear model, and the usual studentized slope statistic has the exact distribution in equation (64).

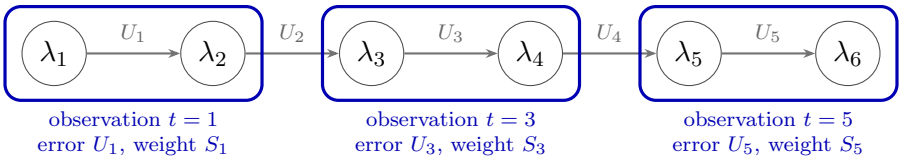


Figure 5: The exact archive test. Vintage slopes form a chain in which each innovation U_t carries slope λ_t into slope λ_{t+1}, so on the full archive every slope serves as regressand of one observation and regressor of the next. Using only the odd observations, boxed, keeps the error set $\{U_1, U_3, U_5, \dots\}$ disjoint from the innovations embedded in the regressors, and the weighted regression with the period-t design sum S_t as weight is then a classical normal linear model with an exactly Student-t slope statistic under the null of no feedback.

The exact procedure uses approximately half of the archive in exchange for finite-sample validity. The estimator in part (i) uses every vintage and measures the magnitude of feedback asymptotically. The test in part (ii) uses alternate vintages and tests the point null of no feedback exactly. The two procedures therefore serve complementary purposes.

Part (iii) measures the same loop gain using variation created by experimental design. The treated slope equals $\rho + k\lambda$, while the holdout slope equals ρ. Their difference therefore identifies $k\lambda$. A larger holdout share improves the precision of the baseline slope but reduces the number of units exposed to the deployed allocation. The asymptotic variance in equation (65) is minimized at $h = \frac{1}{2}$, when statistical precision is the only design objective. Economic costs associated with withholding the deployed allocation may instead favor a smaller holdout group.

The archive and holdout estimators use different sources of variation. The archive relies on sampling fluctuations accumulated over T deployment periods, and its asymptotic precision is governed by $\frac{1-k^2}{T}$. The holdout creates contemporaneous experimental variation within one cross-section, and its precision is of order $1/n$. The archive requires no change to the allocation rule, but it depends on the stability and parametric structure of the retraining environment. The holdout consumes experimental capacity in the current cross-section, while identifying the causal contrast without relying on a long stationary history.

Remark 1 (An overidentifying restriction)**.** *The innovation variance links the autoregressive representation to the structural primitives through* $\sigma_u^2 = \rho(1-\rho)V + \sigma_\varepsilon^2$*. Define the archive-implied residual variance* $\widehat{\sigma}_u^2 \coloneqq (n-3)\,\widehat{V}\,\widehat{\mathrm{Var}}(U_t)$*, where* $\widehat{\mathrm{Var}}(U_t)$ *is estimated from the residuals of the archive autoregression and the scaling converts the estimated innovation variance to the structural scale through* $\mathrm{Var}(U_t) = \sigma_u^2/\big(V(n-3)\big)$*. Since* V *is identified from the cross-sectional distribution of the feature, estimates obtained from the archive must satisfy* $\widehat{\sigma}_u^2 \geq \widehat{\rho}\big(1-\widehat{\rho}\big)\widehat{V}$*. The structural model also restricts the persistence parameter to the unit interval, so a formal specification test considers the joint restrictions* $0 \leq \widehat{\rho} \leq 1$ *and the variance inequality, because the inequality can become mechanically uninformative when the unrestricted estimate of* ρ *lies outside its structural parameter space. The test requires no external estimate of outcome noise. If an independent estimate of* σ_ε^2 *is available, the model imposes the equality restriction* $\widehat{\sigma}_u^2 = \widehat{\rho}\big(1-\widehat{\rho}\big)\widehat{V} + \widehat{\sigma}_\varepsilon^2$*.*

The restriction in Remark 1 tests more than the autoregressive fit of the coefficient archive. A time series may satisfy the reduced-form law in equation (53) while failing the variance decomposition implied by the latent-fundamental model. The overidentifying restriction therefore provides a way to distinguish persistence generated by the allocation loop from an arbitrary autoregressive specification for model coefficients.

4.4 Pooled retraining with decreasing gain

Fresh-sample retraining assigns full weight to the most recent cross-section and discards all earlier observations. Its effective learning gain is therefore constant, which allows sampling disturbances to remain permanently visible in the sequence of deployed coefficients. Many institutions instead re-estimate their models using the complete history of available data. The influence of each new cross-section then declines as the archive grows, and retraining becomes a decreasing-gain learning process.

Suppose that, at the end of period t, the modeler estimates the slope using all cross-sections observed through that date. Period fixed effects remove changes in the cross-sectional intercept, so the pooled estimator is

$$\lambda_{t+1} = \frac{\sum_{s=1}^{t}\sum_{i=1}^{n}\left(\widetilde{x}_{i,s} - \overline{x}_s\right)\left(y_{i,s} - \overline{y}_s\right)}{\sum_{s=1}^{t} S_s}, \qquad S_s = \sum_{i=1}^{n}\left(\widetilde{x}_{i,s} - \overline{x}_s\right)^2. \tag{66}$$

Each observation remains in the estimation sample after it is collected, even though its outcome was generated by the coefficient deployed at that date. The historical sample therefore records a sequence of policy regimes rather than repeated observations from a fixed data-generating process.

Proposition 6 (Decreasing-gain retraining)**.** *Suppose the modeler deploys the pooled OLS coefficient in equation (66) after every period.*

(i) If $k < 1$, then $\lambda_t \to \lambda^*$ almost surely from every nonanticipative initial condition. No auxiliary projection device or exogenous boundedness restriction is required.

(ii) If $k > 1$, then $|\lambda_t| \to \infty$ almost surely from every nonanticipative initial condition, including the repelling rest point $\widetilde{\lambda} = \frac{\rho}{1-k}$. The divergence is polynomial and satisfies

$$|\lambda_t| \asymp t^{\,k-1} \qquad \text{almost surely.} \tag{67}$$

In particular, the probability of convergence to the repelling rest point is zero.

Pooling therefore slows instability without eliminating it. Fresh-sample retraining magnifies deviations at the geometric rate k^t, while pooled retraining magnifies them at the polynomial rate

t^{k-1}. Historical averaging moderates the speed because each new deployment receives a declining weight. It cannot restore stability because the historical observations were themselves generated under coefficients that had already begun to move away from the rest point. Appendix A develops the stochastic-approximation representation behind this result, its connection to decreasing-gain learning and E-stability, and the intermediate cases of rolling windows and escape dynamics.

The amplification mechanism is not confined to linear estimation. If the population retraining operator is a contraction with modulus $\overline{\kappa}_{\text{abs}} < 1$ and every estimated update deviates from the population update by at most $\overline{e}$ uniformly on the feature space, the long-run uniform distance between the deployed predictor and the stable predictor is bounded by $\overline{e}/(1 - \overline{\kappa}_{\text{abs}})$. Appendix A states the formal result, which extends the amplification logic of the exact linear-Gaussian analysis to nonparametric and machine-learning estimation methods.

5 Dynamic Retraining

The static economy isolates the feedback created by a single deployment and a single round of retraining. Persistence and durability add a second channel. A score assigned today changes current investment, the resulting capital survives into future periods, and persistent fundamentals make that inherited capital informative about future outcomes. Retraining therefore responds both to the current score and to the stock of past allocations. The stability problem becomes a random linear system rather than a scalar recurrence.

This section develops the dynamic economy in two steps. The first subsection makes the timing explicit, derives the effective feedback gain directly from the population economy, and shows that an exact two-state representation reproduces the population stability frontier in the large-panel limit. The second subsection lifts the finite-panel system to a higher-dimensional linear recurrence whose random coefficients and additive forcing terms depend only on primitive shocks. This representation yields a contraction theorem based on the top Lyapunov exponent and an expansion result for coefficient products, and it makes precise what is proved, what is conjectured, and what is numerical about the location of the finite-panel frontier.

5.1 The effective gain and the population frontier

Let the centered fundamental follow the stationary process $\widetilde{\theta}_{i,t} = \phi\widetilde{\theta}_{i,t-1} + w_{i,t}$, $\phi \in [0, 1)$, where $w_{i,t} \sim N(0, (1 - \phi^2)\sigma_\theta^2)$. The observed feature is $\widetilde{x}_{i,t} = \widetilde{\theta}_{i,t} + \eta_{i,t}$, $\eta_{i,t} \sim N(0, \sigma_\eta^2)$, and the outcome shock satisfies $\varepsilon_{i,t} \sim N(0, \sigma_\varepsilon^2)$. The shocks are mutually independent across units and dates, and the fundamental process is initialized at its stationary distribution. As in the static economy, define $V = \sigma_\theta^2 + \sigma_\eta^2$ and $\rho = \frac{\sigma_\theta^2}{V}$.

The modeler deploys a slope λ_t on the cross-sectionally demeaned feature $\widetilde{x}_{i,t}^c$. The induced

investment becomes durable capital according to

$$K_{i,t} = \delta K_{i,t-1} + \gamma \lambda_t \widetilde{x}^c_{i,t}, \quad \delta \in [0, 1), \tag{68}$$

and the outcome is

$$y_{i,t} = \theta_{i,t} + \beta K_{i,t} + \varepsilon_{i,t}. \tag{69}$$

The coefficient $k = \beta\gamma$ is the contemporaneous loop gain. Capital durability and fundamental persistence enlarge this gain because an allocation made today remains in future outcomes and remains correlated with future features.

The within-period timing determines which objects are predetermined when period-t shocks arrive and which information the next estimate uses. At the beginning of period t, the inherited capital stock $K_{i,t-1}$ and the previously estimated coefficient λ_t are given. The fundamental $\theta_{i,t}$ and the observed feature $x_{i,t}$ are realized. The model assigns the score $\lambda_t \widetilde{x}^c_{i,t}$. The score induces current investment and determines $K_{i,t}$ through equation (68). The outcome $y_{i,t}$ is realized through equation (69). The modeler runs the period-t cross-sectional regression and obtains λ_{t+1}. Figure 6 displays this sequence within one period, together with the two channels through which the period-t deployment crosses the period boundary.

This timing carries the economic content of the dynamic feedback. A score assigned today changes the borrower's current investment. Capital survives, so the current score affects future outcomes even after the model has been updated. Persistent fundamentals keep the inherited capital correlated with future features, which leads subsequent regressions to attribute part of the legacy allocation to the new score. Conditional-mean statements below are relative to the filtration $\mathcal{F}_t = \sigma\big(\{\theta_{i,s}, \eta_{i,s}, \varepsilon_{i,s}\}_{i \le n,\ s \le t}\big)$ augmented by the initial capital cross-section, so $K_{i,t-1}$ and λ_t are $\mathcal{F}_{t-1}$-measurable.

The strength of the dynamic loop can be derived directly in the population economy. Suppose the cross-section is infinite and a fixed slope λ is deployed in every period, so that capital accumulates according to

$$K_{i,t} = \gamma\lambda \sum_{j=0}^{\infty} \delta^j \widetilde{x}^c_{i,t-j}. \tag{70}$$

The stationary feature autocovariances satisfy $\mathrm{Cov}(\widetilde{x}_{i,t}, \widetilde{x}_{i,t-j}) = \phi^j \rho V$ for $j \ge 1$, so the covariance between the current feature and accumulated capital is

$$\frac{\mathrm{Cov}\big(\widetilde{x}_{i,t}, K_{i,t}\big)}{V} = \gamma\lambda \left(1 + \rho \sum_{j=1}^{\infty} (\delta\phi)^j\right) = \gamma\lambda\psi, \quad \psi = \frac{1 - \delta\phi(1-\rho)}{1 - \delta\phi} = 1 + \frac{\delta\phi\rho}{1 - \delta\phi}. \tag{71}$$

The population regression of the outcome on the demeaned feature therefore returns

$$\lambda' = \rho + k\psi\lambda = \rho + \kappa\lambda, \quad \kappa = k\psi. \tag{72}$$

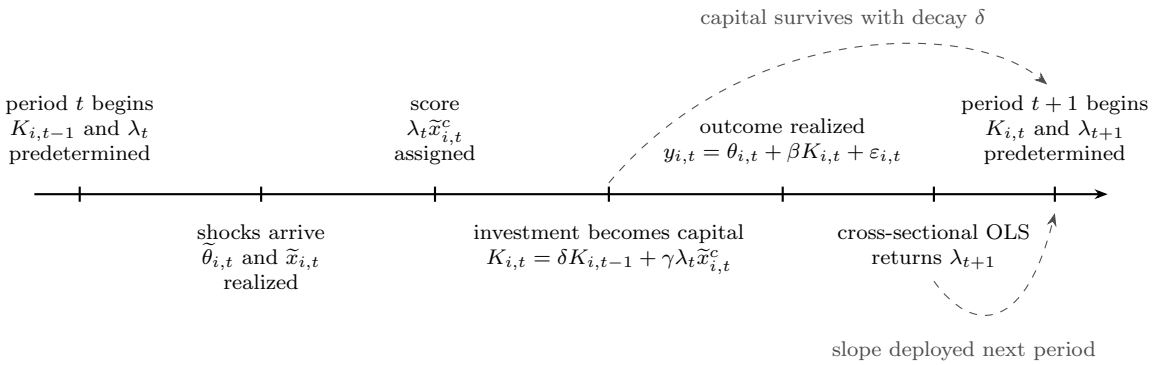


Figure 6: The within-period timing of dynamic retraining. Inherited capital and the previously estimated slope are predetermined when the period-t shocks arrive, the assigned score converts current features into new capital, and the realized outcome feeds the cross-sectional regression that produces the next slope. The dashed arrows mark the two channels through which the current deployment enters the next period, durable capital carried forward with decay δ and the reestimated coefficient carried forward by deployment.

The effective gain κ exceeds the contemporaneous gain k whenever capital is durable, fundamentals are persistent, and the feature contains information about the fundamental. The additional term $\delta\phi\rho/(1-\delta\phi)$ sums the future rounds through which today's score changes capital, capital affects future outcomes, and persistent fundamentals keep the inherited capital correlated with future features. A lender that expands credit to a highly scored borrower affects both current repayment and the sequence of future repayments generated by the resulting capital stock.

Corollary 5 (Destabilization through durability and persistence)**.** *The population retraining map $\lambda \mapsto \rho + \kappa\lambda$ is stable if and only if $\kappa < 1$, equivalently $k < k_c$ with*

$$k_c = \frac{1}{\psi} = \frac{1-\delta\phi}{1-\delta\phi(1-\rho)}. \tag{73}$$

The critical gain is weakly decreasing in durability, persistence, and feature reliability. It is strictly decreasing in δ when $\rho\phi > 0$, in ϕ when $\rho\delta > 0$, and in ρ when $\delta\phi > 0$. Each of the limiting cases $\delta = 0$, $\phi = 0$, and $\rho = 0$ gives $\kappa = k$ and $k_c = 1$, and $k_c \to 0$ as $\delta\phi \uparrow 1$ for fixed $\rho > 0$.

Durability, persistence, and informativeness all lower the amount of contemporaneous feedback required to destabilize retraining. A system that would be stable in a one-period economy can become unstable because earlier allocations remain economically active. As $\delta\phi$ approaches one with $\rho > 0$, arbitrarily weak contemporaneous feedback destabilizes the system because its effects accumulate over a sufficiently long horizon. Dynamic amplification requires all three ingredients, since removing durability, persistence, or informativeness returns the economy to the static frontier. Figure 7 displays the critical gain.

We now return to the finite panel. The modeler re-estimates the slope by cross-sectional OLS in

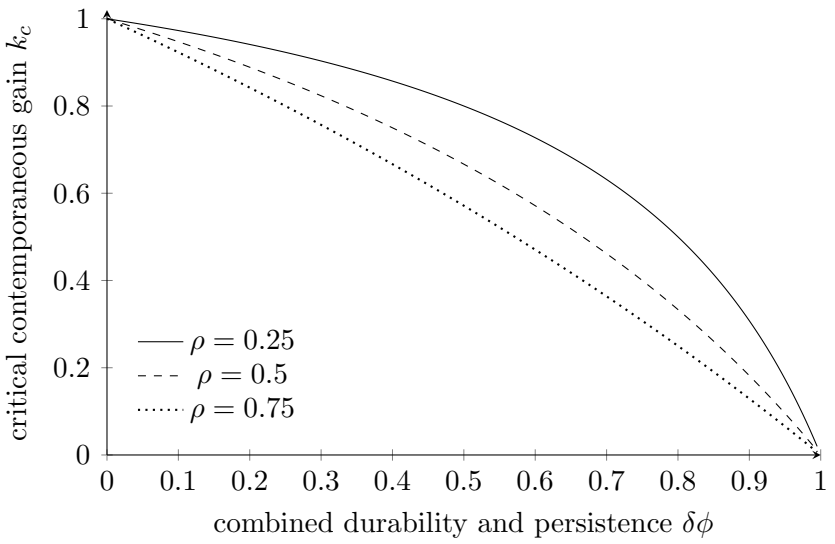


Figure 7: The critical contemporaneous gain $k_c = (1-\delta\phi)/(1-\delta\phi(1-\rho))$ below which the population economy is stable, as a function of the combined durability and persistence $\delta\phi$ for three levels of feature reliability. The static economy corresponds to $\delta\phi = 0$, where $k_c = 1$. The critical gain falls toward zero as $\delta\phi$ approaches one, so arbitrarily weak contemporaneous feedback destabilizes retraining when allocations are durable and fundamentals are persistent.

every period using the same panel of $n \geq 2$ units. Let

$$s_t^{xx} = \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{x}_{i,t}^{c}\right)^2, \quad s_t^{x\theta} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_{i,t}^{c}\widetilde{\theta}_{i,t}^{c}, \tag{74}$$

and define

$$c_t^{xK} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_{i,t}^{c}K_{i,t-1}^{c}, \quad p_t = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\theta}_{i,t}^{c}K_{i,t}^{c}. \tag{75}$$

The variable p_t measures the covariance between current fundamentals and the capital inherited from current and earlier allocations. In a credit application, it records whether borrowers with stronger latent repayment capacity also carry more capital because earlier scores gave them greater access to financing.

The identity

$$c_t^{xK} = \phi p_{t-1} + \zeta_t + \iota_t, \tag{76}$$

where

$$\zeta_t = \frac{1}{n}\sum_{i=1}^{n}w_{i,t}^{c}K_{i,t-1}^{c}, \quad \iota_t = \frac{1}{n}\sum_{i=1}^{n}\eta_{i,t}^{c}K_{i,t-1}^{c}, \tag{77}$$

gives the exact finite-panel recursion

$$\lambda_{t+1} = k\lambda_t + \frac{\beta\delta\phi}{s_t^{xx}} p_{t-1} + \nu_t, \tag{78}$$

$$p_t = \delta\phi p_{t-1} + \gamma s_t^{\theta x}\lambda_t + \delta\zeta_t, \tag{79}$$

where $s_t^{\theta x} = s_t^{x\theta}$ and

$$\nu_t = \frac{s_t^{x\theta} + s_t^{x\varepsilon} + \beta\delta(\zeta_t + \iota_t)}{s_t^{xx}}, \quad s_t^{x\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{x}_{i,t}^{c}\varepsilon_{i,t}^{c}. \tag{80}$$

Conditional on the past, ζ_t and ι_t have mean zero. Their conditional variances depend on the inherited dispersion of capital, which makes the noise in the two-state system state-dependent.

Let

$$Z_t = \begin{pmatrix} \lambda_t \\ p_{t-1} \end{pmatrix}. \tag{81}$$

Equations (78) and (79) can be written as

$$Z_{t+1} = A_t\, Z_t + \xi_t, \tag{82}$$

where

$$A_t = \begin{pmatrix} k & \beta\delta\phi/s_t^{xx} \\ \gamma s_t^{\theta x} & \delta\phi \end{pmatrix}, \quad \xi_t = \begin{pmatrix} \nu_t \\ \delta\zeta_t \end{pmatrix}. \tag{83}$$

This representation separates the contemporaneous score channel from the inherited-capital channel. The upper-left entry is the static loop gain, while the off-diagonal entries transmit past allocations into current retraining and current scores into future capital.

Proposition 7 (Large-panel dynamics and the stability frontier)**.** *As $n \to \infty$, the random coefficient matrix in equation (83) converges in probability to*

$$\overline{A} = \begin{pmatrix} k & \beta\delta\phi/V \\ \gamma\sigma_\theta^2 & \delta\phi \end{pmatrix}. \tag{84}$$

Its characteristic polynomial is

$$\det(\mu I - \overline{A}) = \mu^2 - (k + \delta\phi)\mu + k\delta\phi(1 - \rho). \tag{85}$$

Writing $d = \delta\phi$, the eigenvalues

$$\mu_{1,2} = \frac{k + d \pm \sqrt{(k-d)^2 + 4kd\rho}}{2} \tag{86}$$

are real and nonnegative. When $kd\rho > 0$ they satisfy $\mu_1 > \max\{k, d\}$ and $\mu_2 < \min\{k, d\}$, and

$\mu_2 > 0$ exactly when $kd(1-\rho) > 0$, since $\mu_1\mu_2 = kd(1-\rho)$. The spectral radius of $\overline{A}$ is below one if and only if $\kappa < 1$, with $\kappa = k\psi$ the effective gain in equation (72). In that case, the large-panel deterministic system has the stationary state

$$\overline{\lambda}^* = \frac{\rho}{1-\kappa}, \quad \overline{p}^* = \frac{\gamma\sigma_\theta^2}{1-\delta\phi}\,\overline{\lambda}^*. \tag{87}$$

The eigenvalue ordering gives the two-state system a two-mode reading. The dominant mode combines current feedback with inherited capital and exceeds both the contemporaneous gain and the capital-covariance decay rate, while the smaller mode describes the faster adjustment of the remaining linear combination of the two states. The stationary covariance state $\overline{p}^*$ gives economic content to the second coordinate. Borrowers with stronger fundamentals carry more inherited capital in proportion to the deployed slope, and this covariance is what makes the stationary slope rise steeply as the frontier approaches.

The frontier follows by evaluating equation (85) at one. In particular,

$$\det(I - \overline{A}) = (1-k)(1-\delta\phi) - k\delta\phi\rho = (1-\delta\phi)(1-\kappa). \tag{88}$$

Since the two eigenvalues are nonnegative and $\delta\phi < 1$, this determinant is positive exactly when both eigenvalues lie below one. The population stability frontier is therefore the spectral-radius condition for the large-panel limit of the exact finite-panel dynamics.

Remark 2 (The large-panel matrix is not a mean)**.** *The matrix $\overline{A}$ is the probability limit of A_t and generally differs from $\mathbb{E}[A_t]$, because $\mathbb{E}[1/s_t^{xx}] \neq 1/V$ and the entries of A_t are correlated. The stationary mean of the finite-panel random-coefficient system likewise differs from the fixed point of the deterministic recursion. The simulated gaps documented in Section 5.2.1 reflect this distinction. They are positive and shrink at the rate $1/n$ in the calibrations studied, with a coefficient that grows steeply near the frontier, although an analytical expansion that delivers the sign and the constant remains open.*

Convergence of the coefficient matrix alone does not show that the finite-panel path tracks the deterministic recursion. The following proposition closes this gap over fixed horizons.

Proposition 8 (Finite-horizon large-panel convergence)**.** *Fix a horizon T and suppose the initial pair $Z_0 = (\lambda_0, p_{-1})'$ converges in probability to a deterministic limit z_0 as $n \to \infty$, while the initial capital cross-section has uniformly bounded fourth moments. Then*

$$\max_{0\le t\le T} \|Z_t - z_t\| \xrightarrow{p} 0, \quad z_{t+1} = \overline{A} z_t + \begin{pmatrix} \rho \\ 0 \end{pmatrix}. \tag{89}$$

The proposition upgrades Proposition 7 from convergence of coefficients to convergence of the finite-panel path itself, which justifies reading $\overline{A}$ as the large-panel dynamics of the economy rather

than merely as the probability limit of a coefficient matrix. It does not control long-run behavior, because the horizon is fixed before the panel grows. The stability of the finite-panel process over unbounded horizons is governed by the Lyapunov theory of the next subsection.

The two eigenvalues describe how the effects of a model vintage decay in the large-panel deterministic system. Whether they also govern the autocorrelation of the stationary finite-panel process is not established. The forcing that drives the finite-panel fluctuations is serially correlated because the fundamental follows an AR(1), so the long-lag autocorrelation of the limiting fluctuation system could be governed by primitive persistence as well as by the dominant eigenvalue. In the calibrations reported below the two candidates are numerically close, with $\phi = 0.875$ and $\mu_1 = 0.881$ at $k = 0.30$, so a fitted decay of 0.890 cannot sharply distinguish the coupled mode from primitive persistence. The autocorrelation evidence in Section 5.2.1 is therefore presented as a calibration-specific numerical finding rather than as an identification result, and the finite-panel analysis below replaces the spectrum of the limiting matrix with the growth rate of random matrix products as the relevant stability object.

5.2 Finite-panel dynamics and Lyapunov stability

The two-state system is useful for interpretation, although its disturbances inherit state-dependent variances from lagged capital. A higher-dimensional representation removes this complication. The resulting recurrence is linear, its additive term is exogenous, and both the coefficient matrix and the additive term are generated entirely by primitive shocks.

Let

$$K^c_{t-1} = \begin{pmatrix} K^c_{1,t-1} & \cdots & K^c_{n,t-1} \end{pmatrix}'. \tag{90}$$

Cross-sectional demeaning implies $\mathbf{1}'K^c_{t-1} = 0$. Fix an $n \times (n-1)$ matrix Q with orthonormal columns spanning $\mathbf{1}^\perp$, and define the reduced state

$$\widetilde{W}_t = \begin{pmatrix} \lambda_t \\ Q'K^c_{t-1} \end{pmatrix} \in \mathbb{R}^n. \tag{91}$$

Lemma 2 (Exact linear representation)**.** *The reduced state satisfies the exact stochastic recurrence*

$$\widetilde{W}_{t+1} = \widetilde{B}_t \widetilde{W}_t + \widetilde{E}_t, \tag{92}$$

where

$$\widetilde{B}_t = \begin{pmatrix} k & \dfrac{\beta\delta}{S_t}(\widetilde{x}^c_t)'Q \\ \gamma Q'\widetilde{x}^c_t & \delta I_{n-1} \end{pmatrix}, \quad S_t = \sum_{i=1}^n (\widetilde{x}^c_{i,t})^2, \tag{93}$$

and

$$\widetilde{E}_t = \begin{pmatrix} g_t & 0 & \cdots & 0 \end{pmatrix}', \quad g_t = \frac{\sum_{i=1}^n \widetilde{x}^c_{i,t}\left(\widetilde{\theta}^c_{i,t} + \varepsilon^c_{i,t}\right)}{\sum_{i=1}^n (\widetilde{x}^c_{i,t})^2}. \tag{94}$$

The additive term has the marginal representation

$$g_t - \rho = \frac{\sigma_r}{\sqrt{V(n-1)}}\, T_{n-1}, \quad \sigma_r^2 = \sigma_\theta^2(1-\rho) + \sigma_\varepsilon^2, \tag{95}$$

where T_{n-1} is a Student t variable with $n-1$ degrees of freedom. The mean of g_t is therefore undefined for $n = 2$, equals ρ with infinite variance for $n = 3$, and for $n \geq 4$

$$\mathrm{Var}(g_t) = \frac{\sigma_r^2}{V(n-3)}. \tag{96}$$

The logarithmic moment required below exists for every $n \geq 2$. The sequences $\{\widetilde{B}_t\}$ and $\{\widetilde{E}_t\}$ are measurable functions of the primitive panel process and form a stationary, geometrically β-mixing sequence. They do not depend functionally on deployed scores or inherited capital, although they are dependent on one another contemporaneously and serially dependent through the persistent fundamentals.

Lemma 2 replaces state-dependent regression noise with an exogenous additive forcing term. We call $\widetilde{E}_t$ a forcing term rather than an innovation because its first component has mean ρ for $n \geq 3$ and it is statistically dependent on the contemporaneous coefficient matrix. Serial dependence remains because the fundamental is persistent, so the current coefficient matrix is correlated with earlier matrices and with the capital inherited from earlier periods. This dependence is generated by the primitive AR(1) process rather than by the endogenous state. The distinction permits the use of stationary random-recurrence results.

Lemma 3 (Spectral structure of the coefficient matrix)**.** *For every realization of the feature vector with $S_t > 0$, the matrix $\widetilde{B}_t$ has characteristic polynomial*

$$\det\big(zI - \widetilde{B}_t\big) = z\,(z - k - \delta)\,(z - \delta)^{n-2}. \tag{97}$$

Its eigenvalues are 0, $k + \delta$, and δ with multiplicity $n - 2$, in every period and for every realization.

Three consequences organize the analysis that follows. First, the $n - 2$ eigenvalues at δ belong to capital directions orthogonal to the current feature. These directions decay at the survival rate δ, their number grows with the panel, and they explain why the conjectured large-panel exponent below involves δ rather than $\delta\phi$. The population system tracks the covariance between capital and fundamentals, which decays at $\delta\phi$, while the finite-panel state also carries capital dispersion that is temporarily orthogonal to the current score and decays at δ. Second, the zero eigenvalue makes $\widetilde{B}_t$ singular in every period. Standard multivariate Kesten-type tail theorems and strong irreducibility results assume invertibility or conditions that may fail for this matrix class, so heavy-tail conclusions cannot be imported directly. Third, the coupled eigenvalue $k + \delta$ attaches to the two-dimensional subspace spanned by the score direction and the feature-aligned capital direction. This subspace

rotates with the realized features, which links the exact eigenstructure to the two-state system of Section 5.1.

The geometry of the state space is also economically meaningful. Capital is measured relative to the cross-sectional mean, so the admissible capital vector lies in the hyperplane $\mathbf{1}^{\perp}$. The common-capital direction is an eigenvector of the full system with eigenvalue δ, but neither the economic state nor the forcing term has a component in that direction. The Lyapunov analysis must therefore be conducted on the reduced state in equation (91). Figure C.2 in Appendix C illustrates this geometry.

Define the top Lyapunov exponent on the reduced state space by

$$\gamma_{L,n} = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\Big[\log \Big\|\widetilde{B}_t \cdots \widetilde{B}_1\Big\|\Big], \tag{98}$$

where $\|\cdot\|$ is the Euclidean operator norm. The subscript records that the exponent depends on the panel size through the dimension and the law of the coefficient matrices. For fixed n, every matrix norm yields the same exponent, and fixing the operator norm provides a common convention for comparisons across panel sizes. The limit exists under the logarithmic moment conditions

$$\mathbb{E}\log^{+}\big\|\widetilde{B}_1\big\| < \infty, \quad \mathbb{E}\log^{+}\big\|\widetilde{E}_1\big\| < \infty, \tag{99}$$

which are verified in the appendix. Kingman's subadditive ergodic theorem implies that the normalized logarithm of the matrix-product norm converges almost surely to the same constant.

The Lyapunov exponent governs propagation of the full reduced state, while a supervisor observes the deployed coefficient rather than the Euclidean norm of the cross-sectional capital vector. The deployed slope may forget initial conditions more quickly than the full state when the dominant Lyapunov direction lies primarily in the capital coordinates, as in the weak-loop regime where the conjectured exponent below equals $\log\delta$ and is generated by capital directions whose effect on the estimated slope may be small. Establishing equality between the full-state and slope-specific propagation rates requires an irreducibility condition and remains open, because every coefficient matrix is singular and contains a large capital subspace. For the sign of the exponent the distinction is harmless, since $\delta < 1$ implies that the full-state exponent can become positive only through the coupled score-capital mode. For statements about persistence and observability it matters, and the numerical autocorrelation evidence below does not rely on the equality.

Theorem 7 (Contraction under a negative Lyapunov exponent)**.** *Suppose the logarithmic moment conditions* (99) *hold and* $\gamma_{L,n} < 0$*. Then equation* (92) *has a unique causal stationary and ergodic solution, given by the almost surely convergent series*

$$\widetilde{W}_t^{*} = \widetilde{E}_{t-1} + \sum_{j=1}^{\infty} \widetilde{B}_{t-1} \cdots \widetilde{B}_{t-j}\, \widetilde{E}_{t-j-1}. \tag{100}$$

Any two solutions driven by the same primitive sequence converge to one another almost surely from

arbitrary initial conditions. The vintage-slope process initialized in the invariant law is therefore stationary and ergodic.

The theorem holds under the actual dependence structure of the model, because the random-recurrence results it invokes require only a stationary ergodic sequence of pairs $\{(\widetilde{B}_t, \widetilde{E}_t)\}$ together with the logarithmic moments in equation (99). Serial dependence of the coefficients and contemporaneous dependence between the coefficient matrix and the forcing term are permitted. The series solution has a direct economic reading. The current model vintage aggregates all past regression disturbances after each has been propagated through the subsequent sequence of allocation rules. A negative exponent implies that the effect of every initial condition eventually disappears, even though individual coefficient matrices may be locally expansive. In a lending application, temporary overreaction in one model vintage is reversed by later vintages, and the system remains stable when these expansions are dominated by contraction over long histories.

Proposition 9 (Expansion under a positive Lyapunov exponent)**.** *Suppose the logarithmic moment conditions* (99) *hold and* $\gamma_{L,n} > 0$*. Then the reduced coefficient products satisfy*

$$\lim_{t\to\infty} \frac{1}{t} \log \left\| \widetilde{B}_t \cdots \widetilde{B}_1 \right\| = \gamma_{L,n} > 0 \quad \textit{almost surely.} \tag{101}$$

The proposition establishes an exponentially expanding propagation mechanism rather than divergence of the state itself. Divergence of the affine state additionally requires that the forcing term repeatedly excite the expanding directions, and a primitive nondegeneracy condition guaranteeing this excitation remains open. The simulations below exhibit the required excitation and produce divergence above the frontier. Together, Theorem 7 and Proposition 9 replace the scalar contraction condition of the static economy with the sign of the growth rate of random matrix products. They fall short of a complete finite-panel phase transition, because the positive side concerns coefficient products rather than the state and because the critical case is untreated.

Remark 3 (Moments of the stationary solution)**.** *Four statements are progressively stronger and coincide in the static economy but not here. These are existence of a stationary solution, finiteness of its first moment, finiteness of its second moment, and existence of an autocorrelation function. Theorem 7 delivers only the first. The forcing term has a finite mean only for* $n \geq 3$ *and a finite variance only for* $n \geq 4$ *by Lemma 2, and even then the stationary solution need not inherit finite moments when contraction is slow. Statements about stationary means, variances, and autocorrelations therefore require conditions beyond a negative Lyapunov exponent, and the corresponding comparisons in Section 5.2.1 are numerical findings.*

Remark 4 (The critical case)**.** *The dynamic theory is silent at* $\gamma_{L,n} = 0$*. This contrasts with the static economy, where Theorem 6 shows that the coefficient follows a random walk with drift at the frontier. Whether the critical dynamic system is null recurrent, how its fluctuations scale with time and with the panel size, and whether the running exponent estimate fluctuates around zero at the*

$t^{-1/2}$ *rate are open questions. A near-critical numerical exercise at the population frontier, the value of* k *solving* $\kappa = 1$*, is reported in Section 5.2.1.*

The finite-panel results are stated in terms of the Lyapunov exponent, while the population frontier is stated in terms of the effective gain κ. The next result provides a one-sided bridge between these objects. It shows that every population economy strictly below the frontier eventually generates a contracting finite-panel recurrence as the cross-section grows.

Theorem 8 (Stable side of the large-panel Lyapunov bridge)**.** *Fix the primitives of the dynamic economy and let* $\gamma_{L,n}$ *denote the top Lyapunov exponent of the exact reduced coefficient sequence for a panel of size* n*. Let* $\mu_1(\kappa)$ *be the dominant eigenvalue of the large-panel matrix* $\overline{A}$ *in Proposition 7. Then*

$$\limsup_{n\to\infty} \gamma_{L,n} \leq \log \max \{\mu_1(\kappa), \delta\} . \tag{102}$$

Consequently, if $\kappa < 1$*, there is an* $N < \infty$ *such that* $\gamma_{L,n} < 0$ *for every* $n \geq N$*. Dynamic retraining therefore has the unique causal stationary and ergodic solution characterized in Theorem 7 for every sufficiently large panel whenever the population economy lies strictly below the stability frontier.*

The theorem turns the stable side of the population frontier into a finite-panel stability result. Durability creates many capital directions that decay at the rate δ, while the coupled score-capital direction evolves asymptotically at the rate $\mu_1(\kappa)$. Since both rates are below one when $\kappa < 1$, sufficiently large panels have a negative Lyapunov exponent. The conclusion is uniform over all panel sizes beyond a finite threshold, although the theorem does not provide a matching lower bound on the exponent.

The result leaves two parts of the large-panel problem unresolved. It does not establish that $\gamma_{L,n}$ converges to the upper bound in (102), and it does not show that a population economy above the frontier eventually has a positive finite-panel exponent. These remaining claims are summarized in the following conjecture.

Conjecture 1 (The large-panel Lyapunov frontier)**.** *As* $n \to \infty$ *with the primitives fixed,*

$$\gamma_{L,n} \longrightarrow \log \max \{\mu_1(\kappa), \delta\} . \tag{103}$$

The capital mode governs when $\mu_1(\kappa) < \delta$*, while the coupled score-capital mode governs when* $\mu_1(\kappa) > \delta$*. Since* $\delta < 1$ *and* $\mu_1(\kappa) > 1$ *if and only if* $\kappa > 1$*, the limiting Lyapunov exponent changes sign at the population frontier* $\kappa = 1$*.*

Theorem 8 establishes the upper bound in Conjecture 1 by combining fixed-block subadditivity with a large-panel limit. The remaining difficulty is the matching lower bound. Such a bound requires showing that random products retain sufficient mass along the coupled score-capital direction, while every one-period coefficient matrix is singular by Lemma 3. The numerical evidence below measures this remaining gap and studies the positive side of the conjecture.

5.2.1 Numerical evidence

The numerical analysis checks the exact static formulas, evaluates the conjectured large-panel limit in the dynamic economy, and documents behavior in regions where the theory remains incomplete. The replication code is provided in the supplementary files, and Appendix C collects the point estimates, the diagnostic detail, and the sensitivity analyses summarized in this subsection.

Direct computation of the Lyapunov exponent uses power iteration with renormalization on the reduced state space, with each iterate confined to the centered-capital subspace so that the inadmissible common-capital direction cannot affect the calculation. The baseline calibration sets $n = 200$, $\rho = 0.5$, $\delta = 0.8$, and $\phi = 0.875$, so that $\delta\phi = 0.7$ and the population critical gain is $k_c = 0.4615$. Figure 8 traces the estimated exponent across a dense grid of loop gains for panel sizes from twenty-five to one thousand. The estimates follow the capital mode $\log \delta$ in the weak-loop region and the coupled mode $\log \mu_1$ in the stronger-loop region, the gaps to the conjectured limit shrink approximately linearly in $1/n$, and the interpolated zero crossings fall from 0.475 at $n = 25$ to 0.462 at $n = 1000$, so the estimated finite-panel zero crossing approaches the population frontier as the panel grows. Calibration variations that hold $\delta\phi$ fixed while moving durability and persistence separately confirm the max structure of the conjectured limit, and the point estimates, their batch-mean standard errors, and the variation detail are collected in Appendix C.

At the population frontier $\kappa = 1$, the running exponent estimate at $n = 200$ stays within 10^{-3} of zero and slightly negative, so the simulations describe a long near-critical transient at the population frontier rather than the unresolved case $\gamma_{L,n} = 0$ of Remark 4, which remains uncharacterized. The affine state does not settle on the simulated horizon, and the near-critical growth measurements, together with the caveat that cross-path moments can be dominated by a few exceptional paths, are reported in Appendix C.

Divergence above the frontier is operationally sharp. At $k = 0.50$ every replication crosses successive order-of-magnitude thresholds, with a spacing between decades that matches the estimated exponent, while at $k = 0.44$ no replication crosses even the lowest threshold; the hitting-time distributions and burn-in diagnostics are in Appendix C.

The simulated stationary distributions become more heavy-tailed near the frontier. Kesten-type results for recurrences with independent coefficients often generate power-law tails under expansion, irreducibility, and regularity conditions (Kesten, 1973), while the coefficient sequence here is serially dependent and every coefficient matrix is singular by Lemma 3, so the tail behavior is treated as a numerical finding. A counterfactual decomposition on the exact linear lift attributes most of the near-frontier dispersion to random coefficient propagation and the associated high-dimensional capital dynamics rather than to the marginal tail of the additive regression disturbance; extremes arrive in long clusters, and the Hill estimates are sensitive to the tail fraction and the simulation seed, so they should not be read as measurements of an established power law. The robust conclusion is that the model produces much larger and more persistent coefficient excursions near the frontier even though the process remains stationary, and the decomposition and the full tail diagnostics are

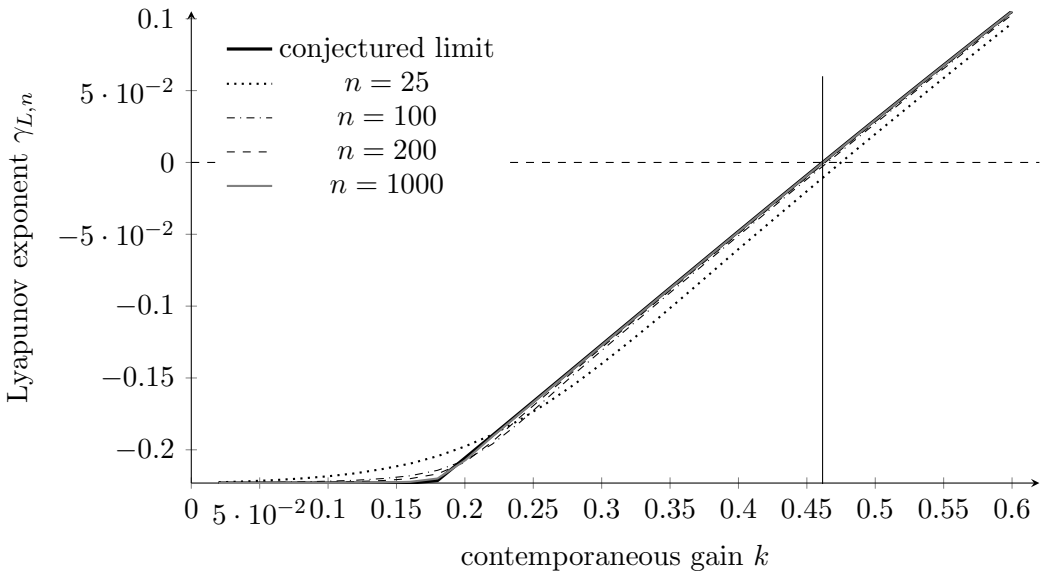


Figure 8: The estimated Lyapunov exponent $\gamma_{L,n}$ as a function of the contemporaneous gain k for four panel sizes ($\rho = 0.5$, $\delta = 0.8$, $\phi = 0.875$), together with the conjectured large-panel limit $\log \max\{\mu_1(\kappa), \delta\}$, the zero level, and the population critical gain k_c. The kink in the limit marks the transition from the capital mode to the coupled score-capital mode. The finite-panel curves approach the limit as the panel grows, and their zero crossings approach k_c.

in Appendix C.

Figure C.3 in Appendix C compares the dynamic vintage autocorrelation function with the dominant eigenvalue of the large-panel system, Figure C.4 in the same appendix reports the survival functions of centered vintage slopes on logarithmic axes, and Table C.1 there compares simulated moments with their theoretical counterparts.

The numerical results support a unified reading of the static and dynamic theories across panel sizes from twenty-five to one thousand. The population frontier organizes the finite-panel behavior, the estimated finite-panel zero crossing approaches the population critical gain as the panel grows, and durability and persistence determine the speed at which model vintages forget past allocations. Close to the frontier, average behavior remains well described by the population fixed point up to a positive $1/n$ difference, while the stochastic system becomes highly persistent and produces occasional large departures from that benchmark. What remains open is the matching lower bound and hence the full large-panel limit, together with the exact critical case and a tail theory for singular and serially dependent coefficient matrices.

6 Concluding Remarks

Predictive models increasingly govern the allocations that shape the outcomes on which they are evaluated and retrained. This feedback is especially consequential for AI systems that operate at scale and are updated repeatedly using data generated after deployment. A credit score affects

financing and repayment, a recommendation changes visibility and demand, and a clinical score directs treatment and health outcomes. Deployment data therefore reflect both the information contained in a prediction and the economic response generated by the associated allocation. We call the resulting gap between measured fit and information self-fulfilling accuracy.

The structure theorem characterizes this mechanism in a general environment. At a stable point, the deployed score is a strictly increasing transformation of the posterior mean of the latent fundamental. Feedback changes the scale of the score while preserving rankings, posterior information, and rank dependence. The stable predictor remains calibrated against the outcomes generated by its deployment. Under monotone amplification, stronger feedback raises explained variance and deployment R^2 because it expands the fitted component of the outcome while leaving the residual distribution unchanged. A predictive or AI system can therefore become more accurate in deployment data without acquiring additional information about fundamentals.

Deployment data alone do not reveal the source of this accuracy. Under a deterministic allocation rule, units with the same observed features receive the same allocation, so the data contain no counterfactual variation at common covariate values. The conditional mean consequently admits observationally equivalent decompositions into fundamental and feedback components. Randomized allocation holdouts restore overlap and identify both components up to a common normalization. Threshold rules identify discrete allocation effects under continuity and functional-form restrictions. Archived model vintages provide another source of variation when changes in the deployed rule are exogenous to fundamentals.

The linear results yield an additional empirical diagnostic. Under Gaussian or elliptical features, score-mediated feedback rescales a stable coefficient vector without rotating it. Changes in coefficient magnitude are therefore consistent with changes in feedback strength, while systematic rotation points toward feature-specific interventions, endogenous features, changes in the information structure, or departures from ellipticity. This restriction can also be applied when a complex AI system feeds into a stable low-dimensional scoring or allocation layer whose coefficients are retained for auditing. When allocation responses depend on private information, the stable score continues to preserve the information in a one-dimensional posterior index under monotone posterior ordering, although the residual invariance underlying the fit result requires the stronger additive environment.

Finite-sample retraining makes feedback visible in the time series of model vintages. In the static linear-Gaussian economy, fresh-sample retraining generates an exact first-order autoregression whose persistence parameter is the loop gain. Below the stability frontier, archived coefficients fluctuate ergodically around the stable value, and their moments identify the feedback gain and feature reliability. At the frontier, the coefficient follows a random walk with drift, while above it the coefficient diverges almost surely. Pooled retraining reduces the effective learning gain and changes geometric divergence into polynomial divergence, but it leaves the location of the frontier unchanged. Automated AI pipelines that preserve model versions and deployment records therefore contain information about both the strength and the stability of the feedback they generate.

Persistence and durability strengthen the mechanism. In the dynamic economy, past allocations affect current outcomes through accumulated capital, and persistent fundamentals transmit this effect across retraining periods. These forces combine with the contemporaneous loop gain to form an effective gain that determines population stability. The finite-panel system is an exact random-coefficient recurrence governed by a top Lyapunov exponent. A negative exponent yields a unique causal stationary and ergodic solution, while a positive exponent produces exponential expansion of the homogeneous coefficient products. Every population economy strictly below the frontier has a negative finite-panel exponent for all sufficiently large panels. The exact large-panel exponent and the affine-state implications of a positive exponent remain open, while the numerical evidence places the estimated finite-panel transition near the population frontier.

The results suggest practical principles for evaluating deployed AI and predictive systems. Calibration and deployment R^2 describe performance within the allocation regime created by the model, but they do not measure information about an unaffected fundamental. Evaluation therefore benefits from variation in the allocation through randomized holdouts, institutional discontinuities, or archived model changes. The archive also records the stability of retraining. Under a stable sampling environment, rising persistence and dispersion in coefficient levels can signal a strengthening allocation loop even while average deployment metrics remain stable or improve.

The central implication is that predictive performance is an equilibrium object. A model's measured success depends on what it knows and on what its predictions cause. As AI systems become more deeply integrated into economic decisions, distinguishing these components becomes essential for interpreting accuracy, identifying feedback, and determining whether repeated retraining converges to a stable prediction rule.

References

Acemoglu, Daron, and Pascual Restrepo. 2018. Artificial intelligence, automation, and work. In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 197–236. Chicago: University of Chicago Press.

Athey, Susan, and Guido W. Imbens. 2017. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives* 31(2): 3–32.

Ball, Ian. 2025. Scoring strategic agents. *American Economic Journal: Microeconomics* 17(1).

Bougerol, Philippe, and Nico Picard. 1992. Strict stationarity of generalized autoregressive processes. *Annals of Probability* 20(4): 1714–1730.

Brandt, Andreas. 1986. The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability* 18(1): 211–220.

Brown, Gavin, Shlomi Hod, and Iden Kalemaj. 2022. Performative prediction in a stateful world. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, PMLR 151: 6045–6061.

Chen, Hui, Winston Wei Dou, and Leonid Kogan. 2024. Measuring "dark matter" in asset pricing models. *Journal of Finance* 79(2): 843–902.

Cho, In-Koo, Noah Williams, and Thomas J. Sargent. 2002. Escaping Nash inflation. *Review of Economic Studies* 69(1): 1–40.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1): 187–199.

D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2021. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics* 221(2): 644–654.

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch. 1997. *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.

Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81: 160–171.

Esponda, Ignacio, and Demian Pouzo. 2016. Berk–Nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica* 84(3): 1093–1130.

Evans, George W., and Seppo Honkapohja. 2001. *Learning and Expectations in Macroeconomics*. Princeton, NJ: Princeton University Press.

Farboodi, Maryam, and Laura Veldkamp. 2023. Data and markets. *Annual Review of Economics* 15: 23–40.

Frankel, Alex, and Navin Kartik. 2022. Improving information from manipulable data. *Journal of the European Economic Association* 20(1): 79–115.

Furstenberg, Harry, and Harry Kesten. 1960. Products of random matrices. *Annals of Mathematical Statistics* 31(2): 457–469.

Hahn, Jinyong. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2): 315–332.

Haidar-Wehbe, Sami, Samuel R. Emerson, Louis J. M. Aslett, and James Liley. 2025. Optimal sizing of a holdout set for safe predictive model updating. *Annals of Applied Statistics* 19(2).

Hardt, Moritz, and Celestine Mendler-Dünner. 2023. Performative prediction: Past and future. arXiv:2310.16608.

Hennessy, Christopher A., and Charles A. E. Goodhart. 2023. Goodhart's law and machine learning: A structural perspective. *International Economic Review* 64(3): 1075–1086.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Kesten, Harry. 1973. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica* 131: 207–248.

Kim, Michael P., and Juan C. Perdomo. 2023. Making decisions under outcome performativity. *Proceedings of the 14th Innovations in Theoretical Computer Science Conference*.

Kingman, J. F. C. 1973. Subadditive ergodic theory. *Annals of Probability* 1(6): 883–899.

Liley, James, Samuel R. Emerson, Bilal A. Mateen, Catalina A. Vallejos, Louis J. M. Aslett, and Sebastian J. Vollmer. 2021. Model updating after interventions paradoxically introduces bias. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, PMLR 130: 3916–3924.

Lucas, Robert E., Jr. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 1: 19–46.

Marcet, Albert, and Thomas J. Sargent. 1989. Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory* 48(2): 337–368.

Mendler-Dünner, Celestine, Frances Ding, and Yixin Wang. 2022. Anticipating performativity by predicting from predictions. *Advances in Neural Information Processing Systems* 35: 31171–31185.

Meyn, Sean P., and Richard L. Tweedie. 1993. *Markov Chains and Stochastic Stability*. London: Springer-Verlag.

Mokkadem, Abdelkader. 1988. Mixing properties of ARMA processes. *Stochastic Processes and their Applications* 29(2): 309–315.

Munro, Evan. 2025. Treatment allocation with strategic agents. *Management Science* 71(1).

Pemantle, Robin. 1990. Nonconvergence to unstable points in urn models and stochastic approximations. *Annals of Probability* 18(2): 698–712.

Perdomo, Juan C. 2025. Revisiting the predictability of performative, social events. *Proceedings of the 42nd International Conference on Machine Learning*.

Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. *Proceedings of the 37th International Conference on Machine Learning*: 7599–7609.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Sargent, Thomas J. 1999. *The Conquest of American Inflation*. Princeton, NJ: Princeton University Press.

Scarone, Bruno, and Ricardo Baeza-Yates. 2026. The illusion of improvement: Reject inference strategies in credit scoring. arXiv:2606.18479.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631: 755–759.

Taylor, John B. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.

Appendix

Appendix A Extensions and Additional Results

This appendix collects results that extend the main analysis. It illustrates threshold identification and its limits, interprets the stable linear coefficient as a Berk–Nash equilibrium, allows the feedback response to interact with private information, develops the stochastic-approximation structure of pooled retraining, and bounds the amplification of estimation error beyond linear scores.

A.1 Threshold identification in detail

The restriction in Proposition 2 that the feedback response changes only through the discrete jump is economically substantive. Suppose a credit score determines approval at a cutoff but also affects the interest rate smoothly on either side of the cutoff. The discontinuity identifies the discrete effect of approval under the continuity condition, while the smooth financing effect remains observationally equivalent to a smooth component of borrower quality. A threshold design therefore recovers the entire feedback response only when the allocation technology is known to contain no additional score-dependent component.

The continuity assumption also concerns the fundamental conditional on the scalar score rather than the fundamental conditional on the full feature vector. This distinction matters when the score compresses multidimensional information. Consider a feature vector $x = (s, z)$ with $m(x) = s$ and $u(x) = z$. The function $u(x)$ is continuous in the full feature vector, but the conditional distribution of z given s may change discontinuously at the cutoff. In that case, $r(s) = \mathbb{E}[z \mid s]$ jumps at c, and the observed discontinuity in outcomes combines the allocation effect with a change in the composition of units near the threshold.

Figure A.1 depicts the identifying configuration. The score-conditional mean of the fundamental passes continuously through the cutoff, so the discontinuity in the observed conditional mean can be attributed to the allocation. When the latent conditional mean also jumps, the same observed contrast admits both a treatment interpretation and a selection interpretation.

A.2 A Berk–Nash interpretation of the stable coefficient

Theorem 4 remains informative when the linear model is misspecified. If f is nonlinear, the conditional expectation of the outcome need not be linear in $\widetilde{x}$, and the stable coefficient is the best linear projection generated by its own deployment. The following lemma connects this fixed point to equilibrium learning under misspecification.

Lemma A1 (Berk–Nash representation)**.** *Suppose the modeler uses the subjective class*

$$\mathcal{M} = \left\{ y \mid \widetilde{x} \sim N\big(\lambda'\widetilde{x}, \sigma^2\big) \;\middle|\; \lambda \in \mathbb{R}^d,\ \sigma^2 > 0 \right\}. \tag{A.1}$$

Assume that the least-squares residual is nondegenerate under every deployment, so that

$$\inf_{b \in \mathbb{R}^d} \mathbb{E}\Big[\big(y^{(\lambda)} - b'\widetilde{x}\big)^2\Big] > 0. \tag{A.2}$$

Let the modeler's action be the deployment of the coefficient associated with the subjective belief.

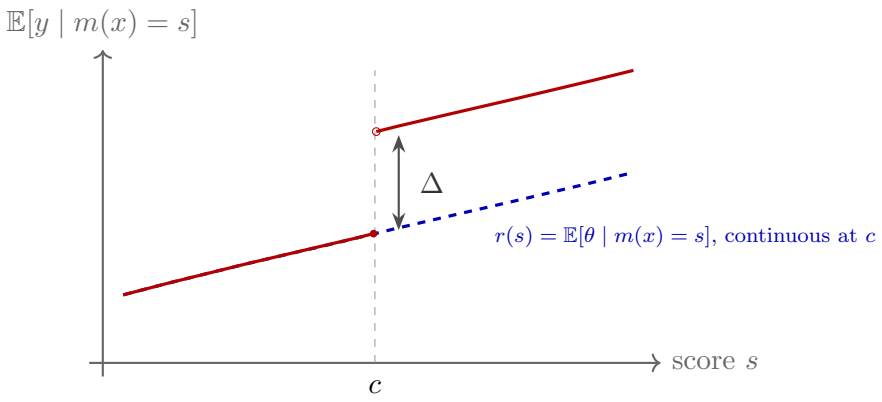


Figure A.1: Identification at a threshold allocation. The observed conditional mean of the outcome given the score, in solid, jumps by the treatment effect Δ at the known cutoff c, and the jump measures Δ alone precisely when the score-conditional fundamental mean r, in dashed, passes continuously through the cutoff. A jump in r, which continuity of the primitives on the feature space does not rule out, would contaminate the contrast with a selection discontinuity.

A coefficient λ is performatively stable if and only if there exists $\sigma^2 > 0$ such that (λ, σ^2) is a Berk–Nash equilibrium in the sense of Esponda and Pouzo (2016). The equilibrium belief maximizes the expected Gaussian log score, or equivalently minimizes the Gaussian quasi-likelihood risk, under the distribution generated by its own deployment. If the true conditional law of y given $\widetilde{x}$ is absolutely continuous and has finite Kullback–Leibler divergence from some member of $\mathcal{M}$, the same belief also minimizes that divergence.

Lemma A1 places performative stability within the theory of equilibrium learning under misspecification. The subjective model may fail to describe the nonlinear conditional mean, yet the fixed point remains the coefficient that is optimal within the linear class under the distribution created by deploying that coefficient. The latent fundamental adds an economic distinction that is absent from a generic misspecified-learning problem because it separates the information carried by the features from the outcome variation created by the allocation.

A.3 Private responses

The additive outcome equation assumes that the action induced by deployment depends on a unit only through its score. This restriction is natural when an institution mechanically assigns an allocation from the score, although many economic responses also depend on private characteristics. A borrower with high productivity may use the same credit line more effectively than a borrower with low productivity, a patient with severe underlying illness may respond differently to the same treatment, and a seller with private inventory information may convert platform visibility into sales differently from other sellers.

Let the induced action be $a = g(s, \theta)$ and define the reduced-form response

$$\psi(s, \theta) \coloneqq b\big(g(s, \theta)\big). \tag{A.3}$$

The outcome becomes

$$y = \theta + \psi\big(m(x), \theta\big) + \varepsilon. \tag{A.4}$$

The score now affects the outcome through an interaction with the latent fundamental. A one-

dimensional posterior mean is generally insufficient to order units once this interaction is allowed, so the analysis imposes a monotone one-dimensional index for the entire posterior distribution.

Assumption A1. *The following conditions hold.*

(i) There is a measurable index $w\colon \mathcal{X} \to \mathbb{R}$ and a family of distributions $\{Q_w\}_{w\in\mathbb{R}}$ such that $\theta \mid x \sim Q_{w(x)}$. The family is strictly increasing in the sense of first-order stochastic dominance, the mapping $w \mapsto Q_w$ is a measurable probability kernel, and

$$\mathbb{E}_{Q_w}[|\theta + \psi(t,\theta)|] < \infty \text{ for every } (t,w). \tag{A.5}$$

(ii) For every w, the map $t \longmapsto \mathbb{E}_{Q_w}[\psi(t,\theta)]$ is differentiable, and its derivative is bounded above by $\kappa^+ < 1$ uniformly in w.

(iii) For every t, the map $\theta \longmapsto \theta + \psi(t,\theta)$ is strictly increasing.

The index $w(x)$ orders posterior distributions rather than posterior means alone. In a lending application, a larger value of w may shift the entire distribution of borrower productivity upward. Part (iii) then ensures that a borrower with a better latent fundamental has a better outcome under any fixed score, even after accounting for differences in how borrowers use credit. Part (ii) preserves the same one-sided feedback restriction used in the additive model.

Remark A1 (Nesting of the two environments)**.** *The additive environment is nested in this formulation. If $\psi(t,\theta) = f(t)$, then the feedback response does not depend on the private fundamental. The posterior index can be chosen as $w = u$, and the fixed-point equation depends on the posterior only through its mean. The stronger posterior-ordering assumptions become necessary only when the response interacts with θ.*

Define

$$H(t,w) \coloneqq \mathbb{E}_{Q_w}[\theta + \psi(t,\theta)]\,. \tag{A.6}$$

For a unit with belief index w, a stable score solves $t = H(t,w)$. The feedback bound makes the fixed-point equation invertible in the score, while stochastic monotonicity and part (iii) make the solution increasing in the belief index.

Proposition A1 (Stable scores with private responses)**.** *Under Assumption A1, there is a unique performatively stable score up to almost-sure equality. It has the representation*

$$m^*(x) = \Phi\big(w(x)\big), \tag{A.7}$$

where Φ is strictly increasing. Consequently, Corollary 1 holds with $u(x)$ replaced by $w(x)$. If $m^*(x)$ and the associated stable outcome are integrable, Corollary 2 also applies. The stable score therefore carries exactly the information contained in the posterior index for every feedback technology satisfying Assumption A1.

Remark A2 (An integrable envelope)**.** *A sufficient condition for the integrability required in Proposition A1 is an integrable envelope $A(w(x))$ together with a constant $c < 1$ such that $\mathbb{E}_{Q_w}\big|\theta + \psi(t,\theta)\big| \le A(w) + c\,|t|$ for all (t,w). The fixed-point equation $m^*(x) = H\big(m^*(x), w(x)\big)$ then gives $|m^*(x)| \le A(w(x)) + c\,|m^*(x)|$, so that $|m^*(x)| \le A(w(x))/(1-c)$, and integrability of $A(w(x))$ delivers integrability of the stable score and, through equation (A.4), of the stable outcome.*

Proposition A1 extends the structure theorem from posterior means to posterior distributions ordered by a scalar index. Two borrowers with different values of w may differ both in expected productivity and in the way productivity affects their use of financing. The stable score still preserves their ordering because the expected outcome under a fixed score rises with the posterior index, and the fixed-point transformation remains increasing. Feedback changes the cardinal scale of the score while leaving the information generated by $w(x)$ unchanged.

The extension changes the fit and identification conclusions. The residual identity in Lemma 1 relied on the feedback term being measurable with respect to x. Under equation (A.4), the centered response $\psi(t,\theta) - \mathbb{E}_{Q_w}[\psi(t,\theta)]$ remains in the conditional residual. Its variance may change with the deployed score, so the monotone fit comparison in Theorem 2 requires additional restrictions on this component. The information-invariance and calibration results survive because they follow from monotonicity and stability rather than from residual invariance.

A randomized holdout also has a more limited structural interpretation in this environment. The treated-minus-holdout contrast identifies the conditional causal effect of the allocation given x, which equals an average of the private response over $Q_{w(x)}$. It does not separately identify the action rule g, the production response b, or their interaction with the latent fundamental without further restrictions. The design therefore recovers the reduced-form feedback relevant for policy while leaving the deeper technology decomposition unidentified.

A.4 Pooled retraining as stochastic approximation

The pooled estimator admits an exact stochastic-approximation representation. Define $G_t = \sum_{s=1}^{t} S_s$ and $\alpha_t = \frac{S_t}{G_t}$. Using the fresh-cross-section slope decomposition from Theorem 5, the pooled update can be written as

$$\lambda_{t+1} = \lambda_t + \alpha_t \left[\rho + (k-1)\lambda_t + U_t\right]. \tag{A.8}$$

The gain α_t is asymptotically of order t^{-1}. Consequently, sampling error receives progressively less weight, while the systematic component of the retraining map continues to accumulate over the entire history.

The associated mean ordinary differential equation is $\frac{d\lambda}{d\tau} = \rho + (k-1)\lambda$ with the unique rest point $\lambda^* = \frac{\rho}{1-k}$. The rest point is locally stable exactly when $k < 1$. The E-stability condition of the learning model therefore coincides with the population stability frontier.

Part (i) of Proposition 6 gives a global version of the E-stability principle developed in Marcet and Sargent (1989) and Evans and Honkapohja (2001). The general theory of self-referential learning often relies on projection devices or boundedness assumptions to keep beliefs within a compact region. The affine retraining map and regime-invariant disturbance make those auxiliary restrictions unnecessary here. Below the frontier, the decreasing gain eventually suppresses sampling variation, and the coefficient converges almost surely to the performatively stable value.

A credit-scoring system illustrates the mechanism. Early cross-sections may produce unusually high or low estimates of the sensitivity of repayment to borrower characteristics. As the lending archive expands, each new estimation error receives less weight, while the accumulated data continue to pull the coefficient toward the stable slope. The model eventually settles at a fixed coefficient even though every historical outcome was affected by the lending policy in place when that outcome was produced.

Part (ii) of Proposition 6 strengthens the usual nonconvergence result for an unstable stochastic-approximation equilibrium (Pemantle, 1990). Above the frontier, the deterministic force in equation (A.8) pushes the coefficient away from the rest point. Conditional Gaussianity and linearity rule out permanent cancellation by the sampling disturbances, so the coefficient diverges almost surely at

the exact polynomial rate in equation (67).

Fresh-sample and pooled retraining describe two limiting approaches to model updating. Under fresh-sample retraining, the learning gain remains constant, and the coefficient fluctuates indefinitely around the stable point when $k < 1$. Under pooled retraining, the gain vanishes, and the coefficient converges to that point. Rolling windows and exponential downweighting lie between these cases. They retain a positive effective gain because older observations eventually leave the estimation sample or receive negligible weight, so persistent coefficient fluctuations remain relevant for such systems.

Near the frontier, a positive effective gain interacts with the heavy-tailed estimation disturbances characterized in Theorem 5. This interaction may produce occasional departures from the neighborhood of the stable coefficient of the kind studied in the escape dynamics literature (Sargent, 1999; Cho et al., 2002). A full characterization of these excursions under rolling or exponentially weighted retraining remains open.

A.5 Estimation error beyond linear scores

The exact autoregressive representation relies on a linear score and a Gaussian sampling environment. The underlying amplification mechanism is more general. Whenever the population retraining operator is a contraction, estimation error enters each update as a disturbance to that contraction and is propagated through subsequent rounds of deployment.

Consider the general environment of Theorem 1, and suppose that

$$\overline{\kappa}_{\text{abs}} = \sup_{t \in \mathbb{R}} |f'(t)| < 1. \tag{A.9}$$

Let $\widehat{m}_t$ denote the predictor deployed in period t. The modeler may estimate the conditional mean using a nonparametric regression, a regularized machine-learning procedure, or another prediction method. Assume that the estimated update satisfies

$$\sup_{x \in \mathcal{X}} |\widehat{m}_{t+1}(x) - (T\widehat{m}_t)(x)| \leq \overline{e}_t,$$

where $\overline{e}_t$ is a uniform bound on the estimation error in period t.

Proposition A2 (Amplification of estimation error)**.** *Suppose that equation* (A.9) *holds and that*

$$\sup_{x \in \mathcal{X}} |\widehat{m}_0(x) - m^*(x)| < \infty. \tag{A.10}$$

Then, for every $t \geq 1$*,*

$$\sup_{x \in \mathcal{X}} |\widehat{m}_t(x) - m^*(x)| \leq \overline{\kappa}_{\text{abs}}^t \sup_{x \in \mathcal{X}} |\widehat{m}_0(x) - m^*(x)| + \sum_{j=0}^{t-1} \overline{\kappa}_{\text{abs}}^{t-1-j} \overline{e}_j. \tag{A.11}$$

If $\overline{e}_t \leq \overline{e}$ *for every* t*, then*

$$\limsup_{t \to \infty} \sup_{x \in \mathcal{X}} |\widehat{m}_t(x) - m^*(x)| \leq \frac{\overline{e}}{1 - \overline{\kappa}_{\text{abs}}}. \tag{A.12}$$

The first term in equation (A.11) measures the effect of the initial deployment. It disappears geometrically because the population operator is contractive. The second term accumulates the

errors made in successive estimation rounds. An error generated many periods earlier receives a smaller weight, but the rate at which it is forgotten declines as the feedback gain approaches one.

The economic interpretation parallels the exact linear model. Suppose a lender estimates a flexible repayment function and makes a persistent approximation error for borrowers in a particular region of the feature space. Deployment translates this error into financing decisions, which affect realized repayment and enter the next training sample. When the allocation loop is weak, later retraining quickly attenuates the original error. Near the frontier, the economic response reinforces the mistaken prediction and allows its effect to persist across many vintages.

The factor $\frac{1}{1-\bar{\kappa}_{\text{abs}}}$ is the worst-case amplification of a uniformly bounded sequence of estimation errors. It reflects the same geometric mechanism that expands the stable score in the population analysis. Stronger feedback makes deployment outcomes more responsive to the prediction and simultaneously makes the retraining path more sensitive to imperfections in estimated models.

The bound concerns worst-case sensitivity rather than the realized distribution of prediction errors. It does not by itself imply that the variance, tail risk, or stochastic ordering of estimation errors increases with the feedback gain. Such conclusions require assumptions on the distribution of the estimation disturbances, their dependence across vintages, and their relation to the deployed predictor. The linear-Gaussian analysis supplies this additional structure and produces the exact variance and tail results in Theorem 6. Proposition A2 identifies the part of that mechanism that survives in general prediction classes.

Appendix B Proofs

Proof of Theorem 1. (i) $h'(t) = 1 - f'(t) \geq 1 - \kappa^+ > 0$, so h is strictly increasing. Since $h(t) \geq h(0) + (1-\kappa^+)t$ for $t \geq 0$ and $h(t) \leq h(0) + (1-\kappa^+)t$ for $t \leq 0$, h is onto $\mathbb{R}$. The inverse function theorem gives (9).

(ii) m is stable iff $m(x) = u(x) + f(m(x))$ for a.e. x, which says $h(m(x)) = u(x)$, which says $m(x) = \Phi(u(x))$.

(iii) For each fixed x the iteration $t_{n+1} = u(x) + f(t_n)$ is a contraction on $\mathbb{R}$ with modulus $\bar{\kappa}_{\text{abs}}$ and fixed point $m^*(x)$. □

Proof of Corollary 1. Φ is a continuous strictly increasing bijection of $\mathbb{R}$, hence a Borel isomorphism onto its range with measurable inverse. Therefore $\sigma(\Phi(u(x))) = \sigma(u(x))$, which gives (i). (ii) follows from applying the data-processing inequality in both directions along Φ and Φ^{-1}. (iii) holds because strictly increasing transformations preserve pairwise orderings. □

Proof of Corollary 2. $\mathbb{E}[y \mid x] = m^*(x)$ at a fixed point, so by the tower property $\mathbb{E}[y \mid m^*(x)] = \mathbb{E}\big[\mathbb{E}[y \mid x] \,\big|\, m^*(x)\big] = m^*(x)$. □

Proof of Lemma 1. Immediate from $y - \mathbb{E}[y \mid x] = \theta - u(x) + \varepsilon$, since $f(m(x))$ is x-measurable. The displayed variance decomposition uses $\mathrm{Var}(\theta + \varepsilon \mid x) = \mathrm{Var}(\theta \mid x) + \mathrm{Var}(\varepsilon \mid x) + 2\mathrm{Cov}(\theta, \varepsilon \mid x)$, and the conditional covariance vanishes because $\mathbb{E}[\varepsilon \mid \theta, x] = 0$ implies $\mathbb{E}[\theta\varepsilon \mid x] = \mathbb{E}[\theta\, \mathbb{E}(\varepsilon \mid \theta, x) \mid x] = 0$ while $\mathbb{E}[\varepsilon \mid x] = 0$. □

Proof of Theorem 2. Fix $\tau_2 > \tau_1 \geq 0$ in $[0, \bar{\tau}]$ and let $t_i = \Phi_{\tau_i}(u)$ for a given u, so that $t_i - \tau_i f_0(t_i) = u$.

Step 1. Monotone amplification is exactly the statement that the transformation between regimes has slope at least one. Define $H := \Phi_{\tau_2} \circ \Phi_{\tau_1}^{-1}$, so that $m^*_{\tau_2}(x) = H\big(m^*_{\tau_1}(x)\big)$. By (9),

$$H'(t_1) \;=\; \frac{\Phi'_{\tau_2}(u)}{\Phi'_{\tau_1}(u)} \;=\; \frac{1 - \tau_1 f_0'(t_1)}{1 - \tau_2 f_0'(t_2)},$$

which is at least one iff $\tau_2 f_0'(t_2) \geq \tau_1 f_0'(t_1)$, the monotone amplification condition at u.

Step 2. Slope at least one implies weakly larger variance. Write $X = m_{\tau_1}^*(x)$ and $D \coloneqq H - \mathrm{id}$, which is nondecreasing by Step 1. Then

$$\mathrm{Var}\big(H(X)\big) = \mathrm{Var}(X) + 2\,\mathrm{Cov}\big(X, D(X)\big) + \mathrm{Var}\big(D(X)\big) \;\geq\; \mathrm{Var}(X),$$

because the identity and D are both nondecreasing functions of the same random variable, so $\mathrm{Cov}(X, D(X)) \geq 0$ by Chebyshev's association inequality. Square-integrability of all terms follows from $\mathrm{Var}(u) < \infty$ and the Lipschitz bounds on Φ_τ. Monotonicity of $R^2(\tau)$ follows since $x \mapsto x/(x + \sigma^2_{\mathrm{res}})$ is increasing and σ^2_{res} is constant by Lemma 1.

Step 3. The sufficient conditions. In case (a), f_0' is a nonnegative constant, so $\tau f_0'$ is nondecreasing in τ, and $\Phi_\tau(u) = (u + \mathrm{const})/(1 - \tau k_0)$ gives the explicit variance and the strictness claims, the one for $R^2(\tau)$ using $\sigma^2_{\mathrm{res}} > 0$. In case (b), it suffices that $\tau \mapsto \Phi_\tau(u)$ is nondecreasing, since then convexity gives $f_0'(t_2) \geq f_0'(t_1) \geq 0$ and hence $\tau_2 f_0'(t_2) \geq \tau_1 f_0'(t_1)$. Subtracting the fixed-point equations,

$$t_2 - t_1 \;=\; \tau_2 f_0(t_2) - \tau_1 f_0(t_1) \;\geq\; \tau_1\big(f_0(t_2) - f_0(t_1)\big) \;\geq\; -\tau_1\big(\sup f_0'\big)\,(t_1 - t_2)_+,$$

where the first inequality uses $f_0 \geq 0$ and $\tau_2 \geq \tau_1$. If $t_2 < t_1$ this gives $t_2 - t_1 \geq \tau_1 \sup f_0'\,(t_2 - t_1)$, and dividing by the negative quantity $t_2 - t_1$ yields $1 \leq \tau_1 \sup f_0' < 1$, a contradiction. Hence $t_2 \geq t_1$. □

Proof of Theorem 3. Under S_c the outcome is $\big(\theta - c(m(x))\big) + f(m(x)) + c(m(x)) + \varepsilon = y$, pathwise, and the law of (x, ε) is unchanged, so the joint law of (x, y) is P. Square-integrability of $c(m(x))$ keeps $\mathbb{E}[\theta_c^2] < \infty$, and $\sigma(\theta_c, x) = \sigma(\theta, x)$ preserves $\mathbb{E}[\varepsilon \mid \theta_c, x] = 0$, so μ_c satisfies the setup moments. The conditional mean under S_c is $u_c = u - c \circ m$, giving (i). For the representation claim in part (i), because the support of x is a Borel subset of the Polish space $\mathcal{X}$ and m is Borel measurable and injective on that support, the Lusin–Souslin theorem implies that $m(\mathrm{supp}\, x)$ is Borel and that the inverse of m on its image is Borel measurable, so every measurable function of x can be written as a measurable function of $m(x)$ on the support. (ii) is immediate since both structures generate the same P. For (iii), the orbit of S under the admissible transformations is contained in the identified set, so a functional that varies along the orbit cannot be identified. Non-invariance in general follows from an example. In the linear-Gaussian economy of Section 3.2 with $0 < \rho < 1$ and $m = u$ deployed, the shift $c(m) = m/2$ halves the informational component of the fundamental, so the mutual information between the score and the shifted fundamental strictly falls. That the orbit need not exhaust the identified set is shown by redecompositions of the latent pair. With x uniform on $\{-1, 1\}$, z standard normal and independent, $y = x + z$, and both m and f constant, the structures $\theta_A = x + z$ with $\varepsilon_A = 0$ and $\theta_B = x$ with $\varepsilon_B = z$ satisfy the setup conditions, generate the same law of (x, y), and share the posterior mean $u(x) = x$, yet $I(u; \theta_A) \neq I(u; \theta_B)$, and no shift $c(m(x))$ relates them, since m is constant while $\theta_A - \theta_B = z$ is random. Invariance under the shift family is therefore necessary but not sufficient for identification. □

Proof of Proposition 1. For (i), independence of D from (θ, x, ε) implies that the conditional law of (θ, ε) given (x, D) equals its conditional law given x, so $\mathbb{E}[y \mid x, D = 1] = u(x) + f(s_0)$ on the full support of x. For (ii), the same argument on the treated arm gives $\mathbb{E}[y \mid x, D = 0] = u(x) + f(m(x))$, and differencing removes u. For (iii), part (ii) identifies the function $s \mapsto f(s) - f(s_0)$ on the range of m; where the support of $m(x)$ contains an interval I, every nonempty open subset of I carries positive probability by the definition of the support, so the function is identified on I and its derivative recovers $k(\cdot) = f'(\cdot)$ on the interior, while the deployment average $\bar{k} = \mathbb{E}[f'(m(x))]$ integrates the identified derivative against the law of $m(x)$ and is therefore identified when the interval covers that law's support up to null sets. □

Proof of Proposition 2. For any structure in the restricted class that is consistent with P, the observed conditional mean decomposes as $\mathbb{E}[y \mid m(x) = s] = r(s) + \Delta\,\mathbf{1}\{s \geq c\} + \text{constant}$, with r continuous at c by the latent restriction. The one-sided limits of the observed function at the cutoff therefore differ by exactly Δ, so every consistent structure carries the same jump, which is the displayed limit. With Δ known, subtracting $\Delta\,\mathbf{1}\{m(x) \geq c\}$ from $\mathbb{E}[y \mid x]$ recovers u up to a single additive constant. □

Proof of Theorem 4. By Stein's lemma for multivariate Gaussians, for Lipschitz $\varphi(\tilde{x}) = f(\lambda'\tilde{x})$,

$$\operatorname{Cov}\big(\tilde{x}, f(\lambda'\tilde{x})\big) = \Sigma\,\mathbb{E}[\nabla\varphi] = \mathbb{E}\big[f'(\lambda'\tilde{x})\big]\,\Sigma\lambda.$$

Hence the fixed-point condition $\Sigma\lambda = q + \operatorname{Cov}(\tilde{x}, f(\lambda'\tilde{x}))$ reads $\big(1 - \bar{k}(\lambda)\big)\Sigma\lambda = q$. Since $q \neq 0$, any solution has $\bar{k}(\lambda^*) \neq 1$, and because $\bar{k} \leq \kappa^+ < 1$, in fact $\bar{k}(\lambda^*) < 1$ and (41) holds. In particular λ^* is proportional to v. For existence, restrict to the ray $\lambda = c\,v$ with $c \in \mathbb{R}$, and define $\xi(c) := c\big(1 - K(c)\big)$ with $K(c) := \mathbb{E}[f'(c\,v'\tilde{x})]$. Then $\xi(0) = 0$, and ξ is continuous on $[0, \infty)$ for the following reason. On $c > 0$ the law of $c\,v'\tilde{x}$ is Gaussian with standard deviation $c\,\sigma_v$, where $\sigma_v^2 = v'\Sigma v$, so $K(c) = \int f'(w)\,\varphi_{c\sigma_v}(w)\,dw$ is continuous in c by L^1-continuity of the scaled Gaussian densities, with no continuity of f' required, since f' exists almost everywhere and is bounded. At $c = 0$, boundedness of f' gives $|\xi(c)| \leq c\,(1 + \bar{\kappa}_{\text{abs}}) \to 0$, so ξ is continuous there as well. Finally $\xi(c) \geq c(1 - \kappa^+) \to \infty$ as $c \to \infty$, so $\xi(c^*) = 1$ for some $c^* > 0$, and $\lambda^* = c^* v$ solves the fixed point. The affine case is direct. □

Proof of Lemma A1. For Gaussian likelihoods with free variance, maximizing the expected log score against the true conditional law of y given $\tilde{x}$ under deployment of λ reduces to minimizing $\mathbb{E}[(y^{(\lambda)} - b'\tilde{x})^2]$ over b, whose unique solution is the population OLS coefficient $T\lambda$, with σ^2 absorbing the residual variance. The Berk–Nash condition that the deployed belief be closest in this sense to the data its own action generates is therefore $\lambda = T\lambda$, which is performative stability. □

Proof of Proposition 3. Elliptical distributions have linear conditional expectations, $\mathbb{E}[\tilde{x} \mid \lambda'\tilde{x} = s] = (\Sigma\lambda/\lambda'\Sigma\lambda)\,s$. Hence

$$\operatorname{Cov}\big(\tilde{x}, f(\lambda'\tilde{x})\big) \;=\; \mathbb{E}\big[\mathbb{E}[\tilde{x} \mid \lambda'\tilde{x}]\,f(\lambda'\tilde{x})\big] \;=\; \frac{\Sigma\lambda}{\lambda'\Sigma\lambda}\operatorname{Cov}\big(\lambda'\tilde{x}, f(\lambda'\tilde{x})\big) \;=\; \bar{k}(\lambda)\,\Sigma\lambda,$$

and the fixed-point condition reads $(1 - \bar{k}(\lambda))\Sigma\lambda = q$ as before, giving collinearity with v. For the bound, let (s, s') be an independent pair with the law of $\lambda'\tilde{x}$ and write $\operatorname{Cov}(s, f(s)) = \frac{1}{2}\mathbb{E}[(s - s')(f(s) - f(s'))] \leq \frac{1}{2}\kappa^+\mathbb{E}[(s - s')^2] = \kappa^+\operatorname{Var}(s)$, using $f' \leq \kappa^+$ pointwise. Existence follows by the ray argument of Theorem 4 with $K(c) = \bar{k}(cv)$, which is bounded by the Lipschitz constant of f and continuous by dominated convergence. The Gaussian case is Stein's lemma. □

Proof of Proposition A1. Retraining returns $(Tm)(x) = \mathbb{E}[y \mid x] = H\big(m(x), w(x)\big)$, so a stable score solves $t = H(t, w)$ pointwise in w. By (ii), $t - H(t, w)$ is strictly increasing in t with slope at least $1 - \kappa^+$, and it is onto, so the solution $t = \Phi(w)$ exists and is unique for each w. By (i) and (iii), $w \mapsto H(t, w)$ is strictly increasing for each t, since it is the expectation of a strictly increasing function of θ under a strictly stochastically increasing family. A monotone-comparison argument then gives Φ strictly increasing. If $w_2 > w_1$ with $t_i = \Phi(w_i)$, then $t_2 - t_1 = H(t_2, w_2) - H(t_1, w_1) > H(t_2, w_1) - H(t_1, w_1) \geq -\kappa^+(t_1 - t_2)_+$ by (ii), which forces $t_2 > t_1$ as in the proof of Theorem 2. For measurability, Φ is a strictly increasing real function, hence Borel measurable, and $m^* = \Phi \circ w$ is a composition of measurable maps, while the kernel measurability in Assumption A1(i) makes H jointly measurable, so the pointwise-selected fixed point is measurable as well. □

Proof of Theorem 5. Substituting $y_i - \bar{y} = (\rho + k\lambda_t)(\tilde{x}_i - \bar{x}) + (u_i - \bar{u})$ into the OLS formula,

$$\lambda_{t+1} = \rho + k\lambda_t + \underbrace{\frac{\sum_i (\tilde{x}_i - \bar{x})(u_i - \bar{u})}{\sum_i (\tilde{x}_i - \bar{x})^2}}_{U_t}.$$

Because (u_i) is independent of $(\tilde{x}_i)$ and both are i.i.d. Gaussian, conditional on the design $U_t \mid (\tilde{x}_1, \ldots, \tilde{x}_n) \sim N\big(0, \sigma_u^2 / \sum_i (\tilde{x}_i - \bar{x})^2\big)$, and $\sum_i (\tilde{x}_i - \bar{x})^2 \sim V\chi^2_{n-1}$ independently of the conditional standard normal. The ratio of a standard normal to $\sqrt{\chi^2_{n-1}/(n-1)}$ is T_{n-1}, giving (54). The distribution of U_t does not involve λ_t because neither the design nor the residuals do. The regime-invariance of u_i is exactly what removes the deployed slope from the sampling error. Independence across t holds because cross-sections are freshly drawn. The variance uses $\mathbb{E}[1/\chi^2_\nu] = 1/(\nu - 2)$. □

Proof of Proposition 4. The decomposition $\lambda_{t+1} = \rho + k\lambda_t + U_t$ and the regime invariance of the residual survive without Gaussianity, because both rest on the covariance identity $\mathrm{Cov}(\tilde{x}, u) = 0$ and the linearity of the outcome equation, neither of which uses normality. The innovation U_t is a fixed functional of the period-t draws alone, which are i.i.d. across periods and independent of the past, and its law does not depend on λ_t because neither the design nor the residuals do. Hence $\{U_t\}$ is i.i.d., and $\mathbb{E}|U_t| \le \big(\mathbb{E}[(\sum_i \tilde{x}_i^c u_i^c)^2]\big)^{1/2} \big(\mathbb{E}[S^{-2}]\big)^{1/2} < \infty$ by Cauchy–Schwarz and the moment conditions, so b_n is well defined. Setting $\tilde{U}_t = U_t - b_n$ gives the displayed recursion. Part (i) is the standard MA(∞) construction for a linear autoregression with i.i.d. mean-zero innovations. Part (ii) follows by the argument in the proof of Proposition 5(i), with the Lindeberg condition supplied by the $2 + \varsigma$ moment. For part (iii), unrolling the recursion as in Theorem 6(iii) expresses $\lambda_t - \tilde{\lambda}$ as k^t times a sum converging almost surely to $(\lambda_0 - \tilde{\lambda}) + M_\infty$. Write the innovation law as the mixture $p\,Q + (1-p)\,R$ with $p > 0$ and Q atomless, and attach to each summand an independent mixture label. With probability one some summand draws from Q; let J be the first such index. Conditional on the labels and on every summand other than the Jth, the limit is a fixed constant plus k^{-J-1} times a draw from the atomless Q, so its conditional law is atomless, and therefore $(\lambda_0 - \tilde{\lambda}) + M_\infty \neq 0$ almost surely. □

Proof of Theorem 6. (i) Existence of the stationary law follows from almost sure absolute convergence of the series $\lambda_\infty = \lambda^* + \sum_{j \ge 0} k^j U_{-j}$, which holds for every $n \ge 2$ by a fractional-moment argument. Choose $0 < p < \min\{1, n-1\}$, so that $\mathbb{E}|U|^p < \infty$ even in the Cauchy case $n = 2$; then $\mathbb{E}\sum_{j \ge 0} |k|^{jp} |U_{-j}|^p < \infty$, so the series of p-th powers converges almost surely and the series itself converges absolutely. Uniqueness and convergence follow by coupling, because two copies driven by the same innovations from initial conditions λ_0 and λ_0' satisfy $|\lambda_t - \lambda_t'| = k^t |\lambda_0 - \lambda_0'| \to 0$ geometrically. Geometric ergodicity in total variation holds by the drift criterion with $V_p(\lambda) = 1 + |\lambda|^p$, since subadditivity of $x \mapsto x^p$ on $[0, \infty)$ for $p \le 1$ gives $\mathbb{E}[V_p(\rho + k\lambda + U) \mid \lambda] \le |k|^p |\lambda|^p + C$ with $|k|^p < 1$, and the everywhere-positive Student-t innovation density makes compact sets small (Meyn and Tweedie, 1993). Symmetry of the stationary law about λ^* holds because the innovations are symmetric and the stationary law is a convergent sum of independent symmetric terms. The mean and variance follow from (54) and independence, and the tail index from the regular variation of T_{n-1} with index $n - 1$, which is preserved under geometrically weighted i.i.d. sums, so moments of order p are finite precisely when $p < n - 1$. (ii) Direct summation, with $t^{-1} \sum_{j<t} U_j \to 0$ a.s. by the strong law, since a Student-t variable with at least two degrees of freedom has a finite mean, which holds for $n \ge 3$. (iii) Unrolling the recursion gives the displayed identity. M_t converges a.s. because, with $0 < p < \min\{1, n-1\}$, $\mathbb{E}\sum_j k^{-(j+1)p} |U_j|^p < \infty$, so the weighted series converges absolutely

almost surely for every $n \geq 2$. The limit $(\lambda_0 - \tilde{\lambda}) + M_\infty$ has a continuous distribution, since it is an independent sum involving U_0, which has a density, so it is nonzero a.s. □

Proof of Corollary 4. In the stationary regime, $\mathrm{Cov}(\lambda_{t+1}, \lambda_t) = \mathrm{Cov}(\rho + k\lambda_t + U_t,\, \lambda_t) = k\mathrm{Var}(\lambda_t)$, because U_t is independent of λ_t, so the lag-one autocorrelation equals k. Taking expectations in (53) at stationarity gives $\mathbb{E}[\lambda] = \rho + k\mathbb{E}[\lambda]$, which is the displayed mean, and inverting the two moments delivers (ρ, k). The stationary variance then ties the archive to the primitives through $\mathrm{Var}(\lambda_\infty)(1 - k^2) = \mathrm{Var}(U) = \sigma_u^2/(V(n-3))$ with $\sigma_u^2 = \rho(1-\rho)V + \sigma_\varepsilon^2$, which is the stated overidentifying restriction. □

Proof of Proposition 5. (i) By Theorem 5 the vintage series is a stationary ergodic AR(1) with i.i.d. innovations of finite variance, so almost-sure convergence is the ergodic theorem. For normality, write $\widehat{k}_T - k = \sum_t(\lambda_t - \bar{\lambda})U_t / \sum_t(\lambda_t - \bar{\lambda})^2$ and decompose the numerator as

$$\sum_t (\lambda_t - \bar{\lambda})U_t \;=\; \sum_t (\lambda_t - \mu_\lambda)U_t \;-\; (\bar{\lambda} - \mu_\lambda)\sum_t U_t, \quad \mu_\lambda \coloneqq \mathbb{E}[\lambda_t].$$

The demeaned terms $(\lambda_t - \bar{\lambda})U_t$ are not themselves martingale differences, because the sample mean $\bar{\lambda}$ depends on the whole path and hence on future innovations. The terms $(\lambda_t - \mu_\lambda)U_t$ do form a stationary ergodic martingale-difference sequence, since U_t is independent of $(\lambda_s)_{s \leq t}$, with variance $\mathbb{E}[(\lambda_t - \mu_\lambda)^2]\,\mathbb{E}[U_t^2] < \infty$ for $n \geq 4$, so the martingale central limit theorem applies to the first sum. The second sum is $o_p(\sqrt{T})$, because $\bar{\lambda} - \mu_\lambda = O_p(T^{-1/2})$ and $\sum_t U_t = O_p(T^{1/2})$ make it $O_p(1)$. The ergodic convergence of the denominator then gives asymptotic variance $\mathrm{Var}(U)/\mathrm{Var}(\lambda_\infty) = 1 - k^2$ by Theorem 6(i). Joint normality with the intercept follows the same way with the standard sandwich.

(ii) Under H_0, $\lambda_{t+1} = \rho + U_t$ with $U_t \mid \mathcal{D} \sim N(0, \sigma_u^2/S_t)$, since U_t is generated from the period-t cross-section, independent across t, where $\mathcal{D}$ is the design σ-algebra. Restricting to odd t makes the error index set $\{U_t \colon t \text{ odd}\}$ and the regressor index set $\{U_{t-1} \colon t \text{ odd}\}$ disjoint, so conditional on $\mathcal{D}$ and on the regressors the errors of the used observations remain independent centered Gaussians with known relative variances. On the full archive this restriction fails, because conditioning on the complete set of time-series regressors conditions on the errors from all but the final observation, as each dependent variable becomes the next observation's regressor. On the odd subsample, the weighted regression is a classical normal linear model with known heteroskedasticity weights, and the studentized slope is exactly t_{m-2} conditional on the design and regressors, hence also unconditionally.

(iii) The treated regression estimates $\rho + k\lambda$ and the holdout regression estimates ρ, each by OLS on independent subsamples of sizes $(1-h)n$ and hn with the same regime-invariant residual variance σ_u^2 and design variance V, so each slope is asymptotically normal with variance $\sigma_u^2/(V \cdot \text{subsample size})$, and independence across subsamples adds the variances. In detail, within each subsample the regression residual is the regime-invariant u_i, independent of the regressor and with variance σ_u^2, so the standard least-squares central limit theorem gives $\sqrt{n_g}\,(\widehat{\beta}_g - \beta_g) \Rightarrow N(0, \sigma_u^2/V)$ for each group g, with subsample fractions $1 - h$ and h; rescaling to the common rate $\sqrt{n}$ multiplies the group variances by $1/(1-h)$ and $1/h$, and the two groups share no units, so their estimators are independent and the variances add. □

Proof of Proposition 6. (i) Substituting $y_{is} - \bar{y}_s = (\rho + k\lambda_s)(\tilde{x}_{is} - \bar{x}_s) + (u_{is} - \bar{u}_s)$,

$$\lambda_{t+1} \;=\; \rho + k\,\mu_t + \nu_t, \quad \mu_t \coloneqq \frac{\sum_{s \leq t} S_s \lambda_s}{A_t}, \quad A_t \coloneqq \sum_{s \leq t} S_s, \quad \nu_t \coloneqq \frac{\sum_{s \leq t} \xi_s}{A_t},$$

where $\xi_s \coloneqq \sum_i (\tilde{x}_{is} - \bar{x}_s)(u_{is} - \bar{u}_s)$. The ξ_s are i.i.d. with mean zero, by independence of u and $\tilde{x}$ within a period and because λ_s is measurable with respect to the past, and they have finite variance $\sigma_u^2 V(n-1)$. By the strong law, $\sum_{s \le t} \xi_s / t \to 0$ and $A_t/t \to V(n-1)$ a.s., whence $\nu_t \to 0$ a.s.

The weighted mean μ_t obeys the exact recursion

$$\mu_{t+1} \;=\; \mu_t + \gamma_{t+1}\big(\lambda_{t+1} - \mu_t\big) \;=\; \mu_t + \gamma_{t+1}\big[\rho + (k-1)\mu_t + \nu_t\big], \quad \gamma_{t+1} \coloneqq \frac{S_{t+1}}{A_{t+1}} \in (0,1).$$

The gains vanish, and their sum diverges, almost surely. Indeed $\gamma_t \to 0$ because $\Pr(S_t > \epsilon t) \le \mathbb{E}[S^2]/(\epsilon t)^2$ is summable. For the divergent sum, note that $1 - \gamma_{t+1} = A_t/A_{t+1}$, so $-\log(1-\gamma_{t+1}) = \log(A_{t+1}/A_t) \ge \gamma_{t+1}$, which bounds the gains from above rather than below by the log increments. Because $\gamma_t \to 0$, eventually $-\log(1-\gamma_t) \le 2\gamma_t$, and therefore

$$\sum_{t_0 \le t \le T} \gamma_t \;\ge\; \tfrac{1}{2}\big(\log A_{T+1} - \log A_{t_0}\big) \;\longrightarrow\; \infty,$$

since $A_T \ge cT$ eventually, almost surely. Setting $e_t \coloneqq \mu_t - \lambda^*$ and using $\rho + (k-1)\lambda^* = 0$,

$$e_{t+1} \;=\; \big(1 - (1-k)\gamma_{t+1}\big)e_t + \gamma_{t+1}\nu_t.$$

Fix $\epsilon > 0$ and, on the almost-sure event above, take t_0 with $|\nu_t| \le (1-k)\epsilon$ and $(1-k)\gamma_t < 1$ for $t \ge t_0$. If $|e_t| > \epsilon$ then $|e_{t+1}| \le |e_t| - (1-k)\gamma_{t+1}(|e_t| - \epsilon)$, and since $\sum \gamma = \infty$ the sequence enters $[-\epsilon', \epsilon']$ for any $\epsilon' > \epsilon$ in finite time. Once $|e_t| \le \epsilon$, $|e_{t+1}| \le (1-(1-k)\gamma_{t+1})\epsilon + \gamma_{t+1}(1-k)\epsilon \le \epsilon$, so the sequence remains there. Hence $\limsup |e_t| \le \epsilon$ for every $\epsilon > 0$, so $\mu_t \to \lambda^*$ a.s., and therefore $\lambda_{t+1} = \rho + k\mu_t + \nu_t \to \lambda^*$ a.s. No projection is needed because the recursion for e_t is globally linear with a stable coefficient. Excursions cannot feed back into the gain or the noise, whose laws do not depend on λ.

(ii) Now set $e_t \coloneqq \mu_t - \tilde{\lambda}$ with $\tilde{\lambda} = \rho/(1-k)$. The recursion is exactly

$$e_{t+1} = \big(1 + (k-1)\gamma_{t+1}\big)e_t + \gamma_{t+1}\nu_t, \quad \Pi_t \coloneqq \prod_{s \le t}\big(1 + (k-1)\gamma_s\big).$$

Step 1. $\Pi_t \to \infty$ at rate t^{k-1}, a.s. Use $x - x^2/2 \le \log(1+x) \le x$ for $x \ge 0$, together with $\sum_s \gamma_s = \infty$ a.s. from part (i) and $\sum_s \gamma_s^2 < \infty$ a.s., which holds because eventually $A_s \ge cs$ while $\sum_s S_s^2/s^2 < \infty$ a.s. by $\mathbb{E}[S^2] < \infty$ and Kolmogorov's criterion. Since $\sum_{s \le t} \gamma_s = \log A_t + O(1) = \log t + O(1)$ a.s., we get $\log \Pi_t = (k-1)\log t + O(1)$ a.s.

Step 2. Normalization. Dividing the recursion by Π_{t+1} and summing,

$$\frac{e_{t+1}}{\Pi_{t+1}} \;=\; e_0 + L_{t+1}, \quad L_t \coloneqq \sum_{s \le t} \frac{\gamma_s\, \nu_{s-1}}{\Pi_s}.$$

The series converges absolutely a.s. Conditional on the design σ-algebra $\mathcal{D} \coloneqq \sigma(\{\tilde{x}_{is}\})$, $\mathrm{Var}(\nu_s \mid \mathcal{D}) = \sigma_u^2/A_s = O(1/s)$, and S_s is independent of ν_{s-1}, so $\mathbb{E}\big[\gamma_s|\nu_{s-1}|\big] = O(s^{-3/2})$. With $\Pi_s^{-1} = O(s^{-(k-1)+\epsilon})$ from Step 1, $\sum_s \gamma_s|\nu_{s-1}|/\Pi_s < \infty$ a.s. for any $k > 1/2$. Write $L \coloneqq \lim L_t$.

Step 3. $e_0 + L \neq 0$ a.s. Conditional on $\mathcal{D}$, the period shocks ξ_r are independent centered Gaussians, by independence of u and $\tilde{x}$ again, and L is a convergent linear combination of them, $L = \sum_r c_r \xi_r$ with $\mathcal{D}$-measurable coefficients $c_r = \sum_{s \ge r+1} \gamma_s/(\Pi_s A_{s-1}) > 0$. Hence $L \mid \mathcal{D}$ is Gaussian with variance at least $c_1^2 \mathrm{Var}(\xi_1 \mid \mathcal{D}) > 0$, so its conditional law is atomless, Condition on $\sigma(\mathcal{D}, e_0)$ rather than on $\mathcal{D}$ alone, which accommodates a random nonanticipative initialization; because e_0 is

independent of the future cross-sectional residuals, L remains conditionally Gaussian with positive variance given $\sigma(\mathcal{D}, e_0)$, so its conditional law is atomless and $\Pr(e_0 + L = 0) = 0$.

Step 4. Divergence. Decompose $e_t = \Pi_t(e_0 + L) - \Pi_t \sum_{s>t} \gamma_s \nu_{s-1}/\Pi_s$. The correction is bounded by $\sum_{s>t} \gamma_s |\nu_{s-1}|$, the tail of a series that converges a.s. because its expected terms are $O(s^{-3/2})$, so it vanishes. On $\{e_0 + L \neq 0\}$, which has probability one, $|e_t| \sim \Pi_t |e_0 + L| \asymp t^{k-1} \to \infty$. Finally $\lambda_{t+1} = \rho + k\mu_t + \nu_t$ with $\nu_t \to 0$, so $|\lambda_t| \to \infty$ at the same rate. □

Proof of Proposition A2. With $d_t := \sup_x |\widehat{m}_t - m^*|$, the triangle inequality and the pointwise contraction give $d_{t+1} \leq \bar{\kappa}_{\text{abs}}\, d_t + \bar{e}$, and iterating yields $d_t \leq \bar{\kappa}_{\text{abs}}^t d_0 + \bar{e}/(1 - \bar{\kappa}_{\text{abs}})$, where d_0 is finite by the initial condition, so the geometric term vanishes. The initial condition is not dispensable. With $u \equiv 0$, affine f with slope $\kappa \in (0, 1)$, exact retraining, and $\widehat{m}_0(x) = x$ on $\mathcal{X} = \mathbb{R}$, the iterates $\widehat{m}_t(x) = \kappa^t x$ converge pointwise to $m^* = 0$ while $\sup_x |\widehat{m}_t(x)|$ is infinite at every t. Define

$$d_t = \sup_{x \in \mathcal{X}} |\widehat{m}_t(x) - m^*(x)|\,. \tag{B.1}$$

Since $Tm^* = m^*$,

$$\begin{aligned} d_{t+1} &\leq \sup_x |\widehat{m}_{t+1}(x) - (T\widehat{m}_t)(x)| + \sup_x |(T\widehat{m}_t)(x) - (Tm^*)(x)| \\ &\leq \bar{e}_t + \bar{\kappa}_{\text{abs}} d_t. \end{aligned} \tag{B.2}$$

Iterating this inequality yields equation (A.11). □

Proof of Corollary 5. The affine map $\lambda \mapsto \rho + \kappa\lambda$ with $\kappa \geq 0$ is stable if and only if $\kappa < 1$, which rearranges to $k < 1/\psi = k_c$. Direct differentiation of $k_c = (1 - \delta\phi)/(1 - \delta\phi(1 - \rho))$ gives

$$\frac{\partial k_c}{\partial \delta} = -\frac{\rho\phi}{\big[1 - \delta\phi(1-\rho)\big]^2}, \quad \frac{\partial k_c}{\partial \phi} = -\frac{\rho\delta}{\big[1 - \delta\phi(1-\rho)\big]^2}, \quad \frac{\partial k_c}{\partial \rho} = -\frac{\delta\phi(1-\delta\phi)}{\big[1 - \delta\phi(1-\rho)\big]^2},$$

all weakly negative on the closed parameter space, and strictly negative exactly when $\rho\phi > 0$, $\rho\delta > 0$, and $\delta\phi > 0$, respectively. The limiting cases follow by substitution, and $k_c \to 0$ as $\delta\phi \uparrow 1$ because the numerator vanishes while the denominator tends to $\rho > 0$. □

Proof of Proposition 7. The off-diagonal product is $(\beta\delta\phi/V) \cdot \gamma\sigma_\theta^2 = k\delta\phi\rho$, giving the characteristic polynomial. Its discriminant $(k + \delta\phi)^2 - 4k\delta\phi(1 - \rho) \geq (k - \delta\phi)^2 \geq 0$ makes the roots real, and nonnegativity of the sum $k + \delta\phi$ and of the product $k\delta\phi(1 - \rho)$ makes them nonnegative, strictly positive when $k\delta\phi(1 - \rho) > 0$. The quadratic formula gives the displayed eigenvalues. For the ordering, write $P(\mu) = \mu^2 - (k+d)\mu + kd(1-\rho)$ with $d = \delta\phi$ and evaluate $P(k) = P(d) = -kd\rho < 0$, so both k and d lie strictly between the roots when $kd\rho > 0$, which gives $\mu_1 > \max\{k, d\}$ and $\mu_2 < \min\{k, d\}$, while $\mu_2 > 0$ follows from the strictly positive product $kd(1 - \rho)$. For the spectral radius, evaluate the upward parabola at one. If the value is negative, the roots bracket one and the radius exceeds one. If the value is positive, the displayed rearrangement gives $\kappa < 1$, which implies $k \leq \kappa < 1$ since $\psi \geq 1$, and together with $\delta\phi < 1$ the vertex $(k + \delta\phi)/2$ lies below one, so both roots lie below one. The stationary state solves $z = \overline{A}z + (\rho, 0)'$. Its second row gives $\overline{p}^* = \gamma\sigma_\theta^2 \overline{\lambda}^*/(1 - \delta\phi)$, and substituting into the first row with $\beta\gamma\sigma_\theta^2/V = k\rho$ gives $\overline{\lambda}^*\big(1 - k - k\delta\phi\rho/(1 - \delta\phi)\big) = \rho$, whose bracket equals $1 - \kappa$. Convergence of the coefficients follows from the law of large numbers for the exogenous panel moments, $s_t^{xx} \to V$ and $s_t^{\theta x} \to \sigma_\theta^2$. □

Proof of Proposition 8. Proceed by induction on $t \leq T$, the base case holding by assumption. Suppose $Z_t \to_p z_t$ and that the empirical capital second moment $n^{-1}\sum_i (K_{i,t-1}^c)^2$ is bounded in

probability, which holds at $t = 0$ by the fourth-moment condition and Chebyshev. The period-t cross-section $(\widetilde{\theta}_{i,t}, \widetilde{x}_{i,t}, \varepsilon_{i,t})$ is i.i.d. across units with moments of every order and a joint law free of n, so the cross-sectional law of large numbers gives $s_t^{xx} \to_p V$, $s_t^{\theta x} \to_p \sigma_\theta^2$, and $s_t^{x\varepsilon} \to_p 0$. For the shock-capital averages, $\mathbb{E}[\zeta_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\iota_t \mid \mathcal{F}_{t-1}] = 0$ and the conditional variances are bounded by a constant times $n^{-2}\sum_i (K^c_{i,t-1})^2$, because $w_{i,t}$ and $\eta_{i,t}$ are independent across units with bounded variances and independent of $\mathcal{F}_{t-1}$, so conditional Chebyshev on the event that the empirical capital moment is below a diverging bound gives $\zeta_t \to_p 0$ and $\iota_t \to_p 0$. Then $\nu_t \to_p \rho$ from equation (80) by the continuous mapping theorem, since the denominator converges to $V > 0$. The exact recursion $Z_{t+1} = A_t Z_t + \xi_t$ with $A_t \to_p \overline{A}$, $Z_t \to_p z_t$, and $\xi_t \to_p (\rho, 0)'$ yields $Z_{t+1} \to_p \overline{A} z_t + (\rho, 0)' = z_{t+1}$, again by continuous mapping. The empirical capital moment stays bounded in probability, because the capital law of motion gives $n^{-1}\sum_i (K^c_{i,t})^2 \le 2\delta^2 n^{-1}\sum_i (K^c_{i,t-1})^2 + 2\gamma^2\lambda_t^2\, s_t^{xx}$ with $\lambda_t \to_p$ a finite limit. The maximum over the fixed horizon of finitely many terms vanishing in probability vanishes in probability. □

Proof of Lemma 2. Substituting $y^c_{i,t} = \tilde{\theta}^c_{i,t} + \beta K^c_{i,t} + \varepsilon^c_{i,t}$ and $K^c_{i,t} = \delta K^c_{i,t-1} + \gamma\lambda_t \tilde{x}^c_{i,t}$ into the OLS formula, the coefficient of λ_t is exactly $\beta\gamma \sum(\tilde{x}^c)^2 / \sum(\tilde{x}^c)^2 = k$, the lagged-capital term appears with first-row weights $\beta\delta\, \tilde{x}^c_{i,t}/S_t$ on the full capital vector, and the remainder is g_t; the rows for $K^c_{i,t}$ are the capital law of motion. This full-space recurrence $W_{t+1} = B_t W_t + E_t$ in $W_t = (\lambda_t, K^c_{t-1})' \in \mathbb{R}^{n+1}$ carries an exact spurious direction. Because the centered features sum to zero, the vector $e = (0, 1, \ldots, 1)'$ satisfies $B_t e = \delta e$ in every period, yet the economic state never has a component along e, since $\mathbf{1}' K^c_{t-1} = 0$ identically and the forcing term also lies in the admissible subspace $\mathcal{H} = \mathbb{R} \times \{K : \mathbf{1}'K = 0\}$, which is invariant under every B_t. Conjugating the restriction of the recurrence to $\mathcal{H}$ by the isometry $K \mapsto Q'K$ produces the displayed reduced system, whose coefficient matrix is $\widetilde{B}_t$, and the top Lyapunov exponent of the full-space products is $\max\{\log\delta,\, \gamma_{L,n}\}$ with $\gamma_{L,n}$ the exponent of the reduced products, so the reduced exponent is the one that governs the realized retraining process. For the marginal representation of g_t, at each date the cross-section $(\tilde{\theta}_{i,t}, \tilde{x}_{i,t})$ is i.i.d. Gaussian across units, and the orthogonal decomposition $\tilde{\theta}^c_i = \rho\, \tilde{x}^c_i + v^c_i$ with $v \perp \tilde{x}$ has residual $v_i + \varepsilon_i$ with variance $\sigma_r^2 = \sigma_\theta^2(1-\rho) + \sigma_\varepsilon^2$. Conditional on the design, $g_t - \rho \sim N(0, \sigma_r^2/S_t)$ with $S_t \sim V\chi^2_{n-1}$ independent of the conditional standard normal, so the ratio is the scaled Student t in equation (95), exactly as in Theorem 5. The mean is undefined for $n = 2$ because T_1 is Cauchy, only the logarithmic moment survives there, and for $n \ge 4$ the variance uses $\mathbb{E}[1/\chi^2_\nu] = 1/(\nu - 2)$, which gives equation (96). Exogeneity and stationarity are immediate. For the mixing rate, the panel fundamentals form an n-dimensional stable Gaussian autoregression, which is geometrically β-mixing (Mokkadem, 1988), and $(\widetilde{B}_t, \widetilde{E}_t)$ is a fixed measurable function of the time-t panel state and of period shocks that are i.i.d. over time, which preserves the mixing rate. □

Proof of Lemma 3. Fix a realization with $S_t > 0$ and write $q = Q'\tilde{x}^c_t \in \mathbb{R}^{n-1}$, so that $q'q = (\tilde{x}^c_t)' QQ' \tilde{x}^c_t = S_t$, because QQ' is the orthogonal projection onto $\mathbf{1}^\perp$ and $\tilde{x}^c_t \in \mathbf{1}^\perp$. Any vector $(0, v')'$ with $v \perp q$ satisfies $\widetilde{B}_t(0, v')' = \delta(0, v')'$, which supplies the eigenvalue δ with multiplicity $n - 2$. The complementary two-dimensional subspace spanned by e_1 and $u = (0, q'/\sqrt{S_t})'$ is invariant, and on it the matrix acts in the basis (e_1, u) as

$$\begin{pmatrix} k & \beta\delta/\sqrt{S_t} \\ \gamma\sqrt{S_t} & \delta \end{pmatrix},$$

since the upper-right entry is $(\beta\delta/S_t)\,(\tilde{x}^c_t)' Q\, q/\sqrt{S_t} = (\beta\delta/S_t)\, S_t/\sqrt{S_t} = \beta\delta/\sqrt{S_t}$ and the lower-left entry collects $\gamma\, Q'\tilde{x}^c_t = \gamma\sqrt{S_t}\, q/\sqrt{S_t}$. The block has trace $k + \delta$ and determinant $k\delta - \beta\gamma\delta = 0$ because $k = \beta\gamma$, so its eigenvalues are 0 and $k + \delta$, which gives the characteristic polynomial. □

Proof of Theorem 7. Given Lemma 2, the statement is the stationary-ergodic random-recurrence theorem of Brandt (1986) and Bougerol and Picard (1992), which requires only stationarity and ergodicity of the sequence of pairs $\{(\widetilde{B}_t, \widetilde{E}_t)\}$, with serial dependence and contemporaneous dependence between the coefficient and the forcing term permitted, together with the logarithmic moment conditions (99) and $\gamma_{L,n} < 0$. The unique causal solution is the series in equation (100), whose almost sure absolute convergence follows because the negative exponent makes $\|\widetilde{B}_{t-1} \cdots \widetilde{B}_{t-j}\|$ decay at the exponential rate $\gamma_{L,n}$ along almost every path while $\log^+ \|\widetilde{E}_{t-j-1}\|$ grows at most linearly by the Borel–Cantelli argument standard in this setting. The logarithmic moment conditions hold by a direct argument, and moment bounds on the entries are neither available nor needed. The first row of $\widetilde{B}_t$ has entries k and $\beta\delta\,(\tilde{x}_t^c)'Q/S_t$, the first column below it is $\gamma\,Q'\tilde{x}_t^c$, the diagonal block is δI_{n-1}, and $\|Q'\tilde{x}_t^c\| \le \|\tilde{x}_t^c\|$ since Q has orthonormal columns, and g_t has the Student-t-type ratio structure of the static theory, so these variables have only finitely many moments, their availability limited by the lower tail of the design sum. The logarithms behave well regardless, since $\log^+$ of each entry is bounded by a constant plus $\log^+ |\tilde{x}_{i,t}^c|$ plus $|\log S_t|$, the panel features are Gaussian with finite logarithmic moments, and S_t is a Gaussian quadratic form that is positive almost surely with $\mathbb{E}|\log S_t| < \infty$ for $n \ge 2$, as in the χ^2 calculation of the static case. Kingman's theorem applies to $\log \|\widetilde{B}_t \cdots \widetilde{B}_1\|$ because the coefficient sequence is stationary and ergodic and the sequence of normalized logarithms is subadditive with an integrable positive part, which also delivers the almost sure limit defining $\gamma_{L,n}$. □

Proof of Proposition 9. The growth statement follows from Kingman's theorem applied to the subadditive sequence $\log \|\widetilde{B}_t \cdots \widetilde{B}_1\|$, which extends the independent-products theorem of Furstenberg and Kesten (1960) to stationary ergodic coefficients (Kingman, 1973). The almost sure limit equals the constant $\gamma_{L,n}$ by ergodicity of the coefficient sequence established in Lemma 2. □

Proof of Theorem 8. The proof has three steps.
Step 1. A balanced representation and a fixed-block limit. Let $D_n = \mathrm{diag}\big(1, n^{-1/2} I_{n-1}\big)$ and define $\widehat{B}_{n,t} = D_n \widetilde{B}_{n,t} D_n^{-1}$. For every fixed n, this deterministic similarity transformation leaves the Lyapunov exponent unchanged, since

$$\left| \frac{1}{t} \log \big\| D_n \widetilde{B}_{n,t} \cdots \widetilde{B}_{n,1} D_n^{-1} \big\| - \frac{1}{t} \log \big\| \widetilde{B}_{n,t} \cdots \widetilde{B}_{n,1} \big\| \right| \le \frac{\log \|D_n\| + \log \|D_n^{-1}\|}{t},$$

which vanishes as $t \to \infty$ for every fixed n. Writing $v_{n,t} = \frac{1}{\sqrt{n}} Q' \widetilde{x}_t^c$ and $s_{n,t} = \frac{S_t}{n} = \|v_{n,t}\|^2$ we obtain

$$\widehat{B}_{n,t} = \begin{pmatrix} k & \dfrac{\beta\delta}{s_{n,t}} v_{n,t}' \\ \gamma v_{n,t} & \delta I_{n-1} \end{pmatrix}. \tag{B.3}$$

The balancing is essential because the untransformed lower-left block has norm of order $\sqrt{n}$.

Fix a block length m. Let $G_{n,m}$ be the $m \times m$ Gram matrix with entries $(G_{n,m})_{st} = v_{n,s}' v_{n,t}$. The cross-sectional law of large numbers gives $G_{n,m} \xrightarrow{p} G_m$, where

$$(G_m)_{st} = \begin{cases} V, & s = t, \\ \rho V \phi^{|s-t|}, & s \ne t. \end{cases} \tag{B.4}$$

For fixed m and every $n \ge m + 1$, the block directions $v_{n,1}, \ldots, v_{n,m}$ are linearly independent almost surely, being jointly Gaussian vectors in general position in $\mathbb{R}^{n-1}$, so $G_{n,m}$ is almost surely nonsingular

and the inverse square root below is well defined; the limiting Toeplitz matrix G_m is positive definite, since it is the covariance matrix of m consecutive observations of the nondegenerate stationary feature process. Let $\mathcal{H}_{n,m} = \operatorname{span}\{v_{n,1}, \dots, v_{n,m}\}$. Its orthogonal complement is invariant under every matrix in the block and is multiplied by δ in each period. On $\mathcal{H}_{n,m}$, orthonormalizing the vectors by $G_{n,m}^{-1/2}$ shows that the restriction of $\widehat{B}_{n,t}$ is a continuous function of $G_{n,m}$. More explicitly, in the orthonormal coordinates generated by the columns of $J_{n,m} = \begin{pmatrix} v_{n,1} & \cdots & v_{n,m} \end{pmatrix} G_{n,m}^{-1/2}$, the restriction is

$$C_t(G_{n,m}) = \begin{pmatrix} k & \dfrac{\beta\delta}{(G_{n,m})_{tt}} e_t' G_{n,m}^{1/2} \\ \gamma G_{n,m}^{1/2} e_t & \delta I_m \end{pmatrix}. \tag{B.5}$$

Therefore, for $n \geq m+2$, $\left\| \widehat{B}_{n,m} \cdots \widehat{B}_{n,1} \right\| = \max\left\{ \left\| C_m(G_{n,m}) \cdots C_1(G_{n,m}) \right\|, \delta^m \right\}$, and hence

$$\left\| \widehat{B}_{n,m} \cdots \widehat{B}_{n,1} \right\| \xrightarrow{p} L_m, \quad L_m = \max\left\{ \left\| C_m(G_m) \cdots C_1(G_m) \right\|, \delta^m \right\}. \tag{B.6}$$

The convergence also holds in logarithmic expectation

$$\left\| \widehat{B}_{n,t} \right\| \leq k + \delta + \frac{|\beta|\delta}{\sqrt{s_{n,t}}} + |\gamma|\sqrt{s_{n,t}}, \tag{B.7}$$

and $s_{n,t}$ has the distribution $V\chi^2_{n-1}/n$. For every $p > 1$, $\sup_{n \geq 2} \mathbb{E}\left|\log s_{n,t}\right|^p < \infty$. It follows from (B.7) that the positive parts of the fixed-block logarithms are uniformly bounded in L^p. Their negative parts are bounded by $m|\log\delta|$, because the orthogonal capital directions give $\left\| \widehat{B}_{n,m} \cdots \widehat{B}_{n,1} \right\| \geq \delta^m$. Thus the fixed-block logarithms are uniformly integrable, and (B.6) implies

$$\mathbb{E} \log \left\| \widehat{B}_{n,m} \cdots \widehat{B}_{n,1} \right\| \longrightarrow \log L_m \quad (n \to \infty). \tag{B.8}$$

Step 2. The growth rate of the deterministic block. Choose a Hilbert space containing vectors $X_1, X_2, \dots$ with Gram matrix given by (B.4). Define deterministic operators on $\mathbb{R} \oplus \mathcal{H}$ by

$$\mathcal{C}_t \begin{pmatrix} \lambda \\ z \end{pmatrix} = \begin{pmatrix} k\lambda + \dfrac{\beta\delta}{V}\langle X_t, z \rangle \\ \gamma X_t \lambda + \delta z \end{pmatrix}. \tag{B.9}$$

The number L_m in (B.6) equals $L_m = \|\mathcal{C}_m \cdots \mathcal{C}_1\|$. The Toeplitz covariance sequence is absolutely summable, so

$$\left\| \sum_{t=1}^{m} a_t X_t \right\|^2 \leq \Gamma \sum_{t=1}^{m} a_t^2, \quad \Gamma = V + 2\rho V \frac{\phi}{1-\phi}. \tag{B.10}$$

Starting from (λ_1, z_1), iteration of the capital equation gives

$$z_t = \delta^{t-1} z_1 + \gamma \sum_{j=1}^{t-1} \delta^{t-1-j} X_j \lambda_j. \tag{B.11}$$

Let $q_{t-1} = \sum_{j=1}^{t-1} d^{t-1-j}\lambda_j$ and $b_t = \frac{\beta}{V}\delta^t \langle X_t, z_1 \rangle$. Using $\langle X_t, X_j \rangle = \rho V \phi^{t-j}$, $(j < t)$, we obtain the

forced two-state recursion

$$\begin{pmatrix}\lambda_{t+1}\\ q_t\end{pmatrix} = A\begin{pmatrix}\lambda_t\\ q_{t-1}\end{pmatrix} + \begin{pmatrix}b_t\\ 0\end{pmatrix}, \quad A = \begin{pmatrix}k & k\rho d\\ 1 & d\end{pmatrix}. \tag{B.12}$$

The matrix A has dominant eigenvalue μ_1. Writing $r = \max\{\mu_1, \delta\}$, the two-dimensional Jordan decomposition gives a constant C such that $\|A^j\| \le C(j+1)r^j$ for every $j \ge 0$. This formulation also covers the degenerate nilpotent case because $\delta > 0$. Moreover, (B.10) implies

$$\sum_{t=1}^{m} |\langle X_t, z_1\rangle|^2 \le \Gamma\|z_1\|^2. \tag{B.13}$$

Equations (B.12)–(B.13) and Cauchy–Schwarz give, uniformly for $t \le m$,

$$|\lambda_t| \le C_1(t+1)^{3/2} r^{t-1} \left(|\lambda_1| + \|z_1\|\right). \tag{B.14}$$

Substituting this bound into (B.11) and using the synthesis bound (B.10) gives

$$\|z_{m+1}\| \le C_2(m+1)^2 r^m \left(|\lambda_1| + \|z_1\|\right). \tag{B.15}$$

Consequently, $L_m \le C_3(m+1)^2 r^m$.

For the matching exponential lower bound, capital directions orthogonal to $\operatorname{span}\{X_1, \dots, X_m\}$ yield $L_m \ge \delta^m$. If $\mu_1 > \delta$, start from $z_1 = 0$ and $\lambda_1 = 1$. The forcing term in (B.12) then vanishes, and the Perron–Frobenius component of $A^m e_1$ is nonzero, so $L_m \ge c\mu_1^m$ for all sufficiently large m. Therefore,

$$\lim_{m\to\infty} \frac{1}{m} \log L_m = \log r = \log \max\{\mu_1, \delta\}. \tag{B.16}$$

Step 3. Subadditivity. For each fixed n, stationarity and subadditivity give

$$\gamma_{L,n} = \inf_{m\ge 1} \frac{1}{m} \mathbb{E} \log \left\| \widehat{B}_{n,m} \cdots \widehat{B}_{n,1} \right\|.$$

Hence, for every fixed m, $\limsup_{n\to\infty} \gamma_{L,n} \le \frac{1}{m} \log L_m$ by (B.8). Letting $m \to \infty$ and using (B.16) proves (102).

The boundary case $\delta = 0$ follows directly from the static recursion and can be added separately. If $\kappa < 1$, Proposition 9 implies $\mu_1 < 1$, while $\delta < 1$ by assumption. Thus $r < 1$ and the right-hand side of (102) is strictly negative. The same conclusion holds uniformly away from the frontier for fixed remaining primitives. This proves the stable-half statement. □

Appendix C Supplementary Figures and Numerical Details

The figures and the table collected here support the numerical discussion in Section 5.2.1, and the following paragraphs collect the numerical detail summarized there.

Calibration and stationary comparisons. Consider a panel with $n = 200$, $\rho = 0.5$, $\delta = 0.8$, and $\phi = 0.875$, so that $\delta\phi = 0.7$. At $k = 0.30$, the effective gain is $\kappa = 0.65$, the eigenvalues of $\overline{A}$ are 0.881 and 0.119, and the fitted geometric decay of the vintage autocorrelation function over lags 10 through 24 is 0.890. At $k = 0.44$, the effective gain is $\kappa = 0.953$, the dominant eigenvalue is

0.9834, and the fitted decay over lags 20 through 39 is 0.9833. The stationary mean slopes are 1.432 and 11.13, compared with population fixed points of 1.429 and 10.71. The gaps are finite-panel differences in the sense of Remark 2. The random coefficient $1/s_t^{xx}$ has expectation above $1/V$ and is correlated with the state, so the stationary mean of the random system need not equal the population fixed point. An analytical expansion that delivers the sign and the constant remains open.

Stationary-mean gaps. Across panel sizes from twenty-five to eight hundred, the gap between the simulated stationary mean and the population fixed point fits $0.95/n$ at $\kappa = 0.65$ and $107/n$ at $\kappa = 0.95$, with intercepts indistinguishable from zero. The difference is therefore positive, vanishes at the rate $1/n$, and its coefficient grows by two orders of magnitude as the frontier approaches, consistent with amplification of small distortions in the retraining operator by powers of $1/(1-\kappa)$.

Lyapunov point estimates. Estimates use 2×10^5 matrix products with batch-mean standard errors of about 10^{-4}. At $n = 200$, the estimates are consistent with Conjecture 1. In the weak-loop region the exponent follows the capital mode, with -0.2229 at $k = 0.05$ against $\log \delta = -0.2231$, while $\log \mu_1 = -0.320$. In the stronger-loop region it follows the coupled mode, with -0.1284 at $k = 0.30$ against $\log \mu_1 = -0.1269$, -0.0177 at $k = 0.44$ against -0.0167, and $+0.0288$ at the supercritical $k = 0.50$ against $+0.0297$. The remaining gaps of about 10^{-3} exceed the simulation standard errors by an order of magnitude, so they are genuine finite-panel deviations rather than numerical noise, and they close as the panel grows. At $n = 1000$ the estimates are -0.2231, -0.1272, -0.0169, and $+0.0295$, within 3×10^{-4} of the conjectured limit in every case.

Zero crossings and calibration variations. For each of $n \in \{25, 100, 200, 1000\}$, the exponent is estimated on a grid of loop gains with step 0.02, and the zero crossing is obtained by interpolation. The crossings are 0.475, 0.465, 0.463, and 0.462, against the population critical gain $k_c = 0.4615$, so the estimated finite-panel zero crossing approaches the population frontier as the panel grows. Figure C.1 plots the gap to the conjectured limit against $1/n$ in the coupled-mode region, and the gap shrinks approximately linearly in $1/n$. Holding $\delta\phi = 0.7$ fixed while varying durability and persistence separately sharpens the interpretation of the two modes. At $(\delta, \phi) = (0.72, 0.972)$, $(0.8, 0.875)$, and $(0.9, 0.778)$, the weak-loop exponent at $k = 0.05$ is -0.3211, -0.2229, and -0.1052, against conjectured limits $\log \max\{\mu_1, \delta\}$ of -0.3204, -0.2231, and -0.1054. The population eigenvalue μ_1 is identical across the three calibrations because it depends only on the product $\delta\phi$, so the movement of the exponent isolates the role of δ and confirms the max structure of the conjecture, including the coupled-mode regime at $\delta = 0.72$ where μ_1 exceeds δ. Near the mode crossing the finite-panel exponent can sit slightly above the conjectured limit and convergence in n is slower, as at $\delta = 0.9$ and $k = 0.30$ where the estimate is -0.0996 against -0.1054.

Near-critical running estimates. At the population frontier $\kappa = 1$, that is $k = k_c$, the running exponent estimate at $n = 200$ is -0.0011, -0.0005, and -0.0010 for product lengths 5×10^3, 2×10^4, and 8×10^4, within 10^{-3} of zero, so the simulated system remains slightly subcritical at this panel size, in line with the finite-panel zero crossings sitting slightly above k_c.

Growth at the population frontier. Over the first 2×10^4 periods, the cross-path mean and variance of the slope grow superlinearly, with log-log growth exponents of about 1.4 and 3.2 at $n = 50$ and 2.3 and 6.0 at $n = 200$, so the near-critical transient at the population frontier depends

strongly on the panel size. Because these cross-path moments can be dominated by a few exceptional paths, medians and fixed quantiles of $|\lambda_t|$ are the more robust summaries for this regime.

Divergence diagnostics. At $k = 0.50$ with five hundred replications of length five thousand, every path crosses the thresholds $|\lambda_t| > 10^3$, 10^4, and 10^5, with median hitting times 170, 249, and 330 periods from $\lambda_0 = 0$ and 154, 232, and 316 from the population fixed point. The roughly eighty-period spacing between successive decades matches the estimated exponent, since $\log 10/\gamma_{L,n} \approx 80$ at $\gamma_{L,n} = 0.0288$. At $k = 0.44$, no replication crosses even the lowest threshold. Burn-in diagnostics at $\kappa = 0.95$ show that stationary means from initial slopes 0, 10.7, and 32.1 agree to three decimals across burn-ins from 10^3 to 5×10^4 periods, while the effective sample size of 1.9×10^5 retained periods is about 1,750, so reported stationary means near the frontier carry Monte Carlo standard errors of roughly 0.1.

Counterfactual tail decomposition. Three mechanisms can generate large slope realizations. The ratio form of the regression disturbance creates Student-type tails even without multiplicative dynamics, persistent coefficients cluster large realizations over time, and random multiplicative propagation can generate Kesten-type power tails. A counterfactual decomposition on the exact linear lift separates these mechanisms at $\kappa = 0.95$. Replacing the regression disturbance g_t with an independent Gaussian of matched variance leaves the stationary dispersion and the tail nearly unchanged, with standard deviation 3.6 against 3.8 and a one-percent Hill index of 4.1 against 4.4. Replacing the random coefficient matrices with the deterministic two-state limit driven by the same disturbance collapses the dispersion by a factor of twenty, to 0.18, and thins the tail to a Hill index above 9. The counterfactual indicates that random coefficient propagation and the associated high-dimensional capital dynamics account for most of the near-frontier dispersion, while the marginal tail of the additive regression disturbance plays a smaller role. The deterministic two-state replacement also removes the rotating feature-aligned directions and the $n - 2$ capital modes, so the comparison bounds the joint contribution of these channels rather than isolating coefficient randomness alone. Persistence concentrates the resulting excursions in clusters.

Tail sensitivity. The Hill estimates drift with the tail fraction and vary with the panel size and the simulation seed, and the full set of tail diagnostics follows Figure C.4 below.

Hill estimates are based on 2×10^5 stationary vintages after a burn-in of 10^4 periods. At $\kappa = 0.65$ the estimates fall from 9.9 to 5.8 as the tail fraction rises from a quarter of one percent to five percent, with a value of 7.8 at one percent. At $\kappa = 0.95$ they fall from 5.0 to 2.8 over the same range, with 3.9 at one percent. Circular block-bootstrap standard errors with blocks of two thousand periods are 0.21 and 0.43 at the one-percent fraction, compared with independent-sampling values of 0.17 and 0.09, so conventional standard errors understate uncertainty by a factor of two to five near the frontier. The runs estimator of the extremal index is 0.36 at $\kappa = 0.65$ and 0.05 at $\kappa = 0.95$, so extreme observations arrive in long clusters. Across panel sizes at $\kappa = 0.95$, the one-percent estimates are 3.1 at $n = 50$, 5.1 at $n = 200$, and 5.4 at $n = 800$, while independent seeds at $n = 200$ give 3.9 and 5.1. The drift across tail fractions, the seed dependence, and the clustering together imply that these values should not be read as measurements of an established Kesten power law.

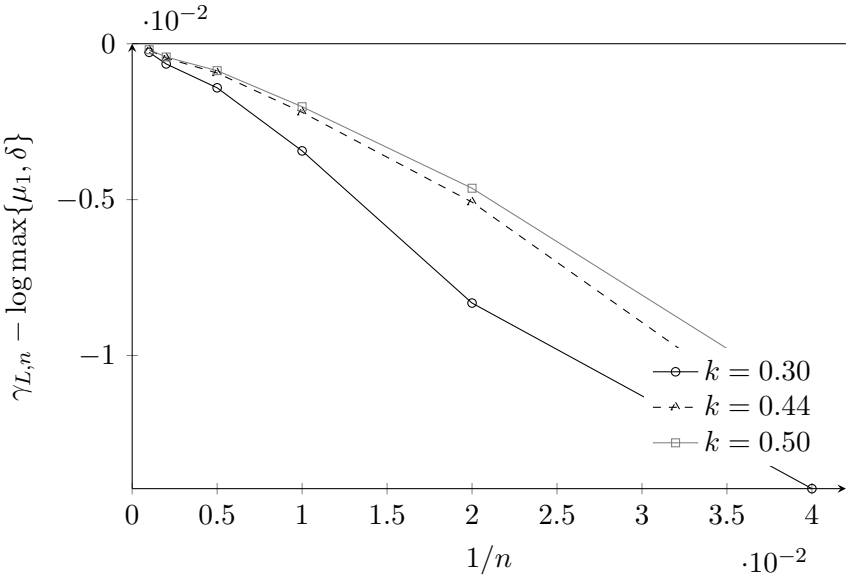


Figure C.1: Panel-size convergence of the Lyapunov exponent in the coupled-mode region. The gap between the estimated exponent $\gamma_{L,n}$ and the conjectured limit is plotted against $1/n$ for n between 25 and 1000. The gaps shrink toward zero approximately linearly in $1/n$.

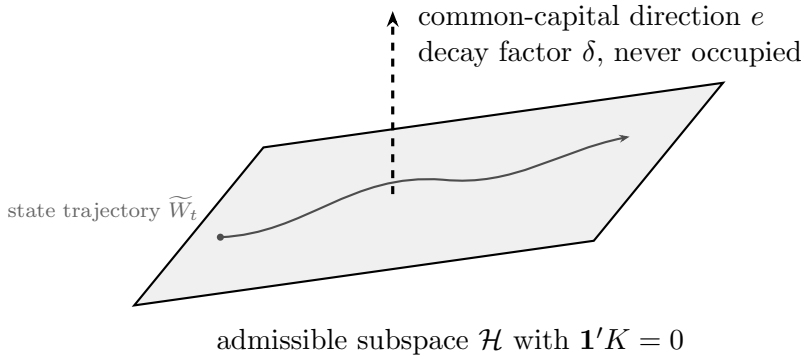


Figure C.2: The admissible state space of the dynamic recurrence. Cross-sectionally demeaned capital lies in the hyperplane $\mathcal{H}$. The coefficient matrices preserve this hyperplane, while the common-capital direction, drawn dashed, is an eigenvector with eigenvalue δ that is never reached by the economic state or the forcing term.

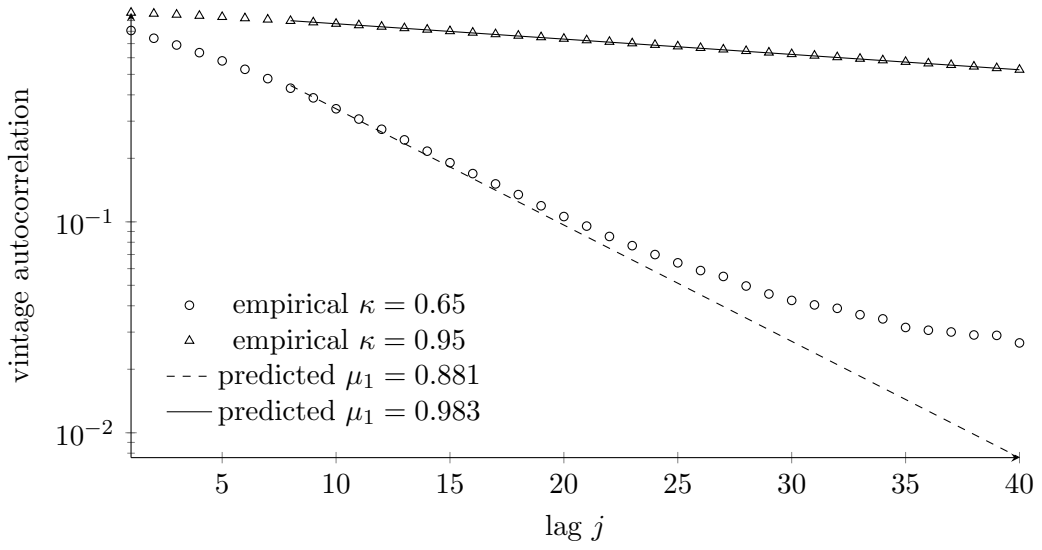


Figure C.3: Vintage autocorrelation of the dynamic panel ($n = 200$, $\delta = 0.8$, $\phi = 0.875$, $\rho = 0.5$) against the predicted decay at the dominant eigenvalue of $\bar{A}$, anchored at lag ten. Markers show the empirical autocorrelation and lines the spectral decay implied by the large-panel dynamics of Proposition 7. The proximity of $\phi = 0.875$ to $\mu_1 = 0.881$ at $\kappa = 0.65$ means the fitted decay cannot sharply distinguish the coupled mode from primitive persistence.

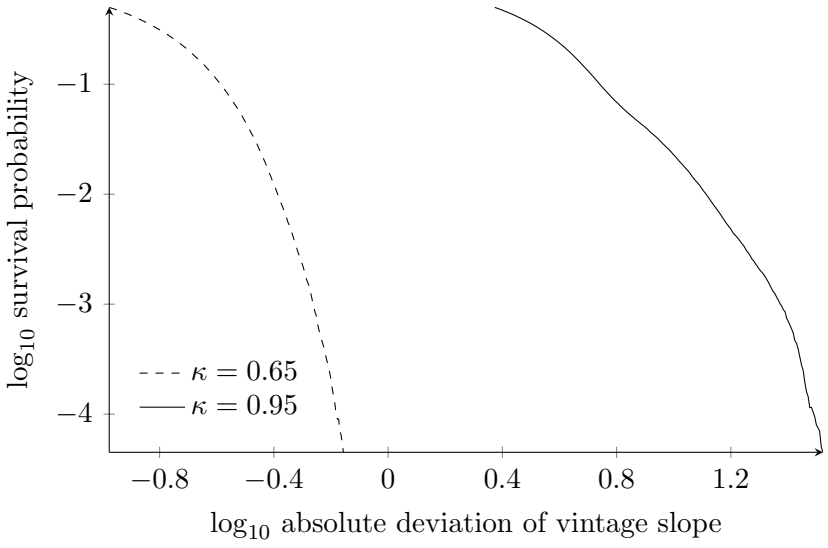


Figure C.4: Survival functions of centered vintage slopes on log-log axes. The near-critical economy displays the flatter, heavier tail, with Hill indices near 8 at $\kappa = 0.65$ and near 4 at $\kappa = 0.95$ at the one-percent tail fraction.

Moment	Simulated	MC s.e.	Benchmark
Static economy, $k = 0.6$, $n = 12$, $\sigma_u^2/V = 0.5$, 2×10^5 periods			
$\mathbb{E}[U_t]$	-0.0006	–	0
$\mathrm{Var}(U_t)$	0.0557	–	0.0556
$\mathbb{E}[\lambda_\infty]$	1.2486	–	1.2500
$\mathrm{Var}(\lambda_\infty)$	0.0876	–	0.0868
$\mathrm{Corr}(\lambda_{t+1}, \lambda_t)$	0.6028	–	0.6000
Excess kurtosis of U_t	0.847	–	0.857
Dynamic panel, $n = 200$, $\delta = 0.8$, $\phi = 0.875$, so $\delta\phi = 0.7$			
Stationary mean, $\kappa = 0.65$	1.432	0.001	1.429
Stationary mean, $\kappa = 0.95$	11.13	0.09	10.71
ACF decay rate, $\kappa = 0.65$	0.891	–	0.881
ACF decay rate, $\kappa = 0.95$	0.9833	–	0.9834
Lyapunov exponent $\gamma_{L,n}$, $\kappa = 0.11$	-0.2229	$< 10^{-4}$	-0.2231
Lyapunov exponent $\gamma_{L,n}$, $\kappa = 0.65$	-0.1284	0.0001	-0.1269
Lyapunov exponent $\gamma_{L,n}$, $\kappa = 0.95$	-0.0177	0.0001	-0.0167
Lyapunov exponent $\gamma_{L,n}$, $\kappa = 1.08$	$+0.0288$	0.0001	$+0.0297$

Table C.1: Simulated moments against theory benchmarks. The static block illustrates Theorems 5 and 6. The dynamic block illustrates Proposition 7, Theorem 7, and Proposition 9, and is consistent with Conjecture 1, with the Lyapunov exponent computed by power iteration on the exact linear lift and its conjectured limit given by $\log \max\{\mu_1, \delta\}$, the capital mode δ governing the weak-loop row. The dynamic stationary means carry the finite-panel differences discussed in the text. The static entries are exact closed forms, while the dynamic benchmark entries in the Lyapunov rows are conjectured large-n limits rather than finite-n closed forms. The Lyapunov standard errors are batch-mean standard errors from 2×10^5 matrix products, and the standard error of the near-frontier stationary mean reflects an effective sample size of about 1,750 induced by serial dependence. The residual Lyapunov gaps of about 10^{-3} are genuine finite-panel deviations that shrink toward the conjectured limit as n grows, as shown in Figure C.1.