THE EXACT RISK OF INSTANCE NORMALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Instance normalization (IN), which removes each input window's statistics before forecasting and restores them afterward, is standard in time-series forecasting, yet there is little theory saying when it helps or fails. We analyze linear forecasters, starting from a characterization of mean-only IN as least squares constrained to weights summing to one, and under ridge training as ridge centered at the uniform vector on that plane. This yields closed forms for IN's cost and benefit under unseen level shift, with a boundary: IN wins when the squared shift exceeds the variance of the best linear estimate of window level. Three consequences follow. First, the choice need not be binary. Interpolating the two forecasters admits an exact risk decomposition around *any* backbone, linear or not: the optimum is interior precisely when price and benefit are both positive, its excess risk is the parallel combination of the two endpoints', and in the linear case it is the shift-to-boundary ratio $\tau^2/(\tau^2 + \delta_\star^2)$, computable before any label is seen. Second, division by the window standard deviation cancels identically in linear prediction, so it cannot amplify test error; it acts through training, reweighting examples by $1/\hat{\sigma}_i^2$, and the scale of that reweighted loss, not weight concentration, separates two catastrophic failures from six benign cases on eight benchmarks, from training data alone. Third, finite-sample estimation noise creates shift fragility even when the population model is shift-insensitive. Three preregistered experiments on PatchTST and iTransformer, outcome texts frozen before any run, mark the mechanism's reach and its limit: the deep statistic $\hat{G} = \hat{S}\hat{\delta}^2$ ranks normalization gains at all four horizons ($\rho = 0.944$–0.965) but is shown *not* to predict their sign, since it estimates the benefit and carries no price term; both regimes realize on transformers; and the interior optimum is found in 8 of 8 deep cells, where, estimated without test labels, it cuts regret against the oracle by $17\times$ relative to the best binary rule in a post-hoc $n{=}8$ comparison.

1 INTRODUCTION

Research on long-horizon time-series forecasting (LTSF) has converged on an unusual position. Single linear layers match or beat far more elaborate deep architectures on the standard benchmark suite (Zeng et al., 2023); most published linear variants are mathematically equivalent to plain linear regression with closed-form solutions (Toner and Darlow, 2024); and carefully tuned ridge regression is broadly competitive, *provided it is combined with local normalization* (Huang et al., 2026). The models are by now well understood; the normalization is not.

Instance normalization, which subtracts each window's own mean (usually also dividing by the standard deviation), forecasts in normalized coordinates and then de-normalizes, entered forecasting as RevIN (Kim et al., 2022) and is now a default across architectures. What is known about it is largely empirical: it helps most on drifting series, sometimes hurts, occasionally fails badly (Fu and Yang, 2025), and its learned affine parameters appear redundant (Berthelier et al., 2026). Recent analyses pose the missing theory directly as an open problem (Toner and Darlow, 2024; Huang et al., 2026; Berthelier et al., 2026).

This paper develops that theory for linear forecasters. The starting point is structural (Remark 1): *mean-only IN is ordinary least squares constrained to effective weights summing to one*; under the ridge objective used in real pipelines, it is exactly ridge *centered at the uniform vector* on that plane (Prop. 1). Everything else follows from constrained versus unconstrained regression, and both sides admit exact formulas: a computable loss of fit on stationary data against computable immunity to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

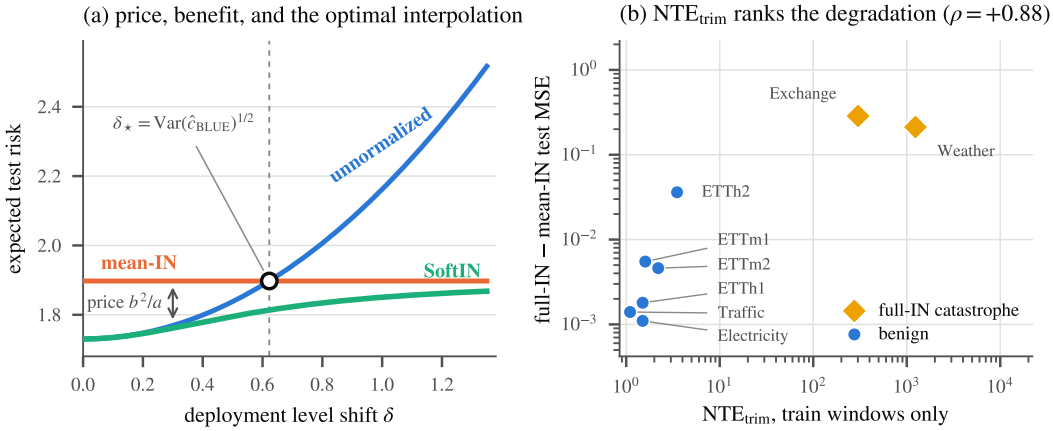

Figure 1: **The price and the benefit of the constraint, both in closed form.** **(a)** Exact population risk against an unseen level shift (AR(1), φ=0.7, L=24, h=3; closed forms, machine-verified). Mean-IN pays a flat premium b^2/a for exact immunity; the unnormalized model pays $\delta^2 b^2$. They cross at $\delta_\star$, the standard error of the window's own level estimate (Cor. 1); interpolating between them, a rule we call SoftIN, beats both at every finite shift, and by the widest margin exactly at the boundary (Prop. 3). **(b)** Why full RevIN sometimes collapses: the train-only statistic NTE_{trim} (Sec. 6) against the realized full-IN − mean-IN test gap. The two catastrophes are the two datasets whose NTE is two orders of magnitude above the rest; weight concentration, the usual explanation, does not separate them at all.

level shift, with the choice between them turning on the size of the shift relative to the precision with which a window determines its own level. One consequence needs no linearity: the mean-IN *wrapper* is exactly shift-equivariant around any deterministic backbone (Thm. 1), so the benefit is architecture-independent by proof and only the raw model's exposure must be measured.

The line between what is classical here and what is new is narrow, so we draw it ourselves. *Classical* is the constraint and its algebra: mean-IN as a restricted least-squares estimator is already in Toner and Darlow (2024), Wang et al. (2026) state the same fact in characteristic-root language (IN forcing a unit root at $r = 1$ is exactly $\mathbf{1}^\top w = 1$), and the GLS/BLUE machinery is Aitken's (Aitken, 1935; Rao, 1973). Those are *representational* statements about what the constrained class can express. Ours is a *risk* statement: what occupying that class costs, what it returns, how far into it one should go, and how to decide before deployment.

Our contributions. **(C1)** An exact theory of the constraint: the price b^2/a, the boundary $\delta_\star^2 = \text{Var}(\hat{c}_{\text{BLUE}})$, the same object for any known nuisance subspace (Thm. 4), the drift/differencing equivalence, and an estimation-induced fragility term strictly positive even when $b = 0$ (Thm. 2), the one result with no classical antecedent. **(C2)** What the standard-deviation division does. It cancels in prediction, so full RevIN is mean-IN under $1/\hat{\sigma}_i^2$-weighted least squares (Thm. 3); the standard account of its failures is ruled out (Cor. 2); and a train-only statistic ranks all eight datasets where the standard account fails. **(C3)** An exact risk decomposition for SoftIN around *any* backbone: its optimum is interior iff price and benefit are both positive, at $\alpha^\star = 1/(1+P/B)$ (Prop. 3). The linear case reduces to the tuning-free $\tau^2/(\tau^2+\delta_\star^2)$; all 8 preregistered deep cell-seeds satisfy the interiority condition, and estimating P and B without test labels cuts regret against the oracle 8–17× relative to the best binary rule in a post-hoc n=8 comparison. **(C4)** Validation. Every formula is machine-checked, the deployment-time diagnostic tracks per-channel gains at $r = 0.90$ at H=96, and across three preregistered experiments the shift-defect statistic ranks mean-IN's benefit to PatchTST and iTransformer at all four horizons but is shown *not* to predict its sign, both regimes realizing on transformers with every failed gate reported in full (Secs. 7–8, App. P).

The paper follows one line: an exact theory of the constraint (Secs. 3–6), a pre-deployment quantity validated on eight benchmarks (Sec. 7), and three preregistered experiments marking where the mechanism reaches in deep forecasters and where it stops (Sec. 8).

2 Related work

Linear models for LTSF. Zeng et al. (2023) showed a linear layer outperforms the transformer generation (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022); recurrent, dense, all-MLP and miniaturized linear designs (Salinas et al., 2020; Oreshkin et al., 2020; Chen et al., 2023; Das et al., 2023; Xu et al., 2024; Lin et al., 2024; Li et al., 2023) then closed most of the gap, PatchTST (Nie et al., 2023) and iTransformer (Liu et al., 2024) are the strong deep baselines with TimesNet (Wu et al., 2023) and channel-independence analyses (Han et al., 2024a) beside them, and Brigato et al. (2026) find no method dominates under fair reruns. Toner and Darlow (2024) prove closed-form equivalences among linear variants, treating IN as a weakly constrained regression but not analyzing its risk; Huang et al. (2026) show tuned ridge plus local normalization is competitive and ask why; Wang et al. (2026) observe via characteristic roots that IN inserts a unit root. This paper supplies the risk analysis those works point toward.

Normalization for non-stationarity. RevIN (Kim et al., 2022); NST (Liu et al., 2022), Dish-TS (Fan et al., 2023), SAN (Liu et al., 2023), FAN (Ye et al., 2024) and SIN (Han et al., 2024b) vary which statistics are removed and how; Fu and Yang (2025) document failure modes; Berthelier et al. (2026) find RevIN's affine parameters redundant. The operator takes its name from vision (Ulyanov et al., 2016), where normalization layers (Ioffe and Szegedy, 2015; Ba et al., 2016; Wu and He, 2018) are studied for optimization rather than forecast risk (Santurkar et al., 2018), and the shift it targets is a covariate shift (Quiñonero-Candela et al., 2009). That literature is architectural and empirical; C1–C4 are what we add to it, and we claim only that in the linear case remove-then-restore operators occupy constraint planes whose price and benefit are computed here. A published report of catastrophic RevIN failure on Electricity (Fu and Yang, 2025) appears to contradict our finding that full-IN is harmless there; App. I.1 shows it is a preprocessing artifact that Thm. 3 predicts.

Classical antecedents. Removing a stochastic level is classical, and so is every ingredient of the trade: differencing in ARIMA (Box and Jenkins, 1970), exponential smoothing (Muth, 1960), forecast failure under structural breaks (Clements and Hendry, 1998), restricted least squares (Aitken, 1935; Rao, 1973), shrinkage toward a prior (Hoerl and Kennard, 1970; James and Stein, 1961), and forecast combination (Bates and Granger, 1969), of which SoftIN is one. That literature never trained L-dimensional windowed ridge regressions, which is where the present results lie; App. Y makes the correspondences precise.

3 Setup: instance normalization as a constraint

Single channel; lookback window $x_t = (y_{t-L+1}, \dots, y_t) \in \mathbb{R}^L$; target y_{t+h} (multi-horizon models fit each horizon column independently). A linear forecaster is $\hat{y} = w^\top x$; the mean-only IN forecaster computes $\hat{\mu} = \frac{1}{L}\mathbf{1}^\top x$ and predicts $\hat{y} = v^\top(x - \hat{\mu}\mathbf{1}) + \hat{\mu}$. Population statements use $\Gamma = \mathbb{E}[xx^\top] \succ 0$ and $\gamma_h = \mathbb{E}[x\, y_{t+h}]$ of the zero-mean stationary component; $u = \mathbf{1}/L$, $\Pi = \mathbf{1}\mathbf{1}^\top/L$.

Remark 1 (IN is a constraint)**.** *The effective weights $w = (I - \Pi)v + u$ satisfy $\mathbf{1}^\top w = 1$ for every v, and every w on that plane is attained. Mean-only IN $\equiv$ least squares over $\{w : \mathbf{1}^\top w = 1\}$ (cf. the constrained-regression view in Toner and Darlow, 2024).*

Proposition 1 (what mean-IN ridge optimizes)**.** *Training the mean-IN forecaster with ridge penalty $\lambda\|v\|^2$ (the standard implementation: row-center the design, penalize v) is exactly $\min_{w:\,\mathbf{1}^\top w=1} \|Y - Xw\|^2 + \lambda\|w - u\|^2$, i.e. ridge* centered at the uniform vector *on the sum-to-one plane; the population version replaces the first term by $R(w) = \mathbb{E}(y - w^\top x)^2$. The optima are $w_{\mathrm{raw}}(\lambda) = (\Gamma + \lambda I)^{-1}\gamma_h$ and $w_{\mathrm{in}}(\lambda) = (\Gamma + \lambda I)^{-1}(\gamma_h + \lambda u + \nu\mathbf{1})$ with ν set by the constraint. As $\lambda \to 0$ both reduce to their OLS counterparts.*

Regularization convention. The implementation penalizes summed squared error, so $\lambda_{\mathrm{pop}} = \lambda_{\mathrm{emp}}/N$ (N = rows in the solve); all population-level λ below are λ_{pop}, which also explains the near-flat benchmark MSE across the numeric grid. Proof: App. B, verified against the training code path to 1.5×10^{-14}. Two consequences matter later: the constraint is architectural, so IN's shift invariance holds *exactly* at every λ; and $b(\lambda)$ grows with λ in our examples (AR(1), $\varphi = 0.7$: $0.66 \to 0.88$ as $\lambda : 0 \to 10$), so regularizing the raw model makes it *more* shift-fragile.

3.1 EXACT EQUIVARIANCE FOR ANY DETERMINISTIC BACKBONE

The constraint reading is about the *normalizer*, not the predictor, so one half of the trade survives without linearity.

Theorem 1 (wrapper equivariance). *Let $g : \mathbb{R}^L \to \mathbb{R}^H$ be any deterministic function and let $F(x) = g\big(x - \mu(x)\mathbf{1}\big) + \mu(x)\mathbf{1}_H$ with $\mu(x) = \frac{1}{L}\mathbf{1}^\top x$ be its mean-IN wrapping. Then $F(x + \delta\mathbf{1}) = F(x) + \delta\mathbf{1}_H$ for every x and every $\delta \in \mathbb{R}$.*

Proof. $\mu(x + \delta\mathbf{1}) = \mu(x) + \delta$, so the centered input $(x + \delta\mathbf{1}) - \mu(x + \delta\mathbf{1})\mathbf{1} = x - \mu(x)\mathbf{1}$ is unchanged; hence g receives an identical argument and $F(x+\delta\mathbf{1}) = g(x-\mu(x)\mathbf{1})+(\mu(x)+\delta)\mathbf{1}_H = F(x) + \delta\mathbf{1}_H$. □

No differentiability, continuity or measurability is used; this is wrapper algebra, and it holds for transformers as a theorem rather than an approximation (verified to $\sim 10^{-14}$ on discontinuous and quantized backbones). Hence $\mathrm{Risk}_F(\delta) = \mathrm{Risk}_F(0)$ for all δ, architecture-independently. The proof leaves exactly three ways for a real system to break it: the backbone also receives the raw level (a skip connection or unshifted exogenous input); inference-time randomness depends on the input; or finite precision, float32 cancellation moving the wrapped risk by $\sim 10^{-7}$ rather than exactly zero. The linear analysis thus supplies the *price* of the constraint and Thm. 1 the *benefit* for any backbone, which is why Sec. 8 need only measure the un-normalized model's exposure, and confirms there that across all 108 wrapped backbones the mean-IN risk curve varies by at most 2.1×10^{-7} relative over $\delta \in [-0.6, 0.6]$ while the same architectures' un-normalized risk moves by a median of 60%.

4 THE PRICE AND THE BENEFIT, EXACTLY

At test time an unseen additive shift acts: $x \mapsto x + \delta\mathbf{1}$, $y \mapsto y + \delta$. For fixed w, zero-mean training statistics give $\mathrm{Risk}(w, \delta) = R(w) + \delta^2(1 - \mathbf{1}^\top w)^2$.

Proposition 2 (level-shift dichotomy). *At any fixed $\lambda \geq 0$, with $b(\lambda)$ as above and $\mathrm{price}(\lambda) = R(w_{\mathrm{in}}(\lambda)) - R(w_{\mathrm{raw}}(\lambda))$: (i) $\mathrm{Risk}_{\mathrm{RAW}}(\delta) = \mathrm{Risk}_{\mathrm{RAW}}(0) + \delta^2 b(\lambda)^2$; (ii) $\mathrm{Risk}_{\mathrm{IN}}(\delta)$ is constant in δ; (iii) at $\lambda = 0$, $\mathrm{price}(0) = b^2/a$ with $b = b(0)$, $a = \mathbf{1}^\top\Gamma^{-1}\mathbf{1}$.*

Corollary 1 (regime boundary). *If $b(\lambda) \neq 0$, IN wins iff $\delta^2 > \mathrm{price}(\lambda)/b(\lambda)^2$; at $\lambda = 0$ this is $\delta^2 > 1/a = \mathrm{Var}(\hat{c}_{\mathrm{BLUE}})$:* IN helps exactly when the shift exceeds the statistical error of estimating the level from one window. *If $b(\lambda) = 0$ the raw estimator is itself shift-invariant and the comparison, now δ-independent, is decided by the sign of* $\mathrm{price}(\lambda)$*: positive, RAW wins at every shift; zero, a tie; negative, IN wins at every shift. At $\lambda = 0$, $b = 0$ forces $\mathrm{price}(0) = 0$, an exact tie: persistence driving $b \to 0$ makes price and benefit vanish together. White noise: price σ^2/L, boundary $\delta_\star^2 = \sigma^2/L$.*

Proofs: App. B, verified to machine precision including the ridge crossing point (Fig. 5).

4.1 THE OPTIMAL NORMALIZATION STRENGTH

Corollary 1 is a *decision*, normalize or do not, but nothing forces the decision to be binary. For a raw predictor f and its mean-IN counterpart F, define **SoftIN**, $\hat{y}_\alpha = (1 - \alpha)f(x) + \alpha F(x)$ for $\alpha \in [0, 1]$. The trade admits an exact decomposition, and it needs neither linearity, Gaussianity nor differentiability.

Proposition 3 (optimal normalization strength). *Let $f, F : \mathbb{R}^L \to \mathbb{R}^H$ be any fixed deterministic predictors. In particular, for any backbone g, $f(x) = g(x)$ and $F(x) = g(x-\mu(x)\mathbf{1})+\mu(x)\mathbf{1}_H$ with $\mu(x) = L^{-1}\mathbf{1}^\top x$ (Thm. 1). Write $\Delta = F - f$ for the* normalization correction, *$e_{\mathrm{RAW}} = y - f(x)$, $e_{\mathrm{IN}} = y - F(x)$, and*

$$B = \mathbb{E}\langle e_{\mathrm{RAW}}, \Delta\rangle, \qquad P = \mathbb{E}\|\Delta\|^2 - B = -\mathbb{E}\langle e_{\mathrm{IN}}, \Delta\rangle, \qquad 0 < \mathbb{E}\|\Delta\|^2 < \infty.$$

Then $\mathrm{Risk}(\alpha) = \mathrm{Risk}_{\mathrm{RAW}} - 2\alpha B + \alpha^2(P + B)$, exactly quadratic. Consequently (i) the minimizer over $[0, 1]$ is interior iff *$P > 0$ and $B > 0$, and then $\alpha^\star = B/(P + B) = 1/(1 + P/B)$; (ii) its excess over $\mathrm{Risk}_{\mathrm{RAW}} - B$ is the* parallel combination *$PB/(P+B)$, below both endpoints; and (iii) its gain over the better endpoint is $\min(P, B)^2/(P+B)$, which for fixed P is maximized at $B = P$, where it equals half the excess.*

For the level-shift model of Prop. 2 at $\lambda = 0$, the normal equations give $P = b^2/a$ and $B = b^2\tau^2$ exactly, so $\alpha^\star = \tau^2/(\tau^2+\delta_\star^2)$: the shift-to-boundary ratio, which needs only train-window moments and *unlabeled* inputs and is therefore available before any label is seen. Away from that model the quadratic is still exact and still free at inference, since three evaluations fix the whole curve with no retraining. That is how Sec. 8 measures $\alpha^\star$ on trained transformers. Calibrating B without labels is open, and we mark it as such. Proof, the linear reduction and machine-precision checks are in App. Q; Fig. 1a plots the linear envelope.

Any known nuisance subspace. The level is one direction, $\mathbf{1}$; everything above is the k=1 case. Let $U \in \mathbb{R}^{L\times k}$ have full column rank, spanning the directions a designer wants invariance to (level, trend, seasonal components), and let ψ be the target response. The normalizer that removes $\hat{c}(x) = (U^\top U)^{-1}U^\top x$ and restores $\psi^\top\hat{c}(x)$ realizes exactly $\{w : U^\top w = \psi\}$, is pathwise invariant to $x \mapsto x+Uc$, $y \mapsto y+\psi^\top c$, and pays $\text{price}_U = b_U^\top A_U^{-1} b_U$ with $A_U = U^\top\Gamma^{-1}U$, $b_U = \psi - U^\top w_{\text{raw}}$ (Thm. 4, App. C). The win region is the complement of a *slab*, not an ellipsoid, and the random-shift criterion $b_U^\top(M_c - A_U^{-1})b_U > 0$ turns on a *second* moment, which is what the benchmark diagnostic estimates. $U = \mathbf{1}$, $\psi = 1$ recovers this section, and A_U^{-1} is again a GLS covariance (Aitken, 1935). The experiments exercise only k=1.

Drift, differencing, self-normalization. Giving the level dynamics, in the local-level model $\mu_t = \mu_{t-1} + \eta_t$, adds three facts (Prop. 4, App. D, verified over a 72-configuration grid to 0.83% of Monte Carlo). A fixed forecaster with weight sum s deployed m steps after training pays an extra $(1-s)^2q^2m$, so shift exposure and staleness multiply. The differencing parameterization $\hat{y} = y_t + u_d^\top Dx$ (D the first-difference matrix) is an affine bijection onto the sum-to-one plane, so *IN and differencing are the same hypothesis class* with identical population optima, differing only as estimators. And $|1 - \mathbf{1}^\top w(V)| = \Theta(V^{-1})$ as level dispersion V grows: unconstrained regression *self-normalizes*, which is why the diagnostic reads near-ties on integrated series.

5 Finite samples: estimation error creates shift fragility

So far shift exposure has been b^2, a property of the population optimum. With estimated weights a second source appears: the fitted weight sum is random, so it misses one even when the population optimum does not. Estimate weights by (constrained) OLS on T windows, idealized as i.i.d. draws $x \sim \mathcal{N}(0,\Gamma)$; σ_r^2, σ_{in}^2 are the population residual variances of the raw and constrained optima.

Theorem 2 (finite sample: estimation noise creates shift fragility)**.** *Let the T training pairs be i.i.d. draws with (x,y) jointly Gaussian, $x \sim \mathcal{N}(0,\Gamma)$, and $T > L+1$. (i) $\mathbb{E}\,\text{Risk}_{\text{RAW}}(\delta,T) = \sigma_r^2 + \delta^2b^2 + \sigma_r^2\frac{L+\delta^2a}{T-L-1}$: the shift coefficient inflates from b^2 to $b^2 + \sigma_r^2a/(T-L-1) > b^2$, positive even when $b = 0$. (ii) $\mathbb{E}\,\text{Risk}_{\text{IN}}(T) = \sigma_{in}^2\big(1 + \frac{L-1}{T-L}\big)$, independent of δ. (iii) The boundary $\delta_\star^2(T)$ is explicit and collapses under persistence: at AR(1) $\varphi = 0.99$, $L = 24$, $h = 1$ it is 45.0 ($T \to \infty$), 12.3 (T=400), and negative at T=100, so the constrained estimator dominates at any shift.*

Proof: App. E; verified at every (T,δ) tested. **Overlap bridge.** Across 60 configurations of overlapping windows from single AR(1) paths the *excess-risk* formulas transfer essentially unchanged (implied effective sample size 0.96–$1.09 \times T$), while the shift-fragility coefficient runs systematically *larger* than the i.i.d. prediction (+8% at φ=0.5, +15% at 0.9, +124% at 0.99), so the empirical boundary sits at or below the theoretical one (App. H); the i.i.d. formula is *conservative* here and we claim no more. This resolves the tension left by Cor. 1: population theory says IN's stakes vanish on persistent data, yet practitioners find it indispensable precisely there. At realistic T the estimation term, not b^2, controls fragility.

6 The standard-deviation division: reweighting, not rescaling

Full RevIN also divides by the window standard deviation $\hat{\sigma}_i$ (we analyze the affine-free version; Berthelier et al., 2026 find the affine parameters redundant).

Theorem 3 (std division: reweighting, not rescaling)**.** *(i) For fixed weights, full-IN and mean-IN predictions are identical: $\hat{\sigma}(v^\top\frac{x-\hat{\mu}\mathbf{1}}{\hat{\sigma}}) + \hat{\mu} = v^\top(x - \hat{\mu}\mathbf{1}) + \hat{\mu}$. (ii) Full-IN ridge training is exactly mean-IN ridge training under weighted least squares with weights $1/\hat{\sigma}_i^2$ (identical normal equations, same penalty). (iii) If windows share a shape law with per-window scale s_i and centered*

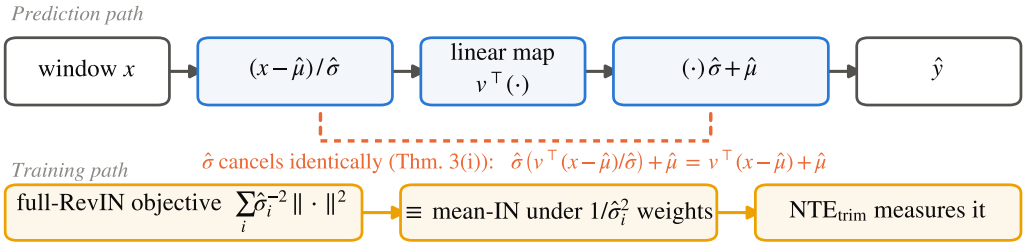

Figure 2: **What the standard-deviation division does** (Thm. 3). *Prediction path:* the division and the later multiplication are inverse operations on the same linear map, so $\hat{\sigma}$ cancels identically, and de-normalization cannot amplify test error, which rules out the standard account of RevIN's failures (Cor. 2). *Training path:* the division survives, as weighted least squares with weights $1/\hat{\sigma}_i^2$, the Aitken/GLS weighting when residual variance scales with the window, a misspecification when it does not.

residual variance $\propto s_i^2$ (multiplicative regime), this weighting is the Aitken/GLS weighting (Aitken, 1935) and full-IN is more efficient; if residual variance is constant (additive regime), the unweighted estimator is efficient (Gauss–Markov) and the weighting is misspecified.

Remark 2 (inert constants, toxic near-constants)**.** *An exactly constant window has centered input* 0*: despite unbounded weight it contributes* nothing *to the normal equations. The dangerous windows are* near*-constant ones whose futures move; their normalized targets explode and dominate the weighted fit.*

This gives a train-only diagnostic, the **trimmed normalized-target energy**

$$\mathrm{NTE}_{\mathrm{trim}} = \mathrm{mean}_{i:\hat{\sigma}_i>\tau}\ \tfrac{1}{H}\|(Y_i-\hat{\mu}_i)/\hat{\sigma}_i\|^2,$$

the scale of the full-IN training objective on non-inert windows (the untrimmed variant is ϵ-guard-dominated and reported separately; the ranking is identical at $\tau \in \{5\times10^{-5}, 10^{-4}, 2\times10^{-4}\}$). Two findings (App. I, Fig. 1b): weight concentration does *not* predict failure, while $\mathrm{NTE}_{\mathrm{trim}}$ ranks all eight datasets' horizon-averaged $\mathrm{MSE}_{\text{full-IN}} - \mathrm{MSE}_{\text{mean-IN}}$ and separates the two catastrophes (Exchange $\approx$300, Weather $\approx$1200) from every other dataset (≤ 3.5). Neither sufficiency nor necessity is claimed.

Corollary 2 (folklore correction)**.** *By (i), the usual explanation of RevIN failures, de-normalization amplifying test error, cannot operate in linear models: the catastrophes are training-side reweighting.*

Proofs: App. F; (i)–(ii) hold to machine precision and both branches of (iii) realize on synthetic data.

7 A DEPLOYMENT-TIME DIAGNOSTIC, TESTED ON EIGHT BENCHMARKS

Protocol. Standard LTSF conventions ($L = 336$, four horizons fit independently, chronological splits, train-statistic z-scoring, one shared channel-independent weight matrix, ridge λ on validation MSE; audit table in App. J). Every estimator is a closed-form solve, so the suite runs on a laptop CPU in under an hour, and test targets never enter fitting or diagnostics. At H=96 the fits sit within 0.005 of published DLinear on ETTh1, Electricity and Traffic; all $8 \times 4 \times 4$ values are in App. K.

Predicted versus realized normalization gain (Fig. 10, App. L)**.** For every channel of *all eight* datasets and every horizon (4,960 channel$\times$horizon observations) we predict the RAW$-$IN test-MSE gap with the exact ridge diagnostic $\hat{G}_\lambda = \hat{b}_\lambda^2\hat{\delta}^2 - \widehat{\mathrm{price}}_\lambda$: $\hat{b}_\lambda$ from the fitted shared raw-ridge weight sums (Prop. 1), $\widehat{\mathrm{price}}_\lambda$ the exact ridge price from train-window moments, and $\hat{\delta}^2$ the mean squared test-*input* window mean. No test targets and no IN training enter it; an automated poisoning test replaces the test targets with noise and verifies every train-side quantity is unchanged. Results: $r = 0.90$ at H=96 (dataset-clustered 95% CI 0.58–0.94), $0.88/0.82/0.62$ at the longer horizons, pooled 0.66, beating both ablations. Calibration is imperfect: slope $\approx$0.4, sign accuracy outside the tie region 0.69–0.74, and predicted ties realize mean $|\mathrm{gap}| \leq 0.008$. On integrated series the diagnostic works through self-normalization: Exchange's fitted weight sums are $\approx$0.93, driving

$\hat{b}_\lambda^2 \approx 0$ and correctly predicting near-ties. We cluster by dataset throughout, since one weight matrix is shared across a dataset's channels.

Decisions and horizons. As a model-selection gate over the 32 cells, a strict train + validation-only rule attains the lowest mean regret against the oracle in our suite (0.0007 against 0.0014 for validation-MSE selection and 0.0355 for always-RAW), lowest-in-suite but not a proven improvement. The RAW−IN gap is monotone in H on all eight datasets, increasing on five and decreasing on three (Table 9), the two regimes the theory admits, compounding shift exposure against compounding price, and an *independent* assignment from train and validation data alone matches on 6 of 8, both mismatches being datasets whose cells are predominantly ties. Full RevIN collapses on exactly the two datasets whose NTE is two orders of magnitude above the rest.

8 A PREREGISTERED TEST ON TRANSFORMERS

Whether the *mechanism*, shift-defect × shift-magnitude, is operative in deep forecasters is an empirical question, so every deep run here was preceded by a frozen preregistration (Nosek et al., 2018): hypotheses, the primary statistic and its clustering, a trivial-baseline gauntlet, hard control gates, and the outcome text for each verdict.[1] Three tiers ran, answering three different questions. **Tier 1** (324 runs: 2 nonlinear backbones × 6 datasets × 3 variants × 2 seeds × 4 horizons, plus a DLinear control) asks whether $\hat{G}$ *ranks* normalization's benefit. **Tier 2** (16 cells) adds the two datasets Tier 1 omits and asks whether it predicts the *sign*. **Tier 3** (20 cells) tests Prop. 3 on the resulting checkpoints. Tier 2 was declared optional in the Tier-1 preregistration, which also forbids aggregating it with the Tier-1 statistic; we honor that, and no verdict is recomputed across tiers.

The predictor is measured on the *un-normalized* model only, since Thm. 1 makes the mean-IN arm exactly shift-invariant, so only the raw arm's exposure need be measured: $s(x) = \nabla f(x)^\top \mathbf{1}$ by one forward-mode JVP, $\hat{S} = \mathbb{E}[(1-s)^2]$ over ≥ 2000 fixed test *input* windows, $\hat{\delta}^2$ the mean squared test-input window mean, $\hat{G} = \hat{S}\hat{\delta}^2$. No test targets enter any predictor, and an automated leakage test verifies it. DLinear is a control carrying zero evidential weight: every verdict criterion uses the two nonlinear backbones only (App. O, App. R).

Tier 1. H1 is STRONG at all four horizons (Table 1, top; Fig. 3). Every preregistered gate passes at every horizon, and in the frozen gauntlet $\hat{G}$ leads, tying only its own monotone reparameterization $\hat{S}\sqrt{\hat{\delta}^2}$ and beating the operational baseline by a factor of two. In the preregistered STRONG wording: a statistic derived from an exactly-solvable linear theory, computed from an un-normalized model and unlabeled deployment inputs, ranks the benefit of mean normalization across PatchTST and iTransformer. We do not claim the linear risk theorem holds for these architectures, only that the mechanism it identifies is operative in them. Two qualifications belong beside that. *The exponent is not identified*: $\hat{S}\hat{\delta}^{2\alpha}$ scores 0.92–0.98 for all $\alpha \in [0.75, 2]$, so we claim the product form, not a power. *The bootstrap is coarse*: with six clusters every upper limit is +1.000, so the CIs establish exclusion of zero, not a sharp interval. A post-hoc battery is in App. S; its weakest entry is +0.85.

Magnitude, not sign. Every Tier-1 cell in Fig. 3a lies above the line, so on that grid $\hat{G}$ is asked how *much* normalization helps, not *whether* it does, a property of the six datasets Tier 1 covers, not of the statistic. Exploratory evidence locates the boundary within Tier 1 (the three datasets whose *linear* gap is negative are exactly the three lowest-$\hat{G}$ datasets for both backbones; App. S), but testing it needs Tier 2.

Tier 2. The negative regime exists, and $\hat{G}$ does not predict it. On Traffic, iTransformer *loses* 0.047 test MSE to mean normalization (Table 1, row 4): in both seeds, in five of five once Tier 3 adds three more (mean −0.0430), and at every horizon tested (−0.0430/ − 0.0387/ − 0.0273 at H=96/192/720). Before this run all 12 seed-averaged nonlinear cells favored normalization. The cell is genuinely price-dominated: re-evaluating RAW at the shift minimizing *its own* risk makes normalization look worse still (−0.0468 → −0.0555), the signature of a premium paid rather than a benefit unmeasured. Note that *architecture, not dataset, decides the sign*; PatchTST gains

[1] Tier 1 frozen 2026-08-13, sha256 `c7ec02740b83...`; Tier 2 `acfdf22c...`; Tier 3 `d85b999c...`. The Tier-2 and Tier-3 hashes were stamped into their run logs with zero result rows in existence; the Tier-1 artifact label `prereg=b9f113e238dd` is not a hash of that file (App. P, D12). All three frozen files ship with the supplement, and every departure is logged gate-by-gate in App. P and App. R.

Table 1: **Three preregistered experiments.** *Top (Tier 1; n=12 seed-averaged nonlinear cells per horizon).* Spearman ρ between $\hat{G}$ and the realized gain $\mathrm{MSE}_{\mathrm{RAW}} - \mathrm{MSE}_{\text{mean-IN}}$, dataset-clustered bootstrap CIs. The preregistered gate asks for $\rho \geq 0.5$, a CI excluding 0, $\rho > 0$ for *both* backbones, and no cap from the frozen five-predictor gauntlet; whose atheoretical operational baseline (RAW validation MSE) is shown alongside. Random-predictor calibration, 20,000 draws: $|\rho|_{95} = 0.58$. *Bottom (Tiers 2–3, H=96, seed-averaged).* Row 4 is the first negative seed-averaged nonlinear cell in this study, and $\hat{G}$ ranks it *highest* of the four, the preregistered P1 failure. Last two columns: the Tier-3 α-sweep, interior in 8 of 8 cell-seeds.

Tier 1. Does $\hat{G}$ rank the benefit?

H	ρ	clustered 95% CI	PatchTST	iTrans.	RAW val. MSE
96	+0.9441	[+0.791, +1.000]	+0.9429	+0.9429	+0.4615
192	+0.9580	[+0.863, +1.000]	+1.0000	+0.8857	+0.4126
336	+0.9650	[+0.863, +1.000]	+1.0000	+0.9429	+0.3846
720	+0.9510	[+0.783, +1.000]	+0.9429	+0.8286	+0.2937

Tier 2. Does it predict the sign? **Tier 3.** Is the binary choice optimal?

backbone	dataset	$\hat{S}$	$\hat{\delta}^2$	$\hat{G}$	gain	$\alpha^\star$	gain at $\alpha^\star$
PatchTST	Electricity	0.041	0.201	0.0082	+0.0046	0.78	+0.0004
PatchTST	Traffic	0.084	0.109	0.0091	+0.0288	0.95	+0.0001
iTransformer	Electricity	0.073	0.201	0.0147	+0.0095	0.83	+0.0003
iTransformer	Traffic	**0.564**	0.109	**0.0616**	**−0.0468**	**0.40**	**+0.0364**

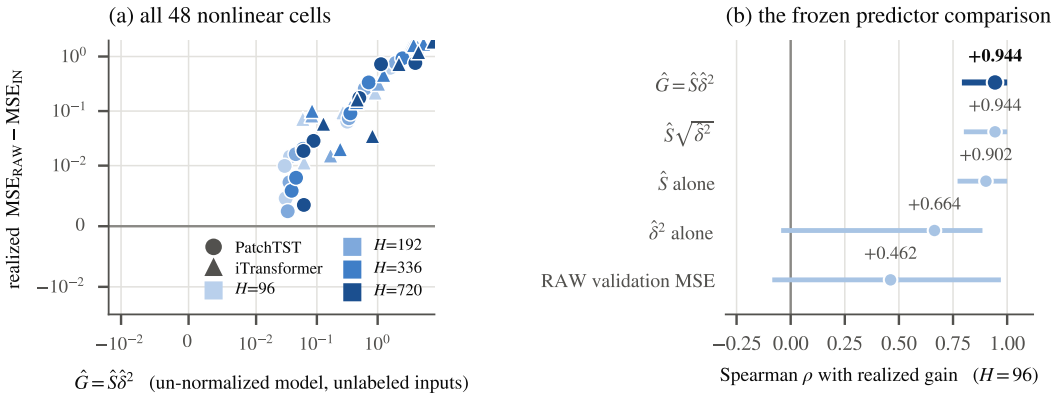

Figure 3: **Tier 1: the preregistered transfer test.** (a) All 48 seed-averaged nonlinear cells: $\hat{G}$, computed from a trained *un-normalized* network and *unlabeled* inputs, against the realized mean-IN gain, monotone across nearly three decades. Both axes are symlog with a linear window around zero, which also makes the central limitation visible: every cell lies above the line. (b) The frozen gauntlet at H=96: $\hat{G}$ leads, ties only its own monotone reparameterization, and beats RAW validation MSE by a factor of two. Bars are dataset-clustered 95% CIs.

+0.0288 on the same data, so no dataset-level statistic can succeed here by construction. Two of the three frozen Tier-2 predictions fail, at their literal rates: P1 predicted $\hat{G}$ for a negative cell would fall below every Tier-1 value for that backbone, and it is instead the highest of the four (0.0616 against the frozen minimum 0.0587), with within-tier $\mathrm{Spearman}(\hat{G}, \text{gain}) = -0.40$; P2 fails on PatchTST/Traffic; P3, that some cell would be non-positive, holds. The failure is structural rather than statistical, and the theory identifies the reason: $\hat{G}$ tracks the benefit B and carries no counterpart for the price P, which Prop. 3 shows is a separate functional of the normalization correction. Ranking within a regime is therefore all $\hat{G}$ can be asked for, and that is what we claim.

Tier 3. The soft rule beats both endpoints, and the decomposition explains why. Sweeping α costs nothing beyond the two prediction sets already computed, and the optimum is interior in 8 of 8 cell-seeds. That is not automatic: Prop. 3(i) makes interiority equivalent to $P > 0$ and $B > 0$, which fails on the synthetic backbones of App. Q. Inverting each measured quadratic recovers the pair, and the ratio reproduces the swept optimum in 8/8 cell-seeds (Table 13): P/B is 0.05–0.33 on the

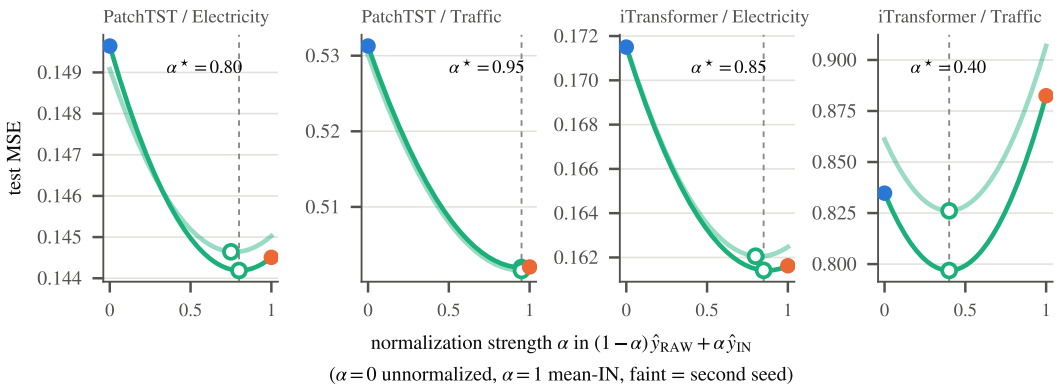

Figure 4: **Tier 3: the soft optimum is interior on every deep cell tested.** Test MSE against normalization strength α, all four Tier-2 cells and both seeds, swept on saved test predictions with no retraining. All eight optima are interior; by Prop. 3(i) that is *equivalent* to $P > 0$ and $B > 0$, which hold in $8/8$ and need not have. The price-dominated cell (right) puts $\alpha^\star$ at 0.40 in both seeds, gaining 0.036 MSE over the better endpoint.

six benefit-dominated cell-seeds and 1.50 on both seeds of the price-dominated one, giving $\alpha^\star = 1/(1+P/B) = 0.40$ exactly, and the identity $P - B = \text{MSE}_{\text{mean-IN}} - \text{MSE}_{\text{RAW}}$ closes to 10^{-4}. *This is a decomposition on the labeled test set, not an out-of-sample prediction.* The improvements track Prop. 3(iii): $+0.0364$ MSE on the cell nearest $P = B$ against $+0.0001$–$+0.0004$ where one side dominates. Tier 3 also confirms Thm. 1 end-to-end (mean-IN risk flat in δ to 3.1×10^{-8}) and resolves the float32 control: the equivariance defect is 1.1×10^{-14} in float64 against 4.6×10^{-4} in float32.

Out-of-sample estimation. The same rule without test labels (post hoc, after all three tiers, not preregistered). P and B are residual inner products, so both can be estimated on the *validation* split. Reloading the Tier-2 checkpoints gives two deployment-time rules: (a) takes both terms from validation; (b) is a hybrid taking P from validation and $B = \hat{G} = \hat{S}\hat{\delta}^2$ from *unlabeled test inputs*. Neither touches a test target.

rule	mean test MSE	mean regret vs. oracle $\alpha^\star$
always RAW	0.42486	0.02001
always mean-IN, and binary chosen on validation	0.42584	0.02099
SoftIN, α from validation (a)	0.40747	0.00261
SoftIN, hybrid $\hat{G}/(P_{\text{val}} + \hat{G})$ (b)	0.40607	**0.00121**

The soft rule cuts regret $8\times$ (a) or $17\times$ (b) against the best binary rule, and the binary rule *chosen on validation* is the worst of the four, because validation prefers mean-IN on every cell, including the one where test prefers RAW. Caveats: n=8, two datasets, and (b) was chosen after seeing (a).

Remaining controls. Two secondary hypotheses land short of their thresholds and one Tier-3 prediction is not evaluable, all pre-committed possibilities; three control gates are exceeded literally, each diagnosed rather than waived. Apps. R.2, P and R give every rate and every gate.

9 Limitations and Conclusion

Thm. 2 is exact only for i.i.d. Gaussian designs; dependence is simulated. Thm. 4 is population-level and needs the nuisance subspace known, fixed and full-rank. NTE does *not* rank degradation for deep backbones, and the linear diagnostic's calibration is rough (slope ≈ 0.4). Prop. 3 holds for any backbone, but the 8-cell agreement in Sec. 8 decomposes labeled test data, and the rules beside it are post hoc over n=8 cell-seeds on two datasets, taking the price from a validation split whose shift statistics need not match deployment's. A fully label-free rule stays open: $\hat{G}$ in place of B transfers in ranking, not in calibration. Learned affine RevIN parameters and cross-channel normalizers are out of scope. **Conclusion.** Instance normalization is a constraint, constraints have prices, and writing this one down turns the choice into a pre-deployment decision that was never binary.

Reproducibility statement

All results regenerate from released code: seeded exact-formula verification scripts, benchmark runners, analyses and figure generation, behind a single entry point `reproduce_paper.py` (package versions and dataset provenance in App. Z). The deep-forecaster experiments ship as a separate harness with all three frozen preregistrations, the 324-row Tier-1 artifact and the Tier-2/3 row artifacts, the risk curves, the probe output, and the analysis scripts that regenerate Table 1 and App. S verbatim. The supplementary archive submitted with this paper contains the Tier-1 frozen preregistration and the two Tier-2/3 prediction files, whose sha256 values recompute to those stamped into their run logs before any result row existed (the Tier-1 artifact label is a fixed string, not a run-time hash; App. P, D12). It also contains the consolidated deviations log, the per-cell CSVs behind Table 1 and Fig. 4, the Tier-2/3 run and analysis scripts, the figure generator, and `verify_soft.py`, which checks Prop. 3 against a grid search and exits non-zero on any failure. Datasets are the public LTSF benchmarks with standard splits.

Use of large language models

Large language models were used solely to assist with language polishing, including improvements to grammar, clarity, and readability. They were not used to develop the research ideas, formulate mathematical claims, construct proofs, design experiments, analyze data, or interpret results. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

References

A. C. Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1935.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. arXiv:1607.06450, 2016.

J. M. Bates and C. W. J. Granger. The combination of forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.

G. Berthelier, T. Nabil, E. Le Naour, R. Niamke, S. Perlaza, and G. Neglia. On the role of reversible instance normalization. arXiv:2603.11869, 2026.

G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

L. Brigato, R. Morand, K. J. Strømmen, M. Panagiotou, M. Schmidt, and S. Mougiakakou. There are no champions in supervised long-term time series forecasting. *Transactions on Machine Learning Research*, 2026. arXiv:2502.14045.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.

S.-A. Chen, C.-L. Li, N. Yoder, S. Ö. Arik, and T. Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023. arXiv:2303.06053.

M. P. Clements and D. F. Hendry. *Forecasting Economic Time Series*. Cambridge University Press, 1998.

A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu. Long-term forecasting with TiDE: Time-series dense encoder. arXiv:2304.08424, 2023.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

W. Fan, P. Wang, D. Wang, D. Wang, Y. Zhou, and Y. Fu. Dish-TS: A general paradigm for alleviating distribution shift in time series forecasting. AAAI, 2023.

F. Fu and Y. Yang. Noise or signal? Deconstructing contradictions and an adaptive remedy for reversible normalization in time series forecasting. arXiv:2510.04667, 2025.

A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, 2nd edition, 2008.

L. Han, H.-J. Ye, and D.-C. Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2024. arXiv:2304.05206.

L. Han, H.-J. Ye, and D.-C. Zhan. SIN: Selective and interpretable normalization for long-term time series forecasting. ICML (PMLR 235), 2024.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

L. Huang, J. Xu, and L. Darlow. How good can linear models be for time-series forecasting? arXiv:2606.27282, 2026.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML (PMLR 37), 2015.

W. James and C. Stein. Estimation with quadratic loss. *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379, 1961.

T. Kim, J. Kim, Y. Tae, C. Park, J.-H. Choi, and J. Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. ICLR, 2022.

L. Kish. *Survey Sampling*. Wiley, 1965.

Z. Li, S. Qi, Y. Li, and Z. Xu. Revisiting long-term time series forecasting: An investigation on linear mapping. arXiv:2305.10721, 2023.

S. Lin, W. Lin, W. Wu, H. Chen, and J. Yang. SparseTSF: Modeling long-term time series forecasting with 1k parameters. ICML, 2024.

Y. Liu, H. Wu, J. Wang, and M. Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. NeurIPS, 2022.

Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long. iTransformer: Inverted transformers are effective for time series forecasting. ICLR, 2024.

Z. Liu, M. Cheng, Z. Li, Z. Huang, Q. Liu, Y. Xie, and E. Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. NeurIPS, 2023.

J. F. Muth. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55(290):299–306, 1960.

B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. The preregistration revolution. *PNAS*, 115(11):2600–2606, 2018.

Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. ICLR, 2023.

B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. ICLR, 2020.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, 2009.

C. R. Rao. *Linear Statistical Inference and Its Applications*. Wiley, 2nd edition, 1973.

D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. How does batch normalization help optimization? NeurIPS, 2018.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

W. Toner and L. Darlow. An analysis of linear time series forecasting models. ICML, 2024.

Z. Wang, K. Zhang, W. Chen, X. Lu, L. Li, and T. Schlagenhauf. Characteristic root analysis and regularization for linear time series forecasting. ICLR, 2026. arXiv:2509.23597.

D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022, 2016.

H. Wu, J. Xu, J. Wang, and M. Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. NeurIPS, 2021.

H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. ICLR, 2023.

Y. Wu and K. He. Group normalization. ECCV, 2018.

Z. Xu, A. Zeng, and Q. Xu. FITS: Modeling time series with 10k parameters. ICLR, 2024.

W. Ye, S. Deng, Q. Zou, and N. Gui. Frequency adaptive normalization for non-stationary time series forecasting. NeurIPS, 2024.

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are transformers effective for time series forecasting? AAAI, 2023.

H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. AAAI, 2021.

T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. ICML, 2022.

APPENDIX

A Notation **15**

B Proofs for Section 4 and the ridge formulation **15**

B.1 The empirical identity behind Proposition 1 15

B.2 Differencing under ridge has a different geometry 16

B.3 Behavior of the ridge price and boundary 16

B.4 Proof of Proposition 2 and Corollary 1 17

B.5 Verification summary for Propositions 1–2 18

C The nuisance-subspace theorem: statement, proofs, extensions, verification **18**

C.1 (i) Architecture ≡ constraint 18

C.2 (ii) Shift risk and the role of the zero-mean law 18

C.3 (iii) The price 19

C.4 (iv) Geometry: a slab, and a second moment 19

C.5 BLUE, generalized ridge, and edge cases 19

D Drift and self-normalization: statement and proofs **19**

E Proof of Theorem 2 **20**

F Proofs for Section 6; the NTE diagnostic **20**

G Synthetic verification details **21**

H Overlapping-window study **21**

I Effective sample size versus NTE **22**

I.1 Resolving the reported Electricity failure 22

J Benchmark protocol audit (linear experiments) **24**

K Complete benchmark results **24**

L Regime-map robustness and per-channel data **26**

M Ridge-grid sensitivity **27**

N Self-normalization and weight profiles **28**

O Deep-forecaster experiment: preregistered protocol **28**

P Preregistration deviations log **29**

P.1 Post-freeze gate run: the step-size cross-check premise, verified 29

P.2 Post-freeze gate run: the leakage (poison) check 29

Q Proposition 3: proof, linear reduction, verification **32**

R Tier-2 sign test and Tier-3 soft-rule test: protocol, verdicts, deviations **33**

R.1 The Tier-3 price/benefit decomposition 34

R.2 Secondary hypotheses and controls, at their literal rates 34

S H1 post-hoc robustness **36**

T Worked examples **38**

U Verification transcripts **38**

V Estimator pseudocode **40**

W Dataset details **41**

X Practical guidance **41**

Y Extended classical background **41**

Z Reproducibility **42**

Reader's guide. The appendix is organized so that no reviewer needs to read all of it. **To check the theory**, App. B–F give every proof, each followed by the machine verification of that proof; App. T works the running examples end to end. **To check the linear experiments**, App. J is the protocol audit, App. K the complete $8 \times 4 \times 4$ results, and App. L the robustness battery and provenance. **To check the preregistered deep experiment**, read App. O (the frozen protocol, including the verdict thresholds fixed in advance), then App. P (every gate that failed, went unexecuted, or was reworded, with the frozen rule quoted verbatim), then App. S (post-hoc robustness, labelled as such throughout). **To use the results**, App. V is estimator pseudocode and App. X the practical guidance.

A NOTATION

$x_t \in \mathbb{R}^L$, y_{t+h}	lookback window; horizon-h target
L, H, T	lookback length; horizon; number of training windows
w, v	effective weights; normalized-space weights ($w = (I - \Pi)v + u$)
$u = \mathbf{1}/L$, $\Pi = \mathbf{1}\mathbf{1}^\top/L$	uniform vector; window-mean projector
Γ, γ_h, γ_0	window autocovariance; cross-covariance; target variance
$R(w)$	population MSE $\mathbb{E}(y - w^\top x)^2$ at $\delta = 0$
$b(\lambda) = 1 - \mathbf{1}^\top w_{\text{raw}}(\lambda)$	raw model's deviation from unit weight sum ($b = b(0)$)
$a = \mathbf{1}^\top \Gamma^{-1}\mathbf{1}$	inverse BLUE variance of the window level
δ; $\delta_\star^2$; $\tau^2 = \mathbb{E}[\delta^2]$	unseen additive level shift; regime boundary; shift second moment
α; $\hat{y}_\alpha$	normalization strength; SoftIN predictor $(1 - \alpha)f + \alpha F$
$\Delta = F - f$	normalization correction (Prop. 3)
$B = \mathbb{E}\langle e_{\text{RAW}}, \Delta\rangle$; $P = \mathbb{E}\|\Delta\|^2 - B$	benefit; price. Linear case: $B = b^2\tau^2$, $P = b^2/a$
m	steps elapsed between training and deployment
q^2, σ^2, V	drift-increment variance; noise variance; level dispersion
$K = q^2 M + \sigma^2 I$, $M_{jk} = \min(j, k)$	local-level window covariance (level removed)
$D \in \mathbb{R}^{(L-1)\times L}$, e_L	first-difference matrix; last-coordinate basis vector
σ_r^2, σ_{in}^2	population residual variances (raw; constrained)
$\hat{\mu}_i$, $\hat{\sigma}_i$, $w_i = \hat{\sigma}_i^{-2}$	window mean/std; full-IN training weight
$T_{\text{eff}} = (\sum_i w_i)^2 / \sum_i w_i^2$	Kish effective sample size of the weights
NTE	normalized-target energy, $\text{mean}_i\|(Y_i - \hat{\mu}_i)/\hat{\sigma}_i\|^2$ (trimmed)
$U \in \mathbb{R}^{L\times k}$, $\psi \in \mathbb{R}^k$	nuisance directions (full column rank); target nuisance response
$b_U = \psi - U^\top w$; $A_U = U^\top \Gamma^{-1} U$	generalized mismatch; nuisance information matrix
Π_U, $u_U = U(U^\top U)^{-1}\psi$	projector onto $\text{range}(U)$; minimum-norm feasible weights
$M_c = \mathbb{E}[cc^\top]$	second moment of the random nuisance shift
deep-forecaster experiment (Sec. 8)	
f, g, F	un-normalized backbone; wrapped backbone's inner map; mean-IN wrapping
$s(x) = \nabla f(x)^\top \mathbf{1}$	JVP of the backbone along the all-ones direction
$\hat{S} = \mathbb{E}[(1 - s)^2]$, $\hat{\delta}^2$	shift-defect statistic; mean squared test-input window mean
$\hat{G} = \hat{S}\hat{\delta}^2$	preregistered H1 predictor
c_1, c_2	first/second-order coefficients of $\text{Risk}(\delta)$ about $\delta = 0$

B PROOFS FOR SECTION 4 AND THE RIDGE FORMULATION

B.1 THE EMPIRICAL IDENTITY BEHIND PROPOSITION 1

The implemented mean-IN estimator row-centers the design and penalizes the normalized-space weights:

$$\hat{v} = \arg\min_v \ \|Y_c - X_c v\|_F^2 + \lambda\|v\|^2, \qquad X_c = X - \hat{\mu}\mathbf{1}^\top = X(I - \Pi), \quad Y_c = Y - \hat{\mu}, \tag{1}$$

with predictions $\hat{Y} = X_c\hat{v} + \hat{\mu}$ (each row's own $\hat{\mu}_i$).

Step 1 ($\hat{v}$ lies in $\mathbf{1}^\perp$). The normal equations are $(X_c^\top X_c + \lambda I)\hat{v} = X_c^\top Y_c$. Since $X_c\mathbf{1} = X(I - \Pi)\mathbf{1} = 0$, we have $X_c^\top X_c\mathbf{1} = 0$ and $\mathbf{1}^\top X_c^\top Y_c = 0$. Multiplying by $\mathbf{1}^\top$ gives $\lambda\mathbf{1}^\top\hat{v} = 0$, hence $\mathbf{1}^\top\hat{v} = 0$ for every $\lambda > 0$.

Step 2 (rewriting on the constraint plane). For $v \in \mathbf{1}^\perp$, put $w = u + v$, so $\mathbf{1}^\top w = 1$. Then $X_c v = Xv - \hat{\mu}(\mathbf{1}^\top v) = Xv$ and $Y_c - X_c v = Y - \hat{\mu} - Xv = Y - Xu - Xv = Y - Xw$, using $\hat{\mu} = Xu$ rowwise; also $\|v\|^2 = \|w - u\|^2$. Hence (1) equals

$$\min_{w:\,\mathbf{1}^\top w = 1} \|Y - Xw\|_F^2 + \lambda\|w - u\|^2. \tag{2}$$

Step 3 (the fiber). Each w on the plane is reached by $v = w - u \in \mathbf{1}^\perp$; general solutions $v = w - u + c\mathbf{1}$ have $\|v\|^2 = \|w - u\|^2 + c^2 L$, so the minimal-penalty representative is $c = 0$.

Step 4 (population KKT). Replacing $\|Y - Xw\|^2$ by $R(w) = \gamma_0 - 2w^\top\gamma_h + w^\top\Gamma w$ in (2), stationarity gives

$$w_{\text{in}}(\lambda) = (\Gamma + \lambda I)^{-1}(\gamma_h + \lambda u + \nu\mathbf{1}), \qquad \nu = \frac{1 - \mathbf{1}^\top(\Gamma + \lambda I)^{-1}(\gamma_h + \lambda u)}{\mathbf{1}^\top(\Gamma + \lambda I)^{-1}\mathbf{1}},$$

with $w_{\text{raw}}(\lambda) = (\Gamma + \lambda I)^{-1}\gamma_h$. As $\lambda \to 0$: $w_{\text{raw}} \to \Gamma^{-1}\gamma_h$, $w_{\text{in}} \to \Gamma^{-1}(\gamma_h + \frac{b}{a}\mathbf{1})$.

Numerical confirmation. (1) $\equiv$ (2) verified against the actual training code path at $\lambda = 7$, with $\max|\mathbf{1}^\top\hat{v}| = 3.3 \times 10^{-13}$ and $\max|W_{\text{impl}} - W_{\text{derived}}| = 1.5 \times 10^{-14}$. The KKT solution agrees with an independent reparameterized solve to 10^{-15}, and with quasi-Newton minimization to solver precision, over 12 random positive-definite covariances and $\lambda \in \{0.01, 0.5, 5\}$ (`verify_ridge.py`).

B.2 Differencing under ridge has a different geometry

The differencing estimator solves $\min_{u_d} \|dY - dX\,u_d\|_F^2 + \lambda\|u_d\|^2$ with $dX = XD^\top$. In effective-weight coordinates $w = e_L + D^\top u_d$, $u_d = (DD^\top)^{-1}D(w - e_L)$, so the penalty is $\lambda(w - e_L)^\top D^\top(DD^\top)^{-2}D(w - e_L)$: a quadratic penalty on the same plane but *centered at* e_L with a non-identity metric. For $\lambda > 0$ this differs from (2). Numerically, at $\lambda = 7$ the fitted effective weights differ by 0.04 in ℓ_∞ while both sum to one to machine precision.

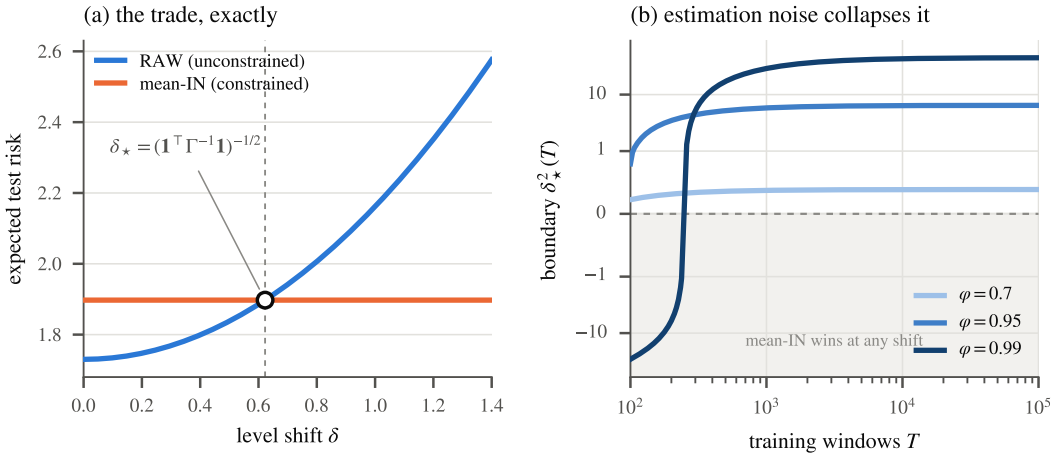

Figure 5: **The population trade-off at $\lambda > 0$, and the finite-sample boundary.** (a) The population picture of Fig. 1a, redrawn with the exact *ridge* price and crossing point (AR(1), $\varphi = 0.7$, $L = 24$, $h = 3$; machine-verified). (b) The finite-sample boundary $\delta_\star^2(T)$ against persistence (Thm. 2): estimation noise collapses it on persistent series, and in the shaded region the constrained estimator wins at *any* shift. Overlapping-window simulations place the true boundary at or below these curves.

B.3 Behavior of the ridge price and boundary

The ridge price has no simple scalar form but is fully computable, with two noteworthy features. First, $b(\lambda)$ increases with λ throughout our examples: *regularizing the raw model increases its shift*

fragility. Second, as $\lambda \to \infty$, $w_{\text{raw}} \to 0$ (risk $\to \gamma_0$) while $w_{\text{in}} \to u$ (risk $R(u)$); whenever the window-mean predictor beats the zero predictor, that is, for any positively persistent series, $\text{price}(\infty) < 0$ and $\delta_\star^2(\lambda)$ becomes negative at finite λ: *over-regularized IN dominates at any shift, because its implicit shrinkage target is a better prior*. At $\varphi = 0.99$, $L = 24$, $h = 3$ the boundary has already turned negative by $\lambda = 10^2$ (Table 2).

Table 2: Ridge price and regime boundary, exact. AR(1) signal, $L = 24$, $h = 3$. $b(\lambda) = 1 - \mathbf{1}^\top w_{\text{raw}}(\lambda)$ grows with λ at every persistence, so regularizing the raw model makes it *more* shift-fragile; and $\delta_\star^2(\lambda) = \text{price}(\lambda)/b(\lambda)^2$ turns negative at finite λ on persistent series, where over-regularized IN wins at any shift.

φ	λ_{pop}	$b(\lambda)$	price(λ)	$\delta_\star^2(\lambda)$
0.7	0	0.6570	+0.16731	+0.3876
0.7	0.1	0.6666	+0.16758	+0.3772
0.7	1	0.7224	+0.17473	+0.3349
0.7	10	0.8751	+0.20384	+0.2662
0.7	100	0.9793	+0.22063	+0.2300
0.7	1000	0.9978	+0.22297	+0.2240
0.95	0	0.1426	+0.13124	+6.4516
0.95	0.1	0.1466	+0.13170	+6.1308
0.95	1	0.1694	+0.14262	+4.9672
0.95	10	0.2594	+0.19964	+2.9678
0.95	100	0.5338	−0.14049	−0.4931
0.95	1000	0.8935	−2.29136	−2.8702
0.99	0	0.0297	+0.03974	+45.0450
0.99	0.1	0.0306	+0.03989	+42.6274
0.99	1	0.0357	+0.04345	+34.0624
0.99	10	0.0560	+0.06349	+20.2160
0.99	100	0.1372	−0.14884	−7.9109
0.99	1000	0.5058	−8.96611	−35.0529

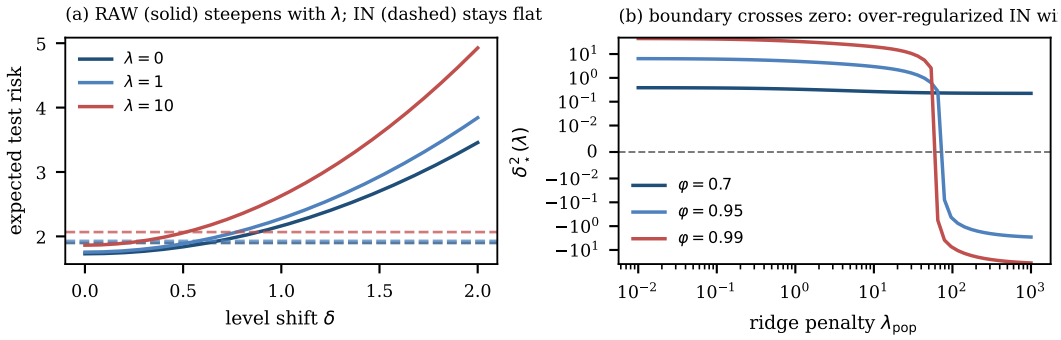

Figure 6: **The effect of regularization on the price and the boundary** (exact formulas; AR(1), $L = 24$, $h = 3$). (a) Expected risk against level shift at $\lambda \in \{0, 1, 10\}$; solid curves are the raw ridge optimum, dashed lines the mean-IN optimum at the same λ. The raw curves steepen as λ grows, because shrinking toward zero pulls the weight sum further from one: *regularizing the raw model makes it more shift-fragile*. The mean-IN lines are flat in δ at every λ, the constraint being architectural. (b) The boundary $\delta_\star^2(\lambda) = \text{price}(\lambda)/b(\lambda)^2$ against λ at three persistence levels. Where it crosses zero, mean-IN has lower risk at *any* shift, its implicit shrinkage target (the window mean) being a better prior than the raw estimator's zero.

B.4 Proof of Proposition 2 and Corollary 1

For fixed w and the shifted pair $(x + \delta\mathbf{1},\, y + \delta)$,

$$\text{Risk}(w, \delta) = \mathbb{E}[(y - w^\top x) + \delta(1 - \mathbf{1}^\top w)]^2 = R(w) + 2\delta(1 - \mathbf{1}^\top w)\mathbb{E}[y - w^\top x] + \delta^2(1 - \mathbf{1}^\top w)^2,$$

the middle term vanishing because x, y are zero-mean under the training distribution. (i) follows with $w = w_{\text{raw}}(\lambda)$; (ii) with any w on the plane. For (iii): minimizing R over the plane gives $w_{\text{in}} = \Gamma^{-1}(\gamma_h + \kappa\mathbf{1})$ with $\kappa = b/a$, and since R is quadratic with Hessian 2Γ and $\nabla R(w_{\text{raw}}) = 0$, $\text{price}(0) = (w_{\text{in}} - w_{\text{raw}})^\top \Gamma (w_{\text{in}} - w_{\text{raw}}) = \kappa^2 \mathbf{1}^\top \Gamma^{-1} \mathbf{1} = b^2/a$. BLUE interpretation: in $x = c\mathbf{1} + s$ with $\text{Cov}(s) = \Gamma$, the BLUE of c is the GLS mean with $\text{Var}(\hat{c}) = 1/a$ (Aitken, 1935); thus $\delta_\star^2 = 1/a$ *is* that variance. White noise: $\Gamma = \sigma^2 I$ gives $a = L/\sigma^2$, price σ^2/L, boundary σ^2/L. □

B.5 Verification summary for Propositions 1–2

check	script	result
implemented fit ≡ (2)	`verify_ridge.py`	1.5×10^{-14}
KKT ≡ reparameterized solve (36 configs)	`verify_ridge.py`	$\leq 10^{-15}$
KKT ≡ quasi-Newton	`verify_ridge.py`	$\leq 3 \times 10^{-6}$ (solver tol.)
$\lambda \to 0$ recovers OLS formulas (12 covs)	`verify_ridge.py`	$\leq 10^{-6}$ at $\lambda = 10^{-9}$
risk formulas vs. 4×10^5-path MC	`verify_core.py`	within MC error ($\sim 2 \times 10^{-3}$)
δ-invariance of IN at $\delta \in \{0, 1, 2, 4\}$	`verify_core.py`	exact
crossing at $\delta_\star(\lambda)$ (MC)	`verify_ridge.py`	RAW = IN within MC error
price identity b^2/a	`verify_core.py`	5×10^{-17}
$b(\lambda)$ increasing (AR(1), $\varphi \in \{.7, .95\}$)	`verify_ridge.py`	monotone over $\lambda \in [0, 10]$
Thm. 1 on pathological backbones	`verify_j3.py`	$\leq 1 \times 10^{-14}$

C The nuisance-subspace theorem: statement, proofs, extensions, verification

This is the general-k form of Sec. 4, summarized in the main text after Prop. 3.

Theorem 4 (nuisance-subspace normalization)**.** *Let $\Gamma \succ 0$ and $U \in \mathbb{R}^{L \times k}$ have full column rank; write $\Pi_U = U(U^\top U)^{-1}U^\top$, $u_U = U(U^\top U)^{-1}\psi$, $A_U = U^\top \Gamma^{-1} U$, $b_U(w) = \psi - U^\top w$, $b_U = b_U(w_{\text{raw}})$.* ***(i)*** *The generalized normalizer realizes exactly $\{w : U^\top w = \psi\}$, with effective weights $w = (I - \Pi_U)v + u_U$ (pure algebra; no distributional assumption). Under the zero-mean training law:* ***(ii)*** *a deployment shift $x \mapsto x + Uc$, $y \mapsto y + \psi^\top c$ gives $\text{Risk}(w, c) = R(w) + (b_U(w)^\top c)^2$ for every fixed w, so on the constraint plane the residual is pathwise unchanged, which is exact invariance;* ***(iii)*** $\text{price}_U := \min_{U^\top w = \psi} R(w) - R(w_{\text{raw}}) = b_U^\top A_U^{-1} b_U$*;* ***(iv)*** *for a fixed shift the raw model loses iff $|b_U^\top c| > \sqrt{\text{price}_U}$, so the win region is the complement of a* slab *and not an ellipsoid; and for a random shift with second moment M_c, normalization lowers expected risk iff $b_U^\top (M_c - A_U^{-1}) b_U > 0$.*

Notation as in Theorem 4; A_U is positive definite since $\Gamma \succ 0$ and U has full column rank. All checks are `verify_nuisance.py`, bracketed [N1]–[N9].

C.1 (i) Architecture ≡ constraint

The generalized normalizer predicts $\hat{y} = v^\top(x - U\hat{c}(x)) + \psi^\top \hat{c}(x)$ with $\hat{c}(x) = (U^\top U)^{-1}U^\top x$, i.e. $\hat{y} = w^\top x$ with $w = (I - \Pi_U)v + u_U$. Feasibility: $U^\top w = U^\top v - U^\top \Pi_U v + U^\top u_U = \psi$ since $U^\top \Pi_U = U^\top$. Conversely, if $U^\top w = \psi$ then $\Pi_U w = u_U$, so $w = (I - \Pi_U)w + u_U$. Fibers are $v + \text{range}(U)$; the minimal-norm representative satisfies $U^\top v = 0$. For k=1, $U = \mathbf{1}$, $\psi = 1$ recovers Remark 1. Verified over 200 random (L, k): feasibility 5.2×10^{-15}, converse 2.2×10^{-15} [N1]. □

C.2 (ii) Shift risk and the role of the zero-mean law

For fixed w and the shifted pair, $y + \psi^\top c - w^\top(x + Uc) = e + b_U(w)^\top c$, an exact pathwise identity; on the constraint plane $b_U(w) = 0$, so the residual itself is unchanged. For general w, $\text{Risk}(w, c) = R(w) + 2\mathbb{E}[e](b_U(w)^\top c) + (b_U(w)^\top c)^2$ and $\mathbb{E}[e] = 0$ under the zero-mean training law. The zero-mean assumption is load-bearing: with $\mathbb{E}[y] = 0.8$ the identity is off by exactly the predicted cross term -0.4031 [N2]; the zero-mean version matches MC over 4×10^5 draws to relative error 1.4×10^{-4}. □

C.3 (III) THE PRICE

Stationarity gives $w_{\text{in}} = \Gamma^{-1}(\gamma_h + U\tilde{\nu})$ and the constraint fixes $\tilde{\nu} = A_U^{-1} b_U$; since R has Hessian 2Γ and $\nabla R(w_{\text{raw}}) = 0$, $\text{price}_U = \tilde{\nu}^\top A_U \tilde{\nu} = b_U^\top A_U^{-1} b_U$. For k=1 this is b^2/a. Verified against an independent null-space solve over 300 random configurations with condition numbers to 5×10^4: worst relative error 2.1×10^{-9} [N3]. □

C.4 (IV) GEOMETRY: A SLAB, AND A SECOND MOMENT

Fixed c. The constrained optimum's shifted risk is $R(w_{\text{raw}}) + \text{price}_U$; the raw optimum's is $R(w_{\text{raw}}) + (b_U^\top c)^2$; normalization wins iff $|b_U^\top c| > \sqrt{\text{price}_U}$, a rank-one condition and not an ellipsoid (10^5 random shifts, zero mismatches [N6]). **Random c.** With $M_c = \mathbb{E}[cc^\top]$, $\mathbb{E}_c\text{Risk}(w, c) = R(w) + b_U(w)^\top M_c b_U(w)$, so the expected gain is $b_U^\top(M_c - A_U^{-1})b_U$ (MC with $\mathbb{E}[c] \neq 0$: +34.74 vs. formula +34.68 [N5]). **The second moment is essential**: substituting $\text{Cov}(c)$ drops $(b_U^\top \mathbb{E}[c])^2$ and understates the benefit by $1.5\times$ in the verified example, and by arbitrarily much as the mean grows. For k=1 the paper's $\hat{\delta}^2$ is already a second moment. □

C.5 BLUE, GENERALIZED RIDGE, AND EDGE CASES

In $x = Uc + s$ with c fixed and $\text{Cov}(s) = \Gamma$, $\hat{c}_{\text{GLS}} = A_U^{-1} U^\top \Gamma^{-1} x$ has $\text{Cov} = A_U^{-1}$ and is optimal in the PSD order (Aitken, 1935; Rao, 1973) (MC: max entrywise error 2.7×10^{-4}; 2000 random unbiased competitors, zero violations [N4]). The implemented generalized normalizer with a ridge penalty on v is exactly $\min_{U^\top w = \psi} \|Y - Xw\|_F^2 + \lambda\|w - u_U\|^2$, which is ridge centered at the minimum-norm feasible point (200 configurations, 1.2×10^{-14} [N7]). If U is rank-deficient the constraint set is nonempty iff $\psi \in \text{range}(U^\top)$; in that consistent case all statements survive with pseudoinverses. For H output columns the aggregate price is $\text{tr}(A_U^{-1} M_b)$ with $M_b = \sum_j b_U^{(j)} b_U^{(j)\top}$ (7×10^{-16} [N9]).

D DRIFT AND SELF-NORMALIZATION: STATEMENT AND PROOFS

We return to the level direction (k=1) and give it dynamics. Let $y_t = \mu_t + \varepsilon_t$, $\mu_t = \mu_{t-1} + \eta_t$, $\text{Var}(\eta) = q^2$, $\text{Var}(\varepsilon) = \sigma^2$ (the local-level model). Relative to the level at the window start, window covariances are $K = q^2 M + \sigma^2 I$ with $M_{jk} = \min(j, k)$. All four claims are verified over a 72-configuration grid ($q/\sigma \in [0.05, 3]$, $L \in \{12, 24, 48\}$, $h \in \{1, 8, 24\}$, $V \in \{0, 100\}$; Table 3), closed forms matching Monte Carlo within simulation error (max 0.83%) and the IN–differencing optima agreeing to 4.8×10^{-14}. The optimal constrained weights decay approximately geometrically in the lookback, consistent with the classical optimality of exponential smoothing here (Muth, 1960), a structural connection, not an equivalence claim.

Proposition 4 (drift). *(i) A fixed forecaster with weight sum s deployed m steps after training incurs additional risk $(1-s)^2 q^2 m$, a linear decay in time since training. (ii) On the sum-to-one plane the level cancels identically; the constrained optimum is closed-form in (K, q, σ, L, h) and independent of level dispersion. (iii) The differencing parameterization $\hat{y} = y_t + u_d^\top Dx$ spans exactly $\{w : \mathbf{1}^\top w = 1\}$: $u_d \mapsto w = e_L + D^\top u_d$ is an affine bijection onto the constraint plane. Hence IN and differencing are the* same hypothesis class *with identical population optima; they differ only as estimators, and under ridge they provably differ, since the induced penalties have different centers (u vs. e_L) and metrics (App. B.2). (iv) The unconstrained optimum under level dispersion V converges to the constrained optimum as $V \to \infty$ (self-normalization), with $|1 - \mathbf{1}^\top w(V)| = \Theta(V^{-1})$ (App. N).*

Setup. Fix a window starting at t_0 with level $c = \mu_{t_0}$. For $j = 1, \ldots, L$: $x_j = c + W_j + \varepsilon_j$ with $W_j = \sum_{i \leq j} \eta_i$, target $y = c + W_{L+h} + \varepsilon_*$. Then $\text{Cov}(x \mid c) = K = q^2 M + \sigma^2 I$, $k_j = q^2 j$, $\text{Var}(y \mid c) = q^2(L + h) + \sigma^2$; c is independent of the increments with $\text{Var}(c) = V$.

(i) With $s = \mathbf{1}^\top w$: $y - w^\top x = (1 - s)c + [(W_{L+h} + \varepsilon_*) - w^\top(W + \varepsilon)]$, independent brackets, so $\text{Risk}(w) = (1-s)^2 V + [\text{Var}(y \mid c) - 2w^\top k + w^\top K w]$. Deploying m steps later replaces V by $V + q^2 m$, adding $(1-s)^2 q^2 m$. □**(ii)** If $s = 1$ the c-terms cancel pathwise; minimizing over the

plane gives $w_{\text{in}} = K^{-1}(k + \kappa\mathbf{1})$ with $\kappa = (1 - \mathbf{1}^\top K^{-1}k)/(\mathbf{1}^\top K^{-1}\mathbf{1})$. □**(iii)** $\Phi(u_d) = e_L + D^\top u_d$ has $\mathbf{1}^\top\Phi = 1$ (rows of D sum to zero); $D^\top$ has full column rank $L-1$, so Φ is injective onto an $(L-1)$-dimensional plane, hence bijective onto it. By summation by parts $\hat{y} = y_t + u_d^\top Dx = \Phi(u_d)^\top x$. □**(iv)** $F_V(w) = [\mathrm{Var}(y \mid c) - 2w^\top k + w^\top Kw] + V(1 - \mathbf{1}^\top w)^2$ is strictly convex; as $V \to \infty$ it epi-converges to the constrained objective, and by Sherman–Morrison $1 - \mathbf{1}^\top w(V) = \Theta(1/V)$, the rate plotted in Fig. 11. □

Exponential-smoothing connection. For the local-level model the optimal one-step predictor in the infinite-past limit is simple exponential smoothing (Muth, 1960); our finite-L constrained optimum is the finite-window analogue, with weights decaying approximately geometrically (fitted rate $\alpha = 0.74$ at $q/\sigma = 0.3$, $L = 24$, $h = 3$). A structural connection, not an equivalence.

Table 3: Local-level grid verification of Prop. 4, all 72 configurations ($q/\sigma \in [0.05, 3]$, $L \in \{12, 24, 48\}$, $h \in \{1, 8, 24\}$, $V \in \{0, 100\}$), 1.5×10^5 Monte Carlo paths each. Relative errors are at Monte Carlo noise level (max 0.83%); the IN–differencing optima coincide to machine precision throughout, confirming Prop. 4(iii). V is the level dispersion the model was trained under, and risks are invariant to it, which is (ii).

q/σ	L	h	V	closed form	MC	rel. err.	$\|w_{\text{D}} - w_{\text{IN}}\|_\infty$	q/σ	L	h	V	closed form	MC	rel. err.	$\|w_{\text{D}} - w_{\text{IN}}\|_\infty$
0.05	12	1	0.0	1.09439	1.0953	0.00084	4.02×10^{-16}	1.0	12	1	0.0	2.61803	2.62399	0.00228	1.01×10^{-15}
0.05	12	1	100.0	1.09439	1.09278	0.00147	4.02×10^{-16}	1.0	12	1	100.0	2.61803	2.63376	0.00601	1.01×10^{-15}
0.05	12	8	0.0	1.11189	1.11654	0.00419	4.02×10^{-16}	1.0	12	8	0.0	9.61803	9.60323	0.00154	1.01×10^{-15}
0.05	12	8	100.0	1.11189	1.1107	0.00107	4.02×10^{-16}	1.0	12	8	100.0	9.61803	9.60396	0.00146	1.01×10^{-15}
0.05	12	24	0.0	1.15189	1.15372	0.00159	4.02×10^{-16}	1.0	12	24	0.0	25.61803	25.50389	0.00446	1.01×10^{-15}
0.05	12	24	100.0	1.15189	1.14763	0.0037	4.02×10^{-16}	1.0	12	24	100.0	25.61803	25.53346	0.0033	1.01×10^{-15}
0.05	24	1	0.0	1.06125	1.05686	0.00414	5.97×10^{-16}	1.0	24	1	0.0	2.61803	2.62669	0.00331	3.59×10^{-15}
0.05	24	1	100.0	1.06125	1.05974	0.00142	5.97×10^{-16}	1.0	24	1	100.0	2.61803	2.63003	0.00458	3.59×10^{-15}
0.05	24	8	0.0	1.07875	1.08186	0.00289	5.97×10^{-16}	1.0	24	8	0.0	9.61803	9.57026	0.00497	3.59×10^{-15}
0.05	24	8	100.0	1.07875	1.08059	0.00171	5.97×10^{-16}	1.0	24	8	100.0	9.61803	9.59977	0.0019	3.59×10^{-15}
0.05	24	24	0.0	1.11875	1.12039	0.00147	5.97×10^{-16}	1.0	24	24	0.0	25.61803	25.50153	0.00455	3.59×10^{-15}
0.05	24	24	100.0	1.11875	1.11881	$5e-05$	5.97×10^{-16}	1.0	24	24	100.0	25.61803	25.6277	0.00038	3.59×10^{-15}
0.05	48	1	0.0	1.0521	1.04813	0.00377	5.41×10^{-16}	1.0	48	1	0.0	2.61803	2.6178	$9e-05$	1.05×10^{-14}
0.05	48	1	100.0	1.0521	1.05857	0.00616	5.41×10^{-16}	1.0	48	1	100.0	2.61803	2.61923	0.00046	1.05×10^{-14}
0.05	48	8	0.0	1.0696	1.06531	0.004	5.41×10^{-16}	1.0	48	8	0.0	9.61803	9.598	0.00208	1.05×10^{-14}
0.05	48	8	100.0	1.0696	1.06946	0.00013	5.41×10^{-16}	1.0	48	8	100.0	9.61803	9.68631	0.0071	1.05×10^{-14}
0.05	48	24	0.0	1.1096	1.1172	0.00685	5.41×10^{-16}	1.0	48	24	0.0	25.61803	25.69091	0.00284	1.05×10^{-14}
0.05	48	24	100.0	1.1096	1.11063	0.00094	5.41×10^{-16}	1.0	48	24	100.0	25.61803	25.70346	0.00333	1.05×10^{-14}
0.3	12	1	0.0	1.34882	1.35414	0.00394	2.78×10^{-16}	3.0	12	1	0.0	10.90833	10.8911	0.00158	3.43×10^{-15}
0.3	12	1	100.0	1.34882	1.34735	0.00109	2.78×10^{-16}	3.0	12	1	100.0	10.90833	10.98301	0.00685	3.43×10^{-15}
0.3	12	8	0.0	1.97882	1.97895	$6e-05$	2.78×10^{-16}	3.0	12	8	0.0	73.90833	73.83809	0.00095	3.43×10^{-15}
0.3	12	8	100.0	1.97882	1.98295	0.00209	2.78×10^{-16}	3.0	12	8	100.0	73.90833	73.62329	0.00386	3.43×10^{-15}
0.3	12	24	0.0	3.41882	3.40157	0.00505	2.78×10^{-16}	3.0	12	24	0.0	217.90833	216.49285	0.0065	3.43×10^{-15}
0.3	12	24	100.0	3.41882	3.43541	0.00485	2.78×10^{-16}	3.0	12	24	100.0	217.90833	217.14747	0.00349	3.43×10^{-15}
0.3	24	1	0.0	1.34836	1.3524	0.003	3.47×10^{-16}	3.0	24	1	0.0	10.90833	10.99932	0.00834	7.40×10^{-15}
0.3	24	1	100.0	1.34836	1.3464	0.00145	3.47×10^{-16}	3.0	24	1	100.0	10.90833	10.9071	0.00011	7.40×10^{-15}
0.3	24	8	0.0	1.97836	1.98026	0.00096	3.47×10^{-16}	3.0	24	8	0.0	73.90833	73.71752	0.00258	7.40×10^{-15}
0.3	24	8	100.0	1.97836	1.98135	0.00151	3.47×10^{-16}	3.0	24	8	100.0	73.90833	74.04724	0.00188	7.40×10^{-15}
0.3	24	24	0.0	3.41836	3.43082	0.00365	3.47×10^{-16}	3.0	24	24	0.0	217.90833	218.18334	0.00126	7.40×10^{-15}
0.3	24	24	100.0	3.41836	3.41852	$5e-05$	3.47×10^{-16}	3.0	24	24	100.0	217.90833	217.27418	0.00291	7.40×10^{-15}
0.3	48	1	0.0	1.34836	1.34974	0.00102	1.39×10^{-15}	3.0	48	1	0.0	10.90833	10.9082	$1e-05$	4.80×10^{-14}
0.3	48	1	100.0	1.34836	1.34557	0.00206	1.39×10^{-15}	3.0	48	1	100.0	10.90833	10.97583	0.00619	4.80×10^{-14}
0.3	48	8	0.0	1.97836	1.97673	0.00082	1.39×10^{-15}	3.0	48	8	0.0	73.90833	74.25418	0.00468	4.80×10^{-14}
0.3	48	8	100.0	1.97836	1.97584	0.00127	1.39×10^{-15}	3.0	48	8	100.0	73.90833	73.76851	0.00189	4.80×10^{-14}
0.3	48	24	0.0	3.41836	3.4255	0.00209	1.39×10^{-15}	3.0	48	24	0.0	217.90833	219.05927	0.00528	4.80×10^{-14}
0.3	48	24	100.0	3.41836	3.41702	0.00039	1.39×10^{-15}	3.0	48	24	100.0	217.90833	217.61408	0.00135	4.80×10^{-14}

E Proof of Theorem 2

Lemma. If rows of $X \in \mathbb{R}^{T\times L}$ are i.i.d. $\mathcal{N}(0, \Gamma)$ with $T > L + 1$, then $\mathbb{E}[(X^\top X)^{-1}] = \Gamma^{-1}/(T - L - 1)$. **Raw.** Conditional on X, $\hat{w} - w_\star \sim \mathcal{N}(0, \sigma_r^2(X^\top X)^{-1})$ and $\mathbb{E}[\hat{w}] = w_\star$, so $\mathbb{E}R(\hat{w}) = \sigma_r^2(1 + \frac{L}{T-L-1})$ and $\mathbb{E}(1 - \mathbf{1}^\top\hat{w})^2 = b^2 + \sigma_r^2 a/(T - L - 1)$, the cross term vanishing by unbiasedness; combining with $\mathrm{Risk}(\hat{w}, \delta) = R(\hat{w}) + \delta^2(1 - \mathbf{1}^\top\hat{w})^2$ gives (i). **Constrained.** Write $w = w_0 + Bz$ with B orthonormal on $\mathbf{1}^\perp$; the constrained regression is unconstrained regression of $y - w_0^\top x$ on $B^\top x$, and at the constrained optimum the residual is uncorrelated with $B^\top x$, hence (joint Gaussianity) independent, with variance σ_{in}^2. Applying the raw argument in dimension $L - 1$ gives (ii). Setting (i)=(ii) yields $\delta_\star^2(T)$; a negative solution means (ii)<(i) already at $\delta = 0$, which is (iii). □

F Proofs for Section 6; the NTE diagnostic

(i) Cancellation. $\hat{\sigma}(v^\top\frac{x-\hat{\mu}\mathbf{1}}{\hat{\sigma}}) + \hat{\mu} = v^\top(x - \hat{\mu}\mathbf{1}) + \hat{\mu}$ for every v, x (verified to 1.5×10^{-13}, floating-point noise from the ϵ guard). **(ii) WLS identity.** The full-IN objective equals $\sum_i \hat{\sigma}_i^{-2}\|(Y_i - \hat{\mu}_i) - v^\top(X_i - \hat{\mu}_i\mathbf{1})\|^2 + \lambda\|v\|^2$: the mean-IN objective with row weights $\hat{\sigma}_i^{-2}$, same parameter, same

penalty, and the normal equations coincide term by term (0.0 difference on real data). With (i), *every* effect of the std division is a training-loss reweighting. **(iii) Efficiency dichotomy.** Multiplicative regime: weighting rows by s_i^{-2} is the Aitken/GLS weighting (Aitken, 1935), efficient among linear estimators. Additive regime: Gauss–Markov makes uniform weighting efficient. Both branches realize on synthetic data (multiplicative: full-IN $5.480 < 5.494$; additive: mean-IN $4.229 < 4.352$).

Remark 2. If $X_i = c\mathbf{1}^\top$ exactly then $X_i - \hat{\mu}_i\mathbf{1} = 0$, so the window contributes zero to both $X_c^\top X_c$ and $X_c^\top Y_c$ regardless of weight. Near-constants contribute design rows of magnitude $O(1)$ but targets of magnitude $\|Y_i - \hat{\mu}_i\|/\hat{\sigma}_i$, which dominate. □

NTE. Throughout, NTE means $\mathrm{NTE}_{\mathrm{trim}}$ at $\tau = 10^{-4}$: $\mathrm{NTE}_{\mathrm{trim}} = \mathrm{mean}_{i:\hat{\sigma}_i > \tau} \frac{1}{H}\|(Y_i - \hat{\mu}_i)/\hat{\sigma}_i\|^2$, pooled across channels exactly as the shared fit pools windows. Train-only, unitless, and precisely the average magnitude of the full-IN training loss at $v = 0$. The trim threshold removes only provably inert windows; doubling or halving it does not change the ranking. Channel-level correlations are weak ($|\rho| < 0.2$) because the fit is shared; the training pool, not the channel, is the causal unit. Per-dataset values: Table 4.

Table 4: Full-IN failure diagnostics per dataset, all train-only, against realized degradation (horizon-averaged $\mathrm{MSE}_{\text{full-IN}} - \mathrm{MSE}_{\text{mean-IN}}$; rows sorted by it). T_{eff}/T *fails* as a diagnostic. Electricity and Traffic are the two most concentrated datasets in the study and the two least degraded, while $\mathrm{NTE}_{\mathrm{trim}}$ ranks all eight (Spearman $+0.88$) and separates Exchange and Weather from the rest by two orders of magnitude.

dataset	T_{eff}/T (pooled)	NTE (trim mean)	NTE (p99.9)	full−mean gap (avg H)
Exchange	0.0054	301.9	62441.8	+0.2865
Weather	0.0298	1239.1	1467.8	+0.2120
ETTh2	0.0203	3.5	112.4	+0.0361
ETTm1	0.2891	1.6	26.6	+0.0055
ETTm2	0.0490	2.2	94.5	+0.0046
ETTh1	0.5683	1.5	36.7	+0.0018
Traffic	0.0000	1.1	9.7	+0.0014
Electricity	0.0044	1.5	25.6	+0.0011

G Synthetic verification details

Every theorem and proposition ships with a seeded verification script. Key numbers: Prop. 4(i) time-decay at $m \in \{0, 100, 400, 1600\}$, predictions $6.20/17.86/64.48$ vs. MC $6.21/17.87/64.47$; Thm. 2 at all $(T, \delta) \in \{100, 400, 1600\} \times \{0, 1\}$ (e.g. T=100, δ=1: theory 2.775, MC 2.776; IN 2.472 vs. 2.470). Total runtime minutes on CPU.

H Overlapping-window study

Design and per-configuration results: Table 5; visual summary Fig. 7. Overlapping windows are drawn from single AR(1) paths, 300 independent paths per configuration; raw and constrained OLS are fitted on the overlapping design and their risks evaluated by the exact population formulas at the fitted weights, so the only sampling noise is in the fit.

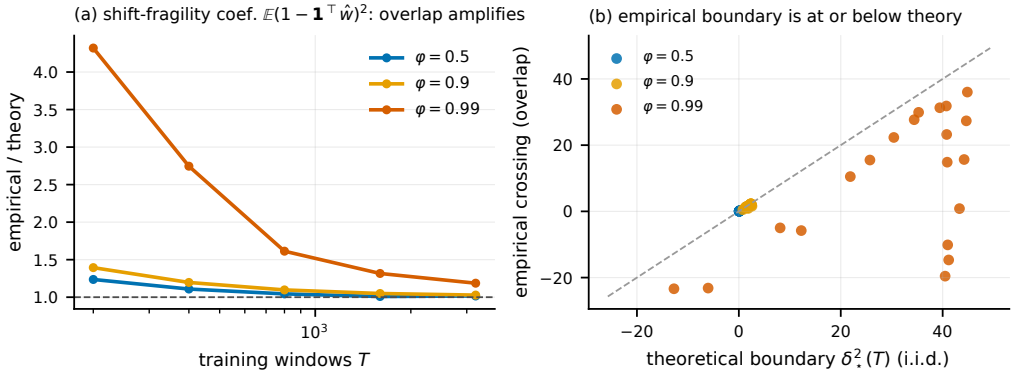

Figure 7: **The i.i.d. formula is conservative under overlap.** (a) Ratio of the empirical shift-fragility coefficient $\mathbb{E}[(1-\mathbf{1}^\top\hat{w})^2]$ to its i.i.d. prediction $b^2+\sigma_r^2 a/(T-L-1)$, averaged over the (L,h) configurations at each T: overlap *amplifies* fragility, by $+8\%$, $+15\%$ and $+124\%$ on average at $\varphi = 0.5, 0.9, 0.99$, decaying toward one as T grows. (b) Empirical against theoretical $\delta^2_\star$: points sit below the diagonal, so IN starts winning at a *smaller* shift than the i.i.d. theory predicts, systematically, and increasingly so with persistence. Per-configuration values in Table 5.

Table 5: Overlapping-window calibration, all 60 configurations: relative deviations of Monte Carlo from the i.i.d. formulas, 300 trials each. A positive *coef. rel.* means dependence *amplifies* shift fragility, so the i.i.d. formula understates it; it is positive in 95–100% of configurations and grows with persistence (mean $+8\%$, $+15\%$, $+124\%$ at $\varphi = 0.5, 0.9, 0.99$). The excess-risk columns stay near zero, which is why the implied effective sample size is 0.96–$1.09 \times T$. Consequently $\delta^2_\star$ emp. sits at or below $\delta^2_\star$ thy.

φ	T	L	h	raw exc.	IN exc.	coef.	$\delta^2_\star$ thy	$\delta^2_\star$ emp	φ	T	L	h	raw exc.	IN exc.	coef.	$\delta^2_\star$ thy	$\delta^2_\star$ emp
0.5	200	24	1	−0.053	−0.061	0.198	0.129	0.105	0.9	800	48	1	−0.043	−0.044	0.083	1.35	1.241
0.5	200	24	12	−0.113	−0.219	0.153	0.158	0.119	0.9	800	48	12	0.105	−0.084	0.198	1.583	1.222
0.5	200	48	1	−0.128	−0.13	0.255	0.053	0.041	0.9	1600	24	1	0.001	0.0	0.039	2.29	2.204
0.5	200	48	12	−0.186	−0.234	0.339	0.084	0.047	0.9	1600	24	12	0.227	−0.144	0.041	2.404	2.254
0.5	400	24	1	−0.008	−0.011	0.124	0.142	0.126	0.9	1600	48	1	−0.006	−0.003	0.028	1.434	1.403
0.5	400	24	12	−0.039	−0.163	0.035	0.156	0.14	0.9	1600	48	12	0.124	−0.082	0.089	1.549	1.364
0.5	400	48	1	−0.087	−0.085	0.115	0.068	0.062	0.9	3200	24	1	0.001	0.003	0.015	2.336	2.303
0.5	400	48	12	−0.094	−0.138	0.161	0.082	0.064	0.9	3200	24	12	0.301	−0.119	0.024	2.392	2.306
0.5	800	24	1	−0.025	−0.023	0.028	0.148	0.144	0.9	3200	48	1	−0.024	−0.019	0.043	1.475	1.421
0.5	800	24	12	−0.017	−0.144	0.063	0.155	0.141	0.9	3200	48	12	0.095	−0.066	0.03	1.532	1.463
0.5	800	48	1	−0.019	−0.011	0.01	0.074	0.075	0.99	200	24	1	0.057	−0.07	2.474	−6.03	−23.17
0.5	800	48	12	−0.019	−0.076	0.075	0.081	0.071	0.99	200	24	12	0.863	−0.138	3.258	41.182	−14.674
0.5	1600	24	1	0.012	0.023	0.006	0.151	0.151	0.99	200	48	1	−0.068	−0.126	2.418	−12.735	−23.375
0.5	1600	24	12	0.077	−0.078	−0.014	0.154	0.153	0.99	200	48	12	0.535	−0.149	5.128	40.483	−19.566
0.5	1600	48	1	−0.021	−0.021	0.029	0.077	0.075	0.99	400	24	1	0.042	−0.02	1.138	12.25	−5.801
0.5	1600	48	12	−0.022	−0.081	0.016	0.081	0.077	0.99	400	24	12	0.603	−0.203	1.629	43.301	0.834
0.5	3200	24	1	−0.0	−0.001	0.023	0.152	0.149	0.99	400	48	1	−0.045	−0.069	1.008	8.104	−4.987
0.5	3200	24	12	0.019	−0.123	0.007	0.154	0.152	0.99	400	48	12	0.683	−0.09	3.207	40.987	−10.124
0.5	3200	48	1	0.002	0.004	0.006	0.079	0.078	0.99	800	24	1	0.007	−0.01	0.415	25.725	15.508
0.5	3200	48	12	0.006	−0.034	0.029	0.08	0.077	0.99	800	24	12	0.657	−0.015	0.739	44.215	15.665
0.9	200	24	1	−0.079	−0.089	0.311	1.65	1.19	0.99	800	48	1	−0.006	−0.018	0.57	21.877	10.506
0.9	200	24	12	0.133	−0.202	0.336	2.573	1.573	0.99	800	48	12	0.216	−0.075	0.728	40.893	14.858
0.9	200	48	1	−0.11	−0.114	0.457	0.777	0.496	0.99	1600	24	1	−0.021	−0.029	0.209	34.41	27.69
0.9	200	48	12	−0.061	−0.216	0.472	1.827	0.884	0.99	1600	24	12	0.606	−0.022	0.344	44.64	27.342
0.9	400	24	1	−0.035	−0.035	0.095	2.018	1.841	0.99	1600	48	1	−0.003	−0.01	0.285	30.415	22.326
0.9	400	24	12	0.181	−0.2	0.214	2.474	1.832	0.99	1600	48	12	0.24	−0.059	0.427	40.787	23.227
0.9	400	48	1	−0.09	−0.089	0.163	1.174	1.024	0.99	3200	24	1	0.001	−0.014	0.229	39.441	31.301
0.9	400	48	12	0.031	−0.141	0.314	1.658	1.078	0.99	3200	24	12	0.341	−0.058	0.175	44.845	36.031
0.9	800	24	1	−0.051	−0.048	0.04	2.199	2.129	0.99	3200	48	1	0.001	−0.004	0.158	35.281	29.908
0.9	800	24	12	0.16	−0.186	0.067	2.427	2.169	0.99	3200	48	12	0.231	−0.01	0.184	40.723	31.828

I Effective sample size versus NTE

I.1 Resolving the reported Electricity failure

One apparent contradiction is worth resolving. Fu and Yang (2025) report RevIN failing catastrophically on Electricity, a DLinear baseline MSE of 13.28 rising to 104.0 (the 683% surge they headline), where we find full-IN harmless there ($0.1404 \to 0.1410$). Model class and 336-step lookback are the same, so the difference is not architectural: their Electricity errors sit two orders of magnitude above the standard protocol's, which z-scores per channel by train statistics and yields ≈ 0.14 (ours 0.1400). Thm. 3 predicts both: full-IN failure tracks the *scale of the reweighted training loss*, a property of the preprocessing pipeline and not of the dataset, and unstandardized inputs make normalized targets extreme, exactly the regime the mechanism flags.

Decisive comparisons: Electricity ($T_{\text{eff}}/T = 0.004$) and Traffic (0.000) show extreme concentration and no degradation (≤ 0.0014); ETTh2 (0.020, NTE 3.5) shows mild degradation (0.036); Exchange and Weather, with NTE $302/1239$, are catastrophic ($0.29/0.21$). Weight concentration is neither necessary nor sufficient; NTE ranks all eight (Spearman $+0.88$, $+0.91$ for the tail version). Per-channel values for the five datasets whose solver produces per-channel test errors are in Table 6.

Table 6: Per-channel weight-concentration and normalized-target-energy diagnostics on train windows, with per-channel full-IN degradation at H=96 where the solver produces per-channel test errors (Weather's are not separable, marked "–"). Channel-level correlations are weak ($|\rho| < 0.2$) because one weight matrix is shared across a dataset's channels, so the training pool and not the channel is the causal unit (App. F). NTE above 10^3 is dominated by near-constant windows whose futures move.

ds	ch	T_{eff}/T	NTE	gap	ds	ch	T_{eff}/T	NTE	gap	ds	ch	T_{eff}/T	NTE	gap
ETTh1	0	0.66241	1.3	0.0007	ETTm1	5	0.21364	1.7	0.00099	Weather	2	0.55596	2.15	–
ETTh1	1	0.77653	1.41	−0.00118	ETTm1	6	0.68981	2.27	−0.00172	Weather	3	0.48847	3.41	–
ETTh1	2	0.63788	1.27	0.00081	ETTm2	0	0.76079	2.46	−0.00733	Weather	4	0.47938	1.71	–
ETTh1	3	0.79288	1.38	−0.00016	ETTm2	1	0.7702	1.97	−0.00337	Weather	5	0.27771	2.37	–
ETTh1	4	0.82312	1.14	−0.00119	ETTm2	2	0.11034	9.4×10^{13}	−0.002	Weather	6	0.2599	3.73	–
ETTh1	5	0.60277	2.34	−0.00027	ETTm2	3	0.81033	1.68	0.00341	Weather	7	0.19352	2.12	–
ETTh1	6	0.79345	1.6	−0.00045	ETTm2	4	0.03768	4.21×10^{12}	−0.00281	Weather	8	0.25031	3.77	–
ETTh2	0	0.71828	2.1	0.01227	ETTm2	5	0.19467	3.36×10^{14}	−0.00022	Weather	9	0.25199	3.76	–
ETTh2	1	0.74367	1.98	0.0064	ETTm2	6	0.42863	1.99	−0.00569	Weather	10	0.62839	2.29	–
ETTh2	2	0.08405	4.26×10^{14}	0.00627	Exchange	0	0.58709	4.27	0.32299	Weather	11	0.54868	915	–
ETTh2	3	0.79648	1.78	0.02951	Exchange	1	0.52187	4.58	0.12567	Weather	12	0.52041	1.67	–
ETTh2	4	0.00524	1.9×10^{12}	−0.00222	Exchange	2	0.4202	4.71	0.11487	Weather	13	0.32942	1.83	–
ETTh2	5	0.05263	1.39×10^{15}	−0.00029	Exchange	3	0.50197	3.73	0.49908	Weather	14	0.37327	1.74×10^{15}	–
ETTh2	6	0.77809	1.5	−0.00222	Exchange	4	0.04328	2.38×10^{3}	0.47244	Weather	15	0.25192	8.3×10^{14}	–
ETTm1	0	0.52851	1.41	0.021	Exchange	5	0.51337	3.85	0.2269	Weather	16	0.08781	1.6	–
ETTm1	1	0.7202	1.43	0.00153	Exchange	6	0.52725	4.84	0.22833	Weather	17	0.10814	1.52	–
ETTm1	2	0.48197	1.46	0.01985	Exchange	7	0.39145	4.7	0.16145	Weather	18	0.09171	1.63	–
ETTm1	3	0.65925	1.44	0.00211	Weather	0	0.29735	4.58	–	Weather	19	0.51613	1.83	–
ETTm1	4	0.68167	1.3	0.00832	Weather	1	0.52	2.1	–	Weather	20	0.1196	2.42×10^{4}	–

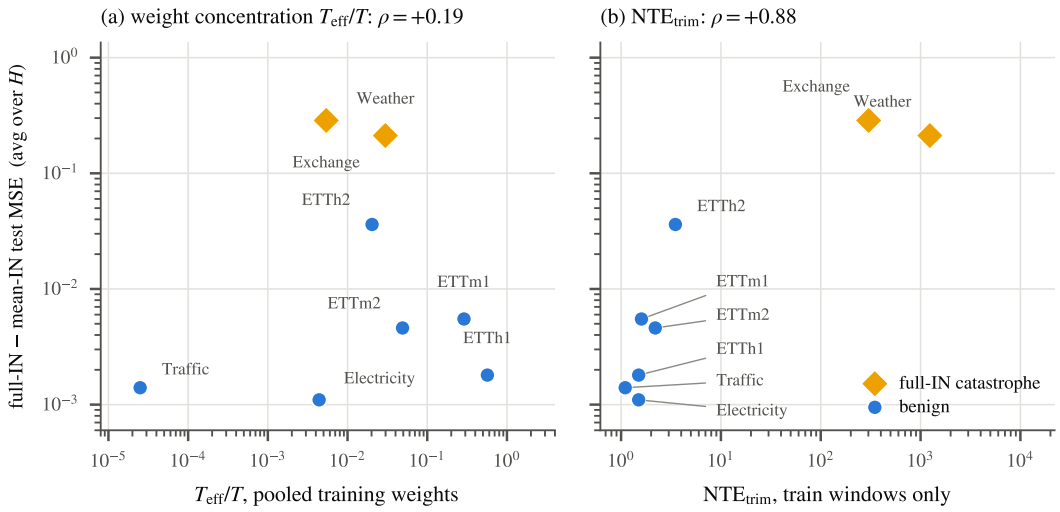

Figure 8: **Weight concentration does not predict full-IN failure; normalized-target energy does.** One point per dataset; y is full-IN minus mean-IN test MSE, averaged over horizons, on a log scale so that the six benign datasets, which span a factor of five, do not collapse onto one point. (a) Weight concentration T_{eff}/T, the usual explanation, is uninformative ($\rho = +0.19$): Electricity and Traffic have the most extreme concentration in this suite and no degradation at all. (b) NTE_{trim}, the scale of the reweighted training loss that Thm. 3 identifies, ranks all eight ($\rho = +0.88$) and separates the two catastrophes from the six benign datasets by two orders of magnitude. Both are train-only statistics.

J BENCHMARK PROTOCOL AUDIT (LINEAR EXPERIMENTS)

Splits	ETTh: 8640/2880/2880 with L-prefix borders; ETTm: 34560/11520/11520; others 7:1:2 chronological.
Scaling	per-channel z-score by train mean/std only.
Windows	$L = 336$; stride 1 (Electricity/Traffic training stride 2 in the Gram solver; stride-1 reruns of 7 of those 8 cells match to within 4×10^{-4} MSE with no sign changes; the remaining cell, Traffic H=720, exceeded our stride-1 compute budget and keeps the disclosed stride-2 protocol).
Architecture	channel-independent: one weight matrix shared across channels, trained on pooled channel windows; per-channel metrics from the shared model.
Estimators	RAW ridge (center 0); mean-IN ridge ($\equiv$ uniform-centered constrained ridge); full-IN ridge ($\equiv 1/\hat{\sigma}_i^2$-weighted mean-IN; $\epsilon = 10^{-8}$ guard); DIFF ridge on difference coefficients, integrated from the last observation.
λ selection	validation MSE over $\{0.1, 1, 10, 10^2, 10^3\}$ ($+10^4, 10^5$ for the Gram solver), per estimator.
Leakage	test targets never used in fitting, tuning, or diagnostics; the regime diagnostic uses test inputs only.
Hardware	all solves closed-form on CPU; full suite under one hour.

K COMPLETE BENCHMARK RESULTS

Table 7: Test MSE at H=96 (z-scored, standard protocol); the full $8 \times 4 \times 4$ grid follows in Table 8. Values track published linear baselines to within a few thousandths except on Exchange (published DLinear ETTh1 ≈0.375, Electricity ≈0.140, Traffic ≈0.410, Exchange ≈0.081).

	RAW	mean-IN	full-IN	DIFF		RAW	mean-IN	full-IN	DIFF
ETTh1	0.3701	0.3702	**0.3699**	0.3834	Exchange	0.0836	0.0832	0.3521	**0.0812**
ETTh2	0.3039	**0.2714**	0.2785	0.2794	Weather	0.1742	0.1753	0.3390	**0.1729**
ETTm1	**0.2992**	0.3005	0.3079	0.3065	Electricity	**0.1400**	0.1404	0.1410	0.1404
ETTm2	0.1723	0.1645	**0.1619**	0.1647	Traffic	0.4102	**0.4099**	0.4109	0.4100

Table 8: All $8 \times 4 \times 4$ test MSE values under the protocol of App. J: eight datasets $\times$ four horizons $\times$ four estimators. Table 7 is the H=96 slice.

Dataset	H	RAW	mean-IN	full-IN	DIFF
ETTh1	96	0.3701	0.3702	0.3699	0.3834
ETTh1	192	0.4036	0.4033	0.4049	0.4050
ETTh1	336	0.4309	0.4256	0.4294	0.4264
ETTh1	720	0.4591	0.4216	0.4236	0.4226
ETTh2	96	0.3039	0.2714	0.2785	0.2794
ETTh2	192	0.4046	0.3335	0.3584	0.3521
ETTh2	336	0.5065	0.3575	0.3969	0.3868
ETTh2	720	0.8035	0.3842	0.4572	0.4006
ETTm1	96	0.2992	0.3005	0.3079	0.3065
ETTm1	192	0.3339	0.3356	0.3418	0.3430
ETTm1	336	0.3682	0.3703	0.3752	0.3772
ETTm1	720	0.4228	0.4262	0.4299	0.4330
ETTm2	96	0.1723	0.1645	0.1619	0.1647
ETTm2	192	0.2412	0.2195	0.2190	0.2199
ETTm2	336	0.3192	0.2732	0.2744	0.2754
ETTm2	720	0.4604	0.3673	0.3876	0.3691
Exchange	96	0.0836	0.0832	0.3521	0.0812
Exchange	192	0.1788	0.1709	0.4048	0.1674
Exchange	336	0.3773	0.3131	0.5896	0.3064

Table 8, continued.

Dataset	H	RAW	mean-IN	full-IN	DIFF
Exchange	720	1.0209	0.8438	1.2105	0.8120
Weather	96	0.1742	0.1753	0.3390	0.1729
Weather	192	0.2152	0.2185	0.4635	0.2165
Weather	336	0.2610	0.2662	0.5083	0.2642
Weather	720	0.3226	0.3347	0.5320	0.3339
Electricity	96	0.1400	0.1404	0.1410	0.1404
Electricity	192	0.1536	0.1545	0.1552	0.1545
Electricity	336	0.1691	0.1710	0.1723	0.1709
Electricity	720	0.2036	0.2096	0.2113	0.2095
Traffic	96	0.4102	0.4099	0.4109	0.4100
Traffic	192	0.4229	0.4229	0.4245	0.4231
Traffic	336	0.4365	0.4356	0.4369	0.4358
Traffic	720	0.4668	0.4639	0.4655	0.4641

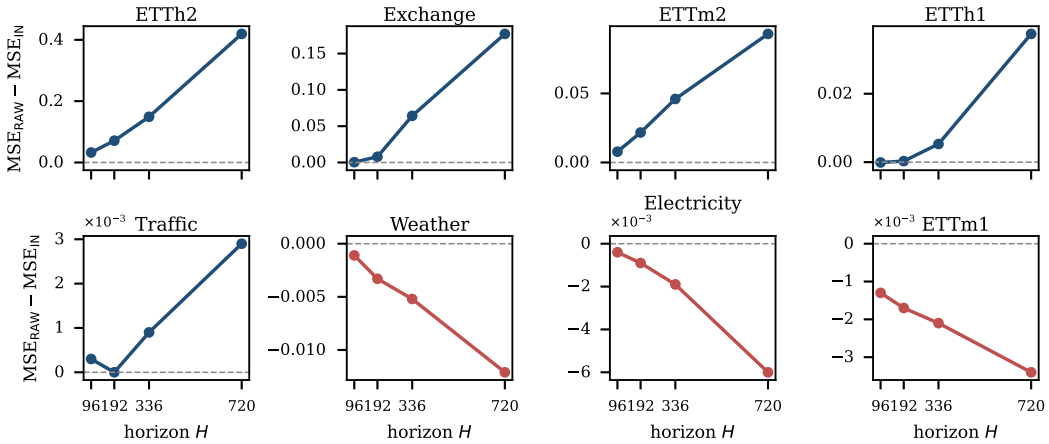

Figure 9: **Both regimes realize, and each is monotone in horizon.** Realized gain $\mathrm{MSE}_{\mathrm{RAW}} - \mathrm{MSE}_{\text{mean-IN}}$ against forecast horizon, one panel per dataset; positive means normalization wins and the dashed line marks zero. Blue (top row): the gain grows with H, the compounding-shift regime. Red (bottom row): the small price compounds instead. Vertical scales differ by panel: ETTh2 spans $+0.03$ to $+0.42$ while Traffic never exceeds $+0.003$. Every dataset is monotone in H (Spearman $|\rho| = 1.0$ on seven, $+0.8$ on Traffic). Regime membership here is an observed outcome; an assignment built from train and validation data alone reproduces 6 of 8, and both mismatches are datasets whose gaps are statistical ties under the block bootstrap.

Per-dataset horizon summary (trend = Spearman of the gap in H):

dataset	gap(96)	gap(720)	trend	regime (observed)
ETTh2	$+0.0325$	$+0.4193$	$+1.0$	drift: IN gains grow
Exchange	$+0.0004$	$+0.1771$	$+1.0$	drift: IN gains grow
ETTm2	$+0.0078$	$+0.0931$	$+1.0$	drift: IN gains grow
ETTh1	-0.0001	$+0.0375$	$+1.0$	drift (weak): IN gains grow
Traffic	$+0.0003$	$+0.0029$	$+0.8$	near-tie, weakly drift
ETTm1	-0.0013	-0.0034	-1.0	stationary-price regime
Electricity	-0.0004	-0.0060	-1.0	stationary-price regime
Weather	-0.0011	-0.0121	-1.0	stationary-price regime

Table 9: Realized normalization gain $\mathrm{MSE}_{\mathrm{RAW}} - \mathrm{MSE}_{\text{IN-mean}}$ at every dataset and horizon (positive = normalization wins). The gap is monotone in H on all eight datasets: it grows on five and shrinks on three, the two regimes the theory admits. "trend" is the Spearman correlation of the gap with H.

dataset	H=96	192	336	720	trend
ETTh1	−0.0001	+0.0003	+0.0053	+0.0375	+1.0
ETTh2	+0.0325	+0.0711	+0.1490	+0.4193	+1.0
ETTm1	−0.0013	−0.0017	−0.0021	−0.0034	−1.0
ETTm2	+0.0078	+0.0217	+0.0460	+0.0931	+1.0
Exchange	+0.0004	+0.0079	+0.0642	+0.1771	+1.0
Weather	−0.0011	−0.0033	−0.0052	−0.0121	−1.0
Electricity	−0.0004	−0.0009	−0.0019	−0.0060	−1.0
Traffic	+0.0003	+0.0000	+0.0009	+0.0029	+0.8

L REGIME-MAP ROBUSTNESS AND PER-CHANNEL DATA

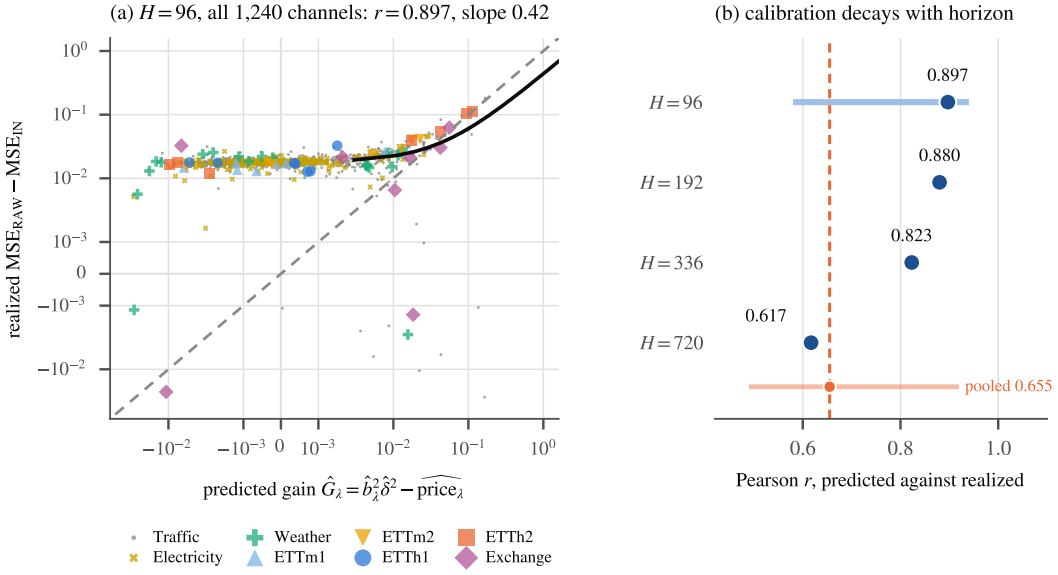

Figure 10: **Predicted versus realized normalization gain.** (a) One dot per channel at H=96, colored by dataset, all 1,240 shown with no trimming. x: the ridge diagnostic $\hat{G}_\lambda$ from the fitted raw model, train moments, and test *inputs* only; y: realized $\mathrm{MSE}_{\mathrm{RAW}} - \mathrm{MSE}_{\mathrm{IN}}$. Both axes are symlog with a linear window of 10^{-3} around zero, which is what separates the several hundred near-tie channels that a linear axis piles onto the origin. Dashed: identity. Solid: the fitted calibration line, drawn over the range where the fit carries mass; its slope 0.42 says magnitudes are over-predicted even where the ordering is informative. (b) The same correlation at each horizon, with the dataset-clustered 95% CI where it was computed, against the pooled value over all 4,960 channel-horizon cells. Calibration decays monotonically with H.

Final diagnostic (all eight datasets). Dataset-clustered percentile bootstrap (8 clusters, 4000 resamples); tie region $|\hat{G}| \le 10^{-3}$ fixed in advance of the all-dataset run. Per horizon: H=96: $r = 0.897$ (CI 0.58–0.94), sign-acc 0.74 (638 confident), ties 602 (mean |gap| 0.0014); H=192: 0.880; H=336: 0.823; H=720: 0.617; pooled $r = 0.655$ (CI 0.49–0.92), slope 0.385. Variants (pooled): exact-ridge 0.655 > OLS-price 0.638 > pure-moment 0.474; strict validation-input variant $r = 0.187$ pooled with sign accuracy 0.68, so signs transfer and magnitudes do not. Robustness: leave-one-dataset-out 0.55–0.82; leave-one-horizon-out 0.63–0.84; 1% winsorization 0.50; within-dataset standardization 0.755; dataset-level aggregation $r = 0.535$, $\rho = 0.763$. Lookback sweep (H=96, smaller datasets): $r = 0.938/0.901/0.901/0.939$ at $L = 96/192/336/720$. Covariance numerics: cond$(\hat{\Gamma} + 10^{-4} I)$ median 2.0×10^4, max 2.6×10^6; jitter $\times/\div 10$ changes r by < 0.001.

Decision protocols (32 cells). Mean regret vs. oracle: strict train+val gate 0.0007; always-IN 0.0012; validation-MSE selection 0.0014; deployment-input gate 0.0017; always-RAW 0.0355.

Block-bootstrap ties in the sense of Diebold and Mariano (1995) (block $L+H$; sensitivity $0.5\times/2\times$): 21–23 of 32 tie; decisive cells 9–11 with gate sign correct on all but one.

Provenance. Every point in Fig. 10 traces to `results/final_master.csv`; per-channel aggregation reproduces the global MSEs by construction and was cross-checked against an independent slow reimplementation (max deviation 5×10^{-6}, identical λ selection). An automated poisoning test passes. An earlier pilot on the five smaller datasets at $H{=}96$ with the OLS-price form gave $r = 0.863$; the numbers above supersede it. Its per-channel quantities are kept for reference in Table 10.

Table 10: Per-channel regime-map quantities and realized RAW−IN gaps at $H{=}96$, all 36 channels of the five stride-1 datasets. This is the *pilot* analysis; the all-dataset, all-horizon numbers in App. L supersede it. $\hat{G}$ is shown in the fitted form used in the paper and in the two moment-based ablations.

dataset	ch	$\hat{\delta}^2$	$\hat{a}$	$\hat{G}$ (fitted)	$\hat{G}$ (OLS-mom.)	$\hat{G}$ (ridge-mom.)	realized gap
ETTh1	0	0.1600	5.0002	−0.0002	−0.0000	−0.0000	−0.0054
ETTh1	1	0.2598	2.3768	−0.0010	−0.0005	−0.0005	−0.0014
ETTh1	2	0.1225	5.7076	−0.0003	−0.0000	−0.0000	−0.0056
ETTh1	3	0.1668	2.1859	−0.0018	−0.0004	−0.0004	−0.0010
ETTh1	4	0.4038	16.8250	0.0021	0.0270	0.0270	−0.0008
ETTh1	5	0.1910	2.0864	−0.0017	−0.0039	−0.0039	−0.0006
ETTh1	6	1.7657	2.0882	0.0078	0.0239	0.0238	0.0143
ETTh2	0	0.4953	3.8332	0.0035	0.0090	0.0088	−0.0063
ETTh2	1	6.6720	3.2943	0.0947	0.0751	0.0924	0.0853
ETTh2	2	0.0844	2.6163	−0.0044	−0.0095	−0.0094	−0.0019
ETTh2	3	8.7686	4.0771	0.1268	0.5619	0.5975	0.0939
ETTh2	4	2.9212	1.4153	0.0329	0.0042	0.0042	0.0358
ETTh2	5	0.0152	1.7416	−0.0083	−0.0093	−0.0077	−0.0003
ETTh2	6	1.2581	1.9330	0.0110	0.0047	0.0007	0.0211
ETTm1	0	0.2192	3.0900	−0.0004	−0.0008	−0.0004	−0.0047
ETTm1	1	0.2776	1.4028	−0.0016	−0.0008	−0.0007	−0.0013
ETTm1	2	0.1950	3.3277	−0.0004	−0.0007	−0.0001	−0.0053
ETTm1	3	0.1822	1.3219	−0.0021	−0.0005	−0.0004	−0.0010
ETTm1	4	0.4229	3.1908	0.0004	0.0032	0.0045	−0.0036
ETTm1	5	0.2009	1.3654	−0.0020	−0.0019	−0.0013	−0.0007
ETTm1	6	1.8481	1.1419	0.0036	0.0007	0.0020	0.0074
ETTm2	0	0.6725	1.7913	0.0005	0.0017	0.0019	−0.0035
ETTm2	1	6.6483	1.9863	0.0257	0.1223	0.1236	0.0221
ETTm2	2	0.1240	1.1260	−0.0032	−0.0007	−0.0007	−0.0017
ETTm2	3	8.7832	1.9235	0.0345	0.2249	0.2254	0.0228
ETTm2	4	3.1173	1.3852	0.0100	0.0009	0.0011	0.0104
ETTm2	5	0.0179	1.0810	−0.0038	−0.0004	−0.0005	−0.0001
ETTm2	6	1.3781	1.1986	0.0023	0.0004	0.0004	0.0048
Exchange	0	3.7277	1.8315	0.0208	0.0239	0.0220	−0.0197
Exchange	1	0.4757	1.6695	−0.0008	−0.0019	−0.0027	0.0142
Exchange	2	1.8570	1.3236	0.0072	0.0001	−0.0015	−0.0118
Exchange	3	8.9101	1.8101	0.0546	0.0269	0.0263	0.0442
Exchange	4	0.8250	2.6999	0.0030	0.0028	0.0018	0.0030
Exchange	5	2.2420	2.9858	0.0125	0.1387	0.1374	−0.0410
Exchange	6	3.7908	1.7574	0.0211	0.0410	0.0400	0.0027
Exchange	7	7.3063	1.6238	0.0437	0.0482	0.0472	0.0120

M RIDGE-GRID SENSITIVITY

Every λ in the selection grid, for RAW and mean-IN at $H{=}96$, with the validation-selected value marked. Test MSE is nearly flat across the grid for the reason given in §3: the code's sum-of-squares convention makes $\lambda_{\text{pop}} = \lambda_{\text{emp}}/N$ tiny against $\hat{\Gamma}$'s eigenvalues, so no conclusion in this paper is an artifact of the grid.

Table 11: Ridge-grid sensitivity at H=96: validation and test MSE at every λ in the grid, for RAW and mean-IN, with the validation-selected λ marked (←). Validation selection is near-optimal on test throughout, so none of the paper's conclusions is an artifact of the grid. Test MSE is nearly flat in λ because the code's sum-of-squares convention makes $\lambda_{\text{pop}} = \lambda_{\text{emp}}/N$ tiny against $\hat{\Gamma}$'s eigenvalues.

dataset	model	λ	val	test	sel.	dataset	model	λ	val	test	sel.
ETTh1	RAW	0.1	0.6515	0.3701	←	ETTm1	IN-mean	0.1	0.3893	0.3005	
ETTh1	RAW	1	0.6515	0.3701		ETTm1	IN-mean	1	0.3893	0.3005	
ETTh1	RAW	10	0.6515	0.3701		ETTm1	IN-mean	10	0.3893	0.3005	
ETTh1	RAW	100	0.6515	0.37		ETTm1	IN-mean	100	0.3893	0.3005	←
ETTh1	RAW	1000	0.652	0.3694		ETTm1	IN-mean	1000	0.3893	0.3004	
ETTh1	IN-mean	0.1	0.67	0.3702	←	ETTm2	RAW	0.1	0.1201	0.1725	
ETTh1	IN-mean	1	0.67	0.3702		ETTm2	RAW	1	0.1201	0.1725	
ETTh1	IN-mean	10	0.67	0.3701		ETTm2	RAW	10	0.1201	0.1725	
ETTh1	IN-mean	100	0.67	0.3701		ETTm2	RAW	100	0.1201	0.1725	
ETTh1	IN-mean	1000	0.6705	0.3695		ETTm2	RAW	1000	0.1199	0.1723	←
ETTh2	RAW	0.1	0.2236	0.3047		ETTm2	IN-mean	0.1	0.1162	0.1647	
ETTh2	RAW	1	0.2236	0.3047		ETTm2	IN-mean	1	0.1162	0.1647	
ETTh2	RAW	10	0.2236	0.3047		ETTm2	IN-mean	10	0.1162	0.1647	
ETTh2	RAW	100	0.2235	0.3046		ETTm2	IN-mean	100	0.1162	0.1647	
ETTh2	RAW	1000	0.2229	0.3039	←	ETTm2	IN-mean	1000	0.1159	0.1645	←
ETTh2	IN-mean	0.1	0.2061	0.2726		Exchange	RAW	0.1	0.1738	0.0836	←
ETTh2	IN-mean	1	0.2061	0.2726		Exchange	RAW	1	0.1738	0.0836	
ETTh2	IN-mean	10	0.2061	0.2726		Exchange	RAW	10	0.1739	0.0835	
ETTh2	IN-mean	100	0.206	0.2724		Exchange	RAW	100	0.1748	0.0833	
ETTh2	IN-mean	1000	0.2053	0.2714	←	Exchange	RAW	1000	0.1797	0.0844	
ETTm1	RAW	0.1	0.3822	0.2993		Exchange	IN-mean	0.1	0.1304	0.0833	
ETTm1	RAW	1	0.3822	0.2993		Exchange	IN-mean	1	0.1304	0.0833	
ETTm1	RAW	10	0.3822	0.2993		Exchange	IN-mean	10	0.1304	0.0832	←
ETTm1	RAW	100	0.3822	0.2992		Exchange	IN-mean	100	0.1305	0.083	
ETTm1	RAW	1000	0.3822	0.2992	←	Exchange	IN-mean	1000	0.132	0.0838	

N SELF-NORMALIZATION AND WEIGHT PROFILES

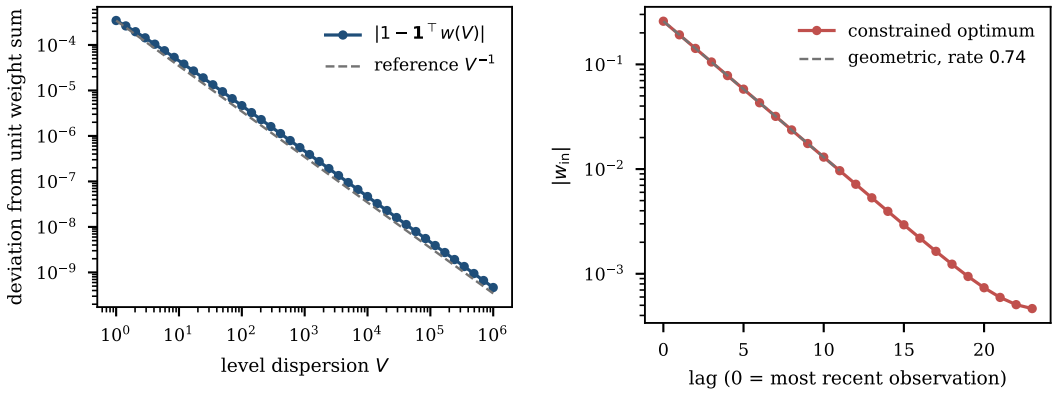

Figure 11: **Self-normalization, and the shape of the constrained optimum** (local-level model, $q/\sigma = 0.3$, $L = 24$, $h = 3$; exact formulas). Left: as level dispersion V grows the *unconstrained* optimum's weight sum approaches one unaided, the deviation $|1 - \mathbf{1}^\top w(V)|$ decaying as V^{-1} (dashed reference), so unconstrained regression self-normalizes in the drift limit (Prop. 4(iv)). Right: the drift-optimal constrained weights decay approximately geometrically in the lookback, fitted rate 0.74 (dashed), the finite-window counterpart of exponential smoothing's optimality for this model.

O DEEP-FORECASTER EXPERIMENT: PREREGISTERED PROTOCOL

The full frozen document (sha256 `c7ec02740b83...`, frozen 2026-08-13 before any deep-model run) ships with the supplement and the code release; this section reproduces the parts a reader needs to judge Sec. 8.

What was open. Three claims were settled before the experiment and are not what it tests: Thm. 1 (wrapper equivariance, exact, architecture-independent); the regularity statement $f(x+\delta\mathbf{1}) - f(x) =$

$\int_0^\delta s(x + t\mathbf{1})\,dt$ for f locally Lipschitz along the shift ray, which covers ReLU/GELU networks ("differentiable" is too strong, "a.e. differentiable" too weak); and the expansion $\mathrm{Risk}(\delta) = R_0 + c_1\delta + c_2\delta^2 + O(\delta^3)$ with $c_1 = 2\mathbb{E}[e(1-s)]$ and $c_2 = \mathbb{E}[(1-s)^2] - \mathbb{E}[e\,s']$, $s' = \nabla s^\top \mathbf{1}$. *The only open question the experiment addresses* is whether $\hat{S} = \mathbb{E}[(1-s)^2]$, measured on a trained un-normalized backbone from test inputs, is empirically associated with how much mean normalization helps that backbone.

Grid. Backbones DLinear, PatchTST, iTransformer in one framework, identical data pipeline, splits and evaluation. Variants per backbone produced by the normalization switch plus an explicit wrapper only (never loss or data edits): RAW (internal norm off), MEAN-IN (norm off + wrapper), FULL-REVIN (standard RevIN, affine off). Datasets ETTh1, ETTh2, ETTm1, ETTm2, Exchange, Weather with the same splits, $L = 336$, per-channel train-statistic z-scoring, stride 1, test MSE on the z-scored scale. Seeds $\{2021, 2022\}$; per-seed and seed-averaged both reported; no seed ever selected. Identical tuning budget across the three variants of a backbone.

Statistic and clustering. $s(x) = \nabla f(x)^\top \mathbf{1}$ via one forward-mode JVP (Griewank and Walther, 2008) with the all-ones tangent; $\hat{S}$ averages $(1 - s)^2$ over output steps and channels on ≥ 2000 fixed test input windows (subsample seed 0). Primary statistic Spearman ρ with a *dataset*-clustered bootstrap 95% CI (Efron, 1979; Cameron et al., 2008); datasets are the independent unit, never channels, never seeds.

Pre-committed verdict thresholds. STRONG: nonlinear-pooled $\rho \geq 0.5$, clustered CI excluding 0, and $\rho > 0$ for both nonlinear backbones. PARTIAL: $\rho \geq 0.5$ with a CI including 0, or exactly one nonlinear backbone $\rho \leq 0$. FAIL: $\rho < 0.3$ pooled, or any nonlinear backbone $\rho < -0.3$. Correction D: DLinear is a control and contributes zero evidential weight; all criteria are computed on nonlinear cells only. Correction E: any grade above FAIL additionally requires the frozen five-predictor gauntlet of Table 1, and if RAW validation MSE matches or beats $\hat{G}$ the verdict is capped at PARTIAL. A random-predictor calibration at $n = |\text{cells}|$ with 20,000 draws must accompany the table. RAW *test* MSE is descriptive only and is never a verdict input; post hoc it scores $\rho = +0.853$ at H=96, and the DLinear control scores $+0.600$ ($n = 6$), neither of which decides anything.

Leakage controls. $\hat{S}$ and $\hat{\delta}^2$ may use test *inputs*; nothing may use test targets; validation targets are used only for early stopping and selection, identically across variants. A poison test (replace all test targets with noise; weights, hyperparameters, $\hat{S}$, $\hat{\delta}^2$, $\mathrm{NTE}_{\mathrm{trim}}$ must be bit-identical) was specified and is one of the three items still open (D8 below).

P PREREGISTRATION DEVIATIONS LOG

P.1 POST-FREEZE GATE RUN: THE STEP-SIZE CROSS-CHECK PREMISE, VERIFIED

The §3.3 step-size cross-check compares a first-order JVP against a finite difference. Its failure was attributed in the deviations table to curvature: a finite difference carries an $O(\delta^2)$ term the JVP cannot match, so the discrepancy should grow *linearly* in the probe shift δ rather than indicate a wrong Jacobian. A post-freeze gate run (3 backbones × 6 datasets × 2 seeds at H=96, probe $\delta \in \{0.0025, 0.005, 0.01\}$; `results/results_v5_rows.jsonl`) tests that premise directly, and Figure 12 confirms it. Across the 24 nonlinear cells the ratio of successive-δ errors has median 1.96 (range 1.32–2.08) against the value 2 that exact $O(\delta)$ scaling predicts. The DLinear control behaves as the same explanation requires in the opposite direction: a linear backbone has no curvature, its JVP is exact, and its residual sits between 10^{-5} and 10^{-6}, pure round-off, which does *not* grow with δ (ratio ≈ 0.5). The premise is therefore verified rather than asserted; the gate measured curvature, not a defect in the equivariance claim, which passes its own primary clause exactly.

P.2 POST-FREEZE GATE RUN: THE LEAKAGE (POISON) CHECK

The preregistered leakage check retrains each of six cells with the test targets replaced by noise and requires every train-side quantity to be unchanged. It was executed in the same post-freeze run (`results/results_v5_poison_rows.jsonl`). Across all six cells `val_mse`, `epochs_run`, `NTE_trim` and the test-*input* statistic $\hat{\delta}^2$ are *bit-identical* to their clean twins, while `test_mse` rises by 2.3–8.3×, as it must. $\hat{S}$ is bit-identical in five of six cells and differs by

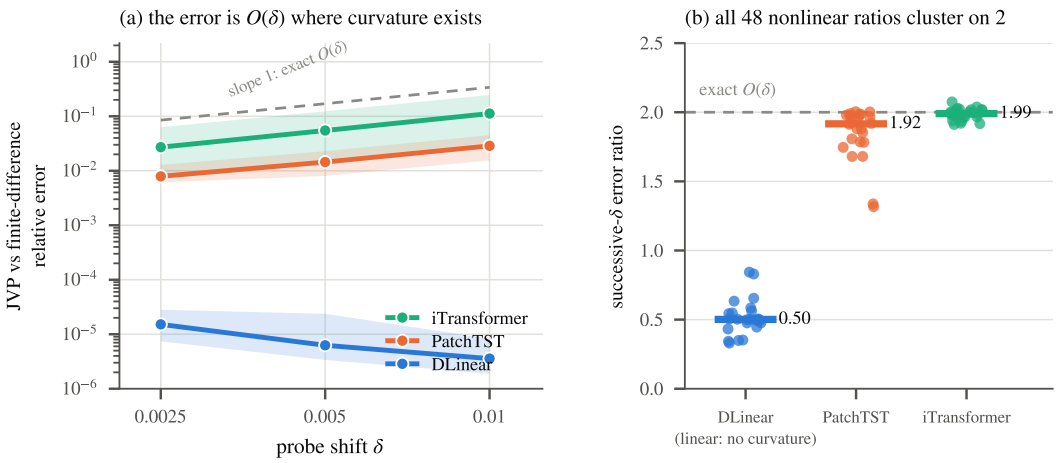

Figure 12: **The §3.3 gate failed because of finite-step curvature.** (a) Median JVP-vs-finite-difference relative error against probe shift, band = min–max over cells. The two nonlinear backbones track slope 1; DLinear sits three to four orders of magnitude lower and *falls* with δ, the signature of round-off rather than curvature. (b) Ratio of successive-δ errors, two per cell, so 48 from the 24 nonlinear cells; bars are medians. They cluster on 2, the value exact $O(\delta)$ scaling predicts, while the linear control sits at 0.5.

4.3×10^{-7} in one (iTransformer/ETTh1), which is GPU non-determinism and not information flow; we report the bound rather than claim bit-identity. The c_1 and c_2 coefficients do move, and should: they are fitted to risk curves evaluated on test targets, which the poisoning destroys by construction. No train-side quantity depends on test data.

Append-only. Every gate is quoted from the frozen document, followed by what the artifact computed. Where a gate failed, the literal outcome is recorded and any explanation is labeled as such.

ID	Gate (frozen rule, abridged)	What was measured	Status
D1	§6: evaluate $\hat{S}$ on "$\geq$ 2000 fixed test input windows"	Initial 324-run grid probed 256 windows, reduced mid-run to 64 by an out-of-memory fix, so cells were *inconsistently* sampled. All 108 RAW cells were re-probed at $\geq$ 2000 windows by chunked JVP.	**Violated in the first artifact; remedied.** Sec. 8 and App. S report only the compliant 108-cell re-probe (shipped as `rows_v4.jsonl`); realized gains come from the 324-run grid (`rows_tier1.jsonl`), which supplies no $\hat{S}$ used anywhere. Rank agreement between the two probes is $\rho = 0.999$, but the earlier $\hat{S}$ values are not preregistered results and are not used. `h1_recompute.py` in the supplement regenerates every $\hat{S}$-derived number in this paper from those two files and prints both probes side by side.
D2	§6, cont.	Exchange's test split contains fewer than 2000 windows at every horizon (1422/1326/1182/798 at $H = 96/192/336/720$); the code takes $\min(2000, \lvert\text{test}\rvert)$. Every other cell is at the full 2000.	**Unattainable**, a property of the dataset. Logged. Dropping Exchange entirely leaves a fully compliant subset, on which $\rho = +0.903/+0.939/+0.939/+0.915$.

continued on the next page

Preregistration deviations log, continued from the previous page.

ID	Gate (frozen rule, abridged)	What was measured	Status
D3	§3.1: mean-IN must show $\mathbb{E}[(1-s)^2] < 10^{-6}$ *and* max defect $< 10^{-4}\lvert\delta\rvert$	Clause 1: $\mathbb{E}[(1-s)^2] = 0.0$ exactly in all 36 cells at $H = 96$ (108 across horizons). Clause 2: exceeded in 18/36 (DLinear 0/12, PatchTST 6/12, iTransformer 12/12; max 3.7×10^{-4}).	**Clause 1 passes exactly; clause 2 partially fails.** Attributed to float32 cancellation in an `.abs().max()` over $\sim 256 \cdot H \cdot C$ elements; Thm. 1 makes the true defect 0, so the gate sits below the float32 noise floor at transformer depth. Also run at $\delta = 0.5$, which the frozen text does not fix.
D4	§3.2a: DLinear's s must be constant across inputs (spread $< 10^{-4}$)	Initially not evaluable: the naive $\mathbb{E}[s^2] - (\mathbb{E}[s])^2$ estimator in float32 has a 1.196×10^{-3} noise floor, above the gate, and read 1.119×10^{-3}, with 8 distinct values repeating across different datasets and seeds. Recomputed with a float64 Welford estimator on the same models: **exactly** 0.0 **in all 12 cells**, one distinct value. Contrast: PatchTST median 2.9×10^{-1}, iTransformer 9.3×10^{-1}.	**CLOSED; the gate passes as measured.** The estimator, not the harness, was at fault. An exact zero also settles D5 below: s constant means f is affine, which the independent δ-scaling evidence ($R(0.1)/R(0.025) = 0.253$ where curvature would give ≈ 4) had only implied.
D5	§3.2b: DLinear $\lvert c_1\rvert < \max(0.1\,\lvert c_1\rvert^{\text{med}}_{\text{PT}}, 0.005) = 0.00907$; failure means the harness leaks nonlinearity, STOP	4/12 cells exceed (ETTh2 $\times 2$ seeds, Weather $\times 2$ seeds).	**Fails literally.** The gate's premise, not the harness, is what breaks. D4 establishes that f is affine by direct measurement, so $c_1 = 2(s-1)\mathbb{E}[e]$ is an *identity*, and the gate's implicit $\mathbb{E}[e] = 0$ is what fails under train/test mean shift: the four exceedances are exactly the cells with the largest $\lvert 1-s\rvert\,\lvert\mathbb{E}[e]\rvert$, Weather carrying by far the largest weight-sum deficit ($\lvert 1-s\rvert = 0.147$, double any other dataset).
D6	§3.3: finite-difference defect $\lVert f(x+\delta\mathbf{1}) - f(x) - \delta s_{\text{JVP}}\rVert$ at $\delta = 0.1$ within 5%	In the first artifact this was never computed, because the code subtracted δ rather than $\delta\, s_{\text{JVP}}$, i.e. it computed §3.1's statistic, at $\delta = 0.5$. Recomputed correctly: **62/108 pass** (DLinear 12/12, PatchTST 42/48, iTransformer 8/48).	**Unexecuted in the first artifact; fails literally in 46/108 cells on re-run.** Recorded as 62/108 and not converted to a pass. Diagnosed as finite-step curvature by the δ-scaling argument in Sec. 8: pass rate $84 \to 75 \to 62$ of 108 as δ goes $0.025 \to 0.05 \to 0.10$ on identical weights and JVP.
D7	§2 H3: c_2 comparisons must use the corrected form with the smoothed-s' estimator at $h \in \{0.05, 0.15, 0.30\}$, all reported	Not emitted by the original harness (§8 schema gap); recomputed and reported at all three bandwidths.	**CLOSED.** The corrected form sits $5.6\times$ closer than the naive one to the independently fitted quadratic (median per-cell $2.1\times$, max $332\times$), median bandwidth sensitivity 3.7%, and coincides with the naive form in $12/12$ DLinear cells as $s' = 0$ requires. No threshold is asserted: the frozen text sets none.

continued on the next page

Preregistration deviations log, continued from the previous page.

ID	Gate (frozen rule, abridged)	What was measured	Status
D8	§7.1 poison test: with noised test targets, weights, hyperparameters, $\hat{S}$, $\hat{\delta}^2$, $\mathrm{NTE}_{\mathrm{trim}}$ bit-identical	Executed on 6 cells. Test MSE rises to 1.38–1.75 as intended. All 13 target-independent fields are bit-identical in 5 of 6 cells; in the sixth (iTransformer/ETTh1) they agree to $\leq 6 \times 10^{-5}$ relative.	**Passes on 5 of 6; the sixth is not evaluable as specified.** Two repetitions of the same *un-poisoned* command on that cell differ by the same magnitude ($\hat{S}$ rel. 8.8×10^{-8}), so its probe is nondeterministic at float32 and bit-identity is unattainable independently of poisoning. DLinear and PatchTST are deterministic and pass exactly. $\hat{S}$'s run-to-run variation is $\sim 10^{-7}$ relative, orders of magnitude below any spacing that could change a Spearman rank.
D9	§8 artifact schema: 22 named columns	`c2_naive` and `c2_corr_h05/h15/h30` missing; `commit_sha` recorded only after the fact, so §0 provenance is absent from the original artifact.	**Partial.** Logged.
D10	§9 STRONG outcome text (verbatim)	One clause amended: "with DLinear serving as a *passing* linear control" → "…a linear control that passes with two documented deviations (App. P)".	**Wording deviation, deliberate.** The frozen phrasing would have been inaccurate given D4 and D5. Recorded here rather than used silently.
D11	§5 seeds: $\{2021, 2022\}$ minimum, $\{2021, 2022, 2023\}$ preferred	$\{2021, 2022\}$ used.	**Compliant.** The frozen protocol requires the third-seed decision *before* unblinding H1; H1 is unblinded, so any seed added now would be exploratory only and none was added.
D12	§0: "record … the md5 of this file" in the artifact's `commit_sha` column	Every artifact row carries `prereg=b9f113e238dd`. That string is a hard-coded literal in `run_variant.py` (line 137): the harness never hashes the preregistration at run time. The frozen document shipped in the supplement hashes to md5 `674fb85d6d21...`, sha256 `c7ec02740b83...`	**Recorded as a label, not a checksum.** The field identifies which protocol a run executed under and cannot verify the file, so we do not present it as verification. What is independently checkable: the frozen document's own §0 table lists the superseded spec at md5 `23034442352e`, and that file still hashes to `23034442352e92e2...`, so its internal provenance table is accurate. The Tier-2 and Tier-3 stamps are genuine run-time hashes and are unaffected.

What remains. D4, D7 and D8 have been executed and are closed above; §0 provenance is now recorded per row (GPU, driver, CUDA, framework and TSLib SHA). Two gates fail literally and stay that way in the record: §3.2b's c_1 bound (4/12, diagnosed as the gate's premise) and §3.3's cross-check (62/108, diagnosed as finite-step curvature). Neither bears on the nonlinear cells that carry H1. The re-execution used a rebuilt software environment; it reproduces the released artifact bit-for-bit on the cell we re-ran as a control (`test_mse` 0.406051, $\hat{S}$ 0.07614080492655437, and the three δ cross-checks all exact), so the closed gates are measured on the same models the verdict rests on. D12 is a provenance correction rather than a gate: the Tier-1 preregistration is shipped in the supplement with its hashes, and the label the harness stamps is described as a label.

Q PROPOSITION 3: PROOF, LINEAR REDUCTION, VERIFICATION

Proof. $\hat{y}_\alpha = f + \alpha\Delta$, so its residual is $e_{\mathrm{RAW}} - \alpha\Delta$ and

$$\mathrm{Risk}(\alpha) = \mathrm{Risk}_{\mathrm{RAW}} - 2\alpha\,\mathbb{E}\langle e_{\mathrm{RAW}}, \Delta\rangle + \alpha^2\mathbb{E}\|\Delta\|^2 = \mathrm{Risk}_{\mathrm{RAW}} - 2\alpha B + \alpha^2(P + B).$$

Since $e_{\mathrm{IN}} = e_{\mathrm{RAW}} - \Delta$, we get $P = \mathbb{E}\|\Delta\|^2 - B = -\mathbb{E}\langle e_{\mathrm{IN}}, \Delta\rangle$. The unconstrained stationary point is $B/(P+B)$; as $P + B = \mathbb{E}\|\Delta\|^2 > 0$ the quadratic is strictly convex, so this point lies in $(0, 1)$ iff $0 < B < P + B$, that is iff $B > 0$ and $P > 0$, and otherwise the minimizer over $[0, 1]$ is an endpoint. Substituting gives excess $B - B^2/(P+B) = PB/(P+B)$ over $\mathrm{Risk}_{\mathrm{RAW}} - B$; the two endpoints' excesses over the same base are B and P, so the gain over the better one is $\min(P, B) - PB/(P+B) = \min(P, B)^2/(P+B)$, which for fixed P increases in B up to $B = P$ and decreases after, equalling $P/2$ there. □

No step uses linearity, differentiability, continuity or a distributional assumption: as an identity between empirical means it holds on any finite sample and any pair of predictors.

Linear reduction. Under $x \mapsto x + \delta\mathbf{1}$, $y \mapsto y + \delta$ with $\mathbb{E}\delta = 0$, $\mathbb{E}\delta^2 = \tau^2$ and δ independent of (x, y), write $e = y - w_{\mathrm{raw}}^\top x$ and $b = 1 - \mathbf{1}^\top w_{\mathrm{raw}}$. Then $\Delta = (w_{\mathrm{in}} - w_{\mathrm{raw}})^\top x + \delta b$ and $e_{\mathrm{RAW}} = e + \delta b$. At $\lambda = 0$ the KKT condition gives $w_{\mathrm{in}} - w_{\mathrm{raw}} = (b/a)\Gamma^{-1}\mathbf{1}$, so independence annihilates the cross terms and

$$\mathbb{E}\|\Delta\|^2 = \tfrac{b^2}{a} + b^2\tau^2, \qquad B = \mathbb{E}\big[e\,(w_{\mathrm{in}} - w_{\mathrm{raw}})^\top x\big] + b^2\tau^2 = b^2\tau^2, \qquad P = \tfrac{b^2}{a},$$

the middle equality because $\mathbb{E}[ex] = \gamma_h - \Gamma w_{\mathrm{raw}} = 0$ is the normal equations. Hence $\alpha^\star = \tau^2/(\tau^2 + 1/a)$ and the boundary $\delta_\star^2 = 1/a$ of Cor. 1 is exactly the shift level at which $P = B$. Because Prop. 3 assumes nothing about how f was fitted, it applies verbatim to ridge-trained predictors and to $H > 1$ outputs; only these two identities are model-specific. A per-horizon α_h is strictly better than one shared α, so the deep gains in Sec. 8 are a lower bound.

Verification. Because the general statement is an identity rather than an approximation, it is checked to machine precision rather than estimated. `verify_soft.py` evaluates $\mathrm{Risk}(\alpha)$ directly at 20,001 values of α on $N = 4000$ samples and compares against $\mathrm{Risk}_{\mathrm{RAW}} - 2\alpha B + \alpha^2(P+B)$, for four backbones chosen to exercise the "any deterministic function" clause:

backbone g	$\max_\alpha$ |identity error|	$\lvert P + \mathbb{E}\langle e_{\mathrm{IN}}, \Delta\rangle\rvert$	$\alpha^\star$ closed form vs. grid
GELU MLP	8.9×10^{-16}	0	-0.0051: boundary minimizer 0
ReLU MLP	5.3×10^{-15}	1.1×10^{-16}	0.455372 vs. 0.455350
quantized (discontinuous)	5.3×10^{-15}	1.7×10^{-16}	0.275846 vs. 0.275850
hard-clip (non-smooth)	3.6×10^{-15}	0	0.348893 vs. 0.348900

The first row matters as much as the others: there $B < 0$, the interiority condition of Prop. 3(i) fails, and the minimizer sits at $\alpha = 0$, so "interior" is a falsifiable property of a backbone and not a consequence of blending two distinct predictors. The linear reduction is likewise exact rather than sampled: computing $B = (w_{\mathrm{in}} - w_{\mathrm{raw}})^\top(\gamma_h - \Gamma w_{\mathrm{raw}}) + b^2\tau^2$ in closed form on the AR(1) running example matches $b^2\tau^2$ to 0 and P matches b^2/a to 1.1×10^{-16}. Finally, the linear closed forms are checked against an independent 2×10^5-point grid search at $\tau^2 \in \{0.05, \delta_\star^2, 0.5, 2\}$ and over 200 random configurations ($L \in [4, 40]$, random SPD Γ): $\alpha^\star$ is interior in $200/200$, matches the grid argmin to the grid step, and the parallel-combination excess matches the grid minimum to 7.3×10^{-7}, that grid's resolution floor.

Deployment quantities. $P + B = \mathbb{E}\|\Delta\|^2$ needs only unlabeled inputs and the two prediction streams, so it is available at deployment for any backbone. In the linear model $B = b^2\tau^2$ is also label-free, which is what makes $\alpha^\star$ tuning-free there. For a nonlinear backbone a calibrated label-free estimate of B is the missing piece; Sec. 8 reports the size of that gap rather than treating the test-set decomposition as a prediction.

R Tier-2 sign test and Tier-3 soft-rule test: protocol, verdicts, deviations

Status of these tiers. The Sec. 8 preregistration declares Tier 2 optional, fixes its grid (Electricity and Traffic, the two datasets absent from Tier 1) and forbids aggregating it with the Tier-1 statistic. Tier 3 was preregistered separately, after Tier 2 results were known and therefore *exploratory relative to* H1 and to the Tier-2 verdict, neither of which is recomputed. Both prediction files were hashed and the hash stamped into the run log with zero result rows in existence: Tier 2 sha256 `acfdf22c3dbf213e8c3c3e037f029f37711d96ff...`, written 2026-08-22 21:14 UTC;

Tier 3 sha256 `d85b999ce2ab6761c4e33194a13cb8cff9c48800...`. Both files, both run logs, the raw row artifacts and their hashes ship with the code.

Grid and protocol. Identical to App. O except for the dataset list. Tier 2: 2 backbones $\times$ 2 datasets $\times$ 2 variants $\times$ 2 seeds at H=96 (16 cells, all complete; integrity check clean, with no duplicates, no extras, batch size constant within dataset, probe $\geq$ 2000 windows on every RAW cell). Tier 3: three further seeds on the negative cell, $H \in \{192, 720\}$ on it, the 8-cell α-sweep, and an 8-cell float64 equivariance probe (20 cells, all complete). Hardware: one RTX 4090 (24 GB) on a rented host; torch 2.8.0+cu128, CUDA 12.8, TSLib pinned at commit `4e938a1`. Positive controls (preregistration §3) pass $9/9$ before any training, with wrapper equivariance exactly 0.0 for mean-IN and full-IN on all three backbones.

Verdicts, verbatim against the frozen rule.

- **P1 fails.** $\hat{G}$ for the negative cell was to fall below every Tier-1 value for that backbone. It is the highest of the four Tier-2 cells: 0.0616 against the frozen Tier-1 minimum 0.0587 (a 5% miss). Within Tier 2, $\mathrm{Spearman}(\hat{G}, \text{gain}) = -0.40$ on n=4.
- **P2 fails.** Tier-2 gains were to be smaller than every Tier-1 gain: 3 of 4 hold; PatchTST/Traffic is $+0.0288$ against the frozen Tier-1 minimum $+0.0046$.
- **P3 holds.** At least one cell with gain ≤ 0: iTransformer/Traffic, both seeds.
- **T1 is not evaluable.** $\hat{S}$ was to halve once ≥ 8 epochs ran. All five iTransformer/Traffic RAW cells early-stopped at exactly 4 epochs ($\hat{S}$ 0.541–0.635), so the specified comparison cannot be formed. The frozen text anticipates this case; we do not substitute another.
- **T2 holds.** New-seed gain negative: seeds 2023/2024/2025 give -0.0157, -0.0623, -0.0434; with 2021/2022, $5/5$ negative, mean -0.0430.
- **T3 holds.** $\geq$ 2 of 4 cells interior and $\alpha^\star$(iTrans/Traffic) $<$ $\alpha^\star$(PatchTST/Traffic): $8/8$ cell-seeds interior; 0.40 against 0.95.
- **T4 holds.** Gain negative at H=192 and H=720: -0.0387 and -0.0273 (seed-averaged).
- **T5 holds (n=8).** Float64 defect $< 10^{-9}$: worst 1.066×10^{-14}, against a worst float32 defect of 4.630×10^{-4}.

Scoring consequence, per the rule fixed in advance. P1 or P2 failing requires the failure to be reported prominently in the body and in the deviations log, leaves the Tier-1 H1 verdict unaffected (it does not use these cells), and requires the transfer claim to be *narrowed to the Tier-1 regime*. We do all three: Sec. 8 leads with the failure, Sec. 8 and the abstract now claim ranking within a regime rather than sign prediction, and the price-side gap is named as the reason.

Declared deviations, both memory-only. *(D4)* The four H=720 cells were killed by the Linux OOM killer in the post-training NTE step, which stacks $\sim$4000 Traffic windows ($4000 \times 720 \times 862 \times 4$ bytes $\approx$9.9 GB for the target array alone, plus temporaries). An `NTE_N` cap of 400 was introduced and the realized count recorded per row (`nte_n_windows`= 401 in all four). This changes a recorded *diagnostic* on those rows; T4 reads only `test_mse`, so the prediction is untouched. *(D5)* Six of eight float64 probes hit CUDA OOM casting a 256-window batch to double (8.6–11.5 GiB); an `F64_CHUNK` loop of 16 windows was introduced. This is *exact*, not an approximation: the reported defect is a maximum over windows, and a maximum is unchanged by chunking. All failed attempts, including an earlier launch that died at import on all 20 cells because the launch script omitted a path export, are preserved in the shipped failure logs.

R.1 The Tier-3 price/benefit decomposition

Inverting each measured quadratic $\mathrm{Risk}(\alpha) = \mathrm{Risk}_{\mathrm{RAW}} - 2\alpha B + \alpha^2(P+B)$ from its three measured points recovers B and P per cell-seed (Prop. 3); $\alpha^\star = 1/(1+P/B)$ reproduces the swept optimum in all eight.

R.2 Secondary hypotheses and controls, at their literal rates

H2 is not supported, a pre-committed possible outcome: $\mathrm{NTE}_{\mathrm{trim}}$ against full-RevIN degradation gives $\rho = +0.382$, CI $[-0.244, +0.717]$, so the NTE account stays scoped to linear models. *H3 is*

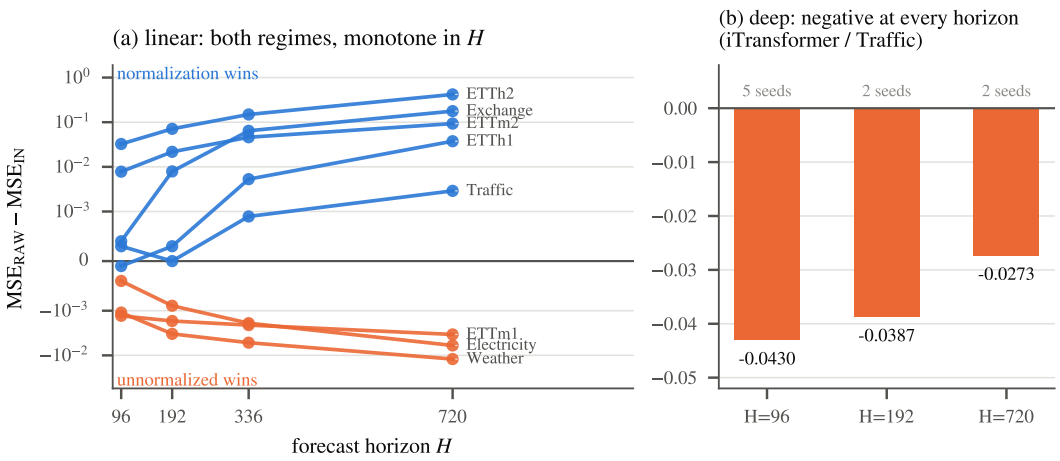

Figure 13: **Both regimes, and the horizon behavior of the negative one.** (a) Linear benchmark: the RAW−IN gap against horizon on all eight datasets, symlog with a linear window around zero. Five datasets increase, three decreasing, matching the two regimes the theory admits (Sec. 7). (b) The deep negative cell (iTransformer/Traffic) at H=96/192/720: negative at every horizon, with the magnitude shrinking rather than compounding, consistent with Traffic being the stationary/periodic end of the suite, where the window level carries persistent predictive information instead of accumulating drift.

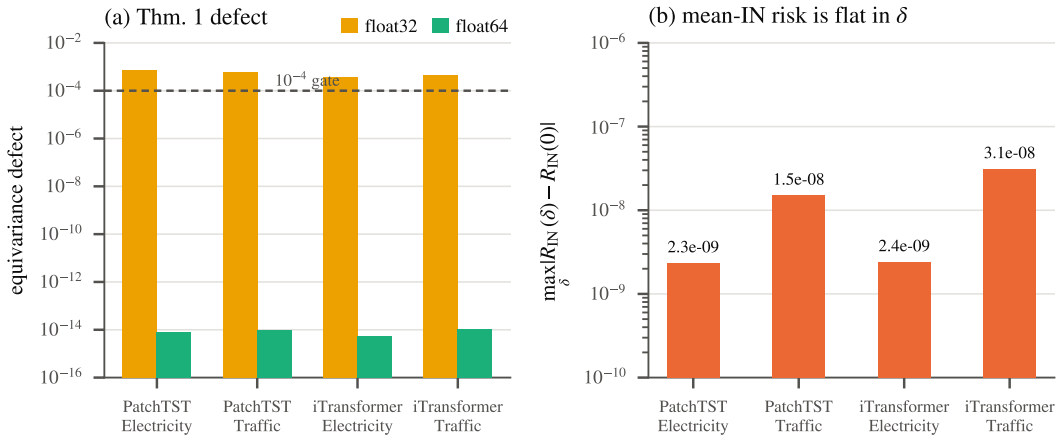

Figure 14: **Theorem 1 is exact; float32 was the whole story.** (a) Worst wrapper equivariance defect per trained cell in float32 and in float64: 4.630×10^{-4} against 1.066×10^{-14}, a factor of $\approx 4 \times 10^{10}$. The 10^{-4} gate that the Sec. 8 control exceeded is a floating-point threshold, not a statement about the theorem. (b) Measured mean-IN risk against level shift on the four trained transformers: flat to 2.3×10^{-9}–3.1×10^{-8} relative, while the same architectures' un-normalized risk moves by tens of percent over the same range.

partially executed: fitted c_1 matches its theoretical sign on 86 of 96 nonlinear cells (90%) against a $2/3$ threshold, the c_2 comparison being reported without a threshold since the protocol sets none. *Tier-3 T1 is not evaluable*: every relevant run early-stopped before the epoch the prediction named, a case the frozen text anticipated. DLinear's s-spread is *exactly* zero in all 12 float64 cells against a 10^{-4} gate, which is also why its c_1 bound is exceeded in 4 of 12: $c_1 = 2(s-1)\mathbb{E}[e]$ is an identity for affine f and the gate's implicit $\mathbb{E}[e] = 0$ fails under train/test mean shift. The JVP cross-check passes in 62 of 108, the exceedances being finite-step curvature; the leakage test passes exactly in 5 of 6 cells, the sixth being float32-nondeterministic at 10^{-7}. App. P and App. R diagnose every gate.

Table 13: Tier-3 price/benefit decomposition, per seed. $P > 0$ and $B > 0$ in $8/8$, which by Prop. 3(i) is exactly the condition for an interior optimum.

backbone	dataset	B	P	P/B	$1/(1+P/B)$	swept $\alpha^\star$
PatchTST	Electricity	.0068/.0060	.0017/.0020	0.25/0.33	.80/.75	.80/.75
PatchTST	Traffic	.0308/.0300	.0016/.0016	0.05/0.05	.95/.95	.95/.95
iTransformer	Electricity	.0119/.0119	.0021/.0030	0.18/0.25	.85/.80	.85/.80
iTransformer	Traffic	.0945/.0873	.1418/.1309	1.50/1.50	.40/.40	.40/.40

S H1 POST-HOC ROBUSTNESS

Provenance of the H1 numbers, and which probe they come from. Two artifacts exist and the distinction matters, so we state it explicitly. The 324-run grid (`rows_tier1.jsonl`) supplies every *realized gain* $\mathrm{MSE}_{\mathrm{RAW}} - \mathrm{MSE}_{\text{mean-IN}}$. It also carries an $\hat{S}$ column, but that column violates the preregistration: the grid probed 256 windows and was reduced mid-run to 64 by an out-of-memory fix, so its cells are inconsistently sampled (deviation D1). The remedy was a 108-cell re-probe of every RAW cell at the preregistered population (`rows_v4.jsonl`), and *that* is the only source of $\hat{S}$ and $\hat{\delta}^2$ used anywhere in this paper. Every ρ in Table 1, in Fig. 3b and in the exponent sweep below is a join of the two: gains from the grid, predictor from the re-probe, seed-averaged to 12 nonlinear cells, `scipy.stats.spearmanr`, clustered percentile bootstrap over the six dataset clusters with 4000 resamples and seed 0.

The two probes rank-agree at $\rho = 0.999$ over the 108 matched cells but are not identical, and the difference is visible in exactly two places: at $H{=}96$ and $H{=}720$ the $\hat{G}$ values of ETTm1 and Weather sit within 1–2% for PatchTST, and the two probes order that one pair differently. That single rank swap moves the pooled ρ from $+0.9441$ to $+0.9510$ at $H{=}96$ and from $+0.9510$ to $+0.9580$ at $H{=}720$, and the PatchTST column from $+0.9429$ to $+1.0000$ at those horizons. **We report the lower, compliant figures.** An intermediate draft of this paper reported the higher ones, having regenerated the table from the 324-row grid alone, and the grid is the only file that can produce a gain, so it is the natural thing to reach for, and the non-compliant $\hat{S}$ rides along unnoticed. Three independent checks pin the compliant artifact as the intended one: the leave-Exchange-out row below ($+0.903/+0.939/+0.939/+0.915$), the deep-versus-linear rank correlation at the end of the counter-shift paragraph ($+0.771$ for PatchTST), and the $H{=}96$ clustered interval $[+0.791, +1.000]$ all reproduce from the re-probe and from nothing else. `h1_recompute.py` in the supplement regenerates every number in this appendix from both probes and prints them side by side, so the choice is auditable rather than asserted. Everything not derived from $\hat{S}$, namely the RAW-validation-MSE control at all four horizons, $\hat{\delta}^2$ alone, and the entire iTransformer column, is identical under both, which localizes the discrepancy to $\hat{S}$ exactly as this account requires.

Horizon-widening of the gauntlet margin (post hoc). $\hat{S}$ alone degrades $0.902 \to 0.650$ across $H{=}96 \to 720$ while $\hat{S}\hat{\delta}^2$ stays flat, so the shift-magnitude factor contributes $+0.04$ at $H{=}96$ and $+0.30$ at $H{=}720$, as the mechanism predicts when longer horizons carry more drift. This was not preregistered and decides nothing; it is recorded because it is the strongest reading of the data and should be visible as such.

c_2 **bandwidth detail.** Median improvement $2.1\times$, max $332\times$, median bandwidth sensitivity 3.7% across $h \in \{0.05, 0.15, 0.30\}$.

Everything in this section is exploratory. The frozen protocol forbids inventing statistics after seeing results; the verdict is the preregistered Spearman with the dataset-clustered CI in Table 1, and nothing below can change it.

concern	check	result
"Exchange is an outlier driving it"	leave-one-dataset-out, all 24 combinations	min $\rho = +0.9030$
	Exchange dropped entirely ($n = 10$)	$+0.903/+0.939/+0.939/+0.915$
"you averaged seeds to smooth it"	per-seed, not seed-averaged	$+0.895$ to $+0.979$
"Spearman is fragile at $n = 12$"	Kendall τ	$+0.849/+0.879/+0.909/+0.879$
	jackknife on cells, H=96	min $+0.927$, max $+0.973$
	all horizons pooled, $n = 48$, clustered	$+0.949$, CI $[+0.808, +0.986]$
"which IN variant you subtract matters"	gap against full-RevIN instead of mean-IN	$+0.895/+0.902/+0.909/+0.909$
"why $\hat{\delta}^2$ and not $\hat{\delta}$ or $\hat{\delta}^4$?"	exponent sweep $\hat{S}\hat{\delta}^{2\alpha}$	see below

The counter-shift test (where does normalization lose?). The preregistered risk curves shift test inputs *and* targets by a common δ. Two things follow. First, the wrapped models' curves are flat to $\leq 2.1 \times 10^{-7}$ relative across all 108 cells, which is Thm. 1 realized on trained networks, against a median 60% variation for the un-normalized models. Second, evaluating each RAW model at the δ that minimizes its own risk approximates its deployment risk with the *common* level offset removed; the per-window dispersion of window means survives, so this is a partial, conservative counterfactual, not a no-shift risk. Under it, mean-IN's advantage falls by 50% on average at H=96, the removed fraction rank-correlates with $\hat{\delta}^2$ ($\rho = +0.41$ at H=96, $+0.41$ pooled over 48 cells, weak at this n and reported as a direction rather than a result), and the advantage *reverses* in the following 8 of 96 cell-seeds:

backbone	dataset	H, seed	gap at δ=0	RAW$_{\min}$−IN
PatchTST	ETTm1	96, 2022	−0.0001	−0.0038
PatchTST	ETTm1	192, 2021 / 2022	+0.0034 / +0.0015	−0.0012 / −0.0060
PatchTST	ETTm1	336, 2022	+0.0051	−0.0008
PatchTST	Weather	336, 2021	+0.0070	−0.0007
PatchTST	Weather	720, 2021 / 2022	+0.0030 / +0.0040	−0.0088 / −0.0073
PatchTST	Exchange	720, 2021	+0.7198	−0.0119

The minimizing δ lies on the grid boundary $|\delta| = 0.6$ in 48 of 96 cell-seeds, so the counter-shift is truncated and the reversal count understates the effect. Every reversal is on ETTm1, Weather or Exchange; ETTm1 and Weather are two of the three datasets whose *linear* gap is negative, and the two lowest-$\hat{G}$ datasets for both nonlinear backbones. Deep $\hat{G}$ against the linear model's signed gap over the six shared datasets: $\rho = +0.771$ (PatchTST), $+0.829$ (iTransformer), $n = 6$.

Exponent sweep. $\hat{G}_\alpha = \hat{S}\hat{\delta}^{2\alpha}$; $\alpha = 0$ is $\hat{S}$ alone, $\alpha = 1$ is the preregistered $\hat{G}$.

α	H=96	H=192	H=336	H=720
0.00 ($\hat{S}$ alone)	+0.9021	+0.8322	+0.8042	+0.6503
0.25	+0.9301	+0.8881	+0.8951	+0.7762
0.50	+0.9441	+0.9371	+0.9441	+0.9021
0.75	+0.9580	+0.9580	+0.9441	+0.9161
1.00 (prereg $\hat{G}$)	+0.9441	+0.9580	+0.9650	+0.9510
1.50	+0.9441	+0.9441	+0.9510	+0.9790
2.00	+0.9441	+0.9510	+0.9510	+0.9790

Two conclusions follow, and they point in opposite directions. *The shift-magnitude factor earns its place, increasingly so with horizon*: $\hat{S}$ alone degrades $0.902 \to 0.650$ across H while $\hat{S}\hat{\delta}^2$ holds $0.944 \to 0.951$, so the factor contributes $+0.04$ at H=96 and $+0.30$ at H=720, which is what "shift-defect × shift-magnitude" predicts if longer horizons carry more level drift. *The exponent itself is not identified*: every $\alpha \in [0.75, 2]$ scores 0.92–0.98, and $\hat{S}\sqrt{\hat{\delta}^2}$ ties $\hat{G}$ exactly in the frozen gauntlet. We therefore claim the product form and not a specific power.

T WORKED EXAMPLES

White noise ($\Gamma = \sigma^2 I$, $h \geq 1$). $\gamma_h = 0$, so $w_{\text{raw}} = 0$, $b = 1$, $a = L/\sigma^2$, $\sigma_r^2 = \sigma^2$. Price $= \sigma^2/L$ and boundary $\delta_\star^2 = \sigma^2/L$: normalizing a pure-noise series costs exactly one window's worth of level-estimation variance and pays off as soon as the shift exceeds it. Finite sample: the shift coefficient becomes $1 + L/(T-L-1)$, so with $L = 24$ and $T = 100$ that is a 32% inflation purely from estimation error.

AR(1), $\varphi = 0.7$, $L = 24$, $h = 3$ (the running example). $b = 0.6570$, $a = 2.5800$, $\text{Risk}_{\text{RAW}}(0) = 1.7301$; price $= b^2/a = 0.16731$ (identity to 5×10^{-17}); boundary $\delta_\star = 0.6226$. Monte Carlo (4×10^5 paths): RAW risk at $\delta \in \{0, 1, 2, 4\}$: $1.7253/2.1574/3.4555/8.6528$ against theory $1.7301/2.1617/3.4567/8.6365$; IN risk 1.897 at every δ. Finite-T boundary $0.3876 \to 0.3892 \to 0.3825$ at $T = \infty/400/100$; the same formulas at $\varphi = 0.99$, $h = 1$ give $45.0 \to 12.3 \to$ negative (Thm. 2(iii)).

Local level, $q/\sigma = 0.3$, $L = 24$, $h = 3$. Constrained-optimal risk: closed form 1.52836, MC 1.52404 (and 1.52800 under level dispersion $V = 5000$, confirming level-invariance). Differencing optimum identical to 3×10^{-16} in the weights. A stationary-belief model ($\varphi = 0.7$ weights, $b = 0.657$) deployed on this drifting world decays at slope $b^2 q^2 = 0.03885$ per step: predicted $2.320/6.205/17.859/64.477$ at $m \in \{0, 100, 400, 1600\}$; simulated $2.322/6.206/17.867/64.471$.

U VERIFICATION TRANSCRIPTS

Verbatim outputs of the seeded verification scripts, regenerated by `reproduce_paper.py` at `--stage verify`.

VERIFY_CORE.PY

```
AR(1) phi=0.7, L=24, h=3; b = 1 - 1'w_raw = 0.657000
 delta | RAW analytic     RAW MC | IN analytic     IN MC
   0.0 |      1.73010    1.72533 |     1.89741   1.89663
   1.0 |      2.16175    2.15743 |     1.89741   1.91229
   2.0 |      3.45670    3.45548 |     1.89741   1.89886
   4.0 |      8.63648    8.65279 |     1.89741   1.89373

[R1] RAW risk growth coef (should equal b^2): 0.43164900000000006 vs b^2 = 0.43164900000000006
[R2] IN risk at delta=0 vs delta=4 (should be equal): 1.8974058139534886 1.8974058139534886
[R3] Price of IN direct vs formula: 0.16730581395348842 vs 0.16730581395348837
[R4] Regime boundary delta* = 0.6225728063646903 (BLUE std of window level; also = crossing
↪  point)
     at delta*, RAW - IN analytic gap (should be ~0): 0.0

Persistence sweep (h=1): price of IN and boundary delta*
  phi=0.00: b= 1.00000 price_of_IN= 0.041667 delta*= 0.2041
  phi=0.50: b= 0.50000 price_of_IN= 0.038462 delta*= 0.3922
  phi=0.90: b= 0.10000 price_of_IN= 0.023810 delta*= 1.5430
  phi=0.99: b= 0.01000 price_of_IN= 0.004505 delta*= 6.7116
```

VERIFY_RIDGE.PY

```
=== [R2,R3,R4,R5] random-covariance sweep ===
  L=6 lam=0.5: |KKT-reparam|=3.2e-16 |KKT-BFGS|=1.6e-08
  ↪  1'w=1.000000000000
  sweep failures: 0

=== [R3,R4] MC vs analytic (lam=1) ===
  b(lam)=0.9133 price(lam)=0.40934 delta*^2(lam)=0.49079
  delta= 0.000: RAW analytic   2.4078 MC 2.4167 | IN analytic                          2.8171
  ↪  MC     2.8220
  delta= 1.000: RAW analytic   3.2418 MC 3.2348 | IN analytic                          2.8171
  ↪  MC     2.8190
  delta= 0.701: RAW analytic   2.8171 MC 2.8179 | IN analytic                          2.8171
  ↪  MC     2.8193   <- crossing

=== [R1] empirical identity (bench code path == derived formulation) ===
  1'v (should be 0): 3.25e-13
  max|W_bench - W_derived|: 1.45e-14
  1'w_diff = 1.000000 (in class) max|w_diff - w_IN| = 0.0412 (nonzero => different estimator)

=== [R6] b(lam) growth under ridge (AR(1), L=24, h=3) ===
```

```
  phi=0.7: b(lam) = 0:0.6570, 0.01:0.6580, 0.1:0.6666, 1:0.7224, 10:0.8751
  ↪  monotone increasing: True
  phi=0.95: b(lam) = 0:0.1426, 0.01:0.1430, 0.1:0.1466, 1:0.1694, 10:0.2594
  ↪  monotone increasing: True

ALL RIDGE CHECKS COMPLETE
```

VERIFY_NUISANCE.PY

```
======================================================================
verify_nuisance.py -- nuisance-subspace theorem, all checks (seed 2027)
======================================================================
[N1] U^T((I-P_U)v+u_U)=psi, 200 random (L,k):        max err 5.22e-15
[N1] converse (feasible w realized with v=w):        max err 2.22e-15
[N2] Risk(w,c) vs R(w)+(b_U^T c)^2 (MC 4e5, zero-mean): rel err 1.40e-04
[N2] nonzero mean: identity off by -0.4031; predicted cross term 2E[e](b_U^Tc) = -0.4031
[N3] price_U = b_U^T A_U^-1 b_U vs null-space solve, 300 configs: worst rel err 2.12e-09
[N4] GLS unbiased (MC 3e5): max |mean-c| 4.13e-04; Cov vs A_U^-1: max err 2.69e-04
[N4] Gauss-Markov: 2000 random unbiased competitors, violations 0, min eig gap 1.23e+00
[N5] E[gain] MC +34.7381 vs b_U^T(M_c-A_U^-1)b_U +34.6822 (rel err 1.6e-03)
[N5] covariance form instead of M_c: benefit term 27.2425 vs 41.9269 -- understates by 1.5x
↪  here
[N6] fixed-c win set == complement of slab |b_U^Tc|<=sqrt(price): mismatches 0/100000
[N7] implemented generalized-IN ridge: feasibility U^Tw=psi max err 2.66e-15
[N7] == ridge centered at u_U (200 random configs, lam 1e-3..1e2): max weight dev 1.20e-14
[N8] k=1: b_U-b = 0.00e+00; A_U-a = 0.00e+00; price - b^2/a = 0.00e+00; A_U^-1 - 1/a =
↪  0.00e+00
[N8] agreement to >= 16 decimal places on the price
[N9] multi-output: sum of per-column prices 162.032016 vs tr(A_U^-1 B) 162.032016 (rel err
↪  7.0e-16)
======================================================================
ALL NUISANCE-SUBSPACE CHECKS COMPLETE
```

V E R I F Y G R I D S . P Y (SUMMARY LINES)

```
[G2] Prop 2 grid: 72 configs; max rel err (MC noise) = 0.0083; median = 0.0021; max IN-DIFF
↪  weight gap = 4
[G3] phi=0.5 done
[G3] phi=0.9 done
[G3] phi=0.99 done
[G3] RAW excess : median rel err +0.055, 90th pct 0.360, worst 0.863
[G3] IN excess : median rel err +0.070, 90th pct 0.187, worst 0.234
[G3] shift coef : median rel err +0.162, 90th pct 1.187, worst 5.128
[G3] boundary   : median rel err +0.191, 90th pct 1.008, worst 2.843
```

VERIFY_DRIFT.PY

```
local level: q=0.3, sig=1.0, L=24, h=3
constrained-optimal risk (formula): 1.528357
constrained-optimal risk (MC):      1.524040
                (with big level c): 1.527996                  <- must be identical [B]

[A] raw weights (trained at zero level spread): b = 0.0013
  Delta |   analytic |         MC
      0 |    1.52836 |    1.52881
    100 |    1.52837 |    1.53098
    400 |    1.52842 |    1.52678
   1600 |    1.52861 |    1.52429

[C] diff-model optimal risk: 1.5283565830                vs constrained optimum: 1.5283565830
↪  (gap 0.00e+00, weights g

[D] unconstrained optimum trained at level-variance V -> constrained optimum:
   V=    0: 1'w = 0.99866     max|w - w*| = 0.00035
   V=    1: 1'w = 0.99965     max|w - w*| = 0.00009
   V=   10: 1'w = 0.99996     max|w - w*| = 0.00001
   V= 1000: 1'w = 1.00000     max|w - w*| = 0.00000

[A'] stationary-belief model on drifting world: b = 0.6570, predicted slope b^2 q^2 = 0.038848
↪  per step
  Delta |   analytic |         MC
      0 |    2.32001 |    2.32228
    100 |    6.20485 |    6.20594
    400 |   17.85937 |   17.86715
   1600 |   64.47747 |   64.47105

w* head (should decay smoothly, most mass on recent points):
[0.058 0.0782 0.1054 0.1421 0.1916 0.2584] (last 6 weights, most recent last)
```

VERIFY_FINITE.PY

```
AR(1) phi=0.7, L=24, h=3: b=0.6570, a=2.5800, R_raw0=1.7301, R_in0=1.8974
population boundary delta*^2 = 0.3876

     T     delta |          RAW F1    RAW MC |        IN F2        IN MC
   100       0.0 |          2.2837    2.2850 |       2.4716       2.4704
   100       1.0 |          2.7749    2.7762 |       2.4716       2.4664
   400       0.0 |          1.8408    1.8401 |       2.0135       2.0127
   400       1.0 |          2.2844    2.2840 |       2.0135       2.0127
  1600       0.0 |          1.7565    1.7565 |       1.9251       1.9252
  1600       1.0 |          2.1909    2.1913 |       1.9251       1.9251

finite-T regime boundary delta*(T)^2              (population value 0.3876):
  T=       100: delta*^2 =    0.3825
  T=       400: delta*^2 =    0.3892
  T=      1600: delta*^2 =    0.3881
  T=1000000000: delta*^2 =    0.3876

overlapping windows from one path (excess risk vs iid-design formula):
  T= 400: RAW excess MC 0.1126 vs iid formula 0.1107      | IN excess MC
  ↪  0.1115 vs   0.1161
  T= 1600: RAW excess MC 0.0276 vs iid formula 0.0264     | IN excess MC
  ↪  0.0272 vs   0.0277

boundary collapse under persistence (phi=0.99, h=1):
  T=       100: delta*(T)^2 =    -29.0422   (population: 45.0450 ->                 -1.6x
  ↪  smaller)
  T=       400: delta*(T)^2 =     12.2504   (population: 45.0450 ->                  3.7x
  ↪  smaller)
  T=      1600: delta*(T)^2 =     34.4103   (population: 45.0450 ->                  1.3x
  ↪  smaller)
  T=1000000000: delta*(T)^2 =     45.0450   (population: 45.0450 ->                  1.0x
  ↪  smaller)
```

VERIFY_STDIN.PY

```
[S1] max |full-IN pred - mean-IN pred| with shared weights: 1.49e-13
[S2] max |W_fullIN - W_WLS(1/sig^2)|: 0.00e+00
[S3] mult regime: in_mean 5.4942 in_full 5.4799    -> in_full wins
```

```
[S3] add regime: in_mean 4.2291 in_full 4.3516    -> in_mean wins
[S4] dataset   T_eff/T   (observed in_full - in_mean MSE gap from bench)
     ETTh1      0.5683   (-0.0003)
     ETTh2      0.0203   (+0.0071)
     ETTm1      0.2891   (+0.0074)
     ETTm2      0.0489   (-0.0026)
     Exchange   0.0054   (+0.2689)
```

VERIFY_GRIDS.PY (SUMMARY LINES)

```
[G2] Prop 2 grid: 72 configs; max rel err (MC noise) = 0.0083; median = 0.0021; max IN-DIFF
↪  weight gap = 4
[G3] phi=0.5 done
[G3] phi=0.9 done
[G3] phi=0.99 done
[G3] RAW excess : median rel err +0.055, 90th pct 0.360, worst 0.863
[G3] IN excess : median rel err +0.070, 90th pct 0.187, worst 0.234
[G3] shift coef : median rel err +0.162, 90th pct 1.187, worst 5.128
[G3] boundary   : median rel err +0.191, 90th pct 1.008, worst 2.843
```

V ESTIMATOR PSEUDOCODE

All estimators operate on pooled channel windows $X \in \mathbb{R}^{N\times L}$, $Y \in \mathbb{R}^{N\times H}$ (z-scored by train statistics) and return $W \in \mathbb{R}^{L\times H}$.

RAW ridge. $W \leftarrow (X^\top X + \lambda I)^{-1} X^\top Y$; predict $\hat{Y} = X_{\text{test}} W$.

Mean-IN ridge. $m_i \leftarrow$ row means; $X_c \leftarrow X - m\mathbf{1}^\top$, $Y_c \leftarrow Y - m$; $V \leftarrow (X_c^\top X_c + \lambda I)^{-1} X_c^\top Y_c$; predict $\hat{Y} = X_{c,\text{test}} V + m_{\text{test}}$. Equivalently $W = V + u\mathbf{1}_H^\top$ solves uniform-centered constrained ridge.

Full-IN ridge. As mean-IN with X_c, Y_c divided rowwise by $s_i = \text{std}(X_i) + \epsilon$; prediction multiplies back. Equivalently mean-IN under row weights $1/s_i^2$.

DIFF ridge. Within-window first differences; solve; predict by cumulative summation from the last observed value.

Normalization diagnostic (linear, deployment time). From the fitted RAW ridge W, train windows, and test *inputs*: $\hat{b}^2 \leftarrow$ mean over horizon columns of $(1 - \mathbf{1}^\top W_{\cdot j})^2$; $\hat{a} \leftarrow \mathbf{1}^\top \hat{\Gamma}^{-1} \mathbf{1}$; $\hat{\delta}^2 \leftarrow$ mean squared test-input window mean; $\hat{G}_\lambda = \hat{b}_\lambda^2 \hat{\delta}^2 - \widehat{\text{price}}_\lambda$ with $\widehat{\text{price}}_\lambda$ the exact ridge price of Prop. 1 at the selected λ (substituting the OLS form $\hat{b}^2/\hat{a}$ is the weaker ablation of App. L); add mean-IN if $\hat{G}_\lambda > 0$ by a margin, treat $|\hat{G}_\lambda| \le 10^{-3}$ as a tie.

Normalization diagnostic (deep backbone). Replace $\hat{b}^2$ by $\hat{S} = \mathbb{E}[(1-s)^2]$ from one forward-mode JVP on the trained un-normalized model over unlabeled deployment windows, and rank cells by $\hat{S}\hat{\delta}^2$ (Sec. 8). Ranking only: no calibrated magnitude is claimed.

Full-IN safety check (train time). Compute NTE on train windows with $s_i > 10^{-4}$; if NTE is orders of magnitude above $O(1)$, do not use the std division.

W Dataset details

dataset	channels	length	sampling	source/notes
ETTh1 / ETTh2	7	17,420	hourly	electricity-transformer temperature; calendar splits
ETTm1 / ETTm2	7	69,680	15 min	same system, finer sampling
Exchange	8	7,588	daily	currency rates; near-integrated
Weather	21	52,696	10 min	meteorological sensors; several near-constant channels
Electricity	321	26,304	hourly	client-level consumption; strongly periodic
Traffic	862	17,544	hourly	road occupancy; strongly periodic

Character relevant to the theory: Exchange is integrated (fitted raw weight sums ≈ 0.93); ETTh2 and ETTm2 carry visible level drift (largest $\hat{\delta}^2$); Electricity/Traffic are the most stationary/periodic (near-tie regime); Weather and Exchange contain near-constant windows whose futures move (largest NTE). The deep experiment uses the first six only.

X Practical guidance

(1) *Always* consider mean-IN on persistent or drifting data: its population price vanishes with persistence (Cor. 1), finite samples push the boundary further down (Thm. 2), dependence pushed it down further still in our simulations, and on deep backbones mean-IN beat the un-normalized model in every cell we ran (Sec. 8). (2) On large, stationary, periodic datasets, raw ridge can be slightly better at every horizon; the price is real but small, so decide with the diagnostic rather than by default. (3) Compute NTE_{trim} before enabling the std division on a linear model; it is not validated for deep backbones. (4) Differencing and mean-IN are the same model class; prefer whichever estimator's shrinkage geometry suits the data, but do not expect a different hypothesis space. (5) Longer horizons raise the stakes in whichever direction the regime points; re-run the diagnostic per horizon.

Y Extended classical background

Local level and exponential smoothing. $y_t = \mu_t + \varepsilon_t$, $\mu_t = \mu_{t-1} + \eta_t$ has reduced form ARIMA$(0,1,1)$; Muth (1960) showed the minimum-MSE one-step predictor is an EWMA whose

rate is set by q^2/σ^2. Our App. D setting differs in three ways that matter for modern practice: a finite window of length L, H horizons fitted jointly, and weights estimated from T windows by ridge-regularized least squares. Props. 4–2 are about exactly those differences.

Differencing, ARIMA, structural breaks. Classical practice removes stochastic levels by differencing (Box and Jenkins, 1970); Prop. 4(iii) shows the modern IN operator spans the same hypothesis space, so the differences are purely estimation-side. The econometric literature attributes most forecast failure to deterministic shifts, precisely our δ, and recommends robust devices whose costs and benefits were discussed qualitatively (Clements and Hendry, 1998); our contribution is the exact finite-window price b^2/a, the exact benefit $\delta^2 b^2$, and the finite-sample interaction.

Covariate shift. The deployment shift $x \mapsto x + \delta\mathbf{1}$ is a covariate shift in the sense of Quiñonero-Candela et al. (2009), but the standard remedy there is importance weighting on a known density ratio (Sugiyama et al., 2007); mean-IN instead removes the shifted direction by construction, which is why its benefit is exact and its price is a variance, not a weighting error.

Restricted least squares and GLS. The sum-to-one constraint makes mean-IN a restricted least-squares estimator; the classical theory gives its algebra (Aitken, 1935; Rao, 1973) but is silent on the forecasting-specific quantities analyzed here. The std-division result is an instance of Aitken efficiency; the Kish effective sample size (Kish, 1965) is the standard survey summary of weight dispersion, and our finding is precisely that it is *not* the operative quantity for this failure mode.

Z Reproducibility

Python 3.11.15; `numpy` 2.4.4, `scipy` 1.17.1, `matplotlib` 3.10.9 (pinned). Seeds fixed in every script. Data: ETT from the ETDataset repository; Exchange/Electricity/Traffic from the standard multivariate benchmark archive; Weather from the Autoformer bundle; retrieval instructions included, raw data not redistributed. Entry point `reproduce_paper.py`, with `--stage` one of `verify`, `bench`, `analysis`, `figures`, `paper`; figures and longtables are generated from result artifacts, with no manually transcribed numbers. The deep-forecaster harness is released separately with the frozen preregistration, the 324-row artifact, the risk curves, the compliant probe CSV, and the two analysis scripts that regenerate Table 1 and App. S verbatim; GPU model, driver, CUDA, framework and library versions are recorded in the run log.