WHAT DOES ATTENTION RETAIN? INFORMATION LOSS AND SUFFICIENT MOMENTS FOR IN-CONTEXT REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

What information must an attention memory retain for accurate in-context prediction? Gaussian regression makes second moments sufficient, so success on this benchmark can hide the cost of discarding other information. We characterize that cost for complete degree-P moments of the prompt: their local information is the squared L^2 norm of the noise score projected onto polynomials of degree $P-1$. A strong limit of the retained statistical experiment makes the resulting risk bound apply to every measurable decoder. These moments are fully locally efficient exactly when the log noise density is a polynomial of degree at most P; in that case, they are also sufficient at every sample size. Yet a smooth density within 5% of Gaussian, matching its first eight moments exactly, can contain more than thirteen times the information accessible to degree-four moments. We connect this hierarchy to attention: common-source affine layers retain only the Gram, while a source-side score operation reaches full-data efficiency. Standard causal Transformers improve Laplace regression risk by 26.6% over a learned Gram control. A separate, task-preserving diagnostic gives positive numerical information-loss witnesses for every Gram decoder under Laplace and quartic noise. Together, the results make information retention a distinct criterion for evaluating attention memories.

1 INTRODUCTION

An attention memory determines which evidence remains available to a learner after the prompt has been aggregated. This creates a basic design question: *does the memory preserve the information needed for the task, even if its decoder has unlimited computational power?* In-context regression is a useful setting for answering it because the target and the available evidence are explicit. Its familiar Gaussian formulation, however, makes second moments sufficient. A model can therefore succeed on Gaussian tasks while discarding information that becomes essential under another noise law.

Changing only the additive noise exposes this problem while keeping the linear regression target fixed. Our sharpest example is a smooth density within 5% of Gaussian at every point, with exactly the same first eight moments, but more than thirteen times its location Fisher information. Every decoder of complete degree-four prompt moments retains only the Gaussian level of local information (Figure 1). For any fixed moment budget, we construct such a fixed noise law with an arbitrarily large asymptotic efficiency loss and an arbitrarily small relative density change. Thus closeness to the Gaussian benchmark in density and low-order moments can coexist with a large information gap.

We establish a hierarchy that explains which information survives. For centered isotropic absolutely continuous designs, complete degree-P moments retain the noise score projected onto polynomials through degree $P-1$: one degree is spent on the task-bearing covariate. We prove a strong limit for the entire retained statistical experiment, so the resulting risk bound covers arbitrary nonlinear and discontinuous decoders. This gives an exact losslessness criterion. A finite degree is fully locally efficient precisely when it can represent the log noise density; the same moments are then sufficient at every sample size. Gaussian noise needs degree two, quartic log density needs degree four, and Laplace noise has no finite lossless degree.

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.

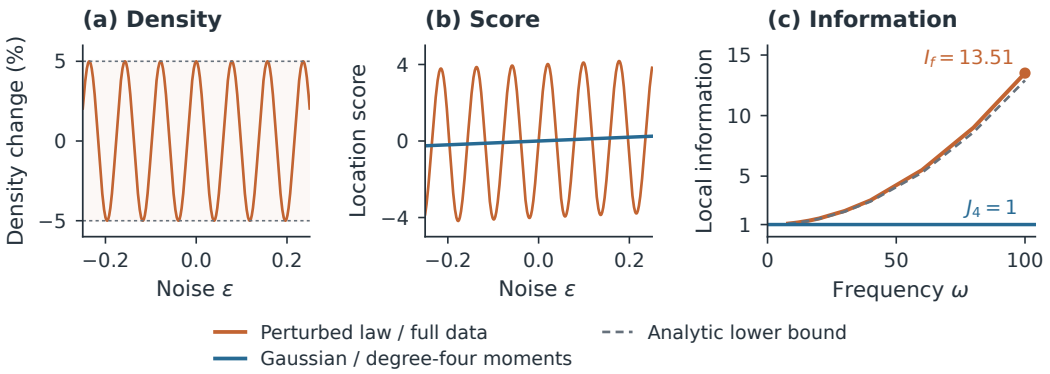

Figure 1: **Near-Gaussian noise can hide information from every moment decoder.** The explicit family in Lemma J.2, with $P = 4$ and $\delta = 0.05$, matches Gaussian moments through order eight exactly. (a–b) Relative density and location score at $\omega = 80$ on a zoomed noise interval. (c) Full-data information grows with frequency, while complete degree-four prompt moments retain $J_4 = 1$. At $\omega = 100$, $I_f \approx 13.508$ and an analytic lower bound is 12.905. Curves show population Fisher information; moment matching is exact by construction.

The hierarchy has a direct architectural interpretation. Common-source affine attention factors through the augmented Gram at every finite depth and width, extending the all-layer recurrences of Vladymyrov et al. (2024). Its optimal integrated risk has leading term $d\sigma^2/N$. A scalar source-side residual-score operation supplies the missing information and reaches the full-data coefficient d/I_f. This identifies both the restriction imposed by aggregation and a concrete way to overcome it before the records are compressed.

We then test information use in standard causal softmax Transformers. A four-block model with rotary positions, RMSNorm, and SwiGLU improves Laplace prediction risk over a learned Gram control in independent two- and four-dimensional studies. To measure the intrinsic cost of the summary, we also construct task-preserving pairs of equal-Gram prompts. Their reversible model drift yields a finite-context lower bound on the Bayes penalty of *every* Gram decoder. Independent calibration and replication produce positive numerical witnesses under Laplace and quartic noise, with a Gaussian negative control.

Together, the moment experiment, source-score construction, and Transformer audit connect information retention to attainable prediction risk, from an exact statistical limit to trained-model behavior.

2 Related Work

In-context computation and attention memories. Attention-based regression analyses identify gradient updates and optimal linear-attention predictors (von Oswald et al., 2023; Ahn et al., 2023; Mahankali et al., 2024; Fu et al., 2024). All-layer linear representations and fast weights expose related aggregated states (Vladymyrov et al., 2024; Schlag et al., 2021). Nonlinear feature extraction, gating, and higher-order memories expand the available computations (Zhang et al., 2024; Sun et al., 2025; Li et al., 2025; Zhang et al., 2025). We characterize the statistical experiment retained by a complete summary and optimize risk over all its decoders. This complements high-dimensional learning curves (Lu et al., 2025).

Non-Gaussian regression and statistical compression. Studies of noisy-label ICL compare laws and estimators (Cheng et al., 2025; Cao et al., 2026); Raj & Reddy (2026) study non-Gaussian task priors. Here the regression target and task prior remain fixed while the noise changes. Sufficiency, experiment comparison, and efficient scores provide the classical foundations (Blackwell, 1953; Le Cam & Yang, 2000; Stein, 1956; Bickel, 1982; van der Vaart, 1998). Score projection is classical in estimating-function theory (Godambe, 1960; Heyde, 1997) and moment information hierarchies

(Jarrett, 1984; Carron, 2011). Carron (2011, Eq. (11), Appendix C) gives the univariate and multivariate score decomposition; Hansen (1982, Theorem 3.2) optimizes covariance over consistent GMM estimators satisfying his moment equations. Our result establishes a strong limit for the actual complete mixed-moment experiment in regression, then transfers it to local and integrated risks over all measurable decoders. The losslessness classification connects local information to exact likelihood factorization (Pitman, 1936; Koopman, 1936). Near-Gaussian high-Fisher examples already appear in Gupta et al. (2022); the additional separation here matches any prescribed finite set of moments exactly and transfers the information loss to every decoder of the retained envelope.

Finite-context information witnesses. Optimal-risk feature quality and model-agnostic variable importance have decision-theoretic foundations (van Rooyen & Williamson, 2014; Williamson et al., 2021). Layerwise Fisher diagnostics use moment-based lower bounds and auxiliary nonlinear expansions (Weimar et al., 2025). Conditional synthetic controls and Metropolis constructions are also established (Candès et al., 2018; Bates et al., 2021). Our information-loss witness combines an exact attention summary, a task-preserving kernel, and the classical Markov-chain drift $g - Kg$ (Dellaportas & Kontoyiannis, 2012).

3 Information Retained by a Moment Memory

3.1 Task, Risk, and Observable Information

Let $D = \{(X_i, Y_i)\}_{i=1}^N$ be a prompt from

$$W \sim \pi, \qquad X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d), \qquad Y_i = X_i^\top W + \varepsilon_i,$$

where $\varepsilon_i \overset{\text{iid}}{\sim} f$, independently of the task and design. The target is the clean value $Z = X_q^\top W$ at an independent $X_q \sim \mathcal{N}(0, I_d)$. Gaussian design is the default architecture and Transformer setting; Theorem 3.1 explicitly relaxes this assumption. We evaluate squared prediction risk $R_N = \mathbb{E}(\widehat{Y}_q - Z)^2$ and its normalized coefficient $C_N = NR_N/d$. Predicting the clean target avoids adding an irreducible query-label noise term. An estimator of W has exactly the same integrated coefficient risk as its linear query predictor.

Write $z_i = (X_i, Y_i)$ and $\widetilde{z}_i = (z_i, 1)$. The augmented Gram is

$$T_N(D) = \frac{1}{N} \sum_i \widetilde{z}_i \widetilde{z}_i^\top. \tag{1}$$

It includes $X^\top X$, $X^\top Y$, $Y^\top Y$, and the covariate and label sums. Thus T_N is the complete degree-two moment summary, including the first-order statistics available to affine maps.

For square-integrable Z and any summary $T = T(D)$, let $m(D, X_q) = \mathbb{E}[Z \mid D, X_q]$. Conditional expectation gives the exact decomposition

$$R_T^* - R_{\text{full}}^* = \mathbb{E}[\text{Var}\{m(D, X_q) \mid T, X_q\}]. \tag{2}$$

This classical identity separates discarded decision information from a particular model's computational error. A summary is sufficient for the squared-loss decision exactly when the posterior mean is constant along its conditional fibers, up to null sets. Our theory identifies an architectural summary and computes the leading cost of that restriction.

3.2 From Second Moments to a Hierarchy of Experiments

Let $\mathcal{M}_{P,N}$ retain every mixed empirical moment in (X, Y) through total degree P. Define

$$v_P(e) = (1, e, \ldots, e^{P-1})^\top, \qquad B_P = \mathbb{E}[v_P(\varepsilon) v_P(\varepsilon)^\top], \qquad a_P = \mathbb{E}[v_P'(\varepsilon)].$$

Theorem 3.1 (The local information of complete moments)**.** *Fix d and an integer $P \geq 2$. The covariates may have any law absolutely continuous with respect to Lebesgue measure, with $\mathbb{E}X = 0$, $\mathbb{E}XX^\top = I_d$, and finite $2P$th moment. For independent i.i.d. mean-zero errors with an everywhere-positive density and finite $2P$th moment, put $J_P = a_P^\top B_P^{-1} a_P$. For every center w_0 and compact $H \subset \mathbb{R}^d$, observing $\mathcal{M}_{P,N}$ under $w = w_0 + h/\sqrt{N}$, $h \in H$, converges in Le Cam distance to observing*

$$Z = h + J_P^{-1/2} G, \qquad G \sim \mathcal{N}(0, I_d).$$

With locally absolutely continuous f, weak location score $\psi = -f'/f \in L^2(f)$, and $x^k f(x) \to 0$ at both tails for $0 \le k \le P-1$, the classical score-projection identity (Carron, 2011) follows by integration by parts:

$$J_P = \|\Pi_{\mathcal{P}_{P-1}}\psi\|^2_{L^2(f)}, \qquad J_2 = \sigma^{-2} \le J_P \le I_f, \tag{3}$$

where $\mathcal{P}_{P-1}$ is the error-polynomial space through degree $P-1$. After whitening, this coefficient depends on the noise and degree budget, but not on higher-order design geometry. The all-measurable local asymptotic minimax coefficient, expanding local cubes after the sample-size limit, is d/J_P. If the design, noise, and a continuous, everywhere-positive prior have moments of every order, the optimal integrated risk is $d/(J_P N) + o(N^{-1})$ (Theorem J.1 and proposition L.1).

Why degree P retains a degree-$(P-1)$ score. The informative coordinates are $X_j \varepsilon^b$, with $0 \le b \le P-1$. Orthogonalizing the design polynomials isolates these coordinates: their mean derivative is a_P, their covariance is B_P, and all remaining blocks become independent Gaussian ancillaries. A total-variation central limit theorem, obtained from a full-rank polynomial sum map, transfers this calculation to the entire observed experiment. This strong convergence is what allows arbitrary measurable decoders in the risk bound.

The characterization also gives a decoder. Let $g_P(e) = a_P^\top B_P^{-1} v_P(e)$ be the projected score and $u = N^{-1}\sum_i X_i Y_i$. The correction

$$\widehat{W}_P = u + \frac{1}{J_P N}\sum_i X_i g_P(Y_i - X_i^\top u) \tag{4}$$

uses only degree-P prompt moments: expanding each polynomial residual multiplies at most $P-1$ record coordinates by one covariate. Under the integrated-risk assumptions above, it attains $d/(J_P N) + o(N^{-1})$. The memory budget therefore determines both a sharp limit and the score component that an efficient decoder should use. Sections J.2 and L.1 give the strong limit and risk proofs.

3.3 When a finite moment budget is lossless

The hierarchy yields a classification, rather than only a sequence of upper and lower bounds. The following statement uses the score and tail conditions above, $0 < I_f < \infty$, and a locally absolutely continuous, everywhere-positive f.

Theorem 3.2 (Local efficiency coincides with exact sufficiency)**.** *Under the conditions of Theorem 3.1 and the stated score regularity, the following conditions are equivalent:*

1. $J_P = I_f$*: degree-P moments retain full local information;*
2. *the score ψ is a polynomial of degree at most $P-1$, f-a.e.;*
3. $f(e) = \exp\{q(e)\}$ *for a polynomial q of degree at most P;*
4. $\mathcal{M}_{P,N}$ *is sufficient for w for every sample size N.*

An integrable density in (3) has an even polynomial degree and negative leading coefficient. Its smallest lossless total degree is $\deg q$.

Why this is an exact classification. Equality in an orthogonal projection holds exactly when the score lies in the projected space. Integrating the score gives $\log f = q$. Then $\sum_i q(Y_i - X_i^\top w)$ is determined by degree-P mixed moments, which factor the entire likelihood at every N. Conversely, exact sufficiency for every N preserves local testing information; the full-data and moment binary local testing limits have information $I_f I_d$ and $J_P I_d$, forcing equality. Section L.2 supplies the argument, including the regularity needed for the full-data local limit.

Gaussian noise is lossless at degree two, quartic log density at degree four, and $p = 8$ exponential-power noise at degree eight. Laplace's sign score is not a polynomial, so no fixed complete moment degree is locally lossless. Since an integrable polynomial log density has even leading degree, an odd budget P introduces no new exactly lossless noise family beyond $P-1$. The noise score therefore specifies the smallest lossless moment memory. Figure 3 in the appendix shows the corresponding analytic risk coefficients.

3.4 A FIXED MOMENT BUDGET CAN MISS ARBITRARILY MUCH INFORMATION

The classification has no uniform approximate converse: being close to a lossless noise law does not guarantee comparable asymptotic efficiency.

Corollary 3.3 (Arbitrarily costly fixed moment budgets)**.** *For every integer $P \geq 2$, every $M > 1$, and $0 < \delta < 1$, there exists a fixed even, smooth, everywhere-positive noise density f with*

$$(1-\delta)\varphi \leq f \leq (1+\delta)\varphi, \qquad \mathbb{E}_f \varepsilon^j = \mathbb{E}_\varphi \varepsilon^j \quad (0 \leq j \leq 2P), \qquad J_P = 1, \quad I_f \geq M,$$

where φ is the standard Gaussian density.

An explicit construction. Let ℓ_j be the Lagrange basis on $\{0^2, 1^2, \ldots, P^2\}$ and set $c_j = e^{(j^2-\omega^2)/2}\ell_j(\omega^2)$ for $\omega > P$. Then

$$h_\omega(e) = \frac{\cos(\omega e) - \sum_{j=0}^{P} c_j \cos(je)}{1 + \sum_{j=0}^{P} |c_j|}, \qquad f_\omega = \varphi(1 + \delta h_\omega).$$

Gaussian Fourier identities and polynomial interpolation cancel all moments through degree $2P$ exactly. The score projection remains e, so $J_P = 1$, whereas $I_{f_\omega} = 1 + \delta^2 \mathbb{E}_\varphi[h_\omega'^2/(1+\delta h_\omega)]$ with $I_{f_\omega} = 1 + (1 - \sqrt{1-\delta^2})\omega^2 + o(\omega^2)$ for fixed P, δ. Thus arbitrarily small density oscillations can carry arbitrarily much score information invisible to the moment budget. Lemma J.2 proves these identities and an explicit finite-frequency bound; Figure 1 illustrates it. For each P, M, δ, one density is fixed before $N \to \infty$. Even the optimal decoder incurs the asymptotic risk-inflation factor $I_f/J_P \geq M$.

4 THE ATTENTION BOUNDARY AND HOW TO CROSS IT

4.1 A SHARP GRAM–FISHER BOUNDARY

In the common-source affine class, each layer applies

$$e_j^+ = Re_j + r + \sum_h O_h G_h (Q_h e_j + q_h), \qquad G_h = \frac{1}{N}\sum_i (V_h e_i + v_h)(K_h e_i + k_h)^\top, \quad (5)$$

using the same demonstrations for every destination; the query is never a source. Initialization and readout are affine. Fixed token-type embeddings are allowed, but tokenwise nonlinearities, source normalization, softmax, demonstration-specific positions, and causal source masks are excluded.

Theorem 4.1 (Information envelope and sharp Gram risk)**.** *Every finite-depth, finite-width model in (5) factors through $(T_N(D), X_q, N)$. This remains true with arbitrary deterministic query-only maps while the source paths remain in the class. Fix d. If π is continuous and everywhere positive and the errors have an everywhere-positive density, zero mean, variance σ^2, and finite fourth moment, then*

$$\inf_{\widehat{W}\ T_N\text{-measurable}} \mathbb{E}_\pi \|\widehat{W} - W\|^2 = \frac{d\sigma^2}{N} + o(N^{-1}). \quad (6)$$

If additionally $\mathbb{E}\|W\|^2 < \infty$, two affine attention layers of width $d+1$ attain this coefficient.

Proof structure. The argument attaches a statistical lower bound to the all-layer closure underlying Vladymyrov et al. (2024). In homogeneous coordinates, every layer applies a shared matrix $M_\ell(\bar{C}_{\ell-1})$, where $\bar{C}_{\ell-1}$ is the source second moment. Consequently,

$$\bar{C}_\ell = M_\ell(\bar{C}_{\ell-1})\bar{C}_{\ell-1}M_\ell(\bar{C}_{\ell-1})^\top.$$

The initial $\bar{C}_0$ is a function of T_N; induction gives the factorization, including query-only computation without source feedback.

For the risk bound, a total-variation limit of the complete recentered Gram has informative shift $h + \sigma G$ and parameter-free ancillary coordinates. Local Bayes comparisons and prior exhaustion transfer this limit to all measurable decoders; OLS supplies the upper bound. The two-layer rule

$(2I - H)u$, with $H = X^\top X/N$ and $u = X^\top Y/N$, attains the coefficient within the architecture under the stated prior moment condition. Full proofs appear in Sections K.1 to K.3.

For regular location noise, let $\psi = -f'/f$ and $I_f = \mathbb{E}\psi(\varepsilon)^2$. Under the explicit smoothness and boundary conditions in Theorem F.4,

$$R^*_{\text{full},N} = \frac{d}{I_f N} + o(N^{-1}), \qquad N(R^*_{\text{Gram},N} - R^*_{\text{full},N}) \longrightarrow d(\sigma^2 - I_f^{-1}). \tag{7}$$

The full-data endpoint follows from a smooth-score construction (Theorem F.4), a nonsmooth Laplace argument (Corollary I.1), or the polynomial-score moment law (Corollary B.1), according to the noise family. The inequality $\sigma^2 I_f \geq 1$ is classical, with equality only for Gaussian noise in the regular class. Thus equal variance does not imply equal recoverable information.

4.2 A MATCHED SOURCE-SIDE ESCAPE

For smooth-score noise under Theorem F.7, or polynomial log-density noise under Corollary B.1, and a Gaussian task prior, one scalar source operation between two affine attention layers implements

$$\widehat{W} = u + \frac{1}{I_f N} \sum_i X_i \psi(Y_i - X_i^\top u), \tag{8}$$

and attains the Fisher coefficient. The first layer forms a pilot and source residuals; the source nonlinearity computes their scores; the second aggregates a correction. The theorem accounts for reuse of the same prompt in the pilot and correction. Under Laplace noise, the corresponding architectural endpoint uses a sharpening $\tanh$ sequence with a controlled gain rate (Corollary I.2).

The construction identifies where the additional statistic enters the architecture: at the source, before aggregation. It also predicts which record-level response is efficient for each noise law. For unit-variance exponential-power noise,

$$f_p(e) = \frac{p}{2\alpha_p \Gamma(1/p)} \exp[-|e/\alpha_p|^p], \qquad \alpha_p^2 = \frac{\Gamma(1/p)}{\Gamma(3/p)}, \tag{9}$$

the score is $p\ \text{sign}(e)|e|^{p-1}/\alpha_p^p$. Laplace ($p = 1$), Gaussian ($p = 2$), and quartic ($p = 4$) tasks therefore have sign, linear, and cubic score responses, respectively; the Laplace construction uses a sharpening $\tanh$ sequence to approximate the sign. Here "score" denotes the statistical location score. The efficient response depends on the noise law: it saturates for Laplace errors and amplifies large residuals for quartic errors.

5 INFORMATION USE IN STANDARD TRANSFORMERS

The exact affine theorem suggests questions for richer models: can they use information absent from the Gram, and is that information useful for prediction? We evaluate frozen trained models on independent tasks, then use a distribution-preserving diagnostic to measure useful information beyond the Gram.

5.1 ARCHITECTURAL PLACEMENT AND STANDARD-MODEL RISK

A controlled placement test. A four-block common-source model places the same learned FFN on either the demonstration or query path. Both use width 64, four heads, $d = 2$, unit-variance $p = 8$ noise, three training seeds, and 5,000 updates. On 20,000 independent tasks at $N = 128$, normalized source/query risks are 0.4683 ± 0.0053 and 1.0238 ± 0.0008 (training-seed SD), a 54.3% reduction. The task-bootstrap 95% interval for query minus source is $[0.5310, 0.5809]$, conditional on the fitted models. A float64 equal-Gram test changes source predictions by RMS 0.1515–0.1630, while query-only changes remain below 4×10^{-16}. The matched placement test shows that transforming records before aggregation unlocks predictive information that query-side computation cannot recover.

A standard causal Transformer. Our four-block, width-128 model uses four-head causal softmax attention, rotary positions (Su et al., 2021), pre-RMSNorm (Zhang & Sennrich, 2019), and SwiGLU

Table 1: **Independent standard-Transformer risk.** Normalized risk NR_N/d at $N = 128$. Learned entries are means $\pm$ sample SD over three training seeds. Intervals are paired task Monte Carlo 95% intervals after averaging losses across fitted seeds; they are conditional on these models. The $d = 4$ Laplace experiment is an independent confirmation.

	$d = 2$			$d = 4$
Predictor	Laplace	Gaussian	Quartic	Laplace
Transformer	$.7436 \pm .0236$	$1.0394 \pm .0005$	$.9796 \pm .0105$	$.8352 \pm .0125$
Gram control	$1.0136 \pm .0001$	$.9970 \pm .0002$	$.9990 \pm .0001$	$1.0181 \pm .0001$
Ridge	1.0136	.9969	.9989	1.0180
Transformer − Gram	$-.2700$	.0424	$-.0194$	$-.1829$
95% interval	$[-.2927, -.2473]$	$[.0348, .0500]$	$[-.2680, .2293]$	$[-.2214, -.1444]$
Asymptotic $1/I_f$	.5000	1.0000	.7295	.5000

(Shazeer, 2020). It receives raw numerical covariate and labeled-record tokens. Separate Laplace, Gaussian, and quartic models use $W \sim \mathcal{N}(0, I_2)$, unit noise variance, three seeds per law, 12,000 updates, and lengths $N \in \{32, 64, 128\}$ after a short-context curriculum. The Gram control augments prior-matched ridge with a residual MLP receiving every augmented-Gram block, query, and length. The Transformer and control have 657,281 and 534,017 parameters. Each law/length uses 16,384 fresh tasks shared across fitted models. Section A gives training and evaluation details.

The Laplace advantage over the learned Gram control is 26.6% at $d = 2$ and 18.0% at $d = 4$, with improvement in all three seeds in each study. Gaussian risk remains above ridge, and the primary quartic risk contrast is inconclusive. At the extrapolation length $N = 256$, seed-averaged Transformer risks exceed their Gram controls under every evaluated law. Sections A.5 and D report all lengths and seeds. These results establish a finite-length Laplace gain. The next diagnostic asks a complementary question: how much prediction-relevant information does the Gram itself discard, even under optimal decoding?

5.2 A DISTRIBUTION-PRESERVING GRAM-FIBER AUDIT

An equal-Gram perturbation isolates information dependence. Arbitrary rotations, however, change a non-Gaussian task law. We correct this using a reversible likelihood test, without training or tuning the model. Let $A = [\mathbf{1}, X]$, write $Y = P_A Y + r$, and sample a unit vector v uniformly orthogonal to both A and r. At fixed angle θ, propose

$$\widetilde{Y} = P_A Y + \cos(\theta) r + \sin(\theta) \|r\| v. \tag{10}$$

The proposed prompt has the same complete augmented Gram. For a generated task with latent coefficient W, accept it with probability

$$a = \min\left\{1, \frac{\prod_i f(\widetilde{Y}_i - X_i^\top W)}{\prod_i f(Y_i - X_i^\top W)}\right\}; \quad Y' = \begin{cases} \widetilde{Y}, & \text{accepted,} \\ Y, & \text{otherwise.} \end{cases} \tag{11}$$

Proposition 5.1 (An exact task-preserving pair)**.** *Starting from an independent task drawn from the specified model, (10)–(11) preserve the joint law of (W, X, Y) and the augmented Gram, with fixed W and X. Both endpoints therefore have the original task distribution after a single step. For Gaussian noise every proposal is accepted.*

Conditionally on (W, X, T_N), the labels lie on an affine sphere with density proportional to their likelihood. The proposal is symmetric on that sphere; the acceptance rule gives detailed balance. The original sample already has this conditional law, so no burn-in is needed. Section L.3 covers the geometry and degenerate cases. The synthetic diagnostic uses latent W to preserve the task law; the predictor receives the prompt and query. Every rejected proposal is retained as a zero-change pair, preserving the original-distribution average.

We first evaluate one stationary pair per task. Writing $\Delta_g = g(D', X_q) - g(D, X_q)$ and $\Delta_* = m(D', X_q) - m(D, X_q)$, we compare $\mathbb{E}(\Delta_g - \Delta_*)^2$ with the zero-change reference $\mathbb{E}\Delta_*^2$ imposed on every Gram predictor.

We fix $\theta = 0.25$ before generating 256 independent original tasks for each $p \in \{1, 2, 4\}$ and $N \in \{32, 128\}$. All nine unchanged checkpoints are evaluated in CPU float64, including every rejected transition. Posterior means use two refinement/domain settings for non-Gaussian laws and analytic ridge for Gaussian noise. All 1,536 pairs pass the specified checks; the maximum between-setting prediction discrepancy is 9.04×10^{-5}.

The all-pair discrepancy contrast is negative for Laplace and quartic noise at both lengths. Gaussian posterior changes are below 10^{-15} RMS, while Transformer responses remain nonzero. Thus the non-Gaussian responses follow posterior-relevant distinctions under the original task law. Section M gives all six cells and pointwise intervals; these pairs also supply the calibration data for the new witness below.

From model responses to an all-decoder information bound. Let $S = (T_N(D), X_q)$ and $B_N = R^*_{S,N} - R^*_{\text{full},N}$. B_N measures the information loss under optimal summary decoding. A task-preserving pair makes this quantity accessible without fitting that optimal decoder.

Theorem 5.2 (Finite-context information witnesses)**.** *Assume $\mathbb{E}Z^2 < \infty$. Let $m = \mathbb{E}[Z \mid D, X_q]$ and let g be any fixed square-integrable predictor. For a stationary same-summary pair, set $a_g = \mathbb{E}[(g' - g)(m' - m)]$ and $b_g = \mathbb{E}(g' - g)^2$. Then*

$$B_N \geq \frac{\mathbb{E}(m' - m)^2}{4} \geq \frac{a_g^2}{4b_g}. \tag{12}$$

If this pair is generated by a summary-preserving kernel K reversible with respect to the original joint task law on the full simulation state, define $d_g = g - Kg$ and $c_g = \mathbb{E}d_g^2$. For every fixed $t \in \mathbb{R}$,

$$B_N \geq ta_g - t^2 c_g, \qquad a_g = 2\mathbb{E}[(m - r)d_g], \qquad c_g \leq b_g, \tag{13}$$

where r is any square-integrable summary-measurable reference. Optimizing gives $B_N \geq a_g^2/(4c_g)$. Ratios with zero denominator are defined as zero.

For (12), subtract $\mathbb{E}[m \mid S]$ at both endpoints and apply Cauchy–Schwarz. For (13), stationarity gives $\mathbb{E}[d_g \mid S] = 0$, while reversibility gives the identity for a_g; expand $\mathbb{E}(m - \mathbb{E}[m \mid S] - td_g)^2 \geq 0$. Sections N.1 and N.2 give full proofs and numerical-error conditions. The variance and Markov drift identities are classical (Dellaportas & Kontoyiannis, 2012); their role is to turn the exact architectural summary and task-preserving kernel into a finite-context information test. A positive witness identifies predictive information missing from the summary, even when the best summary decoder is computationally inaccessible. For target-preserving kernels, a population risk upper bound U on the full-data Bayes risk sharpens any positive population drift witness $0 < \ell \leq a_g^2/(4c_g)$ to $B_N \geq U\ell/(U - \ell)$ (Corollary N.3).

Independent drift evaluation and replication. An inspected 256-task bank calibrates a scalar t for the fixed mean prediction of three Transformer checkpoints. After an initial fresh 1,024-task confirmation per non-Gaussian law, we freeze a further 4,096-task replication per law to check heavy positive tails. Every task uses 16 independent Metropolis branches restarted at the original prompt; rejections remain. Ridge is the reference r, and distinct-branch products estimate c_g without squaring a noisy drift. The original calibration fixes $t = 7.3071$ (Laplace) and 7.7783 (quartic). Only original-prompt posteriors are needed.

Both non-Gaussian raw replication intervals and the pre-fixed upper-capped intervals are positive. The latter reduce favorable extreme contributions without removing any task or negative value, and remain conservative population witnesses by pointwise domination. The quartic result is particularly informative: its positive witness identifies useful information beyond the Gram even though its total-risk comparison is inconclusive. Retaining that information and converting it into a net predictive gain are separate requirements. Sections O and O.1 report both banks, all seeds, bootstrap intervals, influence checks, and posterior sensitivity.

5.3 Checking the moment law beyond Gaussian designs

A classical degree-four posterior-mode decoder checks the moment law beyond Gaussian covariates, separately from model training. At $N = 2048$, Gaussian, standardized Gaussian-mixture, and

Table 2: **Finite-context information-loss witnesses.** Independent replication at $N = 128, d = 2$, normalized by $N/d = 64$. The earlier Gaussian control is shown separately. The uncapped fixed-t mean is primary; the calibration-fixed upper cap retains all negative contributions. Both population expectations are at most $(N/d)B_N$. Brackets are pointwise task Monte Carlo 95% intervals, conditional on models and calibration. Posterior quadrature bias is not included.

Law	Tasks	Uncapped [95% interval]	Upper-capped [95% interval]
Laplace	4,096	$0.0666\ [0.0529, 0.0802]$	$0.0553\ [0.0436, 0.0670]$
Quartic	4,096	$0.0950\ [0.0825, 0.1076]$	$0.0865\ [0.0755, 0.0974]$
Gaussian (earlier)	256	$-0.000013\ [-0.000029, 0.000003]$	—

standardized Laplace designs give normalized quartic risks 0.7281, 0.7315, and 0.7215, near $1/I_f = 0.7295$; their degree-two references are 0.9962, 0.9966, and 0.9940. Section L.4 gives all nine cells across $N = 128, 512, 2048$, totaling 180,000 tasks, and direct-record numerical checks.

6 Discussion

An attention memory should be judged by the information it preserves as well as the computation it supports. Our moment hierarchy identifies the retained score component, its optimal risk, and exactly when the memory is sufficient. Smooth near-Gaussian tasks expose arbitrarily large asymptotic losses outside that class. The architectural results locate a concrete remedy: compute the missing score feature before aggregation.

Trained causal Transformers use information beyond the Gram. Independent Laplace risk gains and Laplace/quartic information witnesses establish complementary parts of this picture. The quartic result shows why both measurements matter: using additional predictive information need not produce a net risk gain when computational error remains.

Scope and next questions. The theory fixes dimension, moment degree, and noise law before the sample-size limit. Its density and moment conditions support an exact information analysis; the near-Gaussian construction gives an asymptotic separation, with no uniform small-sample or derivative bound across its noise family. Architectural closure applies to the stated common-source affine class in exact arithmetic. Standard softmax Transformers and adaptive token memories require their own information analysis. The experiments use numerical regression in small dimensions and three training seeds per law; their observed Laplace gain does not extrapolate to longer contexts. They establish information use without identifying an internal score circuit. The diagnostic relies on latent coefficients available in synthetic tasks, and its Monte Carlo intervals do not include uncertified posterior quadrature bias. A target-preserving kernel also limits how much of the Bayes gap its witness can reveal (Corollary N.3).

These results motivate two concrete directions: memories that adapt the retained score features to an unknown task law, and information audits that remain informative when the optimal decoder cannot be fitted. Both put the adequacy of the retained evidence at the center of attention design and evaluation.

Reproducibility statement

The appendix gives assumptions, proofs, model configurations, and protocols. Sections A and D specify independent task banks, seeds, posterior checks, and controls. Checkpoints, task-level outputs, and analysis scripts are indexed in a hashed evidence archive.

AI use statement

Generative AI assisted research planning, mathematical development, literature review, experiments, analysis, and writing, and executed computational workflows. Proofs, computations, references, and manuscript claims underwent targeted checks against the recorded evidence.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *NeurIPS*, 2023.

Vlad Bally and Lucia Caramellino. Asymptotic development for the CLT in total variation distance. *Bernoulli*, 22(4):2442–2485, 2016. doi: 10.3150/15-BEJ734.

Stephen Bates, Emmanuel Candès, Lucas Janson, and Wenshuo Wang. Metropolized knockoff sampling. *Journal of the American Statistical Association*, 116(535):1413–1427, 2021. doi: 10.1080/01621459.2020.1729163. URL https://arxiv.org/abs/1903.00434.

Peter J. Bickel. On adaptive estimation. *The Annals of Statistics*, 10(3):647–671, 1982. doi: 10.1214/aos/1176345863.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265. URL https://doi.org/10.1111/rssb.12265.

Hoang T. H. Cao, Hai D. V. Trinh, Tho Quan, and Lan V. Truong. Transformers learn robust in-context regression under distributional uncertainty. *arXiv preprint arXiv:2603.18564*, 2026.

Julien Carron. On the incompleteness of the moment and correlation function hierarchy as probes of the lognormal field. *The Astrophysical Journal*, 738(1):86, 2011. doi: 10.1088/0004-637X/738/1/86. URL https://arxiv.org/abs/1105.4467.

Chen Cheng, Xinzhi Yu, Haodong Wen, Jingsong Sun, Guanzhang Yue, Yihao Zhang, and Zeming Wei. Exploring the robustness of in-context learning with noisy labels. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. doi: 10.1109/ICASSP49660.2025.10889281. URL https://arxiv.org/abs/2404.18191.

Petros Dellaportas and Ioannis Kontoyiannis. Control variates for estimation based on reversible markov chain monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):133–161, 2012. doi: 10.1111/j.1467-9868.2011.01000.x. URL https://pages.cs.aueb.gr/~yiannisk/PAPERS/cvmcmcJ.pdf.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *NeurIPS*, 2024.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: A Bayesian Cramér–Rao bound. *Bernoulli*, 1(1–2):59–79, 1995. doi: 10.2307/3318681.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

Shivam Gupta, Jasper C. H. Lee, Eric Price, and Paul Valiant. Finite-sample maximum likelihood estimation of location. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30139–30149, 2022. URL https://arxiv.org/abs/2206.02348.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. doi: 10.2307/1912775.

Christopher C. Heyde. *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer, 1997.

R. G. Jarrett. Bounds and expansions for fisher information when the moments are known. *Biometrika*, 71(1):101–113, 1984. doi: 10.1093/biomet/71.1.101. URL https://doi.org/10.1093/biomet/71.1.101.

Abram Kagan and Lawrence A. Shepp. A sufficiency paradox: An insufficient statistic preserving the fisher information. *The American Statistician*, 59(1):54–56, 2005. doi: 10.1198/000313005X21041.

Douglas Kelker and Ted K. Matthes. A sufficient statistics characterization of the normal distribution. *The Annals of Mathematical Statistics*, 41(3):1086–1090, 1970. doi: 10.1214/aoms/1177696988.

Keith Knight. Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, 26(2):755–770, 1998. doi: 10.1214/aos/1028144858.

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(3):399–409, 1936. doi: 10.1090/S0002-9947-1936-1501854-3.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000. doi: 10.1007/978-1-4612-1168-5.

Yingcong Li, Davoud Ataee Tarzanagh, Ankit Singh Rawat, Maryam Fazel, and Samet Oymak. Gating is weighting: Understanding gated linear attention through in-context learning. *arXiv preprint arXiv:2504.04308*, 2025.

Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *ICLR*, 2024.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, 1982. doi: 10.1002/9780470316559.

E. J. G. Pitman. Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(4):567–579, 1936. doi: 10.1017/S0305004100019307. URL https://doi.org/10.1017/S0305004100019307.

Yu. V. Prokhorov. A local theorem for densities. *Doklady Akademii Nauk SSSR*, 83:797–800, 1952.

Ravin Raj and Gautam Reddy. Adaptive inference and function vectors in deep transformers. *arXiv preprint arXiv:2606.16694*, 2026.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pp. 9355–9366, 2021.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. URL https://arxiv.org/abs/2002.05202.

Charles Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 187–196. University of California Press, 1956. doi: 10.1525/9780520313880-017.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Haoyuan Sun, Ali Jadbabaie, and Navid Azizan. On the role of transformer feed-forward layers in nonlinear in-context learning. *arXiv preprint arXiv:2501.18187*, 2025.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Brendan van Rooyen and Robert C. Williamson. Le cam meets lecun: Deficiency and generic feature learning. *arXiv preprint arXiv:1402.4884*, 2014. URL https://arxiv.org/abs/1402.4884.

Max Vladymyrov, Johannes von Oswald, Mark Sandler, and Rong Ge. Linear transformers are versatile in-context learners. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/file/57a3c602f0a1c8980cc5ed07e49d9490-Paper-Conference.pdf.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, et al. Transformers learn in-context by gradient descent. In *ICML*, 2023.

Maximilian Weimar, Lukas M. Rachbauer, Ilya Starshynov, Daniele Faccio, Linara Adilova, Dorian Bouchet, and Stefan Rotter. Fisher information flow in artificial neural networks. *Physical Review X*, 15:031072, 2025. doi: 10.1103/kn3z-rmm8. URL https://arxiv.org/abs/2509.02407.

Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021. doi: 10.1111/biom.13392. URL https://doi.org/10.1111/biom.13392.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *NeurIPS*, 2017.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1910.07467.

Ruiqi Zhang, Jingfeng Wu, and Peter L. Bartlett. In-context learning of a linear transformer block: Benefits of the MLP component and one-step GD initialization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 18310–18361, 2024. doi: 10.52202/079017-0581.

Yifan Zhang, Zhen Qin, Mengdi Wang, and Quanquan Gu. Higher-order linear attention. *arXiv preprint arXiv:2510.27258*, 2025.

A Experimental setup and evaluation protocols

A.1 Experimental evidence and data organization

The empirical study comprises a controlled $p = 8$ source/query comparison, a nine-model standard-Transformer confirmation, independent posterior diagnostics and a three-seed $d = 4$ confirmation. Each experiment retains its configuration, checkpoints, task identities, and task-level measurements. The original risk protocol was fixed on 16 September 2026 before its confirmation runs. The expanded diagnostics and follow-up tests use separate task banks and are described in Section D.

The completed thin-tail table is drawn from `experiments/llm/results/thintail_p8/holdout_seed26091399.json`. The associated task-loss archives are retained in the same directory. The Gram sensitivity numbers come from `experiments/llm/results/thintail_p8/fiber_float64.json`. Training seeds are 26091201, 26091202, and 26091203; the independent evaluation seed is 26091399. The initial evaluation uses 8,000 tasks per length; the independent evaluation uses 20,000 tasks at each of $N = 128$ and $N = 512$. The banks are analyzed separately. Models share evaluation tasks; statistical intervals use the number of distinct tasks in each bank.

Controlled source/query architecture and optimization. Both models have four common-source affine attention blocks, width 64, and four heads. Each block has a residual GELU FFN with hidden width 256; its update mask selects either demonstration tokens or the query token. A learned affine embedding maps raw numerical records to the hidden state, and an affine query readout predicts the clean target. The implementation is `experiments/confirmation/code/fullrank_architecture/models.py`, with training entry point `experiments/exploratory/train_shape.py`. For $p = 8$, both models use batch size 512 and AdamW with learning rate 10^{-3}, weight decay 0.01, 250 warmup steps, cosine decay, and gradient clipping at norm one. Each update samples a length uniformly from $\{32, 64, 128\}$ and draws fresh tasks with $W \sim \mathcal{N}(0, I_2)$. The loss is squared error on $X_q^\top W$. All runs use 5,000 updates.

On the independent 20,000-task bank, the source/query normalized risks are 0.4683 ± 0.0053 and 1.0238 ± 0.0008 at $N = 128$, and 0.4911 ± 0.0286 and 1.1583 ± 0.0309 at $N = 512$. Here uncertainty is sample SD across training seeds. The paired query-minus-source task-bootstrap 95% intervals are $[0.5310, 0.5809]$ and $[0.6396, 0.6976]$, respectively, conditional on those fitted models. Independent-bank evaluation uses float32 with TF32 disabled; the Gram-invariance diagnostic uses float64.

A.2 Frozen confirmation configuration

The entry point is `experiments/transformer_mechanism/train.py`; analytical references and the residual rotation are in `baselines.py` in the same directory. The following table records the frozen configuration used by all nine completed confirmation runs. The 4,000-step budget was the development starting point.

For a task with N observations and one final query, the token sequence is $(q_1, r_1, q_2, r_2, \ldots, q_N, r_N, q_{N+1})$, with $q_i = (X_{i1}, X_{i2}, 0, 1)$ and $r_i = (X_{i1}, X_{i2}, Y_i, 0)$. Thus all earlier observations are visible at q_i, but its own label is not. Intermediate supervision uses the correct prefix target. The causal check changes current and future labels and future covariates while verifying that protected query predictions are unchanged. The numerical input adapter is learned; no moments, residuals, scores, or estimator outputs enter the Transformer.

The Gram MLP has three hidden layers of width 512 with GELU activations, and a scalar output. Its 15 inputs contain the flattened $X^\top X$, $X^\top Y$, covariate sums, label sum, label square sum, query covariate, ridge coefficients, and a scaled log-length feature. Prefix sums are normalized by their sample counts. The redundant off-diagonal Gram entry is retained by the implementation. A ridge skip uses prior variance one and noise variance one; the output layer of the residual network is initialized to zero. Both learned models receive the same sampled tasks, target positions, optimizer family, and learning-rate schedule. Parameter counts and CUDA-synchronized training times are recorded separately for each model, including forward, backward, and optimizer operations.

Table 3: Frozen protocol used in the completed standard-Transformer confirmation.

Setting	Confirmation configuration
Task dimension / prior	$d = 2$, $W \sim \mathcal{N}(0, I_2)$
Design / query	Independent $\mathcal{N}(0, I_2)$
Noise	Unit-variance $p \in \{1, 2, 4\}$, separate runs
Transformer	4 layers, width 128, 4 attention heads
FFN / positions / normalization	SwiGLU width 256 / RoPE / pre-RMSNorm
RMSNorm stabilization	Explicit $\epsilon = 10^{-6}$
Optimization	AdamW, peak learning rate 3×10^{-4}
Weight decay / gradient clipping	0.01 / norm 1
Updates / batch size	12,000 / 128 tasks
Schedule	100-step warmup, cosine decay to 10% of peak
Curriculum	$N = 32$ for first 500 steps
Subsequent train lengths	Uniform choice from $\{32, 64, 128\}$
Supervision	Last 16 query positions, clean-target MSE
Development evaluation	2,048 tasks per length, $N = 32, 64, 128$
Final evaluation	16,384 tasks per length, $N = 32, 64, 128, 256$
Training seeds, each noise law	26091611, 26091612, 26091613
Final evaluation seed	26091699
Checkpoint selection	Fixed final step 12,000
Arithmetic	CUDA bfloat16 autocast; classical references float64

RMSNorm uses explicit $\epsilon = 10^{-6}$ at every block and the final readout normalization. This keeps the stabilization constant when switching evaluation precision. Earlier development checkpoints that used the implicit bfloat16 epsilon are identified by their saved configuration; the diagnostic restores their original epsilon rather than silently interpreting them as the revised model.

The script saves source hashes, software and hardware information, hostname, process identity, configuration, training losses, development metrics, and per-task final losses. Both learned models' losses and pre-clipping gradient norms are checked for finiteness. Checkpoints are written atomically every 1,000 steps by default and at the requested stop. They contain both models and optimizers, the completed step, CPU/CUDA random states, the task generator, the length generator, Python and NumPy states, and cumulative training times. Resumption uses the latest complete checkpoint and records the restored step and process identity; replayed log entries remain auditable. An absolute early-stop step preserves the full scheduled training budget for continuation tests. Development task counts and evaluation frequency are configurable, and development runs can disable final-bank evaluation. Per-run Monte Carlo intervals condition on one fitted model; aggregate intervals condition on the three fitted models after task-wise averaging. Training variation is reported through the sample standard deviation across seeds.

A.3 Classical references and finite-sample interpretation

OLS uses $(X^\top X)^{-1} X^\top Y$ with no hidden regularization. Ridge uses $(X^\top X + I)^{-1} X^\top Y$, matching the unit Gaussian prior and noise variance. The score reference starts from OLS, not the moment pilot in the two-layer construction:

$$\widehat{w}_{\mathrm{step}} = \widehat{w}_{\mathrm{OLS}} + I_p^{-1} \frac{1}{N} \sum_i X_i \psi_p(Y_i - X_i^\top \widehat{w}_{\mathrm{OLS}}).$$

The pilot reuses the correction samples. At $p = 2$, the correction is zero up to numerical error by the OLS normal equations; the unit tests verify this identity.

For the implemented full-data posterior reference,

$$m(D, x_q) = x_q^\top \frac{\int_{\mathbb{R}^2} w \exp[-\|w\|^2/2 - \sum_i |Y_i - X_i^\top w|^p / \alpha_p^p]\, dw}{\int_{\mathbb{R}^2} \exp[-\|w\|^2/2 - \sum_i |Y_i - X_i^\top w|^p / \alpha_p^p]\, dw}. \tag{14}$$

The solver in `experiments/transformer_mechanism/posterior.py` uses float64 tensor-product composite trapezoidal quadrature. Write $w = c + Lz$, where $c = (X^\top X + I)^{-1} X^\top Y$ and $LL^\top = (X^\top X + I)^{-1}$. Centering at ridge and scaling by this covariance improves resolution

while retaining the exact noise likelihood in the quadrature weights. Log weights are rescaled during summation. Constant Jacobian and grid-spacing factors cancel in the normalized mean. The initial square is $[-8, 8]^2$ in z coordinates, with successive per-axis grid sizes $65, 129, 257, 513, 1025$. A domain expansion to $[-12, 12]^2$ maintains approximately the same mesh spacing. A result passes only when both the maximum coordinate change under refinement and under domain expansion, and the expanded-domain boundary mass fraction, are below tolerance. The output retains the history and failure status of every integration.

Original small diagnostic banks and numerical validation. This subsection reports the initial 24-pair diagnostics. The independent 256-pair expansion in Section D supplies the main directional results; the two banks are analyzed separately. The development bank contains six tasks per law at $N = 32$, using base seed 260916310. Separate held-out banks contain 24 tasks per law at each of $N = 32$ and $N = 128$, using base seed 260916910; the generator adds p to the corresponding base seed. Each length is analyzed as a separate task distribution, with risk evaluated on its own final bank. The angle $\pi/2$ and absolute posterior-prediction threshold 0.01 are recorded in the bank protocols before evaluating learned checkpoints. For each original/rotated pair the coefficient tolerance is $\min\{10^{-4}, 0.01/[40 \max(1, \|x_q\|_1)]\}$. The query-scaled sum of the two integration-change estimates is recorded as a numerical sensitivity estimate.

All 288 original and rotated posterior calculations across the two held-out banks passed these checks, with no failed tasks. A separate refinement and domain-expansion pass, using radii 10 and 14 and grids starting at 129 points per axis, also converged for all 288 calculations. Its maximum change in an individual posterior prediction was 8.57×10^{-5} at $N = 32$ and 4.07×10^{-5} at $N = 128$, well below the directional threshold. These observations establish numerical stability on the recorded banks. For Gaussian noise the posterior is computed by the same numerical solver, and its maximum discrepancy from analytic ridge was 4.00×10^{-15} across the held-out banks. The files `runs/20260916_posterior_heldout_n32/` and `runs/20260916_posterior_heldout_n128/`, relative to the experiment directory, retain the protocol, task arrays, source snapshots, all quadrature histories, summary, and stricter validation results.

Same-Gram directional comparison. The residual rotation in `baselines.py` preserves the entire Gram matrix of $[\mathbf{1}, X, Y]$; the query is held fixed. Across both held-out banks the maximum absolute Gram-entry discrepancy was 2.28×10^{-13}, while OLS and ridge prediction changes were at most 2.67×10^{-15}. The posterior is recomputed separately for each member of a pair. Its prediction-change RMS at $N = 32$ was 0.16485 for Laplace and 0.16612 for quartic noise; at $N = 128$ it was 0.07523 and 0.13611. Respectively 21/24 and 22/24 pairs at $N = 32$, and 20/24 and 22/24 at $N = 128$, exceeded the threshold. Gaussian posterior-change RMS was below 5.0×10^{-16} at both lengths, with no threshold-crossing pairs. Thus the Gaussian directional-agreement statistic is undefined, while its near-zero posterior change is the intended negative control. These values characterize the posterior target and preserved geometry.

For fitted checkpoints, `fiber_diagnostic.py` records the model and posterior change RMS, their squared difference, the squared error of the zero-change reference, and sign agreement on converged pairs whose posterior change exceeds 0.01. A Wilson interval describes the finite number of such directional comparisons conditional on the fitted model. Gram MLP invariance is checked on those same pairs. CPU float64 evaluation provides the primary arithmetic check; a separate bfloat16-autocast pass records prediction and change discrepancies and recomputes direction statistics. The normalization epsilon is held fixed across these passes. The two arithmetic results are saved separately. The target for each rotated prompt is its recomputed posterior mean under the specified law; the original latent W is used only in the original-distribution risk evaluation.

Original small-bank CUDA evaluation. All 18 checkpoint/length evaluations were repeated on the H200 with CUDA bfloat16 autocast, retaining the original CPU float64 arrays unchanged. CPU bfloat16 results are separately archived and are not substituted for CUDA measurements. Main-precision and CUDA summaries are in `runs/20260916_takeover/fiber_analysis_cuda/` under the experiment directory. Complete augmented-Gram checks are also retained after float32 input casting: the maximum relative error is 3.03×10^{-8} with float64 accumulation and 4.62×10^{-7} with local float32 accumulation. The former measures input rounding; the latter additionally includes summation roundoff. The largest float64 Gram discrepancy across the independent

accumulation checks is below 10^{-12}. The largest Gram-MLP prediction discrepancy across all checkpoints is 2.22×10^{-15} (float64) and 1.91×10^{-6} (CUDA).

At $N = 32$, CUDA direction rates are 0.7937 (Laplace) and 0.7576 (quartic). The paired discrepancy contrasts against zero change are -0.016978 with interval $[-0.034683, 0.000726]$ and -0.019981 with interval $[-0.042850, 0.002889]$, respectively. Neither interval excludes zero. Gaussian model-change RMS is 0.02947 at this length, while the posterior is invariant. All posterior failures, numerical checks, task identities, and per-seed predictions are retained; none of the 144 held-out pairs was dropped.

A.4 Confirmation, uncertainty, and model selection

The frozen protocol is saved as `runs/20260916_takeover/confirmation_protocol.json`, relative to `experiments/transformer_mechanism/`. It specifies three training seeds (26091611, 26091612, 26091613) for each $p \in \{1, 2, 4\}$, with 12,000 updates and the final-step checkpoint selected in every run. The independent final evaluation uses base seed 26091699 and 16,384 tasks at each $N \in \{32, 64, 128, 256\}$; the length-specific generator seed is the base seed plus $1000N$. Evaluation batch size is 128. Development uses base seed 26091591 with 2,048 tasks per length. The budget was selected after one bounded development comparison between 4,000 and 12,000 steps. Source snapshots and their hashes are fixed; the only final code difference from development is a `--phase` metadata label, with model, optimization, sampling, and evaluation unchanged. All nine confirmation runs completed; no run is excluded from the risk aggregation. The delivery audit verifies all final checkpoints, per-task losses, configuration and bank identities; the historical confirmation files remain unchanged.

The primary risk contrast is Transformer minus residual Gram MLP at $N = 128$, separately for each law. All lengths and ridge, OLS, and score-step references are retained. Performance is reported by law, length, and seed, together with the sample standard deviation across the three training seeds. Common evaluation tasks support paired differences. For an aggregate contrast, task losses are first averaged across the three fitted models, and a normal Monte Carlo interval is computed over the 16,384 independent tasks. This interval conditions on those fitted models. Training-seed variation and individual-seed contrasts are reported separately.

All nine final checkpoints were evaluated on the two 24-pair diagnostic banks at $N = 32, 128$, using the fixed angle and threshold and separately reported float64 and bfloat16 results. Development selected the training budget; confirmation used the fixed final-step checkpoint. Subsequent mechanism, dimension, and order tests use new banks and retain their separate experiment identities.

A.5 Complete risk results and computational audit

The files `runs/20260916_takeover/risk_analysis/summary.json`, `risks.csv`, `paired.csv`, and `curves.csv` in that analysis directory retain all aggregate and seed-specific risks, task Monte Carlo errors, paired contrasts, and learning curves. Every final length uses 16,384 tasks per law; task-bank hashes match across fitted seeds within each law. No original-distribution posterior risk is inserted from the separate 24-task diagnostic banks.

Each run jointly fits a 657,281-parameter Transformer and a 534,017-parameter Gram MLP on a single NVIDIA H200. Summed over all nine processes, training and development took 1,169.12 seconds and total execution including final evaluation took 1,193.89 seconds. Synchronized model-training times sum to 888.79 seconds for the Transformer and 244.73 seconds for the Gram MLP. These are sums of measured process times across the nine runs. The run-level values, machine identifiers, hashes, and verification status are retained in `runs/20260916_takeover/delivery_audit.json`.

Post hoc tail and precision diagnosis. At $p = 4$, $N = 128$, saved task 15799 (zero-based index) has clean target 15.35890. For seed 26091611, the Transformer predicts 9.68750 in the original bfloat16 evaluation and 9.66651 in float64, while ridge predicts 15.31305. The reconstructed task bank matches its saved hash, and the entire bfloat16 batch reproduces its archived losses exactly. The Transformer's normalized loss on this task is 2,058.54 in bfloat16 and 2,073.81 in float64. Rotary angles retain their explicit float32 implementation even in the otherwise float64 replay. The task was

Table 4: All learned confirmation risks by training seed and length. Each entry is Transformer / Gram MLP normalized risk. No seed or length is omitted. Seeds 11, 12, and 13 abbreviate 26091611, 26091612, and 26091613. Task-level Monte Carlo errors are retained in `risks.csv`.

p	Seed	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	11	0.8621 / 1.0614	0.7960 / 1.0695	0.7553 / 1.0137	1.1200 / 1.0095
1	12	0.8673 / 1.0615	0.8004 / 1.0695	0.7590 / 1.0136	1.2774 / 1.0094
1	13	0.8364 / 1.0614	0.7705 / 1.0695	0.7165 / 1.0136	0.9430 / 1.0095
2	11	1.0720 / 1.0471	1.0880 / 1.0539	1.0398 / 0.9972	1.8058 / 1.0027
2	12	1.0689 / 1.0471	1.0852 / 1.0539	1.0389 / 0.9970	1.7024 / 1.0026
2	13	1.0682 / 1.0471	1.0767 / 1.0539	1.0394 / 0.9968	1.6217 / 1.0027
4	11	0.9523 / 1.0598	0.9042 / 1.0366	0.9907 / 0.9990	1.3011 / 0.9754
4	12	0.9457 / 1.0598	0.8988 / 1.0366	0.9785 / 0.9989	1.4937 / 0.9755
4	13	0.9422 / 1.0599	0.8913 / 1.0366	0.9697 / 0.9991	1.3541 / 0.9755

Table 5: Classical references and the paired Transformer-minus-Gram contrast at every length. Brackets are conditional task Monte Carlo 95% normal intervals; negative differences favor the Transformer. Classical risks are common across fitted seeds.

p	N	Ridge	OLS	Score step	Difference	Paired interval
1	32	1.0615	1.1083	0.9142	-0.2061	$[-0.2280, -0.1843]$
1	64	1.0695	1.0922	0.7947	-0.2806	$[-0.3043, -0.2568]$
1	128	1.0136	1.0213	0.6943	-0.2700	$[-0.2927, -0.2473]$
1	256	1.0095	1.0134	0.6320	0.1040	$[0.0346, 0.1734]$
2	32	1.0471	1.0847	1.0847	0.0226	$[0.0164, 0.0288]$
2	64	1.0539	1.0713	1.0713	0.0294	$[0.0230, 0.0358]$
2	128	0.9969	1.0054	1.0054	0.0424	$[0.0348, 0.0500]$
2	256	1.0025	1.0073	1.0073	0.7073	$[0.6368, 0.7778]$
4	32	1.0598	1.1102	0.9457	-0.1131	$[-0.1305, -0.0957]$
4	64	1.0366	1.0522	0.8520	-0.1385	$[-0.1561, -0.1209]$
4	128	0.9989	1.0060	0.7728	-0.0194	$[-0.2680, 0.2293]$
4	256	0.9754	0.9815	0.7443	0.4075	$[0.3301, 0.4849]$

selected after viewing the final losses for descriptive diagnosis. All tasks remain in the primary risk estimate.

For this fitted architecture the final RMSNorm and linear readout imply an input-independent output range. With width m, readout weight u, normalization weight γ, and bias b,

$$|\widehat{y} - b| = \frac{|(u \odot \gamma)^\top h|}{\sqrt{\|h\|^2/m + \epsilon}} \le \sqrt{m}\, \|u \odot \gamma\|_2.$$

For the three quartic checkpoints the upper endpoints are 9.93337, 9.97174, and 9.97079, respectively, below the identified target 15.35890. These are checkpoint-specific bounds in exact arithmetic, with the weights and normalization parameters fixed. The primary estimates retain the original fitted heads and all task losses. Its source, checkpoint hashes, and task-specific error infima are saved in `risk_analysis/output_range_diagnostic.json`; the precision replay is in `remote/tail_p4_s11/diagnosis.json`, both relative to `runs/20260916_takeover/`.

B Polynomial log-density endpoints

Polynomial scores admit an exact Taylor expansion, giving an efficient source channel for the entire lossless family in Theorem 3.2.

Corollary B.1 (Polynomial log-density noise)**.** *Let $f(e) = \exp\{q(e)\}$ be a mean-zero noise density, where q has even degree $m \ge 2$ and negative leading coefficient. Let $W \sim \mathcal{N}(0, \tau^2 I_d)$ with $\tau > 0$, and let the design and query be independent standard Gaussian vectors. With $\psi = -q'$ and $I_f = \mathbb{E}\psi(\varepsilon)^2$,*

$$R^*_{\mathrm{full},N} = \frac{d}{I_f N} + o(N^{-1}).$$

The moment-pilot estimator in (8) attains this coefficient and has the width-$(d+3)$, two-layer scalar-source-channel realization of Theorem F.7. For even exponential-power noise in (9), this gives

$$\psi_p(e) = \frac{p e^{p-1}}{\alpha_p^p}, \qquad I_p = \frac{p^2 \Gamma(2 - 1/p)}{\alpha_p^2 \Gamma(1/p)}.$$

In particular, a cubic channel covers quartic noise, and a seventh-degree channel covers $p = 8$.

Proof. Polynomial exponential tails give every error moment, and the Gaussian prior has every moment. The score belongs to $\mathcal{P}_{m-1}$, so Theorem J.1 gives $J_m = I_f$. Expanding $\sum_i q(Y_i - X_i^\top w)$ shows that complete mixed moments through degree m determine the likelihood in w. Their posterior therefore equals the full-data posterior, and the integrated moment-risk result gives the Bayes endpoint.

For the explicit estimator, put $\Delta = u - W$. The moment estimates in Section J.2 give $\|\Delta\|_{L^r} = O(N^{-1/2})$ for every fixed finite r. Exact polynomial Taylor expansion gives

$$\frac{1}{N}\sum_i X_i \psi(\varepsilon_i - X_i^\top \Delta) = \frac{1}{N}\sum_i X_i \psi(\varepsilon_i) - A_N \Delta + O_{L^2}(N^{-1}),$$

$$A_N = \frac{1}{N}\sum_i X_i X_i^\top \psi'(\varepsilon_i).$$

Independence and integration by parts give $\mathbb{E}A_N = I_f I_d$; $\|A_N - I_f I_d\|_{L^4} = O(N^{-1/2})$ and the L^4 pilot bound control their product on the same prompt. Substitution yields

$$\widehat{W} - W = I_f^{-1} \frac{1}{N}\sum_i X_i \psi(\varepsilon_i) + O_{L^2}(N^{-1}).$$

The leading term has mean zero and squared norm expectation $d/(I_f N)$; Cauchy–Schwarz bounds the risk remainder by $O(N^{-3/2})$. The layer construction in Theorem F.7 is algebraic and applies to this scalar polynomial score. The specialization follows from the exponential-power moment formula. □

C GUIDE TO THE THEORETICAL APPENDIX

The following sections develop the affine information envelope, its sharp risk bound, finite-sample witnesses, the nonsmooth Laplace endpoint, and the complete moment hierarchy. Scalar-channel uniqueness is stated for the specified moment pilot $u = X^\top Y/N$.

D POSTERIOR DIAGNOSTICS AND GENERALIZATION

The original nine-model confirmation, its final test bank, and the 24-pair diagnostics were retained without alteration. The additional experiments below use new task banks and protocols under `experiments/transformer_mechanism/runs/20260916_acceptance/`. They were specified after observing the original risk results, with independent task banks and separate experiment identities.

D.1 EXPANDED POSTERIOR-DIRECTION PROTOCOL

The initial measured integration costs are recorded before freezing 256 independent pairs per law at each of $N = 32, 128$. Seeds are $260918032 + 1000p$ and $260918128 + 1000p$, respectively. The angle $\pi/2$, threshold 0.01, all nine checkpoints, and two integration settings are fixed before evaluating the new banks. The first setting uses tolerance 10^{-4} and radii 8 and 12; the stricter one uses 10^{-5} and radii 10 and 14, allowing refinement through 2,049 points per axis. Reported posterior predictions use the stricter integration. Each endpoint must converge, and the two predictions must differ by at most $0.01/40 = 0.00025$. All source pairs and both solver histories are retained even if a check fails. Here all 1,536 pairs pass, with zero numerical or geometry failures and maximum discrepancy 0.0001237. Gaussian agreement with ridge is within 5.33×10^{-15}. The numerical stages take 340.8 and 789.2 seconds for the two lengths, using at most two concurrent local CPU processes. No training runs on the local CPU.

Table 6: Expanded independent fiber diagnostics, 256 pairs per row. Direction is mean $\pm$ training-seed SD after $|\Delta_*| > 0.01$; error contrasts use all 256 pairs. Intervals are conditional paired normal Monte Carlo intervals. Gram-control directions are not scored.

p	N	Active	Posterior RMS	Direction	Δ MSE minus zero [95% CI]
1	32	238	0.17608	0.8053 ± 0.0246	-0.01991 $[-0.02834, -0.01148]$
2	32	0	$< 10^{-15}$	undefined	0.00130 $[0.00097, 0.00162]$
4	32	242	0.18681	0.8430 ± 0.0072	-0.02464 $[-0.03226, -0.01702]$
1	128	221	0.11080	0.8160 ± 0.0257	-0.00857 $[-0.01095, -0.00620]$
2	128	0	$< 10^{-15}$	undefined	0.00030 $[0.00024, 0.00036]$
4	128	228	0.10956	0.8465 ± 0.0132	-0.00853 $[-0.01105, -0.00600]$

All 18 checkpoint/length combinations are evaluated in CPU float64 and genuine H200 bfloat16, preserving the former arrays exactly when adding the latter. The maximum Gram-MLP prediction change is 7.11×10^{-15} in float64 and 4.77×10^{-6} in CUDA. Its roundoff-scale response is reported as an invariance check. Gaussian posterior changes also never cross the fixed threshold. Its model change RMS is 0.03604 at $N = 32$ and 0.01734 at $N = 128$, despite a posterior change below 5.3×10^{-16} RMS. Thus geometric sensitivity is separated from posterior-relevant direction. An independent full-Gram precision audit gives relative discrepancies at most 5.67×10^{-8} after casting inputs to float32 and accumulating in float64, or 8.54×10^{-7} with float32 accumulation as well. These values measure input and summation roundoff.

D.2 UNCERTAINTY AND EVIDENCE PRESERVATION

All model predictions, fitted parameters, task-bank hashes, source snapshots, machine information, and elapsed times are retained. At each test task, squared errors or contrasts are first averaged across the three fitted models. Normal Monte Carlo intervals then use the number of independent tasks, with training-seed sample SD reported separately. The same rule applies to permutations: repeated orderings remain within their original task. These are pointwise, approximate intervals conditional on the fitted models. Every control is reported.

D.3 Dimension and training-budget checks

The independent dimension branch changes the input dimension to four, and the corresponding token and Gram-feature dimensions, retaining the four-block width-128 model, task prior $\mathcal{N}(0, I_4)$, and optimizer. A development-only Gaussian run (seed 260916401) uses bank 260916491. The prespecified capability rule requires $C_{128} \leq 1.2$ and $C_{128}^{\mathrm{Transformer}} - C_{128}^{\mathrm{ridge}} \leq 0.2$; the observed values are 1.0670 and 0.0788. The 12,000-step configuration is therefore frozen without the optional 24,000-step extension. A Laplace development run is recorded separately from the Gaussian calibration. Confirmation seeds 260916411/412/413 use final bank 260916499 with 16,384 tasks per length and fixed final-step checkpoints. The dimension-parametrized code is checked for exact agreement with the original implementation at $d = 2$, causality at $d = 4$, and bitwise continuation after a GPU interruption.

Table 7: Independent $d = 4$ Laplace confirmation. Normalized risk is $N\,\mathrm{MSE}/4$. Learned entries show training-seed mean $\pm$ SD; paired intervals condition on the three fitted models and use 16,384 shared tasks; $T - G$ denotes Transformer minus Gram.

Predictor	$N = 32$	$N = 64$	$N = 128$	$N = 256$
Transformer	1.0156 ± 0.0101	0.8891 ± 0.0084	0.8352 ± 0.0125	1.8935 ± 0.1589
Gram MLP	1.1505 ± 0.0000	1.0655 ± 0.0000	1.0181 ± 0.0001	1.0087 ± 0.0001
Ridge	1.1506	1.0655	1.0180	1.0087
OLS	1.2025	1.0864	1.0252	1.0123
Score step	1.0589	0.8561	0.7396	0.6742
$T - G$	-0.1349	-0.1764	-0.1829	0.8848
Paired interval	$[-0.1525, -0.1173]$	$[-0.1942, -0.1586]$	$[-0.2214, -0.1444]$	$[0.8405, 0.9290]$

The $d = 4$ Transformer and Gram control have 657,537 and 544,257 parameters. The three confirmation processes take 402.0 seconds in sum; the two development runs take 258.9 seconds. The shared-task training times attributed separately to the Transformer and Gram network are 298.1 and 80.8 seconds. All configurations, checkpoints, per-task losses, and 136 remotely hashed artifacts were retrieved and checked locally.

Random record order. On a new bank of 8,192 tasks per law at $d = 2, N = 128$, all nine original checkpoints are evaluated in the original order and two independently randomized record orders. Each covariate–label pair stays together and the final query is unchanged. Repeated permutations are averaged within task before computing uncertainty. This test takes 10.4 process seconds. The risk changes, permuted minus original, are -0.00208 $[-0.01249, 0.00833]$ for Laplace, 0.00215 $[-0.00262, 0.00692]$ for Gaussian, and -0.00100 $[-0.00860, 0.00660]$ for quartic noise. Laplace's Transformer-minus-Gram contrast after permutation remains -0.26882 $[-0.30180, -0.23583]$. Individual Transformer predictions change under permutation. The original tasks are exchangeable, and each covariate–label pair is preserved by the transformation. This bank is reported separately from the primary risk bank.

Development comparison of training budgets. The original independent $d = 2$ development seed compared 4,000 with 12,000 updates before confirmation. At $N = 128$, Transformer risks change from 0.8983 to 0.7725 for Laplace, 1.0785 to 1.0101 for Gaussian, and 0.9337 to 0.8578 for quartic noise. The risk benefit is already present for Laplace at the shorter budget, and the longer budget improves all three development curves. These development measurements determined the budget before final evaluation. Full curves and the $d = 4$ calibration are retained in the budget analysis.

D.4 Uncorrected rotations: an off-distribution comparison

A prompt transformation that preserves the complete augmented Gram isolates information dependence beyond this summary. For a fixed design, decompose Y into its projection onto $\mathrm{span}(\mathbf{1}, X)$ and an orthogonal residual r. Choose a unit vector v orthogonal to both that span and r, and define

$$Y' = P_{[\mathbf{1},X]}Y + \cos(\theta)r + \sin(\theta)\|r\|v. \tag{15}$$

Table 8: **Posterior-directed responses at $N = 128$.** H200 bfloat16, three fitted seeds, 256 shared pairs per law. Direction rates are means $\pm$ seed SD on the 221 Laplace and 228 quartic pairs with $|\Delta_*| > 0.01$. Direction intervals use task bootstrap; squared-error contrasts use conditional normal Monte Carlo intervals. Each task is averaged across fitted seeds before resampling, giving 256 independent pairs per law.

Quantity	Laplace	Quartic
Posterior Δ RMS	0.11080	0.10956
Transformer Δ RMS	0.08351	0.09223
Direction agreement	0.8160 ± 0.0257	0.8465 ± 0.0132
Direction interval	$[0.7707, 0.8582]$	$[0.8070, 0.8830]$
$\mathbb{E}(\Delta_{\text{model}} - \Delta_*)^2$	0.003704	0.003478
Zero-change reference $\mathbb{E}\Delta_*^2$	0.012277	0.012004
Error minus zero-change error	-0.008573	-0.008526
Paired interval	$[-0.010948, -0.006198]$	$[-0.011052, -0.005999]$

This changes individual labels while preserving all Gram blocks and the query. The implementation uses float64 projections, checks the residual subspace, and retains zero residuals unchanged.

We measure how closely model responses track the corresponding full-data posterior responses:

$$\Delta_{\text{model}} = \widehat{y}(D', X_q) - \widehat{y}(D, X_q), \qquad \Delta_* = m(D', X_q) - m(D, X_q). \tag{16}$$

For each endpoint, the posterior mean is recomputed under the specified noise law and $\mathcal{N}(0, I_2)$ prior. The functional test uses a prescribed rotation $\theta = \pi/2$ and 256 independent pairs per law at each of $N = 32, 128$, giving 1,536 pairs evaluated on the nine fitted checkpoints. Directional agreement is evaluated on pairs with $|\Delta_*| > 0.01$.

All 3,072 endpoints pass integration refinement and domain expansion checks; the largest prediction discrepancy is 1.24×10^{-4}. Gaussian integration reproduces analytic ridge, and numerical checks verify Gram preservation and the control's invariance. Primary predictions use H200 bfloat16, matching risk evaluation. Solver settings, float64 comparisons, and precision audits are in Section D.

At $N = 128$, directional agreement is 81.6% under Laplace noise and 84.6% under quartic noise. The mean squared error in predicting the posterior change is 0.003704 and 0.003478, respectively, compared with 0.012277 and 0.012004 for a predictor constrained to remain constant on the fiber. Both paired improvement intervals exclude zero. The corresponding $N = 32$ direction rates are 80.5% and 84.3%. Thus the learned responses recover part of the posterior variation removed by the Gram summary.

Gaussian noise supplies the sufficient-statistic control: its posterior change is 5.29×10^{-16} RMS, while the Transformer changes by 0.01734 RMS. No Gaussian pair passes the posterior-change threshold. Figure 2 displays all three laws. Numerical checks, both lengths, and individual seeds are reported in Section D.

E THE EXPERIMENT COMPILED BY COMMON-SOURCE ATTENTION

Let the demonstration token contain $z_i = (x_i, y_i)$ and write $\widetilde{z}_i = (z_i, 1)$ for its homogeneous coordinate. Define the augmented empirical Gram statistic

$$\widetilde{S}(D) := \frac{1}{N} \sum_{i=1}^{N} \widetilde{z}_i \widetilde{z}_i^\top. \tag{17}$$

Without biases, its nonconstant block is

$$S = \frac{1}{N} \sum_i z_i z_i^\top = \begin{pmatrix} H & u \\ u^\top & c \end{pmatrix}, \quad H = \frac{1}{N} X^\top X, \quad u = \frac{1}{N} X^\top y, \quad c = \frac{1}{N} \|y\|^2.$$

Definition E.1 (Affine common-source linear attention)**.** At layer ℓ, every destination j, including the query, uses the same set of N demonstration states as keys and values; the query is not a source.

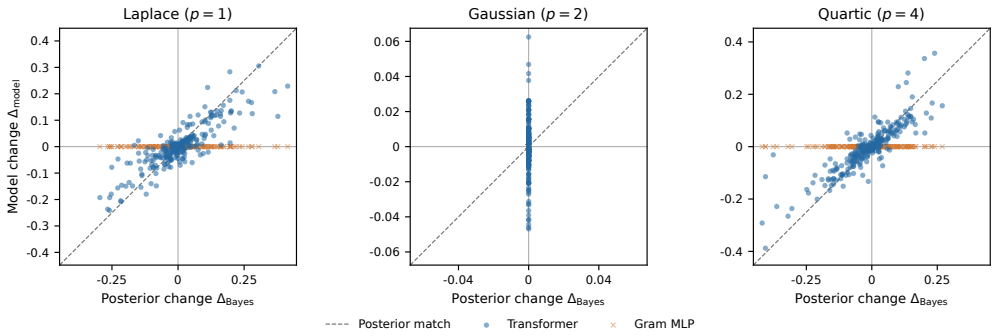

Figure 2: **Prediction changes at fixed augmented Gram.** Posterior versus model changes on 256 independently generated pairs per law at $N = 128$, using the prescribed label rotation in (15). Model changes are averaged over three fitted seeds in CUDA bfloat16. The Gaussian panel provides a sufficient-statistic control.

A layer is

$$e_j^+ = R_\ell e_j + r_\ell + \sum_h O_{\ell h} G_{\ell h} (Q_{\ell h} e_j + q_{\ell h}), \tag{18}$$

$$G_{\ell h} = \frac{1}{N} \sum_{i=1}^{N} (V_{\ell h} e_i + v_{\ell h})(K_{\ell h} e_i + k_{\ell h})^\top.$$

The class has no tokenwise nonlinearity, normalization, softmax, demo-specific positional embedding, or destination-dependent source mask.

To make the experiment language precise, for a task family $\{P_\theta^{(N)} : \theta \in \Theta\}$ and a deterministic prompt statistic T_N, write

$$\mathcal{E}_{T,N} := \{P_\theta^{(N)} \circ T_N^{-1} : \theta \in \Theta\}$$

for the statistical experiment induced by T_N. If T'_N is a deterministic function of T_N, then $\mathcal{E}_{T',N}$ is Blackwell- dominated by $\mathcal{E}_{T,N}$ and cannot have smaller optimal risk for any decision problem (Blackwell, 1953). Our factorization theorem shows that the experiment induced by the output of every model in the class is dominated by the augmented-Gram experiment. We therefore call the latter the class's *information envelope*; this does not assert that one finite network realizes every Gram-measurable rule.

Theorem E.2 (Augmented-Gram factorization)**.** *Initialize every demonstration state by an affine map of z_i and the query state by an affine map of its raw query token; fixed demo/query type embeddings are allowed. For every finite depth, width, and head count in (18), the final prediction factors through $(\widetilde{S}(D), z_q)$. If an explicit length feature such as $\log N$ is provided, the factorization is through $(\widetilde{S}(D), z_q, N)$. Consequently, at fixed N, any two prompts with the same $\widetilde{S}$ are indistinguishable at every query shared by the prompts. More precisely, the prediction is query-affine:*

$$\widehat{y}_q = a_L(\widetilde{S})^\top z_q + b_L(\widetilde{S}),$$

with an additional dependence on N when length is supplied.

The bias-free common-source recursion underlying this statement is closely related to the all-layer linear-transform representation of Vladymyrov et al. (2024). Our claim is the homogeneous-coordinate affine extension and its statistical consequences, not a first characterization of all-depth linear attention.

In the bias-free case with a linear query initialization and readout, there is a matrix polynomial $P_L(S)$ such that the prediction is query-linear,

$$\widehat{y}_q = x_q^\top w_L(S), \qquad \deg_S w_L \le \frac{3^L - 1}{2}.$$

The proof passes to homogeneous coordinates. Every layer then applies one shared matrix $M_\ell(\bar{C}_{\ell-1})$ to all destinations, while the source covariance obeys the closed recursion

$$\bar{C}_\ell = M_\ell(\bar{C}_{\ell-1})\bar{C}_{\ell-1}M_\ell(\bar{C}_{\ell-1})^\top.$$

Since $\bar{C}_0$ is a function of $\widetilde{S}$, induction proves the factorization; tracking polynomial degrees gives the stated bound. The full proof is in Section K.1.

Remark E.3 (Scope)**.** The theorem concerns a canonical theoretical linear-attention class, not a full GPT-style Transformer. A tokenwise MLP can expose higher moments before pooling; a causal mask can give different destinations different source operators. Both operations intentionally leave the theorem's class. A nonlinear readout applied only after the final query state can destroy query affinity, but it cannot recover demonstration information already lost by the Gram factorization.

Two of the commonly cited exclusions are less restrictive than they appear, and it is worth separating what genuinely escapes from what does not. Proposition E.4 shows that dividing the aggregated memory by a data-dependent scalar—normalized linear attention—leaves the factorization intact, so the barrier and its coefficient are inherited verbatim. By contrast a causal mask really does escape: prefix memories differ across destinations, so no single shared operator exists, and two prompts with equal augmented Gram can receive different predictions. The Gram closure requires both a common source operator across destinations and affine evolution of the source states. Scalar normalization as defined below preserves this closure; a shared operator alone does not, since nonlinear source features can expose higher moments.

Proposition E.4 (Scalar normalization preserves the envelope)**.** *Replace the aggregation $G_{\ell h}$ of (18) by $G_{\ell h}/c_{\ell h}$, where $c_{\ell h} = \frac{1}{N}\sum_i \lambda_{\ell h}^\top(K_{\ell h}e_i + k_{\ell h})$ is a scalar affine functional of the source states and $c_{\ell h} \neq 0$ almost surely. Then the conclusion of Theorem E.2 holds unchanged: the prediction factors through $(\widetilde{S}(D), z_q)$. Consequently Theorem F.2 and the coefficient $d\sigma^2$ apply to the normalized class as well. The polynomial degree bound $(3^L-1)/2$ is lost, since the shared maps become rational rather than polynomial in $\widetilde{S}$.*

Proof. Each $c_{\ell h}$ is a linear functional of the first-moment block of $\bar{C}_{\ell-1}$, hence a function of $\bar{C}_{\ell-1}$ alone. Therefore $G_{\ell h}/c_{\ell h}$ is again a function of $\bar{C}_{\ell-1}$, the layer still applies one shared matrix $M_\ell(\bar{C}_{\ell-1})$ to every destination, and the closed recursion $\bar{C}_\ell = M_\ell\bar{C}_{\ell-1}M_\ell^\top$ is unchanged in form. The induction in the proof of Theorem E.2 applies verbatim. Only measurability with respect to $\widetilde{S}$ is used by Theorem F.2, not the degree bound, so the risk statement transfers. □

The factorization is an information containment statement, not a universality claim for the network class. We do not require that one fixed finite affine network realize every measurable function of $\widetilde{S}$. Instead, the all-measurable augmented-Gram experiment is an optimistic information envelope for the architecture: a lower bound that holds even for this stronger envelope automatically holds for every network in the class.

Proposition E.5 (Query-only nonlinearities preserve the bottleneck)**.** *Keep every demonstration path in the affine common-source class of Theorem E.2, but allow an arbitrary deterministic tokenwise map on the query after each attention layer. The query remains excluded from the source set. Then, at fixed length, the final prediction still has the form*

$$\widehat{y}_q = F_L(\widetilde{S}(D), z_q),$$

with an additional argument N when length is supplied. Thus no sequence of query-only nonlinearities can recover demonstration information beyond the augmented Gram.

The source covariance recursion is unchanged because the query is not a source; induction then preserves query measurability through every deterministic query map. See Section K.1.

Because the map is arbitrary and deterministic, this covers destination-side normalization: RMSNorm or LayerNorm applied to the query state between attention layers is such a map, so a normalized query path remains Gram-limited and pays the same coefficient. What the class excludes is normalization on the *source* path, where it would enter the covariance recursion.

F FROM INDUCED EXPERIMENT TO OPTIMAL RISK

We first isolate the decision-theoretic consequence of any deterministic prompt compression.

Lemma F.1 (Bayes-risk decomposition)**.** *Let $Z \in L^2$ and $\mathcal{G} \subseteq \mathcal{F}$. If $R^*_{\mathcal{A}}$ is minimum squared risk among $\mathcal{A}$-measurable predictors, then*

$$R^*_{\mathcal{G}} - R^*_{\mathcal{F}} = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Z \mid \mathcal{F}] \mid \mathcal{G})]\,. \tag{19}$$

F.1 AN ARBITRARY-RULE EFFICIENCY BARRIER

We identify the leading risk coefficient of the *entire* augmented-Gram experiment; there is no regularity, consistency, or asymptotic-linearity assumption on the decision rule. Let

$$\mathcal{S}_N = \sigma\big(X^\top X,\ X^\top Y,\ \mathbf{1}^\top X,\ \|Y\|_2^2,\ \mathbf{1}^\top Y\big)\,,$$

which is exactly the sigma-field of the affine augmented Gram at known N. For a prior density π, write

$$R_{\mathrm{Gram,N}}(\pi) = \inf_{\widehat{w}_N\ \mathcal{S}_N\text{-measurable}} \mathbb{E}_\pi \|\widehat{w}_N - W\|_2^2.$$

Also write

$$R_{\mathrm{full,N}}(\pi) = \inf_{\widehat{w}_N\ \sigma((X_i,Y_i)_{i=1}^N)\text{-measurable}} \mathbb{E}_\pi \|\widehat{w}_N - W\|_2^2.$$

The extra affine Gram blocks contain no direct task term. To see the exact algebra, put $M = X^\top X$, $v = X^\top Y$, $a = \mathbf{1}^\top X$, and $P_X = XM^{-1}X^\top$ on the full-rank event. An invertible residualization of the observed statistic gives

$$\mathbf{1}^\top Y - aM^{-1}v = \mathbf{1}^\top (I - P_X)\varepsilon,$$
$$\|Y\|^2 - v^\top M^{-1} v = \|(I - P_X)\varepsilon\|^2.$$

These residual diagnostics can remain dependent on $X^\top\varepsilon$ at finite N under non-Gaussian noise; no finite-sample independence is assumed. The strong local-experiment proof retains their joint law and shows that they become parameter-free Gaussian ancillary coordinates, independent of the parameter-bearing block in the limit. Thus they cannot improve the leading task-risk coefficient.

Theorem F.2 (Integrated efficiency of the augmented-Gram experiment)**.** *Fix d. Let W have a continuous prior density π that is strictly positive everywhere, let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, and let*

$$Y_i = X_i^\top W + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} f.$$

Assume that W, the design sequence, and the error sequence are mutually independent, and that the errors have an everywhere-positive density, mean zero, variance σ^2, and finite fourth moment. Then, over all measurable augmented-Gram rules,

$$R_{\mathrm{Gram,N}}(\pi) = \frac{d\sigma^2}{N} + o(N^{-1}). \tag{20}$$

The corresponding local-minimax coefficient within the augmented-Gram experiment is also $d\sigma^2$. More precisely, if $H_C = [-C/2, C/2]^d$, then for every fixed center w_0,

$$\lim_{C\to\infty} \liminf_{N\to\infty} \inf_{\widehat{w}_N\ \mathcal{S}_N\text{-measurable}} \sup_{h\in H_C} N\,\mathbb{E}_{w_0+h/\sqrt{N}} \left\| \widehat{w}_N - \left(w_0 + \frac{h}{\sqrt{N}}\right)\right\|^2 = d\sigma^2. \tag{21}$$

For an independent $X_q \sim \mathcal{N}(0, I_d)$, this is exactly the clean-query prediction risk.

The proof, given in Section K.3, is a strong local-experiment argument. It obtains total-variation, rather than merely weak, convergence of the exact Gram statistic by combining a submersion argument with Prokhorov's total-variation CLT criterion (Prokhorov, 1952; Bally & Caramellino, 2016). A compact-cell local Bayes argument then prevents Hodges-type pointwise superefficiency from weakening the integrated claim.

Proposition F.3 (The Gram coefficient is attained inside the architecture)**.** *Under the assumptions of Theorem F.2, additionally assume $\mathbb{E}\|W\|^2 < \infty$. A fixed bias-free common-source affine model with two attention layers, one head per layer, and width $d+1$ predicts $X_q^\top \widetilde{W}_N$, where*

$$\widetilde{W}_N = (2I_d - H)u, \qquad H = \frac{1}{N}X^\top X, \qquad u = \frac{1}{N}X^\top Y. \tag{22}$$

Its clean-query risk satisfies

$$\mathbb{E}\|\widetilde{W}_N - W\|^2 = \frac{d\sigma^2}{N} + O(N^{-2}).$$

Consequently, the optimal integrated risk within the affine architecture itself has coefficient $d\sigma^2$; the all-measurable Gram envelope in Theorem F.2 is sharp for the architecture, not only a lower-bounding relaxation.

The construction preserves x, writes $t_j = X_j^\top u$ in the first layer, and writes $2t_j - X_j^\top Hu$ in the second. For $A = H - I_d$ and $\zeta = N^{-1}X^\top\varepsilon$ its error is $-A^2W + (I_d - A)\zeta$; Gaussian sample-covariance moments give the displayed risk. The reverse inequality follows from Gram measurability and query isotropy. Explicit layer matrices and the calculation are in Section K.2.

The rule in (22) is two unit-step gradient iterations from zero. Linear-attention implementations of gradient steps are well known (von Oswald et al., 2023; Ahn et al., 2023; Mahankali et al., 2024); the point here is not a new optimizer, but that a fixed in-class construction matches the all-measurable information-envelope coefficient and makes the subsequent operation-placement comparison genuinely sharp.

Theorem F.4 (Sharp efficiency law for regular location noise)**.** *In addition to the assumptions of Theorem F.2, suppose that f is absolutely continuous and its location score*

$$\psi = -f'/f, \qquad I_f = \mathbb{E}\psi(\varepsilon)^2 \in (0,\infty),$$

is absolutely continuous. Assume that ψ' has a globally Lipschitz representative, $\mathbb{E}\psi'(\varepsilon)^2 < \infty$, and

$$xf(x) \longrightarrow 0, \qquad \psi(x)f(x) \longrightarrow 0 \quad (|x| \to \infty).$$

Then

$$R_{\mathrm{full,N}}(\pi) = \frac{d}{I_f N} + o(N^{-1}), \qquad N\{R_{\mathrm{Gram,N}}(\pi) - R_{\mathrm{full,N}}(\pi)\} \longrightarrow d(\sigma^2 - I_f^{-1}). \tag{23}$$

The limiting penalty is nonnegative and equals zero if and only if f is the centered Gaussian density with variance σ^2.

The conditions are sufficient rather than maximal. They include Gaussian, logistic, and Student-t location noise with more than four degrees of freedom. The full-prompt coefficient, one-step score construction, adaptivity of location estimation, and the information inequality $\sigma^2 I_f \geq 1$ are classical (Stein, 1956; Bickel, 1982; Bickel et al., 1993; van der Vaart, 1998). The novelty is their architectural coupling: affine common-source attention has the sharp variance coefficient inside the network class, independently of additional depth and width, while a precisely placed source operation reaches the Fisher coefficient.

The coefficient gap is the squared-loss shadow of a strict ordering between the local statistical experiments. We use $\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$ and $\delta(\mathcal{E},\mathcal{F}) = \inf_K \sup_{h\in H}\|KP_h - Q_h\|_{\mathrm{TV}}$, where K ranges over Markov kernels from the sample space of $\mathcal{E}$ to that of $\mathcal{F}$. Thus $\delta(\mathcal{E},\mathcal{F})$ measures the error in simulating $\mathcal{F}$ from $\mathcal{E}$. Write Φ for the standard normal distribution function.

Theorem F.5 (Strict reverse deficiency of the augmented-Gram experiment)**.** *Fix a center w_0 and distinct $h_0, h_1 \in \mathbb{R}^d$. Under local parameter $h \in \{h_0, h_1\}$, let*

$$Y_i = X_i^\top(w_0 + h/\sqrt{N}) + \varepsilon_i,$$

and let $\mathcal{E}_N^{\mathrm{full}}$ observe the raw prompt while $\mathcal{E}_N^{\mathrm{Gram}}$ observes the complete statistic defining $\mathcal{S}_N$. Under the design and noise assumptions of Theorem F.4 (no prior is needed here), at every N,

$$\delta(\mathcal{E}_N^{\mathrm{full}}, \mathcal{E}_N^{\mathrm{Gram}}) = 0.$$

Writing $k = h_1 - h_0$, the pairwise total variations satisfy

$$\left\| P^{\text{full}}_{N,h_1} - P^{\text{full}}_{N,h_0} \right\|_{\text{TV}} \longrightarrow 2\Phi\left(\frac{\sqrt{I_f}\,\|k\|}{2} \right) - 1, \tag{24}$$

$$\left\| P^{\text{Gram}}_{N,h_1} - P^{\text{Gram}}_{N,h_0} \right\|_{\text{TV}} \longrightarrow 2\Phi\left(\frac{\|k\|}{2\sigma} \right) - 1. \tag{25}$$

Consequently,

$$\liminf_{N\to\infty} \delta(\mathcal{E}^{\text{Gram}}_N, \mathcal{E}^{\text{full}}_N) \geq \Phi\left(\frac{\sqrt{I_f}\,\|k\|}{2} \right) - \Phi\left(\frac{\|k\|}{2\sigma} \right). \tag{26}$$

The bound is strictly positive for every non-Gaussian density in scope and vanishes for the centered Gaussian density with variance σ^2.

Thus the full prompt is weakly more informative for every bounded local decision problem and, for non-Gaussian f, strictly more informative for at least one binary test.

Theorem F.6 (Le Cam limit of the induced experiments)**.** *Fix a center w_0 and a compact set $H \subset \mathbb{R}^d$ with nonempty interior. Under the design and noise assumptions of Theorem F.4, the local experiments*

$$\mathcal{E}^{\text{full}}_N = \{P^{\text{full}}_{N,h} : h \in H\}, \qquad \mathcal{E}^{\text{Gram}}_N = \{P^{\text{Gram}}_{N,h} : h \in H\}$$

converge in Le Cam distance to the Gaussian shift experiments

$$\mathcal{G}^{\text{full}} = \{\mathcal{N}(h, I_f^{-1} I_d) : h \in H\}, \qquad \mathcal{G}^{\text{Gram}} = \{\mathcal{N}(h, \sigma^2 I_d) : h \in H\},$$

that is, $\Delta(\mathcal{E}^{\text{full}}_N, \mathcal{G}^{\text{full}}) \to 0$ and $\Delta(\mathcal{E}^{\text{Gram}}_N, \mathcal{G}^{\text{Gram}}) \to 0$. Consequently the reverse deficiency converges:

$$\delta(\mathcal{E}^{\text{Gram}}_N, \mathcal{E}^{\text{full}}_N) \longrightarrow \delta(\mathcal{G}^{\text{Gram}}, \mathcal{G}^{\text{full}}).$$

Moreover,

$$\delta(\mathcal{G}^{\text{Gram}}, \mathcal{G}^{\text{full}}) \geq \sup_{k \in H - H} \left\{ \Phi\left(\frac{\sqrt{I_f}\,\|k\|}{2} \right) - \Phi\left(\frac{\|k\|}{2\sigma} \right) \right\},$$

which is strictly positive for every non-Gaussian noise law in scope and vanishes for the centered Gaussian density with variance σ^2. Thus the Gram experiment is strictly deficient relative to the full experiment as a whole compact-indexed experiment, not only along a single fixed direction.

This closes the gap between the pairwise statement and the headline claim. The pairwise floor of Theorem F.5 is the specialization of Theorem F.6 to a single direction $k = h_1 - h_0$; the experiment-level statement additionally transfers the strict ordering to some bounded decision problem at the experiment level, which is exactly what a positive deficiency certifies, and justifies the term "strict experiment ordering" used throughout.

At the nonsmooth Laplace boundary, Corollary I.1 gives the exact integrated-Bayes coefficients $R_{\text{Gram,N}} = 2db^2/N + o(N^{-1})$ and $R_{\text{full,N}} = db^2/N + o(N^{-1})$, hence a limiting scaled gap db^2.

F.2 A TWO-LAYER SOURCE-SCORE ESCAPE

The lower bound identifies not only what is lost, but also a precise place where a single nonlinear operation can restore first-order efficiency.

Theorem F.7 (Two attention layers and one score channel suffice)**.** *Assume the noise conditions of Theorem F.4, let $W \sim \mathcal{N}(0, \tau^2 I_d)$, and let $X_q \sim \mathcal{N}(0, I_d)$ be independent. There is a width-$(d+3)$ common-source model with two affine linear-attention layers and one scalar tokenwise score channel between them whose prediction is $X_q^\top \widehat{W}_N$, where*

$$u_N = \frac{1}{N} \sum_{i=1}^N X_i Y_i, \qquad \widehat{W}_N = u_N + \frac{1}{I_f N} \sum_{i=1}^N X_i \psi(Y_i - X_i^\top u_N). \tag{27}$$

It satisfies

$$\mathbb{E}\|\widehat{W}_N - W\|_2^2 = \mathbb{E}(X_q^\top \widehat{W}_N - X_q^\top W)^2 = \frac{d}{I_f N} + o(N^{-1}). \tag{28}$$

By Proposition E.5 and Theorem F.2, every model whose source paths remain affine, even with arbitrary query-only nonlinearities, satisfies

$$\liminf_{N\to\infty} N\,\mathbb{E}(\widehat{Y}_{q,N} - X_q^\top W)^2 \geq d\sigma^2.$$

For logistic noise with scale s,

$$\psi(r) = s^{-1}\tanh(r/(2s)), \qquad \sigma^2 = \frac{\pi^2 s^2}{3}, \qquad I_f = \frac{1}{3s^2}.$$

Thus the score channel is exactly one standard tanh *unit after absorbing the input gain $1/(2s)$ and output scale $1/s$ into adjacent affine maps. The source-score coefficient is $3ds^2$, and every query-only extension retains a lower coefficient of at least $d\pi^2 s^2/3$.*

The same scalar primitive also reaches the nonsmooth boundary: Corollary I.2 proves that a sharpening tanh channel with $k_N \to \infty$ and $k_N = o(N^{1/4})$ attains $db^2/N + o(N^{-1})$ under Laplace noise. A fixed gain is not claimed sufficient for this sharp limit.

The construction is exact in the layer convention of (18). Write the hidden state as $e = (x, y, p, a) \in \mathbb{R}^{d+3}$. The first layer preserves (x, y), writes the pilot $p = u_N^\top x$, and writes the residual $a = y - p$; the sole scalar channel overwrites a by $\psi(a)$. The second layer aggregates $N^{-1}\sum_i \psi(a_i)X_i^\top$ and adds I_f^{-1} times its query output to the preserved pilot. All attention offsets are zero, both source sets are exactly the demonstrations, and no position, length feature, matrix inverse, extra affine sublayer, or nonlinear query readout is used. The explicit matrices and the L^2 proof, including same-prompt pilot reuse, are in Section K.6.

Additional consequences. The appendix gives explicit bounded contamination separations, including a full-dimensional strong-label adversary, and records the broader fixed-(N, d) DeepSets-style scope result. These support the information boundary but are not part of the main efficiency chain.

F.3 The score is the unique efficient source channel

Theorem F.7 exhibits one channel attaining the Fisher coefficient. We now show it is the *unique* scalar source channel to do so within the same architectural template: two affine layers with the specific moment pilot $u_N = X^\top Y/N$ and one scalar tokenwise residual channel between them.

Theorem F.8 (Efficiency spectrum of the one-source-channel architecture)**.** *Assume the noise conditions of Theorem F.4, let $W \sim \mathcal{N}(0, \tau^2 I_d)$ with $\tau^2 > 0$, and let ϕ be an admissible scalar channel: ϕ is $C^{1,1}$ with $\mathbb{E}\phi(\varepsilon)^2 < \infty$, $0 < \mathbb{E}\phi'(\varepsilon) < \infty$, and $\phi(x)f(x) \to 0$ as $|x| \to \infty$. Write $a = \mathbb{E}\phi'(\varepsilon)$, $v = \mathbb{E}\phi(\varepsilon)^2$, $b = \mathbb{E}[\varepsilon\,\phi(\varepsilon)]$, and $\kappa = \sigma^2 + \tau^2(d+1)$. Consider the common-source predictor with Gram pilot u_N and one scalar source residual channel ϕ,*

$$\widehat{W}_N^\phi(\gamma) = u_N + \frac{\gamma}{N}\sum_{i=1}^N X_i\,\phi(Y_i - X_i^\top u_N).$$

Then, for the canonical (influence-function) gain $\gamma = 1/a$, the clean-query risk satisfies

$$N\,\mathbb{E}(X_q^\top \widehat{W}_N^\phi - X_q^\top W)^2 \longrightarrow \frac{d\,v}{a^2}, \qquad \frac{v}{a^2} \geq \frac{1}{I_f}, \quad \text{equality iff } \phi \propto \psi \text{ in } L^2(f).$$

For an arbitrary fixed gain γ, the leading coefficient is instead $d\,F(\gamma)$ with

$$F(\gamma) = v\gamma^2 + \kappa(1-\gamma a)^2 + 2b\gamma(1-\gamma a),$$

and $\min_\gamma F(\gamma) \geq 1/I_f$ with equality iff $\phi \propto \psi$. Consequently, within this moment-pilot one-source-channel template, the location score ψ is the unique admissible channel, up to positive scaling, that attains the Fisher coefficient d/I_f, with or without gain tuning.

The uniqueness is specific to the chosen pilot and scalar-channel template. The moment pilot $u_N = X^\top Y/N$ carries the design-error term $(H_N - I_d)W$, of squared norm $\tau^2 d(d+1)/N$; it can be removed only by setting the prefactor $(1-\gamma a)$ of the pilot error to zero. Once that prefactor vanishes, the channel itself must be score-proportional to reach d/I_f. An OLS pilot $H_N^{-1}u_N$ would remove the design term and can admit non-score-proportional mixtures of ψ and ε; the stated uniqueness therefore depends on this specific moment pilot, not on Gram measurability alone. OLS is itself Gram-measurable. A fixed $C^{1,1}$ channel does not cover the nonsmooth Laplace score, which the appendix handles separately by the sharpening sequence of Corollary I.2.

Beyond the Gram: a finite information budget. The same perspective yields a graded boundary for complete mixed moments through degree $P \geq 2$. Under the conditions of Theorem J.1, their local information is

$$J_P = \|\Pi_{\mathcal{P}_{P-1}} \psi_f\|^2_{L^2(f)}, \qquad J_2 = \sigma^{-2} \leq J_P \leq I_f,$$

where $\mathcal{P}_{P-1}$ is the space of scalar error polynomials of degree at most $P-1$. The optimal local squared-risk coefficient is d/J_P; the stated all-moment prior and error assumptions also give the integrated risk $d/(J_P N) + o(N^{-1})$. Thus quartic noise requires the cubic score direction, while Gaussian noise needs only the linear one. A complementary construction matches any prescribed finite set of Gaussian moments with smooth, positive, nearly Gaussian noise whose Fisher information is arbitrarily large (Lemma J.2). No fixed moment budget is uniformly efficient across these fixed noise laws. Section J gives the full statements and proofs. This is a sharp result for the complete moment experiment; it does not supply a matching lower bound for every restricted source-feature bank.

G Finite-sample witnesses

The asymptotic experiment ordering has a finite-sample full-rank counterpart. We record it after the sharp Gram-to-Fisher boundary because it is a complementary witness rather than the source of the leading coefficient. The classical Gaussian characterization and a simpler scalar continuous-prior gap are deferred to Section H.

Let $X \in \mathbb{R}^{N \times d}$ be the design matrix and, for fixed X, define

$$T_X(y) = (\mathbf{1}^\top y, X^\top y, \|y\|_2^2).$$

The augmented Gram $\widetilde{S}$ is a function of $(X, T_X(y))$. Thus an oracle given $(X, T_X(y))$ is strictly stronger than the architecture.

Theorem G.1 (Full-rank Gaussian designs with Laplace noise)**.** *Fix $d \geq 1$, $N \geq d+3$, and $\tau, b > 0$. Draw independently*

$$W \sim \mathcal{N}(0, \tau^2 I_d), \qquad X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d), \qquad \varepsilon_i \overset{\text{iid}}{\sim} \text{Laplace}(0, b), \qquad Y_i = X_i^\top W + \varepsilon_i.$$

Then $m(X,Y) = \mathbb{E}[W \mid X, Y]$ is not measurable with respect to $\sigma(X, T_X(Y))$, and hence not with respect to $\sigma(\widetilde{S})$. If an independent $X_q \sim \mathcal{N}(0, I_d)$ is queried, every Gram-measurable predictor has strict excess Bayes risk

$$\mathbb{E}\left\| m(X,Y) - \mathbb{E}[m(X,Y) \mid \widetilde{S}] \right\|_2^2 > 0. \tag{29}$$

The design has rank d almost surely. The proof in Section K.10 perturbs the scalar fiber witness to a full-rank design, then promotes its nonzero posterior derivative to an open set of Gaussian designs. The sample-size condition makes the regular label fiber have dimension $N - d - 2 \geq 1$; it is sufficient for this construction and is not claimed necessary. Under Gaussian observation noise, by contrast, the posterior mean is a function of $(X^\top X, X^\top y)$, so the same Gram statistic is sufficient.

H Additional finite-sample witnesses

H.1 The Gaussian blind spot

For intercept-only regression, $x_i = x_q = 1$, the label-dependent part of $\widetilde{S}$ is equivalent, at fixed N, to

$$T_N(Y) = \left(\sum_i Y_i, \sum_i Y_i^2 \right).$$

Theorem H.1 (Classical Gaussian characterization Kelker & Matthes, 1970)**.** *Let $f : \mathbb{R} \to (0, \infty)$ be a density with $\log f \in C^2$. For three i.i.d. samples from the location family $f(y - \theta)$, T_3 is sufficient for every $\theta \in \mathbb{R}$ if and only if f is a nondegenerate Gaussian density.*

The characterization is classical and is not counted as a contribution. Its architecture corollary is immediate and broad.

Corollary H.2 (Every regular non-Gaussian family has a Gram gap)**.** *Under the assumptions of Theorem H.1, if f is not Gaussian, then there exist $\theta_0 \neq \theta_1$ such that, under their equal two-point prior and three noisy demonstrations, every T_3-measurable predictor has strictly greater Bayes risk than the full-prompt posterior mean. The same lower bound applies to every architecture in Theorem E.2.*

H.2 A NATURAL CONTINUOUS GAP

The preceding corollary selects a two-point prior. The next theorem uses the standard continuous Gaussian task prior.

Theorem H.3 (Gaussian tasks with Laplace observation noise)**.** *For every $N \geq 3$ and $\tau, b > 0$, let*

$$\Theta \sim \mathcal{N}(0, \tau^2), \qquad Y_i = \Theta + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \text{Laplace}(0, b).$$

If $m_N(Y) = \mathbb{E}[\Theta \mid Y]$, then

$$\delta_N(\tau, b) := \mathbb{E}[\text{Var}(m_N(Y) \mid T_N(Y))] > 0. \tag{30}$$

Thus every common-source affine linear-attention predictor incurs at least $\delta_N(\tau, b)$ excess risk over the full-prompt Bayes predictor.

The proof, in Section K.9, constructs a smooth curve along which both power sums are exactly fixed but m_N has nonzero directional derivative. An almost-everywhere-to-fiber lemma promotes this local variation to positive conditional variance; the argument does not rely on conditioning on a probability-zero collision.

I NONSMOOTH LAPLACE EFFICIENCY

The smooth-score assumptions in the main efficiency theorem exclude Laplace noise. The exact Gram–full gap and the matched architectural escape nevertheless persist at this boundary.

Corollary I.1 (Laplace coefficient and full-prompt separation)**.** *Under the assumptions of Theorem F.2, take $\varepsilon_i \sim \text{Laplace}(0, b)$. Then*

$$R_{\text{Gram,N}}(\pi) = \frac{2db^2}{N} + o(N^{-1}). \tag{31}$$

If $R_{\text{full,N}}(\pi)$ denotes the full-prompt Bayes coefficient risk, then

$$R_{\text{full,N}}(\pi) = \frac{db^2}{N} + o(N^{-1}), \tag{32}$$

and consequently

$$N\{R_{\text{Gram,N}}(\pi) - R_{\text{full,N}}(\pi)\} \longrightarrow db^2. \tag{33}$$

The proof supplies an explicit split one-step full-prompt estimator with risk $db^2/N + o(N^{-1})$ and a vector van Trees lower bound that matches it for every stated prior. The displayed equalities are integrated-Bayes statements. The Gram coefficient also has the local-minimax form stated in Theorem F.2; no pointwise lower bound for every estimator sequence is asserted. Thus Theorem F.2 supersedes a regular-rule-only calculation, while Theorem G.1 remains the complementary, finite-N strict-sufficiency separation.

Corollary I.2 (A standard $\tanh$ channel is Laplace-efficient)**.** *Let W be independent of the design and noise with $\mathbb{E}\|W\|^4 < \infty$, let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, and let the independent query be $X_q \sim \mathcal{N}(0, I_d)$. If $\varepsilon_i \overset{\text{iid}}{\sim} \text{Laplace}(0, b)$, choose any deterministic $k_N \to \infty$ with $k_N = o(N^{1/4})$. With the same moment pilot, define*

$$c_k = \mathbb{E}_{\varepsilon \sim \text{Laplace}(0,b)}\left[k \operatorname{sech}^2(k\varepsilon)\right]$$

and

$$\widehat{W}_N^{\tanh} = u_N + \frac{c_{k_N}^{-1}}{N} \sum_{i=1}^{N} X_i \tanh\big(k_N(Y_i - X_i^\top u_N)\big)\,. \tag{34}$$

Then the resulting sequence of width-$O(d)$ two-layer models, each with one $\tanh$ *unit of gain k_N, satisfies*

$$\mathbb{E}\|\widehat{W}_N^{\tanh} - W\|^2 = \frac{db^2}{N} + o(N^{-1}).$$

Thus the same standard scalar activation attains the exact full-prompt coefficient in the nonsmooth Laplace boundary case; a fixed gain is not claimed sufficient for the sharp limit.

J FINITE MOMENT BUDGETS AND HIDDEN FISHER INFORMATION

J.1 A SHARP LAW FOR FINITE MOMENT BUDGETS

The Gram barrier is the first member of an information hierarchy. What matters is which score directions survive aggregation: a moment of total degree P can access at most a degree-$(P-1)$ error feature, because one factor is needed to retain the task-bearing design direction.

For $P \geq 2$, define the complete mixed-moment statistic

$$\mathcal{M}_{P,N} = \sigma\left\{ N^{-1}\sum_{i=1}^{N} X_i^{\alpha} Y_i^{b} : 1 \leq |\alpha| + b \leq P \right\}.$$

Let $v_P(e) = (1, e, \ldots, e^{P-1})^{\top}$, and write

$$B_P = \mathbb{E}[v_P(\varepsilon)v_P(\varepsilon)^{\top}], \qquad a_P = \mathbb{E}v_P'(\varepsilon), \qquad J_P = a_P^{\top} B_P^{-1} a_P.$$

The inverse exists for any everywhere-positive noise density with the stated moments. Let $\mathcal{P}_k$ denote the scalar polynomials of degree at most k.

Theorem J.1 (Sharp moment-order law)**.** *Fix d and $P \geq 2$. Let $X_i \sim \mathcal{N}(0, I_d)$ and $Y_i = X_i^{\top} w + \varepsilon_i$, with independent, mean-zero errors having an everywhere-positive density and finite $2P$th moment. For every center w_0 and compact $H \subset \mathbb{R}^d$, the experiment observing $\mathcal{M}_{P,N}$ under $w = w_0 + h/\sqrt{N}$, $h \in H$, converges in Le Cam distance to*

$$Z = h + J_P^{-1/2} G, \qquad G \sim \mathcal{N}(0, I_d).$$

Its all-measurable local asymptotic minimax squared-risk coefficient, with expanding local cubes, is d/J_P: writing $H_C = [-C/2, C/2]^d$,

$$\lim_{C\to\infty} \liminf_{N\to\infty} \inf_{\widehat{w}_N \ \mathcal{M}_{P,N}\text{-measurable}} \sup_{h\in H_C} N\,\mathbb{E}_{w_0+h/\sqrt{N}} \left\| \widehat{w}_N - \left(w_0 + \frac{h}{\sqrt{N}} \right) \right\|^2 = \frac{d}{J_P}.$$

The preceding fixed-compact experiment convergence does not assert this minimax risk on each fixed bounded parameter set. If the errors and a continuous, everywhere-positive task prior π have moments of every order, then

$$\inf_{\widehat{w}_N \ \mathcal{M}_{P,N}\text{-measurable}} \mathbb{E}_{\pi}\|\widehat{w}_N - W\|^2 = \frac{d}{J_P N} + o(N^{-1}).$$

If, moreover, f is locally absolutely continuous, its weak location score $\psi = -f'/f$ lies in $L^2(f)$, and $\lim_{x\to\pm\infty} x^k f(x) = 0$ for $k = 0, \ldots, P-1$, then

$$J_P = \|\Pi_{\mathcal{P}_{P-1}}\psi\|^2_{L^2(f)}, \qquad \frac{1}{\sigma^2} = J_2 \leq J_3 \leq \cdots \leq I_f. \tag{35}$$

Equality $J_P = I_f$ holds exactly when $\psi \in \mathcal{P}_{P-1}$.

The moment assumptions for the experiment limit are finite and explicit; all moments are only a convenient sufficient condition for the polynomial attaining rule's integrated risk. The theorem concerns arbitrary measurable decoders, not merely moment estimators with a regular asymptotic expansion.

The degree shift has statistical content. After decomposing the design into Gaussian Hermite coordinates, the only parameter-bearing block is $\{X_j \varepsilon^b : 0 \leq b \leq P-1\}$, for each j. Every other design coordinate is a Gaussian ancillary in the limiting experiment, even though its finite-sample law may depend on the task. This identifies both the retained score projection and all information lost by the complete statistic.

Connection to a source bank. For the read-once architecture of Proposition K.11, with the moment pilot $u = N^{-1}\sum_i X_i Y_i$, suppose the scalar residual bank contains the identity and consists of polynomials of degree at most $D \geq 1$. Its complete accessible statistic is then measurable in $\mathcal{M}_{2D,N}$. Indeed, expanding $X_i \Phi_j(Y_i - X_i^\top u)$ uses raw monomials of degree at most $D+1$, and expanding $\Phi_j \Phi_k(Y_i - X_i^\top u)$ uses degree at most $2D$; their coefficients depend on the pilot u, which is itself measurable in $\mathcal{M}_{2,N}$. Subsequent affine common-source layers cannot escape this envelope. Thus the theorem applies to any depth and width after this fixed source feature budget. This is an envelope lower bound; sharpness for a particular bank requires its own accessible weighted feature span.

J.2 Proof of the sharp moment-order law

Proof of Theorem J.1. Let $m(x, e)$ list all distinct nonconstant monomials of total degree at most P, and put $\mu = \mathbb{E}m(X, \varepsilon)$ and $\Sigma = \mathrm{Cov}(m(X, \varepsilon))$. Its covariance is finite and positive definite. Otherwise a nonconstant polynomial would be constant almost surely under an everywhere-positive density, which is impossible.

A strong CLT for the complete moment vector. A finite convolution of the law of $m(X, \varepsilon)$ has a density. To verify this, observe that the ranges of $Dm(z)$ as z varies jointly span the entire output space: an annihilator c would make $\nabla(c^\top m) = 0$ everywhere, contradicting independence of the listed polynomials modulo constants. Select finitely many points whose derivative ranges span. At this tuple the sum map $\sum_{i=1}^s m(z_i)$ has a nonzero full-rank Jacobian minor. That minor is a nonzero polynomial, so its zero set has Lebesgue measure zero. The inputs have a joint density; the sum map is therefore a submersion almost surely. The coarea formula, applied on a countable collection of submersion charts, gives absolute continuity of the pushforward. Prokhorov's multivariate total-variation CLT criterion (Prokhorov, 1952; Bally & Caramellino, 2016) now yields

$$Z_N = N^{-1/2} \sum_i \{m(X_i, \varepsilon_i) - \mu\} \longrightarrow_{\mathrm{TV}} G \sim \mathcal{N}(0, \Sigma). \tag{36}$$

The local experiment is an exact shear. Replacing Y_i by $Y_i - X_i^\top w_0$ is an invertible linear transformation of the complete list of empirical moments. The substitution preserves total degree. More generally, an invertible matrix $R(t)$, polynomial in t, obeys

$$m(x, e + x^\top t) = R(t)m(x, e), \quad R(0) = I, \quad R(t)^{-1} = R(-t).$$

Under $w = w_0 + h/\sqrt{N}$, the normalized, recentered moment statistic has the exact representation

$$Z_N^{(h)} = R(h/\sqrt{N})Z_N + \sqrt{N}\{R(h/\sqrt{N})\mu - \mu\}.$$

Let D be the derivative at zero of $t \mapsto R(t)\mu$. Uniformly on compact sets of h, $R(h/\sqrt{N}) \to I$ and the displayed mean term tends to Dh. By TV contraction, (36) allows Z_N to be replaced by G. The resulting Gaussian means and positive-definite covariances converge uniformly, giving

$$\sup_{h \in H} \left\| \mathcal{L}(Z_N^{(h)}) - \mathcal{N}(Dh, \Sigma) \right\|_{\mathrm{TV}} \longrightarrow 0. \tag{37}$$

The normalization is invertible and does not depend on h. Identity kernels in both directions thus give convergence of the entire moment experiment in Le Cam distance. This approximation is uniform over centers w_0 as well: after the observable recentering, the exact local law does not involve w_0.

Hermite coordinates isolate the information. Let H_α be the probabilists' multivariate Hermite polynomial, so $\mathbb{E}[H_\alpha(X)H_\gamma(X)] = \alpha!\,\mathbf{1}\{\alpha = \gamma\}$ and $H_{e_j}(X) = X_j$. Changing from monomials to $H_\alpha(X)\varepsilon^b$, $|\alpha| + b \leq P$, is invertible and triangular by degree, with the known constant coordinate removed. The limiting Gaussian covariance is block diagonal in α. For $\alpha \neq 0$ the corresponding covariance block is

$$\alpha!\,\mathbb{E}\left[(1, \varepsilon, \ldots, \varepsilon^{P-|\alpha|})(1, \varepsilon, \ldots, \varepsilon^{P-|\alpha|})^\top\right],$$

and the $\alpha = 0$ block is the covariance of centered error powers. These blocks are positive definite.

The local mean derivative of a coordinate is

$$\frac{d}{dt}\mathbb{E}\left[H_\alpha(X)(\varepsilon + tX^\top h)^b\right]\Big|_{t=0} = b\,\mathbb{E}\varepsilon^{b-1}\mathbb{E}[H_\alpha(X)X]^\top h.$$

By Hermite orthogonality this vanishes unless $\alpha = e_j$. Thus every other block is an independent parameter-free Gaussian ancillary. The remaining blocks are independent across j and have distributions

$$T_j \sim \mathcal{N}(a_P h_j, B_P), \qquad 1 \le j \le d.$$

The sufficient scalar $J_P^{-1} a_P^\top B_P^{-1} T_j$ has law $\mathcal{N}(h_j, J_P^{-1})$. Its Gaussian residual is independent and parameter-free. Projecting out, or independently regenerating, these ancillaries proves the stated Gaussian-shift equivalence.

An attaining rule in the moment experiment. Put $\beta = B_P^{-1} a_P$ and $g(e) = \beta^\top v_P(e)$. The first row of $B_P \beta = a_P$ gives $\mathbb{E} g = 0$, and

$$\mathbb{E} g' = \beta^\top a_P = J_P, \qquad \mathbb{E} g^2 = \beta^\top B_P \beta = J_P.$$

Define

$$\widehat{w}_{P,N} = u + J_P^{-1} N^{-1} \sum_i X_i g(Y_i - X_i^\top u), \qquad u = N^{-1} \sum_i X_i Y_i. \tag{38}$$

It is measurable in $\mathcal{M}_{P,N}$: expanding each term $X_i(Y_i - X_i^\top u)^k$ uses raw monomials of degree at most $k+1 \le P$, with coefficients polynomial in the observed pilot u.

Let $s = u - w$. Exact polynomial Taylor expansion gives

$$\widehat{w}_{P,N} - w = J_P^{-1} \overline{X g(\varepsilon)} - J_P^{-1} \{ \overline{X X^\top g'(\varepsilon)} - J_P I \} s + J_P^{-1} \sum_{k=2}^{\deg g} \frac{(-1)^k}{k!} \overline{X (X^\top s)^k g^{(k)}(\varepsilon)}.$$

Under the all-moment assumptions, $s = O_{L^q}(N^{-1/2})$ for every fixed q; centered empirical polynomial averages are $O_{L^q}(N^{-1/2})$, and noncentered polynomial envelopes are $O_{L^q}(1)$. These facts follow from the iid centered-sum moment inequality and Hölder's inequality, uniformly on compact task sets and after integration over the stated prior. Hölder then bounds the total remainder by $O_{L^2}(N^{-1})$. The leading term has covariance $(J_P N)^{-1} I_d$, so

$$\mathbb{E} \|\widehat{w}_{P,N} - w\|^2 = \frac{d}{J_P N} + O(N^{-3/2}).$$

All-measurable risk lower bounds. On a fixed local cube, projecting the action to the cube bounds the squared loss. Equation (37) therefore transfers its local Bayes risk to the Gaussian-shift experiment. As the cube side tends to infinity, its flat-prior Gaussian Bayes risk tends to d/J_P, by the flat-cube lemma used for the Gram theorem. This proves the local minimax lower bound. For its matching upper bound under the finite-moment assumptions alone, apply the reverse Le Cam kernel to the Gaussian estimator projected to the local cube; bounded-loss transfer applies, and conditional expectation removes any randomization.

For integrated risk, partition a fixed compact task region into cubes of side $C/\sqrt{N}$. On each cube lower-bound the prior by its infimum and lower-bound the integrated risk of any rule by the local flat-prior Bayes risk divided by N. The approximation is uniform over cube centers by the exact recentering above. Uniform continuity of the prior makes the sum of cube volumes times density infima tend to the prior mass of the region. Let the compact region exhaust $\mathbb{R}^d$ and then $C \to \infty$; this gives $\liminf_N N R_{P,N}(\pi) \ge d/J_P$. Equation (38) gives the matching upper bound under the stated moment assumptions.

Finally, integration by parts gives $\mathbb{E}[v_P \psi] = \mathbb{E} v_P' = a_P$, identifying g as the score's orthogonal projection onto $\mathcal{P}_{P-1}$. For $P = 2$, $B_P = \mathrm{diag}(1, \sigma^2)$ and $a_P = (0,1)^\top$. Nested projection spaces give monotonicity, and equality with I_f holds precisely when the score belongs to the polynomial span. □

J.3 An explicit Fourier construction with exactly Gaussian moments

Lemma J.2 (Explicit moment-matched Gaussian perturbations)**.** *Fix an integer $P \ge 2$, $0 < \delta < 1$, and a frequency $\omega > P$. Let $s_j = j^2$, $j = 0, \dots, P$, and define the Lagrange polynomials*

$$\ell_j(s) = \prod_{\substack{0 \le k \le P \\ k \ne j}} \frac{s - s_k}{s_j - s_k}.$$

Set

$$c_j(\omega) = \exp\{(j^2 - \omega^2)/2\}\ell_j(\omega^2), \qquad D_\omega = 1 + \sum_{j=0}^{P} |c_j(\omega)|.$$

$$h_\omega(x) = D_\omega^{-1} \left\{ \cos(\omega x) - \sum_{j=0}^{P} c_j(\omega) \cos(jx) \right\}, \tag{39}$$

$$f_\omega(x) = \varphi(x)\{1 + \delta h_\omega(x)\}. \tag{40}$$

Then f_ω is an even, smooth, everywhere-positive density satisfying

$$(1-\delta)\varphi \le f_\omega \le (1+\delta)\varphi, \qquad \mathbb{E}_{f_\omega} X^r = \mathbb{E}_\varphi X^r \quad (0 \le r \le 2P). \tag{41}$$

In particular it has mean zero, variance one, and all moments, with $\|f_\omega - \varphi\|_{\mathrm{TV}} \le \delta/2$. Its score $\psi_\omega = -f_\omega'/f_\omega$ obeys

$$\psi_\omega(x) = x - \frac{\delta h_\omega'(x)}{1 + \delta h_\omega(x)}, \qquad I_{f_\omega} = 1 + \delta^2 \mathbb{E}_\varphi \frac{h_\omega'(X)^2}{1 + \delta h_\omega(X)}. \tag{42}$$

Consequently, writing $E_\omega = \mathbb{E}_\varphi h_\omega'(X)^2$,

$$1 + \frac{\delta^2}{1+\delta} E_\omega \le I_{f_\omega} \le 1 + \frac{\delta^2}{1-\delta} E_\omega, \qquad \frac{E_\omega}{\omega^2} \longrightarrow \frac{1}{2}. \tag{43}$$

For fixed P, δ, the leading Fisher growth is explicitly

$$I_{f_\omega} = 1 + \big(1 - \sqrt{1-\delta^2}\big)\omega^2 + o(\omega^2). \tag{44}$$

For every fixed ω, the score derivative is globally Lipschitz and all score moments and previously stated score boundary conditions hold. The complete degree-P moment experiment has $J_P = 1$, whereas the full-data Fisher information tends to infinity as $\omega \to \infty$.

Proof. The Lagrange interpolation identity gives, for $0 \le r \le P$,

$$\sum_{j=0}^{P} \ell_j(\omega^2) j^{2r} = \omega^{2r},$$

where the constant polynomial for $r = 0$ is understood to equal one at every node, including zero. For the probabilists' Hermite polynomial He_{2r}, Gaussian Fourier differentiation gives

$$\mathbb{E}_\varphi[\mathrm{He}_{2r}(X)\cos(\nu X)] = (-1)^r \nu^{2r} e^{-\nu^2/2}.$$

This follows by applying Gaussian integration by parts $2r$ times to $\cos(\nu x)$. The definition of c_j and the interpolation identity therefore imply

$$\mathbb{E}_\varphi[\mathrm{He}_{2r}(X) h_\omega(X)] = \frac{(-1)^r e^{-\omega^2/2}}{D_\omega} \left\{ \omega^{2r} - \sum_j \ell_j(\omega^2) j^{2r} \right\} = 0.$$

Odd polynomial moments vanish by symmetry. Hermite polynomials form a triangular basis of the ordinary polynomials, proving matching of all raw moments through degree $2P$, including normalization. The triangle inequality gives $|h_\omega(x)| \le 1$ for every x and every allowed frequency, proving positivity and the Gaussian-relative bounds.

Differentiating the log density gives the score formula. Its Fisher information expands as

$$I_{f_\omega} = \mathbb{E}_{f_\omega} X^2 - 2\delta \mathbb{E}_\varphi[X h_\omega'(X)] + \delta^2 \mathbb{E}_\varphi \frac{h_\omega'(X)^2}{1 + \delta h_\omega(X)}.$$

The first term is one. Gaussian integration by parts and the zeroth and second moment constraints give

$$\mathbb{E}_\varphi[X h_\omega'(X)] = \mathbb{E}_\varphi[(X^2 - 1) h_\omega(X)] = 0.$$

This proves (42); bounding its denominator between $1-\delta$ and $1+\delta$ proves the two Fisher bounds.

For fixed P, each $\ell_j(\omega^2) = O_P(\omega^{2P})$ as $\omega \to \infty$. Consequently,

$$c_j(\omega) = O_P(\omega^{2P} e^{-\omega^2/2}), \qquad D_\omega \to 1.$$

Moreover

$$h'_\omega(x) = D_\omega^{-1} \left\{ -\omega \sin(\omega x) + \sum_{j=0}^{P} j c_j(\omega) \sin(jx) \right\}.$$

The second term is $o(\omega)$ in $L^2(\varphi)$, while $\mathbb{E}_\varphi \sin^2(\omega X) = (1 - e^{-2\omega^2})/2 \to 1/2$. This proves the derivative-energy limit and unbounded Fisher information.

For fixed P and δ, the coefficient bounds also give $\|h_\omega - \cos(\omega\cdot)\|_\infty \to 0$ and $\|h'_\omega/\omega + \sin(\omega\cdot)\|_\infty \to 0$. Since $1 + \delta h_\omega \geq 1 - \delta > 0$, the Fisher identity yields

$$\frac{I_{f_\omega} - 1}{\omega^2} = \mathbb{E}_\varphi F_\delta(\omega X) + o(1), \qquad F_\delta(t) = \frac{\delta^2 \sin^2 t}{1 + \delta \cos t}.$$

Uniform trigonometric approximation and $\mathbb{E} e^{ik\omega X} = e^{-k^2\omega^2/2}$ imply convergence to the period average. Using $F_\delta(t) = 1 - \delta\cos t - (1-\delta^2)/(1+\delta\cos t)$ and $(2\pi)^{-1}\int_0^{2\pi}(1+\delta\cos t)^{-1}\,dt = (1-\delta^2)^{-1/2}$, this average is $1 - \sqrt{1-\delta^2}$.

For each fixed frequency, h_ω and every derivative are bounded, and $1 + \delta h_\omega \geq 1 - \delta > 0$. The correction $\psi_\omega - x$ and all its derivatives are bounded. In particular ψ''_ω is bounded, so ψ'_ω is globally Lipschitz. Gaussian-relative tails give all required moments and make $x f_\omega(x)$ and $\psi_\omega(x) f_\omega(x)$ vanish at both tails.

Finally, for every polynomial q of degree at most $P-1$, integration by parts and moment matching give

$$\langle \psi_\omega - x, q \rangle_{L^2(f_\omega)} = \mathbb{E}_{f_\omega} q'(X) - \mathbb{E}_{f_\omega}[X q(X)] = \mathbb{E}_\varphi q'(X) - \mathbb{E}_\varphi[X q(X)] = 0.$$

The score projection is exactly x, with squared norm one. Hence $J_P = 1$ by the moment information theorem. □

A closed-form finite-frequency lower bound. The derivative energy requires no numerical differentiation or quadrature. Let $\lambda = (\omega, 0, 1, \ldots, P)$ and $u = (1, -c_0, -c_1, \ldots, -c_P)$. Then

$$E_\omega = \frac{1}{2D_\omega^2} \sum_{a,b=0}^{P+1} u_a u_b \lambda_a \lambda_b \left\{ e^{-(\lambda_a - \lambda_b)^2/2} - e^{-(\lambda_a + \lambda_b)^2/2} \right\}. \tag{45}$$

This follows from $\mathbb{E}\sin(aX)\sin(bX) = \{e^{-(a-b)^2/2} - e^{-(a+b)^2/2}\}/2$. Substitution into (43) gives explicit upper and lower Fisher bounds at any selected frequency. The associated matrix is positive semidefinite, since it is a second-moment matrix of the derivative sine features.

For a prescribed efficiency separation $M > 1$, choose one sufficiently large frequency for which the explicit lower bound exceeds M. That density is then fixed before applying the sample-size asymptotics. Neither the construction nor its finite-frequency Fisher bound asserts a uniform finite-sample risk approximation as ω grows with N.

Relation to earlier oscillatory examples. The fact that oscillations can make Fisher information large while keeping a density close to a Gaussian is classical in spirit and appears explicitly in the Gaussian–sawtooth example of Gupta et al. (2022). The extra ingredients here are exact matching of any prescribed finite set of moments, smooth Gaussian-tailed densities, and the strong local experiment of an architecture-induced moment envelope. Our conclusion concerns the asymptotic efficiency of every measurable decoder of that envelope.

K Proofs for the statistical barrier

This appendix records the measure-theoretic steps used by Theorems H.1 and H.3 and corollary H.2.

Proof of Lemma F.1. Conditional expectation is orthogonal projection in L^2. Write $\eta_{\mathcal{F}} = \mathbb{E}[Z \mid \mathcal{F}]$ and $\eta_{\mathcal{G}} = \mathbb{E}[Z \mid \mathcal{G}]$. The Pythagorean identity gives

$$R^*_{\mathcal{G}} - R^*_{\mathcal{F}} = \mathbb{E}(\eta_{\mathcal{F}} - \eta_{\mathcal{G}})^2.$$

The tower property gives $\eta_{\mathcal{G}} = \mathbb{E}[\eta_{\mathcal{F}} \mid \mathcal{G}]$, yielding (19). □

K.1 Architecture factorization proofs

Proof of Theorem E.2. Pass every hidden state to homogeneous coordinates $\bar{e} = (e, 1)$. Affine maps are now linear. If

$$\bar{C}_{\ell-1} = \frac{1}{N} \sum_{i=1}^{N} \bar{e}_i^{(\ell-1)} \bar{e}_i^{(\ell-1)\top},$$

then (18) has a homogeneous matrix $M_\ell(\bar{C}_{\ell-1})$, affine in $\bar{C}_{\ell-1}$ and with last row $(0, \dots, 0, 1)$, such that

$$\bar{e}_j^{(\ell)} = M_\ell(\bar{C}_{\ell-1}) \bar{e}_j^{(\ell-1)}$$

for every destination. The common-source condition is exactly what makes M_ℓ independent of j. Therefore

$$\bar{C}_\ell = M_\ell(\bar{C}_{\ell-1}) \bar{C}_{\ell-1} M_\ell(\bar{C}_{\ell-1})^\top.$$

At initialization, $\bar{C}_0 = \bar{E}_D \widetilde{S} \bar{E}_D^\top$, so induction makes every source covariance a function only of $\widetilde{S}$, while

$$\bar{e}_q^{(L)} = \left(\prod_{\ell=L-1}^{0} M_\ell(\bar{C}_\ell) \right) \bar{e}_q^{(0)}.$$

Thus the query state and affine readout depend only on $(\widetilde{S}, z_q)$ and remain affine in z_q. A shared deterministic length feature simply adds N to this conditioning argument.

For the bias-free specialization, set $C_0 = ESE^\top$ and write $e_j^{(L)} = P_L(S) E z_j$. If $\Gamma_\ell = \deg_S P_\ell$ and $\Delta_\ell = \deg_S C_\ell$, then $\Gamma_0 = 0$, $\Delta_0 = 1$, $\Gamma_\ell \le \Gamma_{\ell-1} + \Delta_{\ell-1}$, and $\Delta_\ell \le 3\Delta_{\ell-1}$. Hence $\Gamma_L \le \sum_{r=0}^{L-1} 3^r = (3^L - 1)/2$. □

Proof of Proposition E.5. The demonstration covariance recursion in the proof of Theorem E.2 is unchanged, because the query is not a source. Hence every shared attention matrix remains a function only of $(\widetilde{S}, N)$. If the query state before a layer already factors through $(\widetilde{S}, z_q, N)$, applying that shared matrix and then an arbitrary deterministic query map preserves the factorization. Induction proves the claim. □

K.2 Matched affine attainment

Proof of Proposition F.3. Use states in $\mathbb{R}^{d+1}$, initialize each demonstration as $e_i^{(0)} = (X_i, Y_i)$ and the query as $e_q^{(0)} = (X_q, 0)$. In the first layer, let

$$P_x = \mathrm{diag}(I_d, 0), \qquad P_y = e_{d+1} e_{d+1}^\top.$$

For both layers take $K = Q = P_x$, $V = P_y$, and zero offsets. In layer one take $R = P_x$ and $O = I_{d+1}$; its state is (X_j, t_j) with $t_j = X_j^\top u$. In layer two take $R = P_x + 2P_y$ and $O = -I_{d+1}$; its final scalar is $2t_j - X_j^\top H u$. At the query this is exactly (22). Both layers use the same demonstrations as sources, the query is never a source, and all biases are zero.

For the risk, put $A = H - I_d$ and $\zeta = N^{-1} X^\top \varepsilon$. Then

$$\widetilde{W}_N - W = -A^2 W + (I_d - A)\zeta.$$

The cross term vanishes after conditioning on (X, W). For fixed d, Gaussian sample-covariance moments give $\mathbb{E}\|A\|_{\mathrm{op}}^k = O(N^{-k/2})$ for $k = 2, 3, 4$, hence by independence of W and X,

$$\mathbb{E}\|A^2 W\|^2 \le \mathbb{E}\|A\|_{\mathrm{op}}^4 \, \mathbb{E}\|W\|^2 = O(N^{-2}).$$

Moreover, $\mathbb{E}[\zeta\zeta^\top \mid X] = \sigma^2 H/N$, so

$$\begin{aligned}\mathbb{E}\|(I_d - A)\zeta\|^2 &= \frac{\sigma^2}{N}\mathbb{E}\,\mathrm{tr}\{(I_d - A)(I_d + A)(I_d - A)\} \\ &= \frac{\sigma^2}{N}\mathbb{E}\,\mathrm{tr}(I_d - A - A^2 + A^3) = \frac{d\sigma^2}{N} + O(N^{-2}),\end{aligned}$$

where $\mathbb{E}A = 0$. This proves the upper bound. Conversely, by Theorem E.2, every model in the class has clean-query output

$$\widehat{Y}_q = a_N(\mathcal{S}_N)^\top X_q + b_N(\mathcal{S}_N).$$

Since X_q is independent of $(W, \mathcal{S}_N)$, with mean zero and covariance I_d,

$$\mathbb{E}(\widehat{Y}_q - X_q^\top W)^2 = \mathbb{E}\|a_N(\mathcal{S}_N) - W\|^2 + \mathbb{E}b_N(\mathcal{S}_N)^2 \geq R_{\mathrm{Gram,N}}(\pi).$$

Together with Theorem F.2, this gives the matching architecture lower coefficient. □

K.3 Arbitrary augmented-Gram rules

This section proves the quantitative statement for the complete affine augmented Gram. The proof uses total variation because a weak CLT alone does not transfer risks of arbitrary, possibly discontinuous decision rules.

Lemma K.1 (Flat-cube risk in a Gaussian shift)**.** *Let $H_C \sim \mathrm{Unif}([-C/2, C/2]^d)$ and observe $Z = H_C + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ is independent. If r_C is the Bayes risk for estimating H_C under squared Euclidean loss, then*

$$r_C \longrightarrow d\sigma^2 \qquad (C \to \infty).$$

Proof. The prior and channel factor by coordinate, so it is enough to consider one dimension and write its risk as $r_C^{(1)}$. Put $a = C/2$ and $\delta_C(Z) = \mathbb{E}[H_C \mid Z]$. Orthogonality of the conditional mean, compared with the estimator Z, gives

$$\sigma^2 = \mathbb{E}(H_C - Z)^2 = r_C^{(1)} + \mathbb{E}\{\delta_C(Z) - Z\}^2. \tag{46}$$

Conditional on $Z = z$, the variable $\xi = z - H_C$ is a $\mathcal{N}(0, \sigma^2)$ variable truncated to $[z - a, z + a]$. Fix $M < a/2$ and let $A = \{|Z| \leq a - M\}$. On A, the truncation interval contains $[-M, M]$, so

$$|\mathbb{E}[\xi \mid Z = z]| \leq u_M := \frac{\mathbb{E}[|\xi|\mathbf{1}\{|\xi| > M\}]}{\Pr(|\xi| \leq M)}.$$

On A^c, conditional Jensen and Cauchy–Schwarz give

$$\mathbb{E}\big[\mathbb{E}(\xi \mid Z)^2 \mathbf{1}_{A^c}\big] \leq (\mathbb{E}\xi^4)^{1/2} \Pr(A^c)^{1/2}.$$

If $|H_C| \leq a - 2M$ and $|\xi| \leq M$, then A occurs; hence

$$\Pr(A^c) \leq \frac{4M}{C} + \Pr(|\xi| > M).$$

Choose $M = M_C \to \infty$ with $M_C/C \to 0$. Then $u_{M_C} \to 0$ and the last display tends to zero. Therefore $\mathbb{E}\{\delta_C(Z) - Z\}^2 = \mathbb{E}\{\mathbb{E}[\xi \mid Z]\}^2 \to 0$. Equation (46) gives $r_C^{(1)} \to \sigma^2$ in one dimension, and product structure gives the stated $d\sigma^2$ limit. □

Proof of Theorem F.2. Fix a local center w_0. From the five nonconstant blocks of $\mathcal{S}_N$, make the invertible, w_0-dependent transformation

$$A_i = X_i X_i^\top - I_d, \tag{47}$$
$$B_i = X_i(Y_i - X_i^\top w_0), \tag{48}$$
$$C_i = (Y_i - X_i^\top w_0)^2 - \sigma^2, \tag{49}$$
$$D_i = X_i, \tag{50}$$
$$E_i = Y_i - X_i^\top w_0. \tag{51}$$

Here $A_i, \dots, E_i$ are one-observation quantities, and their sums are used in the normalized statistic. For example,

$$\sum_i B_i = X^\top Y - X^\top X w_0, \qquad \sum_i C_i = \|Y\|^2 - 2w_0^\top X^\top Y + w_0^\top X^\top X w_0 - N\sigma^2,$$

and the reverse transformation is immediate. Thus no block of the affine Gram, including $\mathbf{1}^\top X$, has been discarded.

At w_0, put $U = (X, \varepsilon) \in \mathbb{R}^p$, $p = d + 1$, and let $V = (A, B, C, D, E)$, representing the symmetric block A by its distinct entries (the nonredundant vech coordinates). Up to centering and an invertible rearrangement, V is the vector of all first and distinct symmetric second monomials of U, with dimension $p + p(p+1)/2$. Its covariance Σ is nonsingular: otherwise a nonzero affine-quadratic polynomial in U would be constant almost surely. Subtracting that constant would give a nonzero polynomial vanishing on the open support of U, which is impossible.

We next verify the hypothesis needed for a total-variation CLT. For $u_1, \dots, u_{p+1} \in \mathbb{R}^p$, define

$$G(u_1, \dots, u_{p+1}) = \left(\sum_j u_j,\ \sum_j u_j u_j^\top \right).$$

At affinely spanning points this map is a submersion. Indeed, an annihilator (a, H) of its differential, with H symmetric, obeys

$$a + 2Hu_j = 0 \quad \text{for every } j.$$

The differences $u_j - u_1$ span $\mathbb{R}^p$, so $H = 0$ and then $a = 0$. The law of U has an everywhere-positive density, and $p + 1$ independent draws affinely span with probability one. More explicitly, the nonspanning set is the zero set of the affine-volume minors and hence is Lebesgue-null. On its complement G is a submersion. If A_0 is a Lebesgue-null output set and B_R is a radius-R ball in the domain, the coarea formula gives

$$\int_{G^{-1}(A_0) \cap B_R} J_m G(u)\, du = \int_{A_0} \mathcal{H}^{p(p+1)-m}(G^{-1}(v) \cap B_R)\, dv = 0, \qquad m = p + p(p+1)/2.$$

Since $J_m G > 0$ on the regular set and the critical set is null, $G^{-1}(A_0) \cap B_R$ is null; letting $R \uparrow \infty$ shows that $G^{-1}(A_0)$ is null. Thus the pushforward of the product density is absolutely continuous. Equivalently, the $(p+1)$-fold convolution of the law of V is absolutely continuous.

The vector V is centered, has finite second moment by the fourth-moment assumption on ε, and has nonsingular covariance. After the fixed whitening map $\Sigma^{-1/2}$, the absolutely continuous convolution and these properties are precisely the hypotheses of the multivariate total-variation CLT stated by Bally & Caramellino (2016, Theorem 2.2). Transforming back gives

$$Z_N := N^{-1/2} \sum_{i=1}^N V_i \ \longrightarrow_{\mathrm{TV}}\ \mathcal{N}(0, \Sigma). \tag{52}$$

That multivariate criterion traces its one-dimensional origin to Prokhorov (1952); the displayed submersion argument supplies its absolute-continuity condition directly. Bally–Caramellino use the doubled test-function normalization of total variation; convergence to zero is unchanged, while quantitative bounds here use $\sup_A |P(A) - Q(A)|$.

Consider now $w = w_0 + h/\sqrt{N}$, coupled through the same (X_i, ε_i), and let $Z_N^{(h)}$ be the normalized statistic after the transformation (51). Direct expansion gives the exact affine map

$$Z_A^{(h)} = Z_A, \tag{53}$$

$$Z_B^{(h)} = Z_B + h + N^{-1/2} Z_A h, \tag{54}$$

$$Z_C^{(h)} = Z_C + 2N^{-1/2} h^\top Z_B + N^{-1/2}\|h\|^2 + N^{-1} h^\top Z_A h, \tag{55}$$

$$Z_D^{(h)} = Z_D, \tag{56}$$

$$Z_E^{(h)} = Z_E + N^{-1/2} h^\top Z_D. \tag{57}$$

For fixed h this is an affine map $T_{N,h}(z) = M_{N,h}z + b_{N,h}$. If H is compact and J inserts h only in the B block, direct inspection gives

$$\sup_{h\in H} \|M_{N,h} - I\| = O_H(N^{-1/2}), \qquad \sup_{h\in H} \|b_{N,h} - Jh\| = O_H(N^{-1/2}).$$

Total variation contracts under measurable maps. Combining this fact with (52), the Gaussian KL formula, and Pinsker's inequality yields

$$\sup_{h\in H} \left\| \mathcal{L}_{w_0+h/\sqrt{N}}(Z_N^{(h)}) - \mathcal{N}(Jh, \Sigma) \right\|_{\mathrm{TV}} \longrightarrow 0 \tag{58}$$

for every compact H. The approximation error is the same after the center-dependent bijection at every fixed w_0; this does not require the estimator to know an unknown center.

The $B = X\varepsilon$ block is orthogonal to all remaining blocks:

$$\mathrm{Var}(B) = \sigma^2 I_d, \qquad \mathrm{Cov}(B, A) = \mathrm{Cov}(B, C) = \mathrm{Cov}(B, D) = \mathrm{Cov}(B, E) = 0.$$

These identities use precisely the stated independence, $\mathbb{E}X = 0$, and $\mathbb{E}\varepsilon = 0$. It is worth recording explicitly how the residualized diagnostics of Section F sit inside these blocks. Writing $M = N(I_d + Z_A/\sqrt{N})$, $X^\top\varepsilon = \sqrt{N}\, Z_B$, and $\mathbf{1}^\top X = \sqrt{N}\, Z_D$,

$$N^{-1/2}\mathbf{1}^\top(I - P_X)\varepsilon = Z_E - \tfrac{1}{\sqrt{N}} Z_D^\top (I_d + Z_A/\sqrt{N})^{-1} Z_B = Z_E + o_P(1), \tag{59}$$

$$N^{-1/2}\big\{\|(I - P_X)\varepsilon\|^2 - N\sigma^2\big\} = Z_C - \tfrac{1}{\sqrt{N}} Z_B^\top (I_d + Z_A/\sqrt{N})^{-1} Z_B = Z_C + o_P(1), \tag{60}$$

so at the $\sqrt{N}$ scale the two diagnostics coincide with the ancillary coordinates Z_E and Z_C; centering the second at $(N - d)\sigma^2$ rather than $N\sigma^2$ shifts it by only $d\sigma^2/\sqrt{N}$. This makes precise the claim that they carry no leading task information while leaving their exact finite-N dependence intact. Hence the Gaussian limit in (58) is equivalent to

$$Z_B = h + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_d), \tag{61}$$

plus parameter-free independent noise. The limit information is therefore $\sigma^{-2} I_d$.

For clarity, we give the step that turns this local experiment into a Bayes bound for arbitrary rules. Let $H_C = [-C/2, C/2]^d$, let q_C be its uniform law, and let r_C be the Bayes risk for estimating h in (61) under q_C. By Lemma K.1,

$$r_C \longrightarrow d\sigma^2 \qquad (C \to \infty). \tag{62}$$

Fix a compact cube K and tile its interior by disjoint cubes $K_{N,j} = w_j + N^{-1/2} H_C$. The uncovered boundary strip has width $O(N^{-1/2})$ and hence prior mass $o(1)$. Write $p_{N,j} = \Pr_\pi(W \in K_{N,j})$, and let $\rho_{N,j}$ be the minimum conditional value of $N\mathbb{E}\|\widehat{w}_N - W\|^2$ on cell j, allowing a separate oracle rule in each cell. Dropping the boundary and allowing cell-specific rules can only lower risk, so

$$N R_{\mathrm{Gram,N}}(\pi) \geq \sum_j p_{N,j}\rho_{N,j}. \tag{63}$$

Uniform continuity and strict positivity of π on K make the rescaled conditional prior in every interior cell converge uniformly in total variation to q_C; call the uniform error $\eta_N \to 0$. In cell j, rescale the action to $\sqrt{N}(\widehat{w}_N - w_j)$ and project it onto H_C, which cannot increase squared loss for parameters in that cell. The rescaled loss is then bounded by dC^2. Since (57) and its convergence rate do not depend on w_j, (58) has a uniform error $e_N \to 0$ over all cells. The total-variation decision bound therefore gives

$$\rho_{N,j} \geq r_C - dC^2(e_N + \eta_N) \quad \text{uniformly in } j.$$

Multiplying by $p_{N,j}$ before summing is essential: the transfer error is weighted by $\sum_j p_{N,j} \leq 1$ and does not grow with the number of cells. Since $\sum_j p_{N,j} \to \pi(K)$, (63) gives

$$\liminf_{N\to\infty} N R_{\mathrm{Gram,N}}(\pi) \geq r_C\, \pi(K). \tag{64}$$

First let $C \to \infty$ in (64); then exhaust $\mathbb{R}^d$ by compact cubes. This proves $\liminf_N NR_{\mathrm{Gram,N}}(\pi) \geq d\sigma^2$. The same bounded-loss total-variation transfer on the local cube H_C gives, for every fixed w_0,

$$\liminf_{N\to\infty} \inf_{\widehat{w}_N \ \mathcal{S}_N\text{-measurable}} \sup_{h\in H_C} N\,\mathbb{E}_{w_0+h/\sqrt{N}} \left\| \widehat{w}_N - \left(w_0 + \frac{h}{\sqrt{N}}\right)\right\|^2 \geq r_C.$$

Letting $C \to \infty$ and using (62) proves the lower half of (21); the OLS calculation below is uniform in h and gives the matching upper half. This is the local-asymptotic-minimax conclusion; compare Le Cam & Yang (2000) and van der Vaart (1998, Theorem 8.11). The direct transfer is why the argument does not make a false pointwise assertion about a superefficient sequence.

For the matching upper bound, OLS is measurable from $(X^\top X, X^\top Y)$, a substatistic of $\mathcal{S}_N$. Since the errors are independent of X, for $N > d+1$ the Wishart inverse identity gives, at every w,

$$\mathbb{E}_w \|\widehat{w}_{\mathrm{OLS}} - w\|_2^2 = \sigma^2 \mathbb{E}\,\mathrm{tr}(X^\top X)^{-1} = \frac{d\sigma^2}{N-d-1}.$$

It also holds after integrating over π. Finally, this argument does not require $W \in L^2$. The residual $U_N = W - \widehat{w}_{\mathrm{OLS}}$ is in L^2, so the optimal Gram coefficient rule can be written as

$$\widehat{w}_{\mathrm{OLS}} + \mathbb{E}[U_N \mid \mathcal{S}_N].$$

More generally, allowing every measurable rule $F(\mathcal{S}_N, X_q)$ and an independent $X_q \sim \mathcal{N}(0, I_d)$ gives

$$\inf_F \mathbb{E}\left(F(\mathcal{S}_N, X_q) - X_q^\top W\right)^2 = \mathbb{E}\,\|U_N - \mathbb{E}[U_N \mid \mathcal{S}_N]\|^2,$$

where $U_N = W - \widehat{w}_{\mathrm{OLS}} \in L^2$ as above; when $W \in L^1$ this equals $\mathbb{E}\|W - \mathbb{E}[W \mid \mathcal{S}_N]\|^2$, but the U_N form requires no integrability of W itself and so covers every continuous positive prior permitted by the theorem. Indeed, conditionally on $(\mathcal{S}_N, X_q)$ the optimal action is $X_q^\top\{\widehat{w}_{\mathrm{OLS}} + \mathbb{E}[U_N \mid \mathcal{S}_N]\}$, since $\widehat{w}_{\mathrm{OLS}}$ is $\mathcal{S}_N$-measurable, and isotropy removes X_q after taking its conditional second moment. In particular, for any Gram coefficient rule and an independent $X_q \sim \mathcal{N}(0, I_d)$,

$$\mathbb{E}_{X_q}\left[(X_q^\top(\widehat{w} - W))^2 \mid \widehat{w}, W\right] = \|\widehat{w} - W\|^2.$$

Thus optimal coefficient and clean-query prediction risks are identical, which completes the proof. □

Lemma K.2 (Gaussian inverse-design remainder bounds)**.** *Fix d and let $n \geq d+9$. Let $X_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_d)$, $G = \sum_{i=1}^n X_i X_i^\top$, and let Δ be independent of the design. Put $a_i = X_i^\top \Delta$.*

(i) *If $|r(a)| \leq Ka^2$, then*

$$\mathbb{E}\left\| G^{-1} \sum_{i=1}^n X_i r(a_i)\right\|^2 \leq C\,\mathbb{E}\|\Delta\|^4.$$

(ii) *If, conditional on $(\Delta, X_1, \ldots, X_n)$, the D_i are independent and centered, with*

$$\mathbb{E}[D_i^2 \mid \Delta, X_1, \ldots, X_n] \leq K(|a_i| \wedge 1),$$

then

$$\mathbb{E}\left\| G^{-1} \sum_{i=1}^n X_i D_i\right\|^2 \leq \frac{C}{n}\,\mathbb{E}(\|\Delta\| + \|\Delta\|^2).$$

(iii) *If instead*

$$\mathbb{E}[D_i^2 \mid \Delta, X_1, \ldots, X_n] \leq K(a_i^2 + a_i^4),$$

then

$$\mathbb{E}\left\| G^{-1} \sum_{i=1}^n X_i D_i\right\|^2 \leq \frac{C}{n}\,\mathbb{E}(\|\Delta\|^2 + \|\Delta\|^4).$$

Here C depends only on (d, K).

Proof. We use only the standard trace moments of a Wishart matrix. If $W \sim W_d(\nu, I_d)$ and $\nu > d+3$, then

$$\mathbb{E}\,\mathrm{tr}(W^{-1}) = \frac{d}{\nu - d - 1}, \qquad \mathbb{E}\,\mathrm{tr}(W^{-2}) = \frac{d(\nu-1)}{(\nu-d)(\nu-d-1)(\nu-d-3)} = O_d(\nu^{-2}) \quad (65)$$

(Muirhead, 1982, Chapter 3).

Condition first on $\Delta = \delta$, write X for the design matrix, $q_i = r(a_i)$, $\beta = G^{-1}X^\top q$, and $G_{-i} = \sum_{j\neq i} X_j X_j^\top$. Since $XG^{-1}X^\top$ is an orthogonal projection,

$$\|\beta\|^2 \le \lambda_{\min}(G)^{-1}\beta^\top G\beta \le \lambda_{\min}(G)^{-1}\sum_i q_i^2.$$

Moreover $G \succeq G_{-i}$, while G_{-i} and X_i are independent. Thus

$$\begin{aligned}\mathbb{E}_X\|\beta\|^2 &\le \sum_i \mathbb{E}\lambda_{\min}(G_{-i})^{-1}\,\mathbb{E}q_i^2 \\ &\le K^2\sum_i \mathbb{E}\,\mathrm{tr}(G_{-i}^{-1})\,\mathbb{E}(X_i^\top\delta)^4 \le C_d\|\delta\|^4,\end{aligned}$$

where (65) and $\mathbb{E}(X_i^\top\delta)^4 = 3\|\delta\|^4$ give the last line. Integrating over Δ proves (i).

For (ii), conditional centering and independence remove all cross terms. If $v_i = \mathbb{E}[D_i^2 \mid \Delta, X]$ and $A_i = G_{-i}^{-1}$, then

$$\mathbb{E}_D\left[\left\|G^{-1}\sum_i X_i D_i\right\|^2 \,\middle|\, \Delta, X\right] = \sum_i v_i\|G^{-1}X_i\|^2.$$

Sherman–Morrison gives

$$G^{-1}X_i = \frac{A_iX_i}{1+X_i^\top A_iX_i}, \qquad \|G^{-1}X_i\|^2 \le X_i^\top A_i^2 X_i.$$

Since $v_i \le K|X_i^\top\delta|$ and A_i is independent of X_i,

$$\mathbb{E}\left[v_iX_i^\top A_i^2X_i \mid A_i, \Delta=\delta\right] \le C_{d,K}\|\delta\|\,\|A_i\|_{\mathrm{op}}^2.$$

Finally, (65) gives $\mathbb{E}\|A_i\|_{\mathrm{op}}^2 \le \mathbb{E}\,\mathrm{tr}(A_i^2) = O_d(n^{-2})$. Summing over i yields the stronger bound $Cn^{-1}\|\delta\|$; integrating over Δ proves (ii).

For (iii), the same conditional-variance and Sherman–Morrison calculation reduces the ith term to

$$K\|A_i\|_{\mathrm{op}}^2\mathbb{E}_{X_i}\left[\|X_i\|^2\{(X_i^\top\delta)^2 + (X_i^\top\delta)^4\}\right] \le C_{d,K}\|A_i\|_{\mathrm{op}}^2(\|\delta\|^2 + \|\delta\|^4).$$

Summing and using $\mathbb{E}\,\mathrm{tr}(A_i^2) = O_d(n^{-2})$ proves the stated bound. □

Lemma K.3 (Smooth subdensity exhaustion)**.** *If π is a continuous everywhere-positive density on $\mathbb{R}^d$, then there are nonnegative $\varphi_k \in C_c^\infty(\mathbb{R}^d)$ such that*

$$s_k := \varphi_k^2 \le \pi \quad \textit{pointwise}, \qquad m_k := \int s_k(w)\,dw \longrightarrow 1.$$

Consequently $q_k = s_k/m_k$ is a probability density with finite prior Fisher information $J(q_k) = 4m_k^{-1}\int\|\nabla\varphi_k\|^2$.

Proof. Choose nested compact balls $K_k \subset \mathrm{int}\,K_k'$ with $\int_{K_k}\pi \to 1$, and write $f = \sqrt{\pi}$. On K_k', the function f is continuous and has positive minimum. Choose a smooth h_k with $\|h_k - f\|_{\infty,K_k'} < \eta_k$, where $3\eta_k < \min_{K_k'} f$, and a smooth cutoff χ_k equal to one on K_k and supported in K_k'. Then

$$\varphi_k = \chi_k(h_k - 2\eta_k)$$

is nonnegative and satisfies $\varphi_k \le f$ everywhere. Taking η_k sufficiently small makes $\int_{K_k}\varphi_k^2$ arbitrarily close to $\int_{K_k}\pi$, so the sequence can be chosen with $m_k \to 1$. The displayed expression for $J(q_k)$ follows from $q_k = (\varphi_k/\sqrt{m_k})^2$, with the score defined q_k-almost everywhere. □

Lemma K.4 (Full-experiment integrated lower bound)**.** *Suppose the error density f is positive and absolutely continuous, with location score $\psi = -f'/f$ and $I_f = \mathbb{E}\psi(\varepsilon)^2 \in (0, \infty)$. Under the Gaussian design and every continuous everywhere-positive prior π,*

$$\liminf_{N\to\infty} N R_{\mathrm{full,N}}(\pi) \geq \frac{d}{I_f}.$$

Proof. First let $q(w) = \varphi(w)^2$ be a normalized smooth compactly supported density and put

$$J(q) = \int \|\nabla \log q(w)\|^2 q(w)\, dw = 4 \int \|\nabla \varphi(w)\|^2\, dw < \infty.$$

For one observation (X, Y), the score for w is

$$\dot{\ell}_w(X, Y) = X\psi(Y - X^\top w), \qquad \mathbb{E}_w \dot{\ell}_w \dot{\ell}_w^\top = I_f I_d.$$

The data score is centered under the stated assumptions. Indeed, $f' = -\psi f \in L^1$ by Cauchy–Schwarz. Absolute continuity then gives finite limits of f at both tails, and integrability of the density forces both limits to be zero. Consequently $\mathbb{E}\psi(\varepsilon) = -\int_{\mathbb{R}} f'(u)\, du = 0$. To justify integration by parts under only absolute continuity, let $p_w(x, y)$ be the one-observation joint density. For each coordinate j it has weak derivative

$$\partial_{w_j} p_w(x, y) = p_w(x, y) x_j \psi(y - x^\top w) \quad \text{a.e.},$$

whose L^1 norm is at most $\mathbb{E}|X_j|\, \mathbb{E}|\psi(\varepsilon)| < \infty$. Hence both terms in the weak derivative of the N-sample joint density $p_w^{(N)}(z) q(w)$ are integrable. Fubini and coordinatewise integration by parts have no boundary term because q is compactly supported. For every estimator T_N of finite Bayes risk, with $S_N = \sum_i \dot{\ell}_w(X_i, Y_i)$, truncation and Cauchy–Schwarz therefore give

$$d = \mathbb{E}_q[(T_N - W)^\top \{S_N + \nabla \log q(W)\}].$$

The prior score is understood q-almost everywhere and extended by zero in q-weighted expressions. Since the data score is conditionally centered,

$$\mathbb{E}_q \|S_N + \nabla \log q(W)\|^2 = N d I_f + J(q).$$

Thus the vector van Trees inequality gives

$$R_{\mathrm{full,N}}(q) \geq \frac{d^2}{N d I_f + J(q)} \tag{66}$$

(Gill & Levit, 1995).

Apply Lemma K.3 to write $s_k = m_k q_k \leq \pi$ with $m_k \to 1$. For every estimator,

$$\int \mathbb{E}_w \|T_N - w\|^2 \pi(w)\, dw \geq m_k \mathbb{E}_{q_k} \|T_N - W\|^2.$$

Taking the infimum, then first $N \to \infty$ and afterward $k \to \infty$ in (66), proves the claim. □

Proof of Theorem F.4. All integrations by parts below are absolutely justified. Specifically, $f' = -\psi f \in L^1$ and $f(\pm\infty) = 0$ as in Lemma K.4, while

$$\int |\psi' | f \leq \{\mathbb{E}\psi'(\varepsilon)^2\}^{1/2}, \qquad \int |\psi f'| = I_f, \qquad \int |u f'(u)|\, du = \mathbb{E}|\varepsilon \psi(\varepsilon)| \leq \sigma \sqrt{I_f}.$$

Together with the two displayed tail conditions in the theorem, the three boundary terms are therefore respectively $[f]_{-\infty}^{\infty}$, $[\psi f]_{-\infty}^{\infty}$, and $[u f(u)]_{-\infty}^{\infty}$, all zero. Thus

$$\mathbb{E}\psi(\varepsilon) = 0, \qquad \mathbb{E}\psi'(\varepsilon) = I_f, \qquad \mathbb{E}[\varepsilon \psi(\varepsilon)] = 1.$$

Let L be a global Lipschitz constant for ψ', and put $m(a) = \mathbb{E}\psi(\varepsilon - a)$. Differentiation under the expectation is valid because $|\psi'(u - a)| \leq |\psi'(u)| + L|a|$. Hence $m(0) = 0$, $m'(0) = -I_f$, and

$$|m(a) + I_f a| \leq \frac{L}{2} a^2. \tag{67}$$

Moreover, the fundamental theorem of calculus gives

$$\mathbb{E}\{\psi(\varepsilon - a) - \psi(\varepsilon)\}^2 \le 2a^2\mathbb{E}\psi'(\varepsilon)^2 + \frac{L^2}{2}a^4. \tag{68}$$

Split the sample into $n_0 = \lceil N^{2/3} \rceil$ and $n_1 = N - n_0$ observations. Let $\widetilde{w}$ be OLS on the first split, $\Delta = \widetilde{w} - w$, and $G_1 = \sum_{i \in I_1} X_i X_i^\top$. The Wishart trace identities in (65) and the finite fourth error moment give

$$\mathbb{E}\|\Delta\|^2 = O(n_0^{-1}), \qquad \mathbb{E}\|\Delta\|^4 = O(n_0^{-2}).$$

Define the score one-step estimator

$$\widehat{w}_{\text{score}} = \widetilde{w} + I_f^{-1} G_1^{-1} \sum_{i \in I_1} X_i \psi(Y_i - X_i^\top \widetilde{w}). \tag{69}$$

As before, define it arbitrarily on the finitely many small-N or singular design events.

For $a_i = X_i^\top \Delta$, write

$$r(a) = m(a) + I_f a, \qquad D_i = \psi(\varepsilon_i - a_i) - m(a_i) - \psi(\varepsilon_i).$$

The D_i are conditionally independent and centered. Equations (67)–(68) and Lemma K.2(i),(iii) give

$$\mathbb{E}\left\| G_1^{-1} \sum_i X_i r(a_i) \right\|^2 \le C\mathbb{E}\|\Delta\|^4 = O(N^{-4/3}),$$

$$\mathbb{E}\left\| G_1^{-1} \sum_i X_i D_i \right\|^2 \le \frac{C}{n_1}\mathbb{E}(\|\Delta\|^2 + \|\Delta\|^4) = O(N^{-5/3}).$$

The linear term $-I_f a_i$ cancels Δ exactly, so

$$\widehat{w}_{\text{score}} - w = I_f^{-1} G_1^{-1} \sum_{i \in I_1} X_i \psi(\varepsilon_i) + o_{L^2}(N^{-1/2}).$$

The leading term has conditional covariance $I_f^{-1} G_1^{-1}$. Wishart moments and Cauchy–Schwarz for all cross terms therefore yield, at every w,

$$\mathbb{E}_w\|\widehat{w}_{\text{score}} - w\|^2 = \frac{d}{I_f(n_1 - d - 1)} + o(N^{-1}) = \frac{d}{I_f N} + o(N^{-1}).$$

The distribution of $\widehat{w}_{\text{score}} - w$ does not depend on w; hence the remainder is uniform in the location parameter and the same upper bound holds after integration under any stated prior π. Together with Lemma K.4, this proves the full-risk equality. Combining it with Theorem F.2 proves (23).

Finally, the third information identity and Cauchy–Schwarz give $1 \le \sigma\sqrt{I_f}$. Equality holds exactly when $\psi(\varepsilon) = \varepsilon/\sigma^2$ almost surely. Positivity of f then implies $-f'/f = x/\sigma^2$ almost everywhere, whose normalized solution is the centered Gaussian density. The converse is immediate. □

K.4 Pairwise reverse deficiency

Proof of Theorem F.5. The augmented Gram is a deterministic measurable function of the raw prompt, so the corresponding deterministic Markov kernel reproduces its law exactly:

$$\delta(\mathcal{E}_N^{\text{full}}, \mathcal{E}_N^{\text{Gram}}) = 0.$$

We first compute the full-prompt pairwise limit. Put $k = h_1 - h_0$ and $\ell = \log f$. Under the theorem assumptions, versions may be chosen so that $\ell' = -\psi$, $\ell'' = -\psi'$, and, if L is a global Lipschitz constant for ψ',

$$\ell(u - a) - \ell(u) = a\psi(u) - \frac{a^2}{2}\psi'(u) + r(u, a), \qquad |r(u, a)| \le \frac{L}{6}|a|^3. \tag{70}$$

The boundary assumptions and integration by parts give

$$\mathbb{E}\psi(\varepsilon) = 0, \qquad \mathbb{E}\psi'(\varepsilon) = I_f.$$

Under P^{full}_{N,h_0}, the log likelihood ratio of h_1 against h_0 has the same distribution as

$$\Lambda_N(k) = k^\top \Delta_N - \frac{1}{2} k^\top J_N k + R_N(k), \tag{71}$$

under the centered model, where

$$\Delta_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \psi(\varepsilon_i), \qquad J_N = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top \psi'(\varepsilon_i).$$

Equation (70) yields

$$|R_N(k)| \le \frac{L\|k\|^3}{6N^{3/2}} \sum_{i=1}^{N} \|X_i\|^3 = o_P(1).$$

In fact $\mathbb{E}|R_N(k)| = O(N^{-1/2})$. The multivariate CLT and law of large numbers therefore give, with $\rho^2 = I_f \|k\|^2$,

$$\Lambda_N(k) \Rightarrow \mathcal{N}(-\rho^2/2, \rho^2). \tag{72}$$

Let $L_N = dP^{\text{full}}_{N,h_1}/dP^{\text{full}}_{N,h_0}$ and $L = \exp\{\rho Z - \rho^2/2\}$ for $Z \sim \mathcal{N}(0,1)$. Then $L_N \Rightarrow L$, while $\mathbb{E}L_N = \mathbb{E}L = 1$. For every M,

$$\mathbb{E}(L_N - M)_+ = 1 - \mathbb{E}(L_N \wedge M) \longrightarrow 1 - \mathbb{E}(L \wedge M) = \mathbb{E}(L - M)_+,$$

Given $\eta > 0$, first choose M_0 so that $\mathbb{E}(L - M_0)_+ < \eta/2$. The preceding convergence gives N_0 such that $\sup_{N \ge N_0} \mathbb{E}(L_N - M_0)_+ < \eta$. Increasing the threshold if necessary handles the finitely many $N < N_0$, and monotonicity preserves the bound for $N \ge N_0$. Thus (L_N) is uniformly integrable. Hence

$$\left\| P^{\text{full}}_{N,h_1} - P^{\text{full}}_{N,h_0} \right\|_{\text{TV}} = \frac{1}{2} \mathbb{E}_{h_0} |L_N - 1| \tag{73}$$

$$\longrightarrow \left\| \mathcal{N}(h_1, I_f^{-1} I_d) - \mathcal{N}(h_0, I_f^{-1} I_d) \right\|_{\text{TV}}. \tag{74}$$

For equal-covariance Gaussian shifts,

$$\left\| \mathcal{N}(m_0, s^2 I_d) - \mathcal{N}(m_1, s^2 I_d) \right\|_{\text{TV}} = 2\Phi\left(\frac{\|m_1 - m_0\|}{2s} \right) - 1. \tag{75}$$

This proves the first pairwise limit.

For the Gram experiment, use the same invertible transformation centered at the fixed w_0 for both h_0 and h_1. Together with the compact-uniform total-variation limit (58), this reduces the parameter-bearing block to

$$Z_B = h + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_d),$$

with independent parameter-free Gaussian ancillary noise. Total variation is invariant under measurable bijections and unchanged by adjoining a common independent ancillary. If Q_h denotes this transformed limit law and $e_{N,h}$ the corresponding total-variation approximation error, then

$$\left| \|P^{\text{Gram}}_{N,h_1} - P^{\text{Gram}}_{N,h_0}\|_{\text{TV}} - \|Q_{h_1} - Q_{h_0}\|_{\text{TV}} \right| \le e_{N,h_1} + e_{N,h_0} \longrightarrow 0.$$

The triangle inequality and (75) therefore prove the second pairwise limit.

Finally, for every Markov kernel K from the Gram sample space to the full sample space, contraction of total variation and the triangle inequality give

$$\left\| P^{\text{full}}_{N,h_1} - P^{\text{full}}_{N,h_0} \right\|_{\text{TV}} \le 2 \max_{j \in \{0,1\}} \left\| K P^{\text{Gram}}_{N,h_j} - P^{\text{full}}_{N,h_j} \right\|_{\text{TV}} + \left\| P^{\text{Gram}}_{N,h_1} - P^{\text{Gram}}_{N,h_0} \right\|_{\text{TV}}.$$

Taking the infimum over K, then the limit inferior, proves (26).

The information identity $\mathbb{E}[\varepsilon\psi(\varepsilon)] = 1$ and Cauchy–Schwarz yield $\sigma^2 I_f \ge 1$. Equality forces $\psi(\varepsilon) = \varepsilon/\sigma^2$ almost surely; positivity of f then gives $f'/f = -x/\sigma^2$ almost everywhere, whose unique normalized solution is the centered Gaussian density. Thus the deficiency floor is strictly positive exactly in the non-Gaussian case covered by the theorem. □

K.5 Le Cam limit of the induced experiments

Lemma K.5 (Compact upgrade for LAN families). *Let $\mathcal{E}_N = \{P_{N,h} : h \in H\}$ and $\mathcal{G} = \{Q_h : h \in H\}$ be experiments on Polish sample spaces, indexed by a compact H. Suppose (i) for every finite $H_0 \subset H$ the restricted experiment converges, $\Delta(\mathcal{E}_N|_{H_0}, \mathcal{G}|_{H_0}) \to 0$; and (ii) for some $C < \infty$,*

$$\sup_N \|P_{N,h} - P_{N,g}\|_{\mathrm{TV}} \le C\|h-g\|, \qquad \|Q_h - Q_g\|_{\mathrm{TV}} \le C\|h-g\|,$$

for all $h, g \in H$. Then $\Delta(\mathcal{E}_N, \mathcal{G}) \to 0$.

Proof of Lemma K.5. For $\eta > 0$ take a finite η-net $H_\eta \subset H$ and let $\pi_\eta : H \to H_\eta$ map each h to a nearest net point. Compose each experiment with π_η: write $\mathcal{E}_N \circ \pi_\eta$ for the experiment whose law at h is $P_{N,\pi_\eta(h)}$, and similarly for $\mathcal{G} \circ \pi_\eta$. The identity kernel gives $\Delta(\mathcal{E}_N, \mathcal{E}_N \circ \pi_\eta) \le C\eta$ and $\Delta(\mathcal{G} \circ \pi_\eta, \mathcal{G}) \le C\eta$ by (ii). The triangle inequality for Δ yields

$$\Delta(\mathcal{E}_N, \mathcal{G}) \le 2C\eta + \Delta(\mathcal{E}_N \circ \pi_\eta, \mathcal{G} \circ \pi_\eta).$$

The last term equals $\Delta(\mathcal{E}_N|_{H_\eta}, \mathcal{G}|_{H_\eta})$ because π_η collapses each fiber to its net point; by (i) it tends to 0. Hence $\limsup_N \Delta(\mathcal{E}_N, \mathcal{G}) \le 2C\eta$, and $\eta \downarrow 0$ proves the claim. □

Proof of Theorem F.6. We argue the two convergences separately, then transfer them to the deficiency.

Full experiment. By (71), under $P^{\mathrm{full}}_{N,h_0}$ the log likelihood ratio of $h = h_0 + k$ against h_0 satisfies, for the centered model,

$$\Lambda_N(k) = k^\top \Delta_N - \frac{1}{2} k^\top J_N k + R_N(k), \qquad |R_N(k)| \le \frac{L\|k\|^3}{6N^{3/2}} \sum_{i=1}^N \|X_i\|^3.$$

For k confined to the compact set $H - H$, the factor $\|k\|^3$ is bounded and $\frac{1}{N}\sum_i \|X_i\|^3 = O_P(1)$, so $\sup_{k \in H-H} |R_N(k)| = o_P(1)$. Together with $J_N \to I_f I_d$, the finite-dimensional likelihood-ratio vectors converge. Fixing h_0 as dominating base, each limiting likelihood ratio has expectation one, so the uniform-integrability argument of (72) applies coordinatewise; the standard finite-experiment criterion—convergence of likelihood-ratio vectors with uniformly integrable limits of expectation one—then gives convergence in Le Cam distance on every finite restriction, $\Delta(\mathcal{E}^{\mathrm{full}}_N|_{H_0}, \mathcal{G}^{\mathrm{full}}|_{H_0}) \to 0$ for every finite $H_0 \subset H$ (Le Cam & Yang, 2000). The centering at the fixed w_0 is exact and $\mathrm{Var}(\Delta_N) \to I_f I_d$.

It remains to verify the uniform equicontinuity (ii). Let $q = \sqrt{f}$. Then $q' = -\frac{1}{2}\psi q$, so $\|q'\|_2^2 = I_f/4$. For the single observation (X_i, Y_i) with density $p_h(x,y) = \varphi(x) f(y - x^\top h/\sqrt{N})$,

$$\int \left(\sqrt{p_h} - \sqrt{p_g}\right)^2 = \int \varphi(x) \left\| q(\cdot - x^\top h/\sqrt{N}) - q(\cdot - x^\top g/\sqrt{N}) \right\|_2^2 dx \le \frac{I_f}{4N}\|h-g\|^2,$$

using $\int \varphi(x)\, xx^\top\, dx = I_d$ and the Minkowski estimate $\|q(\cdot - a) - q(\cdot - b)\|_2 \le \|q'\|_2 |a-b|$. Summing the independent observations gives the product Hellinger bound $\|P^{\mathrm{full}}_{N,h} - P^{\mathrm{full}}_{N,g}\|_{\mathrm{TV}} \le H_N(h,g) \le (\sqrt{I_f}/2)\|h-g\|$, uniformly in N. Here we use throughout the unnormalized convention $H^2(P,Q) = \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$, for which $\|P-Q\|_{\mathrm{TV}} \le H(P,Q)$; under the normalized convention the same final constant results from $\|P - Q\|_{\mathrm{TV}} \le \sqrt{2}\, H$. The Gaussian limit satisfies the same bound by (75). Lemma K.5 therefore applies and gives

$$\Delta(\mathcal{E}^{\mathrm{full}}_N, \mathcal{G}^{\mathrm{full}}) \to 0.$$

Gram experiment. Equation (58) states exactly that, for every compact H,

$$e_N := \sup_{h \in H} \left\| \mathcal{L}_{w_0 + h/\sqrt{N}}(Z_N^{(h)}) - \mathcal{N}(Jh, \Sigma) \right\|_{\mathrm{TV}} \longrightarrow 0,$$

where the parameter-bearing block of the limit is the shift $Z_B = h + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$, of (61), and the remaining blocks are parameter-free. Both laws live on a common Euclidean sample space, so the identity kernel in either direction gives $\delta(\mathcal{E}^Z_N, \mathcal{Q}) \le e_N$ and $\delta(\mathcal{Q}, \mathcal{E}^Z_N) \le e_N$, hence $\Delta(\mathcal{E}^Z_N, \mathcal{Q}) \le e_N$.

The original Gram statistic and $Z_N^{(h)}$ are related by one fixed, parameter-uniform measurable bijection, so their experiments are equivalent; and $\mathcal{Q}$ equals $\mathcal{G}^{\mathrm{Gram}}$ up to a parameter-free independent Gaussian ancillary, which a projection kernel removes and an independent-randomization kernel regenerates. Therefore

$$\Delta(\mathcal{E}_N^{\mathrm{Gram}}, \mathcal{G}^{\mathrm{Gram}}) \to 0.$$

Transfer to deficiency. For any two experiments $\mathcal{E}, \mathcal{F}$ on Polish sample spaces, Le Cam's deficiency satisfies the continuity estimate

$$|\delta(\mathcal{E}, \mathcal{F}) - \delta(\mathcal{E}', \mathcal{F}')| \le \Delta(\mathcal{E}, \mathcal{E}') + \Delta(\mathcal{F}, \mathcal{F}').$$

Indeed, the deficiency $\delta(\mathcal{E}, \mathcal{F})$ is the error of the best Markov-kernel simulation of $\mathcal{F}$ from $\mathcal{E}$; composing an approximately optimal kernel for $(\mathcal{E}, \mathcal{F})$ with the kernels realizing $\Delta(\mathcal{E}', \mathcal{E})$ and $\Delta(\mathcal{F}, \mathcal{F}')$ changes the simulation error by at most the two Le Cam distances. Applying this with $\mathcal{E} = \mathcal{E}_N^{\mathrm{Gram}}$, $\mathcal{F} = \mathcal{E}_N^{\mathrm{full}}$ and the two limit experiments gives

$$\left|\delta(\mathcal{E}_N^{\mathrm{Gram}}, \mathcal{E}_N^{\mathrm{full}}) - \delta(\mathcal{G}^{\mathrm{Gram}}, \mathcal{G}^{\mathrm{full}})\right| \le \Delta(\mathcal{E}_N^{\mathrm{Gram}}, \mathcal{G}^{\mathrm{Gram}}) + \Delta(\mathcal{E}_N^{\mathrm{full}}, \mathcal{G}^{\mathrm{full}}) \to 0,$$

proving the convergence of the reverse deficiency.

Strict positivity. Deficiency is bounded below by the testing distance between the two limit experiments. For any direction $k \in H - H$, the two experiments $\mathcal{G}^{\mathrm{full}}$ and $\mathcal{G}^{\mathrm{Gram}}$ restricted to the binary alternative $\{h_0, h_0 + k\}$ have total-variation separation

$$2\Phi\left(\frac{\sqrt{I_f}\,\|k\|}{2}\right) - 1 \quad \text{versus} \quad 2\Phi\left(\frac{\|k\|}{2\sigma}\right) - 1,$$

respectively, by (75). Because $\sigma^2 I_f \ge 1$, the first is at least the second, with strict inequality unless $k = 0$ or the noise is Gaussian. Since $\delta(\mathcal{G}^{\mathrm{Gram}}, \mathcal{G}^{\mathrm{full}})$ is at least the supremum over directions of half the excess testing separation (positive deficiency certifies that some bounded decision problem witnesses a positive comparison gap; zero deficiency is the statement of weak dominance for all bounded decision problems), we obtain the displayed lower bound, which is strictly positive for non-Gaussian f because H has nonempty interior and $\sigma^2 I_f > 1$ in that case. This proves the theorem. □

K.6 Inversion-free source-score correction

Proof of Theorem F.7. We first prove the risk claim. Put

$$H_N = \frac{1}{N}\sum_i X_i X_i^\top, \qquad g_N = \frac{1}{N}\sum_i X_i \varepsilon_i, \qquad \Delta = u_N - W = (H_N - I_d)W + g_N.$$

Fixed-dimensional Gaussian moment bounds and a conditional fourth-moment expansion of g_N give

$$\mathbb{E}\|\Delta\|^2 = O(N^{-1}), \qquad \mathbb{E}\|\Delta\|^4 = O(N^{-2}). \tag{76}$$

The Gaussian prior is used here only through its finite fourth moment.

The theorem assumptions give a $C^{1,1}$ version of ψ. For $a_i = X_i^\top \Delta$, Taylor's theorem therefore gives pointwise, without any independence between a_i and the ith observation,

$$\psi(\varepsilon_i - a_i) = \psi(\varepsilon_i) - \psi'(\varepsilon_i)a_i + \rho_i, \qquad |\rho_i| \le \frac{L}{2}a_i^2,$$

where L is the Lipschitz constant of ψ'. Define

$$S_N = \frac{1}{N}\sum_i X_i \psi(\varepsilon_i), \qquad A_N = \frac{1}{N}\sum_i X_i X_i^\top \psi'(\varepsilon_i), \qquad Q_N = \frac{1}{N}\sum_i X_i \rho_i.$$

Since $\mathbb{E}\psi'(\varepsilon) = I_f$, direct substitution into (27) yields the exact decomposition

$$\widehat{W}_N - W = I_f^{-1} S_N - I_f^{-1}(A_N - I_f I_d)\Delta + I_f^{-1} Q_N. \tag{77}$$

The Lipschitz property implies $|\psi'(x)| \le |\psi'(0)| + L|x|$, so the finite fourth error moment gives $\mathbb{E}|\psi'(\varepsilon)|^4 < \infty$. A fourth-moment bound for the average of the centered i.i.d. matrices $X_i X_i^\top \psi'(\varepsilon_i) - I_f I_d$ therefore gives

$$\mathbb{E}\|A_N - I_f I_d\|_{\mathrm{op}}^4 = O(N^{-2}).$$

Together with (76) and Cauchy–Schwarz, this proves

$$\mathbb{E}\|(A_N - I_f I_d)\Delta\|^2 = O(N^{-2}), \tag{78}$$

without treating the two factors as independent.

It remains to bound the Taylor remainder under the same dependence. Let $M_{3,N} = N^{-1}\sum_i \|X_i\|^3$. Deterministically,

$$\|Q_N\| \le \frac{L}{2} M_{3,N}\|\Delta\|^2.$$

Conditioning on the complete design and expanding the fourth moment of the independent error sum gives

$$\mathbb{E}[\|\Delta\|^4 \mid X] \le C\left\{\|H_N - I_d\|_{\mathrm{op}}^4 + N^{-2}(\operatorname{tr} H_N)^2\right\}.$$

Here the second term follows from

$$\mathbb{E}\left[\left\|\sum_i X_i\varepsilon_i\right\|^4 \,\middle|\, X\right] \le C\left(\sum_i \|X_i\|^2\right)^2,$$

which uses only the fourth error moment. Gaussian design moments satisfy

$$\sup_N \mathbb{E}M_{3,N}^4 < \infty, \qquad \mathbb{E}\|H_N - I_d\|_{\mathrm{op}}^8 = O(N^{-4}), \qquad \sup_N \mathbb{E}[M_{3,N}^2(\operatorname{tr} H_N)^2] < \infty.$$

Consequently,

$$\mathbb{E}\|Q_N\|^2 = O(N^{-2}). \tag{79}$$

The leading term in (77) has mean zero and exact squared norm

$$\mathbb{E}\|I_f^{-1} S_N\|^2 = \frac{d}{I_f N}.$$

Equations (78) and (79) make the total remainder $O(N^{-1})$ in L^2. Its cross term with the $O(N^{-1/2})$ leading term is $O(N^{-3/2})$, and all remainder–remainder terms are $O(N^{-2})$. This proves the coefficient-risk equality. Independence and isotropy of X_q give the prediction-risk equality.

We now give the exact architecture. Write a width-$(d+3)$ state as $e = (x, y, p, a)$, and initialize

$$e_i^{(0)} = (X_i, Y_i, 0, 0), \qquad e_q^{(0)} = (X_q, 0, 0, 0).$$

Let $E_x = [I_d\ 0\ 0\ 0]$, and let e_y, e_p, e_a select the three scalar slots. In the first layer choose

$$V_1 = e_y^\top, \quad K_1 = Q_1 = E_x, \quad O_1 = e_p - e_a, \quad R_1(x, y, p, a) = (x, y, 0, y),$$

with every affine offset equal to zero. The common memory is $G_1 = N^{-1}\sum_i Y_i X_i^\top = u_N^\top$, so every destination becomes

$$e_j^{(1)} = (X_j, Y_j, u_N^\top X_j, Y_j - u_N^\top X_j),$$

where $Y_q = 0$. Thus the first permitted affine attention layer itself preserves (X_j, Y_j), writes the pilot, and forms the residual; no hidden affine sublayer is used.

Apply the sole scalar tokenwise channel

$$\mathcal{A}_\psi(x, y, p, a) = (x, y, p, \psi(a)).$$

Its value on the query is immaterial because the query is not a source and the final readout ignores the scratch slot. In the second layer choose

$$R_2 = I_{d+3}, \quad V_2 = e_a^\top, \quad K_2 = Q_2 = E_x, \quad O_2 = I_f^{-1} e_p,$$

again with zero offsets. Its common memory is

$$G_2 = \frac{1}{N}\sum_i \psi(Y_i - X_i^\top u_N)X_i^\top.$$

The affine readout selecting the query's p slot is therefore

$$e_p^\top e_q^{(2)} = X_q^\top u_N + \frac{1}{I_f N}\sum_i X_q^\top X_i \psi(Y_i - X_i^\top u_N) = X_q^\top \widehat{W}_N.$$

Both memories use exactly the demonstrations as sources. The construction has one head per layer and uses no position, length input, matrix inverse, or nonlinear query readout.

Finally, the scale-s logistic density is $f_s(r) = (4s)^{-1}\operatorname{sech}^2(r/(2s))$, so its score is the displayed scaled $\tanh$. Its variance is $\pi^2 s^2/3$. If U is its cdf evaluated at a logistic draw, then $2U - 1 = \tanh(\varepsilon/(2s))$ is uniform on $[-1, 1]$; hence $I_f = s^{-2}\mathbb{E}(2U-1)^2 = 1/(3s^2)$. □

K.7 Efficiency spectrum and uniqueness of the source channel

Proof of Theorem F.8. The canonical-gain claim is the proof of Theorem F.7 with ψ replaced by ϕ and I_f replaced by $a = \mathbb{E}\phi'(\varepsilon)$ throughout. The decomposition (77) becomes

$$\widehat{W}_N^\phi(1/a) - W = a^{-1}S_N^\phi - a^{-1}(A_N^\phi - aI_d)\Delta + a^{-1}Q_N^\phi, \tag{80}$$

where $S_N^\phi = \frac{1}{N}\sum_i X_i\phi(\varepsilon_i)$, $A_N^\phi = \frac{1}{N}\sum_i X_i X_i^\top \phi'(\varepsilon_i)$, and $Q_N^\phi = \frac{1}{N}\sum_i X_i\rho_i$ with $|\rho_i| \le \frac{L_\phi}{2}(X_i^\top\Delta)^2$. Every bound in that proof uses only the $C^{1,1}$ property of ϕ, $0 < a < \infty$, $\mathbb{E}\phi^2 < \infty$, and the finite fourth error moment, so $\mathbb{E}\|A_N^\phi - aI_d\|_{\mathrm{op}}^4 = O(N^{-2})$ and $\mathbb{E}\|Q_N^\phi\|^2 = O(N^{-2})$ transfer verbatim. The linear pilot term is cancelled by the exact prefactor $(1 - a^{-1}a) = 0$; the pilot error Δ survives only through $(A_N^\phi - aI_d)\Delta$ (linearly) and through the quadratic remainder Q_N^ϕ. The leading term has

$$\mathbb{E}\|a^{-1}S_N^\phi\|^2 = \frac{dv}{a^2 N},$$

and, as in Section K.6, every leading–remainder cross term is $O(N^{-3/2})$ and every remainder–remainder term $O(N^{-2})$. This proves

$$N\,\mathbb{E}\|\widehat{W}_N^\phi(1/a) - W\|^2 \longrightarrow dv/a^2,$$

and isotropy of X_q transfers it to the clean-query prediction risk.

For the inequality and equality case, integration by parts gives the Stein identity $\mathbb{E}[\phi(\varepsilon)\psi(\varepsilon)] = \mathbb{E}\phi'(\varepsilon) = a$: the boundary term vanishes because $\phi f \to 0$ and $f \to 0$ at infinity, and $\mathbb{E}|\phi\psi| \le \sqrt{vI_f} < \infty$ justifies the integrals. Cauchy–Schwarz then yields

$$a^2 = \big(\mathbb{E}[\phi\psi]\big)^2 \le \mathbb{E}\phi^2\,\mathbb{E}\psi^2 = vI_f,$$

so $v/a^2 \ge 1/I_f$, with equality iff ϕ and ψ are linearly dependent in $L^2(f)$, i.e. $\phi \propto \psi$.

For the arbitrary-gain claim, write $t = 1 - \gamma a$ and $h = \gamma\phi + t\,\varepsilon$. The leading coefficient of the estimator $\widehat{W}_N^\phi(\gamma)$ is $dF(\gamma)$ with

$$F(\gamma) = v\gamma^2 + \kappa t^2 + 2b\gamma t, \qquad \kappa = \sigma^2 + \tau^2(d+1),$$

because the three leading second moments are $\mathbb{E}\|S_N^\phi\|^2 = dv/N$, $\mathbb{E}\|\Delta\|^2 = d\kappa/N$, and $\mathbb{E}[S_N^{\phi\top}\Delta] = db/N$ (only the diagonal $i = j$ contributes to the last, and the design term $(H_N - I_d)W$ has zero cross expectation with S_N^ϕ under the centered independent W). Since

$$\mathbb{E}[h\psi] = \gamma\mathbb{E}[\phi\psi] + t\mathbb{E}[\varepsilon\psi] = \gamma a + t = 1,$$

the information inequality and the centered-score identity give

$$F(\gamma) = \mathbb{E}[h^2] + \tau^2(d+1)t^2 \ge \frac{1}{I_f} + \tau^2(d+1)t^2.$$

With $\tau^2 > 0$, equality forces both $t = 0$ and $h = \psi/I_f$ in $L^2(f)$; hence $\gamma a = 1$ and $\gamma\phi = \psi/I_f$, i.e. $\phi \propto \psi$. Conversely a score-proportional channel with $\gamma = 1/a$ attains $1/I_f$. This proves the free-gain uniqueness, including that the canonical gain is not in general the risk-optimal scalar gain (the optimizer of the finite-N quadratic converges to the minimizer of F). □

Proof of Corollary I.2. Put $\phi_k(r) = b^{-1}\tanh(kr)$ and define

$$J_k = \mathbb{E}\phi_k'(\varepsilon), \qquad V_k = \mathbb{E}\phi_k(\varepsilon)^2, \qquad \gamma_k = J_k^{-1}, \qquad \varepsilon \sim \text{Laplace}(0, b).$$

The change of variables $t = kx$ gives

$$J_k = \frac{1}{2b^2}\int_{\mathbb{R}} \operatorname{sech}^2(t) e^{-|t|/(bk)}\, dt,$$

and

$$V_k = \frac{1}{b^2}\left\{1 - \frac{1}{2bk}\int_{\mathbb{R}} \operatorname{sech}^2(t) e^{-|t|/(bk)}\, dt\right\}.$$

Using $1 - e^{-u} \le u$ and $\int_{\mathbb{R}} |t| \operatorname{sech}^2(t)\, dt = 2\log 2$ yields

$$J_k = b^{-2} + O(k^{-1}), \qquad \gamma_k^2 V_k = b^2 + O(k^{-1}). \tag{81}$$

Since $c_k = bJ_k$, the correction in (34) is exactly $\gamma_k N^{-1}\sum_i X_i\phi_k(\cdot)$.

Use the pilot notation $\Delta = u_N - W$ from (76). Taylor expansion gives

$$\phi_k(\varepsilon_i - X_i^\top\Delta) = \phi_k(\varepsilon_i) - \phi_k'(\varepsilon_i)X_i^\top\Delta + \rho_{i,k}, \qquad |\rho_{i,k}| \le C_b k^2 (X_i^\top\Delta)^2,$$

because $\sup_x |\phi_k''(x)| \le C_b k^2$. Let

$$S_{k,N} = \frac{1}{N}\sum_i X_i\phi_k(\varepsilon_i), \quad A_{k,N} = \frac{1}{N}\sum_i X_i X_i^\top \phi_k'(\varepsilon_i),$$

write $B_{k,N} = A_{k,N} - J_k I_d$, and let $Q_{k,N} = N^{-1}\sum_i X_i\rho_{i,k}$. Equation (34) becomes

$$\widehat{W}_N^{\tanh} - W = \gamma_k S_{k,N} - \gamma_k B_{k,N}\Delta + \gamma_k Q_{k,N}. \tag{82}$$

The leading term is centered and, by (81),

$$\mathbb{E}\|\gamma_k S_{k,N}\|^2 = \frac{d\gamma_k^2 V_k}{N} = \frac{db^2}{N} + O((Nk)^{-1}).$$

Moreover,

$$\mathbb{E}|\phi_k'(\varepsilon)|^2 = O(k), \qquad \mathbb{E}|\phi_k'(\varepsilon)|^4 = O(k^3),$$

so the fourth-moment expansion for centered matrix averages gives

$$\mathbb{E}\|B_{k,N}\|_{\text{op}}^4 = O(k^2N^{-2} + k^3N^{-3}).$$

Combining this with (76) by Cauchy–Schwarz yields

$$\mathbb{E}\|B_{k,N}\Delta\|^2 = O(kN^{-2})$$

for $k = o(N)$. Finally, the design-conditioning argument used for (79), with the Taylor constant now of order k^2, gives

$$\mathbb{E}\|Q_{k,N}\|^2 = O(k^4N^{-2}).$$

Hence, when $k = k_N \to \infty$ and $k_N = o(N^{1/4})$, every remainder in (82) is $o(N^{-1/2})$ in L^2. Its cross term with the leading term is $o(N^{-1})$, and all remainder squares and mutual cross terms are also $o(N^{-1})$. This proves the claim. □

Proof of Corollary I.1. For Laplace$(0, b)$ noise, $\sigma^2 = 2b^2$, so (31) follows immediately from Theorem F.2. We first give a direct full-prompt upper bound, so no separate uniform-integrability assertion for LAD is needed, and then match it by a vector van Trees lower bound. Split the sample into I_0, I_1 of sizes $n_0 = \lceil N^{2/3}\rceil$ and $n_1 = N - n_0$. Put $G_j = \sum_{i\in I_j} X_iX_i^\top$, let $\widetilde{w} = G_0^{-1}\sum_{i\in I_0} X_iY_i$ be OLS on I_0, and define the one-step rule

$$\widehat{w}_{\text{one}} = \widetilde{w} + bG_1^{-1}\sum_{i\in I_1} X_i \operatorname{sgn}(Y_i - X_i^\top\widetilde{w}). \tag{83}$$

Take $\operatorname{sgn}(0) = 0$. On a singular G_0 or G_1 event, set the output to zero; under the Gaussian design this event has probability zero once the corresponding split exceeds d. For the finitely many N with

n_0 or $n_1 \le d+9$, define the rule arbitrarily. None of these conventions affects the asymptotic claim. Write $\Delta = \widetilde{w} - w$. A conditional fourth-moment expansion and (65) give

$$\mathbb{E}\|\Delta\|^4 = O(n_0^{-2}). \tag{84}$$

Indeed, conditional on the pilot design, expanding the fourth moment of $\Delta = G_0^{-1}\sum_{i\in I_0} X_i\varepsilon_i$ bounds it by a constant times

$$(\operatorname{tr} G_0^{-1})^2 + \sum_i \|G_0^{-1}X_i\|^4 \le 2(\operatorname{tr} G_0^{-1})^2 \le 2d\operatorname{tr}(G_0^{-2}).$$

Taking expectations and using (65) proves the stated order. For $\varepsilon \sim \mathrm{Laplace}(0,b)$,

$$m(a) := \mathbb{E}\operatorname{sgn}(\varepsilon - a) = -\operatorname{sgn}(a)(1-e^{-|a|/b}) = -a/b + r(a), \qquad |r(a)| \le \frac{a^2}{2b^2}.$$

Conditional on (Δ, X_i), the centered sign increment

$$D_i = \operatorname{sgn}(\varepsilon_i - X_i^\top\Delta) - m(X_i^\top\Delta) - \operatorname{sgn}(\varepsilon_i)$$

has mean zero and second moment at most $2\min\{|X_i^\top\Delta|/b, 1\}$. These bounds follow directly by separating the event on which shifting the Laplace residual by $X_i^\top\Delta$ changes its sign. Substituting these two displays into (83), and conditioning first on the pilot, yields

$$\widehat{w}_{\mathrm{one}} - w = bG_1^{-1}\sum_{i\in I_1} X_i\operatorname{sgn}(\varepsilon_i) + o_{L^2}(N^{-1/2}). \tag{85}$$

For completeness, the two remainders are $bG_1^{-1}\sum_i X_i r(X_i^\top\Delta)$ and $bG_1^{-1}\sum_i X_i D_i$. Since the second-split Gaussian design is independent of Δ, Lemma K.2 gives

$$\mathbb{E}\left\|G_1^{-1}\sum_i X_i r(X_i^\top\Delta)\right\|^2 \le C\mathbb{E}\|\Delta\|^4, \qquad \mathbb{E}\left\|G_1^{-1}\sum_i X_i D_i\right\|^2 \le \frac{C}{n_1}\mathbb{E}\{\|\Delta\| + \|\Delta\|^2\}.$$

Together with (84), these are respectively $O(n_0^{-2})$ and $O((n_1\sqrt{n_0})^{-1})$, both $o(N^{-1})$; this verifies (85) including the design–inverse dependence. The leading term has squared L^2 norm $O(N^{-1})$, so Cauchy–Schwarz also makes every leading–remainder and remainder–remainder cross term $o(N^{-1})$.

The leading term in (85) has conditional covariance $b^2G_1^{-1}$, because $\operatorname{sgn}(\varepsilon_i)$ has variance one. The Wishart identity therefore gives

$$\mathbb{E}_w\|\widehat{w}_{\mathrm{one}} - w\|_2^2 = \frac{db^2}{n_1 - d - 1} + o(N^{-1}) = \frac{db^2}{N} + o(N^{-1}).$$

The law of $\widehat{w}_{\mathrm{one}} - w$ is independent of w, so this upper bound also integrates unchanged under π. The leading covariance agrees with the classical asymptotic-distribution theory for L_1 regression (Knight, 1998); the preceding argument separately proves the required L^2 risk upper bound.

Laplace density is positive and absolutely continuous, with location Fisher information $I_f = b^{-2}$. Thus Lemma K.4 matches the one-step upper and proves (32); subtracting it from (31) proves (33). □

Lemma K.6 (Almost-everywhere factorization implies fiber stationarity)**.** *Let $U \subset \mathbb{R}^n$ be open, $T : U \to \mathbb{R}^k$ be C^1, and $h \in C^1(U)$. If $\operatorname{rank} DT(x_0) = k$ and $h = g \circ T$ Lebesgue-a.e. in a neighborhood of x_0 for a Borel g, then $\nabla h(x_0) \in \operatorname{row} DT(x_0)$.*

Proof. In constant-rank coordinates (s,t), s is a reparameterization of T and t is a fiber coordinate. Local diffeomorphisms preserve null sets. Fubini therefore implies that for almost every s, the continuous function $t \mapsto h(s,t)$ equals one constant for almost every, hence every, t. Thus $\partial_t h = 0$ for a dense set of s. Continuity of the derivative makes it zero throughout the chart. Transforming back proves the claim. □

K.8 Classical Gaussian characterization

Self-contained proof of Theorem H.1. The Gaussian converse is immediate. For the forward direction, let $\ell = \log f$ and, for fixed $a \in \mathbb{R}$, set

$$r_a(x) = \ell(x-a) - \ell(x), \qquad H_a(y) = \sum_{i=1}^{3} r_a(y_i).$$

Every location density is equivalent to Lebesgue measure. The likelihood- ratio criterion for sufficiency gives $H_a(y) = G_a(T_3(y))$ Lebesgue-a.e. for a Borel G_a.

The two rows of DT_3 are $(1,1,1)$ and $2(y_1,y_2,y_3)$, so the rank is two away from $y_1 = y_2 = y_3$. By Lemma K.6, for every distinct triple,

$$\det \begin{pmatrix} 1 & y_1 & r_a'(y_1) \\ 1 & y_2 & r_a'(y_2) \\ 1 & y_3 & r_a'(y_3) \end{pmatrix} = 0.$$

Fixing two points and varying the third shows that r_a' is affine, hence r_a is a polynomial of degree at most two. For $a \neq 0$, the quadratic polynomials r_a/a converge pointwise as $a \to 0$ to $-\ell'$. Values at three fixed nodes determine their coefficients, so ℓ' is quadratic: $\ell'(x) = Ax^2 + Bx + C$. Thus

$$\ell(x) = \frac{A}{3}x^3 + \frac{B}{2}x^2 + Cx + D.$$

If $A \neq 0$, one tail makes e^{ℓ} nonintegrable. Hence $A = 0$; integrability then forces $B < 0$. Completing the square proves that f is Gaussian. □

Proof of Corollary H.2. If f is not Gaussian, T_3 is not sufficient. The likelihood-ratio criterion gives $\theta_0 \neq \theta_1$ such that $L = p_{\theta_1}/p_{\theta_0}$ is not T_3-measurable. Under the equal prior,

$$m(Y) = \mathbb{E}[\Theta \mid Y] = \frac{\theta_0 + \theta_1 L(Y)}{1 + L(Y)}.$$

This is an invertible strictly monotone function of $L > 0$, so m is not T_3-measurable. It is bounded and square integrable. Applying Lemma F.1 yields a strictly positive gap. □

K.9 Laplace noise with a Gaussian task prior

The detailed analytic proof of Theorem H.3 is given next.

Proof of Theorem H.3. Let $q_b(u) = (2b)^{-1}e^{-|u|/b}$. The posterior mean is

$$m_N(y) = \frac{\int_{\mathbb{R}} \theta \phi_\tau(\theta) \prod_{i=1}^{N} q_b(y_i - \theta)\, d\theta}{\int_{\mathbb{R}} \phi_\tau(\theta) \prod_{i=1}^{N} q_b(y_i - \theta)\, d\theta}, \tag{86}$$

where ϕ_τ is the $\mathcal{N}(0,\tau^2)$ density. The coordinatewise Laplace score is bounded by $1/b$, the posterior has no atom at $\theta = y_i$, and the Gaussian prior has all moments. Differentiation and dominated convergence therefore give $m_N \in C^1$ and

$$\partial_{y_i} m_N(y) = -\frac{1}{b}\operatorname{Cov}\big(\Theta, \operatorname{sgn}(y_i - \Theta) \mid Y = y\big). \tag{87}$$

For $M > 0$, consider

$$y^{(M)} = (0, M, 2M, 0, \ldots, 0).$$

After removing a factor $e^{-3M/b}$, its posterior has unnormalized density

$$w_M(\theta) = \phi_\tau(\theta) \exp\left(\frac{3M - (N-2)|\theta| - |M-\theta| - |2M-\theta|}{b}\right). \tag{88}$$

Pointwise as $M \to \infty$,

$$w_M(\theta) \longrightarrow w_\infty(\theta) := \phi_\tau(\theta) \exp\left(\frac{2\theta - (N-2)|\theta|}{b}\right).$$

Indeed, $3M - |M - \theta| - |2M - \theta| \le 2\theta$, so $w_M \le w_\infty$. The Gaussian factor makes both w_∞ and $|\theta| w_\infty$ integrable. Hence the normalized posteriors and their first moments converge by dominated convergence.

Write $g_i(M) = \partial_{y_i} m_N(y^{(M)})$. For the first zero coordinate, (87) yields

$$g_1(M) \longrightarrow c_N := -\frac{1}{b} \operatorname{Cov}_{\Pi_\infty}\big(\Theta, \operatorname{sgn}(-\Theta)\big) > 0,$$

where Π_∞ is the normalized w_∞ distribution. Strict positivity follows because w_∞ is positive on both half-lines: if $p_\pm$ and $\mu_\pm$ are their probabilities and conditional means, then

$$\operatorname{Cov}_{\Pi_\infty}(\Theta, \operatorname{sgn}(-\Theta)) = 2p_- p_+ (\mu_- - \mu_+) < 0.$$

For the coordinates equal to M and $2M$, the signs in (87) converge pointwise to one. The same domination gives

$$g_2(M) \longrightarrow 0, \qquad g_3(M) \longrightarrow 0.$$

For sufficiently small t, define the exact fiber curve

$$\gamma_M(t) = \left(t, \frac{3M - t - \sqrt{M^2 + 6Mt - 3t^2}}{2}, \frac{3M - t + \sqrt{M^2 + 6Mt - 3t^2}}{2}, 0, \dots, 0 \right). \quad (89)$$

Direct calculation gives

$$\sum_i \gamma_{M,i}(t) = 3M, \qquad \sum_i \gamma_{M,i}(t)^2 = 5M^2,$$

and $\gamma_M'(0) = (1, -2, 1, 0, \dots, 0)$. Therefore

$$\frac{d}{dt} m_N(\gamma_M(t))\Big|_{t=0} = g_1(M) - 2g_2(M) + g_3(M) \longrightarrow c_N > 0.$$

For all sufficiently large finite M, m_N thus has a nonzero derivative along an exact T_N-fiber.

The marginal density of Y is strictly positive everywhere, and DT_N has rank two at $y^{(M)}$. If $m_N(Y)$ were $T_N(Y)$-measurable, the Doob–Dynkin lemma would give $m_N = h \circ T_N$ almost surely for a Borel h. Positivity of the density converts this to Lebesgue-a.e. equality. Then Lemma K.6 would force every fiber derivative at $y^{(M)}$ to vanish, a contradiction. Hence m_N is not T_N-measurable.

Finally, $\Theta \in L^2$ implies $m_N(Y) \in L^2$ by Jensen. Applying Lemma F.1 with $\mathcal{F} = \sigma(Y)$ and $\mathcal{G} = \sigma(T_N(Y))$ gives (30). □

K.10 Full-rank Gaussian designs

Proof of Theorem G.1. For fixed (X, y), the posterior density of W is proportional to

$$\phi_{\tau,d}(w) \exp\big(-b^{-1}\|y - Xw\|_1\big). \quad (90)$$

The likelihood factor is bounded by one and the Gaussian prior has all moments. Hence, for $v \in \mathbb{R}^N$, differentiation under the integral gives

$$D_v m(X, y) = -\frac{1}{b} \operatorname{Cov}\left(W, \sum_{i=1}^N v_i \operatorname{sgn}(y_i - X_i^\top W) \,\middle|\, X, y \right). \quad (91)$$

Residual hyperplanes have posterior measure zero whenever $X_i \neq 0$. Dominated convergence in (90) and (91) also shows that $(X, y, v) \mapsto D_v m(X, y)$ is continuous locally around every point used below.

Let $y^{(M)} = (0, M, 2M, 0, \dots, 0)^\top$ and $v = (1, -2, 1, 0, \dots, 0)^\top$. At the degenerate design

$$X^{(0)} = \mathbf{1} e_1^\top,$$

the posterior of W_1 is exactly the scalar posterior in Section K.9, while all other coordinates retain their centered prior. The limits established there imply that, for some sufficiently large finite M,

$$D_v m_1(X^{(0)}, y^{(M)}) \neq 0. \quad (92)$$

Moreover, $\mathbf{1}^\top v = (y^{(M)})^\top v = 0$.

Since $N \geq d+3$, the space $v^\perp$ has dimension at least $d+2$. Choose $q_1, \ldots, q_d \in v^\perp$ so that

$$\{\mathbf{1}, q_1, \ldots, q_d, y^{(M)}\}$$

is linearly independent, and define

$$X^{(\epsilon)} = [\,\mathbf{1} + \epsilon q_1, \epsilon q_2, \ldots, \epsilon q_d\,].$$

For every $\epsilon \neq 0$, this design has column rank d, $[\mathbf{1}, X^{(\epsilon)}, y^{(M)}]$ has rank $d+2$, and $(X^{(\epsilon)})^\top v = 0$. By continuity from (92), fix a sufficiently small $\epsilon > 0$ such that, with $X^\star = X^{(\epsilon)}$,

$$D_v m_1(X^\star, y^{(M)}) \neq 0. \tag{93}$$

The witness persists on an open set of designs. For X near $X^\star$, set

$$A_X = [\,\mathbf{1}, X, y^{(M)}\,], \qquad P_X = I - A_X(A_X^\top A_X)^{-1}A_X^\top, \qquad v_X = P_X v.$$

The rank remains $d+2$, so P_X and v_X are continuous. At $X^\star$, $v_X = v$. After shrinking to a nonempty open neighborhood $\mathcal{U}$,

$$\mathbf{1}^\top v_X = X^\top v_X = (y^{(M)})^\top v_X = 0, \qquad D_{v_X} m_1(X, y^{(M)}) \neq 0 \tag{94}$$

for every $X \in \mathcal{U}$.

Suppose now that $m_1(X, Y)$ were measurable with respect to $(X, T_X(Y))$. The joint density of (X, Y) is strictly positive everywhere:

$$\left(\prod_i \phi_d(X_i)\right) \int_{\mathbb{R}^d} \phi_{\tau,d}(w) \prod_i q_b(y_i - X_i^\top w)\, dw > 0.$$

Thus the corresponding factorization holds Lebesgue-a.e. in (X, y). Fubini gives, for almost every fixed X,

$$m_1(X, y) = h_X(T_X(y)) \quad \text{for Lebesgue-a.e. } y.$$

Choose such an X in the open set $\mathcal{U}$. The Jacobian of T_X at $y^{(M)}$ has rows $\mathbf{1}^\top, X^\top, 2(y^{(M)})^\top$ and rank $d+2$. By Lemma K.6, every derivative tangent to this fiber must vanish. This contradicts (94). Hence $m(X, Y)$ is not measurable with respect to $(X, T_X(Y))$. Since $\widetilde{S}$ is a function of this stronger oracle statistic, it is not $\widetilde{S}$-measurable either.

Finally, conditional Jensen gives $m \in L^2$. With $\bar{m} = \mathbb{E}[m \mid \widetilde{S}]$ and an independent $X_q \sim \mathcal{N}(0, I_d)$, the full and Gram Bayes predictors are $X_q^\top m$ and $X_q^\top \bar{m}$. Therefore

$$R^*_{\widetilde{S}} - R^*_{\text{full}} = \mathbb{E}\big(X_q^\top (m - \bar{m})\big)^2 = \mathbb{E}\|m - \bar{m}\|_2^2 > 0,$$

where strictness follows from nonmeasurability. Independent mean-zero query noise adds the same irreducible risk to both predictors. □

Remark K.7. The condition $N \geq d+3$ makes the regular fixed-design label fiber have dimension $N - d - 2 \geq 1$. It is sufficient for this differential construction, not asserted necessary. The proof is qualitative: it establishes a strict population gap on a genuinely full-rank i.i.d. Gaussian design, but not a useful closed-form lower bound or its scaling in (N, d).

K.11 Additional finite-sample separations

Proposition K.8 (Bounded arbitrary-fraction separation)**.** *Let $\epsilon = k/N < 1/2$, $B > 0$, $x_i = x_q = 1$, and $\Theta \in \{+\epsilon B, -\epsilon B\}$ be equiprobable. Randomly permute $N - k$ clean labels equal to Θ and k replaced labels equal to*

$$-\frac{1-\epsilon}{\epsilon}\Theta.$$

All labels lie in $[-B, B]$, both signs have the same complete augmented Gram, and every Gram-measurable predictor has MSE at least $\epsilon^2 B^2$. The raw median recovers Θ exactly. Moreover, a biased width-two ReLU applied tokenwise, followed by one common-source mean-attention layer and a linear readout, also recovers Θ exactly.

Proof. The two sign classes satisfy

$$\frac{1}{N}\sum_i y_i = 0, \qquad \frac{1}{N}\sum_i y_i^2 = \epsilon(1-\epsilon)B^2.$$

They are therefore indistinguishable from the augmented Gram, and the optimal common prediction for the symmetric two-point prior is zero. Since a strict majority of labels equal Θ, the median is exact. Put $\gamma = \epsilon B$ and define

$$s(y) = \frac{\mathrm{ReLU}(y+\gamma) - \mathrm{ReLU}(y-\gamma)}{\gamma} - 1.$$

On the finite support, $s(y) = \mathrm{sgn}(y)$ for every demonstration and $s(0) = 0$ in the query label slot. Constant-key mean attention returns $(1-2\epsilon)\,\mathrm{sgn}(\Theta)$; multiplying by $\epsilon B/(1-2\epsilon)$ recovers Θ. □

Theorem K.9 (Full-dimensional bounded strong contamination)**.** *Let $p = d+1$, $\phi = (x, 1)$, $\|\phi\| \le L$, and $\lambda_{\min}(\mathbb{E}\phi\phi^\top) \ge \lambda > 0$, with a continuous design. Fix a replaced-label set C of size $k \ge p$, let G have size $m = N - k$, and draw Θ uniformly from $\{+\beta, -\beta\}$. There is a strong label-replacement strategy with labels bounded by B, provided*

$$\|\beta\| \le \frac{B}{\max\{L, 2mL^3/(k\lambda)\}},$$

such that every augmented-Gram predictor at an independent query with second moment Q has risk at least

$$\big(1 - p\exp\{-k\lambda/(8L^2)\}\big)_+ \beta^\top Q\beta.$$

If $N > 2k + d$, maximum consensus recovers Θ almost surely from the raw prompt. In particular, when $Q \succeq qI$ and the design and contamination constants are fixed, this gives an $\Omega(\epsilon^2 B^2)$ separation at $\epsilon = k/N$.

K.12 Full-dimensional strong contamination

Proof of Theorem K.9. Write Φ_G and Φ_C for the clean and replaced augmented-design matrices and set

$$A = \Phi_G^\top \Phi_G, \qquad C_0 = \Phi_C^\top \Phi_C.$$

The clean labels are $y_G = \Phi_G\Theta$. On the event

$$\mathcal{E} = \{\lambda_{\min}(C_0) \ge k\lambda/2\},$$

the adversary, after observing the design and clean sample, replaces the labels in C by

$$y_C = -\Phi_C C_0^{-1} A\Theta. \tag{95}$$

Off $\mathcal{E}$, it uses any bounded fallback labels, for example zero. This is a strong label-replacement adversary: covariates are unchanged and only the designated k responses are selected after the clean sample is observed.

On $\mathcal{E}$,

$$\Phi^\top y = A\Theta - C_0 C_0^{-1} A\Theta = 0$$

and

$$\|y\|^2 = \Theta^\top\{A + AC_0^{-1}A\}\Theta.$$

The design block is unchanged, the design–label cross moment is zero, and the label second moment is invariant under $\Theta \mapsto -\Theta$. Hence the complete augmented Gram is identical for the two task signs at every fixed design in $\mathcal{E}$.

Because $\phi_i\phi_i^\top \preceq L^2 I_p$ and $\lambda_{\min}(\mathbb{E}[\phi\phi^\top]) \ge \lambda$, matrix Chernoff gives the conservative finite-sample bound (Tropp, 2012)

$$\Pr(\mathcal{E}) \ge 1 - p\exp\{-k\lambda/(8L^2)\}.$$

On this event, clean labels obey $|y_i| \le L\|\beta\|$, while every replaced label obeys

$$|y_i| \le \|\phi_i\|\|C_0^{-1}\|\|A\|\|\beta\| \le \frac{2mL^3}{k\lambda}\|\beta\|.$$

The stated restriction on β therefore bounds all labels by B; the off-event fallback is bounded by construction.

Conditioned on any X in $\mathcal{E}$, the same observed Gram arises under the equiprobable parameters $+\beta$ and $-\beta$. At an independent query ϕ_q the minimum common squared-loss prediction is zero and its conditional risk is $\mathbb{E}(\phi_q^\top \beta)^2 = \beta^\top Q \beta$. Averaging over $\mathcal{E}$ proves the displayed Gram-risk lower bound.

It remains to verify raw recovery. The true parameter fits all m clean rows. By continuous general position, a distinct affine parameter agrees with it on at most d clean rows almost surely, and it can fit at most all k replaced rows. If $m > k + d$, equivalently $N > 2k + d$, the true parameter is the unique maximum-consensus fit, regardless of the bounded fallback used off $\mathcal{E}$.

Finally, at fixed $\epsilon = k/N < 1/2$ and fixed (L, λ, q, d), choose β in any fixed direction with norm a sufficiently small constant times ϵB. The label constraint holds, and $Q \succeq qI$ converts the directional floor into $\Omega(\epsilon^2 B^2)$ whenever the Chernoff event has a fixed positive probability. □

K.13 Generic fixed-length nonlinear scope

Theorem K.10 (Tokenwise nonlinear processing breaks the barrier)**.** *Fix N and d. On every compact prompt set, a model consisting of a tokenwise MLP, common-source mean attention, and a query MLP can uniformly approximate every continuous permutation-invariant predictor. For the Gaussian-prior/Laplace models in Theorems G.1 and H.3, for every $\epsilon > 0$ such a constant-depth model can achieve risk at most $R^*_{\text{full}} + \epsilon$.*

Proof sketch. For scalar tokens, the first N power sums identify the multiset by Newton's identities. More generally, standard DeepSets universality represents a continuous invariant function on a compact set as $\rho(\sum_i \phi(x_i, y_i), x_q)$ to arbitrary accuracy (Zaheer et al., 2017). MLP approximation gives the desired tokenwise map and readout. For the noncompact Gaussian/Laplace models, truncate to a high-probability compact set, truncate the square-integrable posterior mean, and clip the network output. The tail errors can be made arbitrarily small before applying compact approximation. □

K.14 A graded hierarchy for read-once source feature banks

Theorems F.7 and F.8 treat a single scalar source channel. We record here what happens for a *bank* of source features, since it shows that the Gram and Fisher coefficients are the two endpoints of a graded family rather than an isolated dichotomy. Everything in this subsection is an upper bound and an exact structural statement; we do not claim a matching all-measurable lower bound for a general bank, and Remark K.14 states precisely what is missing.

Let $\Phi = (\Phi_0, \dots, \Phi_m)$ be scalar features with the mandatory identity $\Phi_0(r) = r$, each locally absolutely continuous with $\Phi_j \in L^2(f)$, $\mathbb{E}|\Phi_j'| < \infty$, and $\Phi_j(x) f(x) \to 0$ as $|x| \to \infty$. Any learned affine preactivation is absorbed by taking $\Phi_j = \rho(a_j r + b_j)$ as a feature. Write $\mu = \mathbb{E}\Phi$, $\bar{\Phi} = \Phi - \mu$, $V = \mathbb{E}[\bar{\Phi}\bar{\Phi}^\top]$, and $a = \mathbb{E}[\Phi']$; integration by parts gives $a = \mathbb{E}[\bar{\Phi}\psi]$. Define the Godambe-type information and the optimal combination

$$J_\Phi = a^\top V^+ a, \qquad \beta = V^+ a, \qquad g_\Phi = \beta^\top \bar{\Phi}. \tag{96}$$

Since $\bar{\Phi}_0 = \varepsilon$ and $\langle \psi, \varepsilon \rangle_{L^2(f)} = 1$, the classical projection geometry of optimal estimating functions (Godambe, 1960; Heyde, 1997) gives $a \in \operatorname{range}(V)$, identifies g_Φ as the $L^2(f)$ projection of ψ onto $\bar{\mathcal{V}} = \overline{\operatorname{span}}\{\bar{\Phi}_j\}$, and yields

$$\frac{1}{\sigma^2} \le J_\Phi = \|\Pi_{\bar{\mathcal{V}}} \psi\|^2_{L^2(f)} \le I_f, \qquad \mathbb{E}[g_\Phi'] = \operatorname{Var}(g_\Phi) = J_\Phi, \tag{97}$$

with J_Φ monotone in nested banks and $J_\Phi = I_f$ exactly when $\psi \in \bar{\mathcal{V}}$. This geometry is classical and is used here only as a tool.

The architecturally new content is that a bank of this kind has an *exactly* characterizable accessible statistic, and that the extra blocks it exposes are first-order uninformative.

Proposition K.11 (Exact statistic of a read-once source bank)**.** *Consider a model that (i) forms the moment pilot $u = N^{-1}X^\top Y$ and the residual slot $r_i = Y_i - X_i^\top u$ by affine common-source layers, (ii) applies one tokenwise map to the scalar slot r_i only, writing $\Phi_0(r_i), \dots, \Phi_m(r_i)$ into fresh slots, and (iii) is otherwise an arbitrary finite composition of affine common-source layers with an affine readout. Then its prediction factors through z_q and*

$$\mathcal{T}_N = \sigma\Big(\tfrac{1}{N}\textstyle\sum_i X_i X_i^\top,\ \tfrac{1}{N}\sum_i X_i,\ u;\ \big\{\tfrac{1}{N}\sum_i X_i\Phi_j(r_i)\big\}_j;\ \big\{\tfrac{1}{N}\sum_i \Phi_j(r_i)\big\}_j;\ \big\{\tfrac{1}{N}\sum_i \Phi_j(r_i)\Phi_k(r_i)\big\}_{j\le k}\Big). \tag{98}$$

In particular no accessible block has the form $N^{-1}\sum_i X_i\Phi_j(r_i)\Phi_k(r_i)$.

Proof. After step (ii) the minimal source state is $\zeta_i = (1, X_i, \Phi_0(r_i), \dots, \Phi_m(r_i))$, since every retained raw-label or pilot coordinate is an affine function of ζ_i with u-dependent coefficients: $X_i^\top u = u^\top X_i$ and $Y_i = r_i + u^\top X_i$, whence, writing $B_j = N^{-1}\sum_i X_i\Phi_j(r_i)$, $M_j = N^{-1}\sum_i \Phi_j(r_i)$, $Q_{jk} = N^{-1}\sum_i \Phi_j\Phi_k(r_i)$, $H = N^{-1}X^\top X$ and $\bar X = N^{-1}\sum_i X_i$,

$$\overline{Y\Phi_j(r)} = Q_{0j} + u^\top B_j, \quad \overline{Y^2} = Q_{00} + 2u^\top B_0 + u^\top H u, \quad \bar Y = M_0 + u^\top \bar X.$$

By the recursion in the proof of Theorem E.2, every later source state is $M_\ell(C)\zeta_i$ with $C = N^{-1}\sum_i \zeta_i\zeta_i^\top$, and its Gram is $M_\ell(C)CM_\ell(C)^\top$. Hence all accessible empirical averages are averages of *pairwise* products of the coordinates of ζ_i, whose nonconstant blocks are exactly those listed in (98); the singleton means M_j and $\bar X$ appear because the homogeneous coordinate contributes the products $1\cdot\Phi_j(r_i)$ and $1\cdot X_i$. A block $N^{-1}\sum_i X_i\Phi_j\Phi_k(r_i)$ is cubic in the coordinates of ζ_i and is therefore not of that form; affine layers cannot multiply two slots. □

Two qualifications delimit the statement. If a product $\Phi_j\Phi_k$ lies in the affine span of the declared bank, then its x-weighted moment is already one of the B_ℓ; that enlarges the declared bank, and hence J_Φ, rather than escaping the description. And a *second* nonlinear source read would place products into token coordinates and thereby expose their x-weighted moments, so the read-once restriction is essential rather than cosmetic.

The reduction shows that a source bank exposes strictly more than the augmented Gram: besides the x-weighted blocks B_j, it reveals the unweighted moments M_j and Q_{jk} of the product closure. The next lemma explains why those extra blocks cannot change the leading coefficient. Write $s = u - W$, so that $r_i = \varepsilon_i - X_i^\top s$.

Lemma K.12 (Scalar residual features are first-order parameter-blind)**.** *Let $X \sim \mathcal{N}(0, I_d)$ be independent of $\varepsilon \sim f$ and let g satisfy the regularity above. Then*

$$\mathbb{E}\big[X\, g(\varepsilon - X^\top s)\big] = -s\, A_g(\|s\|), \qquad A_g(\|s\|) = \mathbb{E}\big[g'(\varepsilon - X^\top s)\big], \quad A_g(0) = \mathbb{E}g'. \tag{99}$$

Moreover $\mathbb{E}[g(\varepsilon - X^\top s)]$ depends on s only through $\|s\|$; if in addition $\sup_{0\le t\le t_0}\mathbb{E}[Z^2\sup_{\lambda\in[0,1]}|g''(\varepsilon - \lambda t Z)|] < \infty$ for $Z\sim\mathcal{N}(0,1)$, then

$$\mathbb{E}[g(\varepsilon - X^\top s)] - \mathbb{E}g = O(\|s\|^2). \tag{100}$$

If moreover the product $\bar\Phi_j g$ satisfies the same regularity as g and

$$\sup_{\|s\|\le t_0}\Big|\mathbb{E}(\bar\Phi_j g)'(\varepsilon - X^\top s) - \mathbb{E}\Phi_j'(\varepsilon - X^\top s)\,\mathbb{E}g(\varepsilon - X^\top s)\Big| < \infty, \tag{101}$$

then $\mathrm{Cov}(X\bar\Phi_j(\varepsilon - X^\top s), g(\varepsilon - X^\top s)) = O(\|s\|)$, vanishing exactly at $s = 0$.

Proof. Condition on X and set $G_g(t) = \mathbb{E}_\varepsilon g(\varepsilon - t)$, so that $\mathbb{E}[Xg(\varepsilon - X^\top s)] = \mathbb{E}[XG_g(X^\top s)]$. Gaussian integration by parts gives $\mathbb{E}[XG_g(X^\top s)] = s\,\mathbb{E}[G_g'(X^\top s)]$, and $G_g'(t) = -\mathbb{E}g'(\varepsilon - t)$, which is (99). Since $X^\top s \sim \mathcal{N}(0, \|s\|^2)$ is independent of ε, the law of $\varepsilon - X^\top s$, hence every expectation of a fixed function of it, depends on s only through $\|s\|$. Writing $X^\top s = \rho Z$ with $\rho = \|s\|$, Taylor's theorem with integral remainder gives the exact identity

$$\mathbb{E}[g(\varepsilon - \rho Z)] - \mathbb{E}g(\varepsilon) = \rho^2\,\mathbb{E}\Big[Z^2\int_0^1 (1-t)\,g''(\varepsilon - t\rho Z)\,dt\Big],$$

since the nominal linear term is $-\rho\,\mathbb{E}Z\,\mathbb{E}g'(\varepsilon) = 0$; the stated domination makes the right side $O(\rho^2)$, which is (100). Note that $\mathbb{E}|g'| < \infty$ alone would give only a vanishing first derivative, not a quadratic rate. For the covariance, apply (99) to the product $\bar{\Phi}_j g$ and to $\bar{\Phi}_j$:

$$\mathrm{Cov}\big(X\bar{\Phi}_j(r_s), g(r_s)\big) = -s\Big[\mathbb{E}(\bar{\Phi}_j g)'(r_s) - \mathbb{E}\Phi_j'(r_s)\,\mathbb{E}g(r_s)\Big],$$

which is $O(\|s\|)$ by the envelope (101) and is exactly zero at $s = 0$ because $\mathbb{E}X = 0$ and X is independent of ε. □

Thus only x-weighted statistics respond linearly to a local shift of W: a scalar function of the residual sees it at order $\|s\|^2$, because the design is centered and independent of the noise. This is the same mechanism by which $\|Y\|^2$ and $\mathbf{1}^\top Y$ act as ancillary coordinates in Theorem F.2. The lemma establishes the absence of a first-order mean shift; by itself it does not prove that the unweighted blocks are uninformative for arbitrary measurable rules, which would additionally require the total-variation limit discussed in Remark K.14.

Proposition K.13 (Graded attainment by a read-once source bank)**.** *Assume the noise conditions of Theorem F.4, let $W \sim \mathcal{N}(0, \tau^2 I_d)$, and suppose g_Φ of (96) has a $C^{1,1}$ representative with $\mathbb{E}g_\Phi'(\varepsilon)^2 < \infty$. With $u = N^{-1}X^\top Y$ set*

$$\widehat{W}_{\Phi,N} = u + \frac{1}{J_\Phi N}\sum_{i=1}^{N} X_i\, g_\Phi\big(Y_i - X_i^\top u\big). \tag{102}$$

Then $\mathbb{E}\|\widehat{W}_{\Phi,N} - W\|^2 = d/(J_\Phi N) + O(N^{-3/2})$, and (102) is realized by a model of the form in Proposition K.11 with two attention layers, one head per layer, width $d + 3 + m$, and no matrix inversion.

Proof. Identical to the proof of Theorem F.7 in Section K.6 with ψ replaced by g_Φ and I_f by $\mathbb{E}g_\Phi' = J_\Phi$. Writing $\Delta = u - W$, $S_N = N^{-1}\sum_i X_i g_\Phi(\varepsilon_i)$ and $A_N = N^{-1}\sum_i X_i X_i^\top g_\Phi'(\varepsilon_i)$, the same expansion gives $\widehat{W}_{\Phi,N} - W = J_\Phi^{-1}S_N - J_\Phi^{-1}(A_N - J_\Phi I_d)\Delta + J_\Phi^{-1}Q_N$ with $\|Q_N\| \le \frac{L}{2}M_{3,N}\|\Delta\|^2$. By (97), $\mathrm{Var}(g_\Phi) = J_\Phi$, so $\mathbb{E}\|J_\Phi^{-1}S_N\|^2 = d\,\mathrm{Var}(g_\Phi)/(J_\Phi^2 N) = d/(J_\Phi N)$, and the remainder bounds of Section K.6 apply verbatim. For the architecture, replace the single scalar channel by the bank $(x, y, p, r) \mapsto (x, y, p, \Phi_0(r), \dots, \Phi_m(r))$ followed by the affine contraction $\beta^\top(\Phi(r) - \mu)$, and take the second-layer output scale J_Φ^{-1}. □

Both existing endpoints are special cases. For $\Phi = (\Phi_0)$ one has $g_\Phi = \varepsilon/\sigma^2$, $J_\Phi = 1/\sigma^2$, and (102) reduces to $(2I_d - H)u$, which is exactly (22); when $\psi \in \bar{\mathcal{V}}$ one has $J_\Phi = I_f$ and recovers Theorem F.7. So the Gram and Fisher coefficients are the two ends of one graded family, and Theorem F.8 is the case $m = 1$ of the statement that $\beta = V^+ a$ is the unique optimal channel combination modulo $\ker V$. Note that $J_\Phi < I_f$ holds precisely when $\psi \notin \bar{\mathcal{V}}$; equality already occurs at finite width for Gaussian noise with the identity channel and for logistic noise with one $\tanh$ channel. Absent the matching lower bound, $J_\Phi < I_f$ is a limitation of the exhibited rule rather than an architecture-wide barrier, and $J_{\Phi_m} \uparrow I_f$ requires a nested bank whose centered span is dense in $L^2(f)$, not merely a widening one.

Remark K.14 (What is not claimed)**.** Proposition K.13 is an attainability result, so d/J_Φ is an upper bound on the optimal risk of the experiment (98). We do *not* claim a matching all-measurable lower bound for a general bank. Establishing one would require a total-variation limit for the triangular array of summands $X_i\bar{\Phi}(\varepsilon_i - X_i^\top s_N)$ that is uniform over $\|s_N\| \to 0$, together with a nondegeneracy condition replacing the polynomial submersion argument of Section K.3, and a treatment of the same-sample pilot's dependence, for which the exact affine local map (57) has no analogue. Two further restrictions are load-bearing in Proposition K.11: the tokenwise map must read the residual slot only and not X_i, since a feature $G(X_i, \varepsilon_i)$ has sensitivity $\mathbb{E}[XX^\top \partial_r G]$ and covariance $\mathbb{E}[XX^\top G^2]$ that need not be multiples of I_d; and there must be a single nonlinear read, since a second one would expose the product closure of the closure. Discrete features such as sign make the unweighted blocks lattice-valued and lie outside any total-variation statement of this type.

L General designs, exact sufficiency, and stationary diagnostics

L.1 Complete moments under non-Gaussian designs

Proposition L.1 (Design-universal moment experiment)**.** *Fix $P \geq 2$ and $d \geq 1$. Let the covariate law be absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, with mean zero, covariance I_d, and finite $2P$th absolute moment. Let independent errors have an everywhere-positive density, mean zero, and finite $2P$th absolute moment. Then the local experiment and local asymptotic minimax conclusions of Theorem J.1 hold with the same J_P. If the design, errors, and continuous everywhere-positive task prior have moments of every order, the integrated Bayes conclusion also holds. The score-projection identity requires exactly the error-score and tail conditions stated there.*

Proof. Let $m(x, e)$ contain all nonconstant monomials through total degree P, $\mu = \mathbb{E}m(X, \varepsilon)$, and $\Sigma = \mathrm{Cov}(m(X, \varepsilon))$. Independence and the $2P$th moments make Σ finite. It is positive definite because a nonzero polynomial has a Lebesgue-null zero set, and the joint law of (X, ε) is absolutely continuous.

For completeness, the strong central limit argument does not use Gaussian design. The linear spans of the ranges of $Dm(z)$, over $z \in \mathbb{R}^{d+1}$, fill the output space: an annihilator would give a nonconstant polynomial with identically zero gradient. Select a finite tuple of points for which the sum map $F(z_1, \ldots, z_s) = \sum_i m(z_i)$ has a nonzero maximal Jacobian minor. This minor is a nonzero polynomial, so DF has full rank Lebesgue-almost everywhere, hence almost surely under the absolutely continuous product input law. The coarea formula on bounded submersion charts therefore makes F's pushforward absolutely continuous. Thus a finite convolution of the law of $m(X, \varepsilon)$ has a density. The multivariate total-variation CLT criterion (Bally & Caramellino, 2016, Theorem 2.2) yields

$$Z_N = N^{-1/2} \sum_i \{m(X_i, \varepsilon_i) - \mu\} \longrightarrow_{\mathrm{TV}} \mathcal{N}(0, \Sigma).$$

The polynomial shear $m(x, e + x^\top t) = R(t)m(x, e)$ is invertible, with $R(t)^{-1} = R(-t)$, for every design. After the observable recentering at w_0, the local law is exactly

$$R(h/\sqrt{N})Z_N + \sqrt{N}\{R(h/\sqrt{N})\mu - \mu\}.$$

Replacing Z_N by its Gaussian limit contracts total variation. On any fixed compact set of h, the resulting means and nonsingular covariances converge uniformly. Hence the complete experiment converges in Le Cam distance to $\{\mathcal{N}(Dh, \Sigma) : h \in H\}$, where $D = \partial_t R(t)\mu|_{t=0}$. The approximation is independent of the center w_0 after its invertible recentering.

It remains to identify its information without Hermite polynomials. Apply Gram–Schmidt to the design monomials in nondecreasing total degree. Choose the first basis elements to be $q_0 = 1$ and $q_j = X_j$, $1 \leq j \leq d$, which are already orthonormal. Absolute continuity of the design law and the nullity of nonzero polynomial zero sets ensure that all subsequent pivots are nonzero. Index the remaining orthonormal polynomials by ν, with degree k_ν. The products

$$q_\nu(X)\varepsilon^b, \qquad 0 \leq b \leq P - k_\nu,$$

with the constant coordinate removed, give an invertible change of basis of the complete moment space. Center these products for computing their covariances. For $\nu \neq 0$ their means are zero, and different ν blocks are orthogonal by independence and design orthonormality. The $\nu = 0$ block is the covariance of centered error powers. In the Gaussian limit these orthogonal blocks are independent.

The local derivative of a coordinate is

$$\frac{d}{dt}\mathbb{E}\big[q_\nu(X)(\varepsilon + tX^\top h)^b\big]\Big|_{t=0} = b\,\mathbb{E}\varepsilon^{b-1}\,\mathbb{E}[q_\nu(X)X]^\top h.$$

The derivative is zero for $\nu = 0$ by centering, and for all degree-at-least-two basis elements by their orthogonality to the linear polynomials. For $q_j = X_j$ it equals $b\mathbb{E}\varepsilon^{b-1}h_j$. Consequently the informative blocks are independent $T_j \sim \mathcal{N}(a_P h_j, B_P)$, with $B_P = \mathbb{E}[v_P v_P^\top]$. Each is equivalent to $J_P^{-1}a_P^\top B_P^{-1}T_j \sim \mathcal{N}(h_j, J_P^{-1})$ plus an independent parameter-free Gaussian residual. This proves the claimed experiment limit.

For all-measurable local minimax risk, project actions onto each fixed local cube to bound the loss. Total-variation transfer and the flat-cube Gaussian Bayes lemma give the lower bound as the cube expands. Conversely, use the reverse Le Cam kernel followed by the Gaussian observation projected onto that cube; its risk is at most d/J_P up to the bounded-loss transfer error. Conditional expectation removes auxiliary randomization. No pointwise lower bound on arbitrary estimator sequences is used.

For integrated risk, tile a fixed compact task cube by cells of side $C/\sqrt{N}$. Allowing a separate oracle estimator in each cell only lowers risk. Uniform continuity and positivity of the prior make every rescaled cell prior converge uniformly to the flat-cube prior. The experiment error is uniform across cells by the exact shear. Multiplication by the cell prior masses before summation prevents error accumulation. This gives $\liminf_N N R^*_{P,N}(\pi) \geq r_C \pi(K)$; expanding C and K gives d/J_P.

For its upper bound let $g = \beta^\top v_P$, $\beta = B_P^{-1} a_P$, and use

$$\widehat{w} = u + J_P^{-1} N^{-1} \sum_i X_i g(Y_i - X_i^\top u), \qquad u = N^{-1} \sum_i X_i Y_i.$$

The rule is moment-measurable. Since $\mathbb{E} X X^\top = I_d$, $u - w = N^{-1} \sum_i [(X_i X_i^\top - I_d) w + X_i \varepsilon_i]$ is $O_{L^q}(N^{-1/2})$ for every fixed q, both on compact task sets and after integrating the stipulated prior. The centered empirical polynomial averages have the same bound, by the centered-sum moment inequality; uncentered polynomial envelopes are $O_{L^q}(1)$. Exact polynomial Taylor expansion and Hölder's inequality therefore give

$$\widehat{w} - w = J_P^{-1} N^{-1} \sum_i X_i g(\varepsilon_i) + O_{L^2}(N^{-1}).$$

The leading covariance is $(N J_P)^{-1} I_d$, since $\mathbb{E} g^2 = \mathbb{E} g' = J_P$. Thus its integrated risk is $d/(N J_P) + O(N^{-3/2})$, proving the matching upper bound. Finally, the integration-by-parts identification of g as the projected noise score involves only f, and is unchanged. □

Role of the design assumptions. Gaussianity and independence between design coordinates are unnecessary. Centered and whitened uniform laws on bounded sets with nonsingular covariance are included. Centering, however, removes the error-only moment block's local mean derivative; the statement is not asserted for uncentered designs. For a known positive-definite covariance Ω, invertible whitening preserves the complete degree budget and changes the information matrix to $J_P \Omega$.

L.2 Local losslessness and exact sufficiency

Proof of Theorem 3.2. All identities involving the score are understood almost everywhere with respect to f. Take the continuous locally absolutely continuous version of the everywhere-positive density. Integration by parts is justified by the displayed tail conditions and Cauchy–Schwarz: both $\mathbb{E}|\varepsilon^k \psi(\varepsilon)|$ and $\mathbb{E}|k\varepsilon^{k-1}|$ are finite for $k \leq P-1$. Thus the moment theorem identifies J_P with the squared norm of the orthogonal projection of ψ onto $\mathcal{P}_{P-1}$. Equality with $I_f = \|\psi\|_2^2$ holds exactly when ψ is in that polynomial space. This proves (1)⇔(2).

Positivity and continuity bound f away from zero on each compact interval, so $\log f$ is locally absolutely continuous with derivative $-\psi$. Integrating a polynomial score proves (2)⇒(3), including the integration constant in q. Differentiation proves the reverse implication. A nonconstant polynomial exponent is integrable on the whole real line exactly when its degree is even and its leading coefficient is negative; constant exponents are not integrable. Hence its degree is at least two.

Under (3), the joint likelihood for the observed design and labels is

$$\prod_{i=1}^N p_X(x_i) \, \exp\left\{ \sum_{i=1}^N q(y_i - x_i^\top w) \right\}.$$

Expanding the exponent involves only complete empirical mixed monomials of total degree at most $\deg q \leq P$, and constants known from N. The design-density factor is independent of w. The factorization criterion therefore proves exact sufficiency at every N.

For clarity, the reverse implication requires only differentiability in quadratic mean of the full regression experiment, not a globally Lipschitz derivative of the score. Put $a = \sqrt{f}$. Local absolute continuity, positivity, and finite Fisher information give $a' = -\psi a/2 \in L^2(\mathbb{R})$ and $\|a'\|_2^2 = I_f/4$. The Sobolev translation identity implies

$$\|a(\cdot - s) - a + sa'\|_2 = o(|s|), \qquad \|a(\cdot - s) - a + sa'\|_2 \le 2|s|\|a'\|_2.$$

Indeed, write the remainder as an integral of translated differences of a', and use continuity of translations in L^2. For a regression perturbation t, substitute $s = X^\top t$ and integrate over X. After division by $\|t\|^2$, the squared remainder is bounded by $4\|X\|^2\|a'\|_2^2$, which is integrable. Its integrand tends to zero for every fixed X, so dominated convergence proves joint quadratic-mean differentiability, with score $X\psi(\varepsilon)$ and information $I_f I_d$. The score is centered because $\mathbb{E}X = 0$ and $\mathbb{E}|\psi| < \infty$.

The iid DQM-to-LAN theorem (van der Vaart, 1998, Theorem 7.2) gives, for any fixed nonzero h, under the parameter w_0,

$$\log \frac{dP^{\text{full}}_{N,w_0+h/\sqrt{N}}}{dP^{\text{full}}_{N,w_0}} \;\Rightarrow\; \mathcal{N}(-\rho^2/2, \rho^2), \qquad \rho^2 = I_f\|h\|^2.$$

The limiting likelihood ratio is lognormal with expectation one. The finite-sample likelihood ratios also have expectation one; their nonnegativity, weak convergence, and equality of limiting expectations imply uniform integrability. Therefore their L^1 distance from one converges, giving the binary testing limit

$$\|P^{\text{full}}_{N,w_0+h/\sqrt{N}} - P^{\text{full}}_{N,w_0}\|_{\text{TV}} \longrightarrow 2\Phi(\sqrt{I_f}\|h\|/2) - 1.$$

The already proved strong moment experiment gives the analogous limit

$$2\Phi(\sqrt{J_P}\|h\|/2) - 1$$

for the complete statistic. Exact sufficiency supplies a parameter-independent conditional kernel reconstructing the full observation from the statistic, so data processing in both directions makes these two finite-sample total variations identical. Strict monotonicity of Φ forces $J_P = I_f$. This proves (4)⇒(1) and closes the equivalence.

Finally, if $\deg q = m$, its score has degree exactly $m - 1$ and cannot coincide almost surely with a lower-degree polynomial under an everywhere-positive density. Hence the smallest lossless total degree is m. □

Quantifiers and interpretation. Exact sufficiency is asserted for the experiment observing both random design and labels, and holds for every sample size. Sufficiency at one small sample size would not imply the classification: a sufficiently large complete moment vector may identify all records when N is small. The equality of the first-order information and exact sufficiency is a property of this complete polynomial-summary family, not a general equivalence for arbitrary summaries. Fisher preservation alone need not imply sufficiency (Kagan & Shepp, 2005).

L.3 Stationary interventions on Gram fibers

Proposition L.2 (Stationary conditional-fiber intervention)**.** *Generate (W, X, Y) from the stated regression model, with independent errors having an everywhere-positive density f. Hold W and the design matrix X fixed. Write $A = [\mathbf{1}\ X]$, let P_A be the orthogonal projector onto its column space, and put $y_\| = P_A Y$, $r = (I - P_A)Y$, $R = \|r\|$, $m = N - \operatorname{rank}(A)$. When $m \ge 2$ and $R > 0$, draw a uniform unit vector v in $\operatorname{col}(A)^\perp \cap r^\perp$ and propose*

$$r' = \cos\theta\, r + \sin\theta\, Rv, \qquad Y' = y_\| + r', \qquad 0 < \theta < \pi.$$

The angle can be fixed or sampled independently conditional on the invariants $(W, X, P_A Y, R)$. Accept this proposal with probability

$$\alpha(Y, Y') = 1 \wedge \frac{\prod_i f(Y_i' - X_i^\top W)}{\prod_i f(Y_i - X_i^\top W)},$$

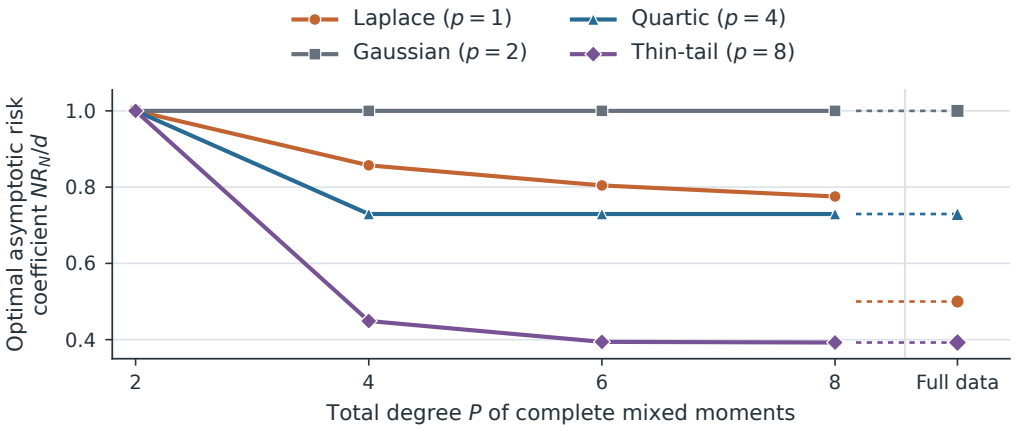

Figure 3: **The complete-moment information hierarchy.** Analytic values of $1/J_P$, the leading normalized risk coefficient, for unit-variance exponential-power laws. Gaussian inference is already complete at $P = 2$; quartic and $p = 8$ reach their Fisher endpoints once the relevant polynomial score is retained. Laplace's discontinuous sign score is approached but not captured by any fixed polynomial degree. Values are computed analytically from the noise moments and the score-projection identity.

and otherwise set $Y^+ = Y$; on acceptance set $Y^+ = Y'$. For $m = 1$, use a symmetric proposal that flips r to $-r$ with any fixed probability, with the same acceptance rule. For $m = 0$ or $R = 0$, keep Y. Then $T_N(X, Y^+) = T_N(X, Y)$ pathwise and

$$(W, X, Y^+) \overset{d}{=} (W, X, Y).$$

One step is sufficient for this invariance; no burn-in assertion is needed.

Proof. Holding X fixed, the augmented Gram constrains precisely $A^\top Y$ and $\|Y\|^2$. Equality of $A^\top Y$ fixes $P_A Y$ even if A is rank deficient, and the norm constraint then fixes R. Thus its label fiber is the sphere

$$\{y_\| + r : r \in \mathrm{col}(A)^\perp, \|r\| = R\}.$$

Use an orthonormal basis for $\mathrm{col}(A)$ and its orthogonal complement. In these coordinates, Lebesgue measure disintegrates into ordinary measure on the parallel coordinates and polar measure $R^{m-1}\, dR\, d\sigma(u)$ on the residual coordinates, where σ is uniform surface measure. Equivalently, the coarea Jacobian for the parallel coordinates and squared residual norm is constant on every positive-radius fiber. Conditional on $W, X, P_A Y, R$, the directional density relative to σ is consequently proportional to

$$L_W(y_\| + Ru) = \prod_i f(y_{\|,i} + Ru_i - X_i^\top W).$$

For almost every conditioning value its normalizing constant is finite and positive, as also follows from this disintegration of a probability density. For $m = 1$ the reference measure is uniform on the two points; the remaining degenerate cases are singletons.

The fixed-angle proposal on the residual unit sphere is reversible with respect to uniform surface measure. One direct construction draws a uniform orthonormal pair (u, v) and sets $u' = \cos\theta\, u + \sin\theta\, v$. The joint law of (u, u') is invariant under exchanging its two entries, by rotational invariance of the orthonormal frame. This proves reversibility even though a fixed-angle transition need not have a density relative to the full surface measure. Mixing angles using only fiber invariants preserves reversibility. The usual Metropolis acceptance rule therefore makes the transition reversible with respect to the conditional density proportional to L_W. The sign-flip case has the same property. Each proposal preserves the parallel part and residual norm, and hence the Gram.

The original generated observation already has precisely this conditional law given (W, X, T_N). A single invariant transition leaves it unchanged; integrating over (W, X, T_N) proves the joint marginal

identity. Independent query covariates and the clean query target may be adjoined without affecting the conclusion. □

What the diagnostic licenses. For any fixed fitted predictor, both endpoints have the original test distribution and the same clean target. Their expected prediction risks are identical. Nevertheless, their predictions and their full-data posterior means can differ pathwise, while every deterministic Gram-measurable predictor agrees exactly on the endpoints. The paired statistic can therefore measure whether a predictor responds to decision-relevant variation within a Gram fiber. It is not an estimate of risk gained from the intervention, nor does response alignment alone prove posterior optimality or identify an internal mechanism.

Rejected proposals must remain in the original-distribution analysis as zero-change pairs. Restricting to accepted or large-displacement pairs conditions on a data-dependent event and forfeits the marginal identity. Several independently restarted proposals can improve power; repeated steps with a predetermined count also preserve stationarity, but their within-task dependence must enter uncertainty estimates. The latent W is used only by the offline simulator to construct the diagnostic, never as an input to the fitted predictor or its posterior-reference computation.

L.4 A COMPLETE DEGREE-FOUR DECODER UNDER THREE DESIGN LAWS

This numerical check tests a sufficient-summary decoder, separately from neural-network training. We draw $W \sim \mathcal{N}(0,1)$ and unit-variance quartic errors, independently of the covariates. The three design laws are standard Gaussian; $(G + 0.8S)/\sqrt{1.64}$ with $G \sim \mathcal{N}(0,1)$ and an independent uniform sign S; and Laplace with scale $1/\sqrt{2}$. All have everywhere-positive densities, zero mean, unit variance, and all moments. Each design/length cell uses 20,000 fresh independent tasks.

The degree-two reference is $\sum_i X_i Y_i/(1+\sum_i X_i^2)$. The degree-four rule is the exact posterior mode under the known quartic law and Gaussian prior. Its objective is strictly convex since

$$L''(w) = 1 + 12\alpha_4^{-4} \sum_i X_i^2 (Y_i - X_i w)^2 > 0.$$

Its derivative is a cubic determined by degree-four moments. We evaluate that cubic around the ridge pilot to avoid cancellation, bracket its unique root, and use 50 bisection steps. This is a change of polynomial coordinates of the same complete statistic. Every computed root is also checked using the direct raw-record gradient. The largest absolute gradient is 3.06×10^{-12}; its ratio to local curvature is at most 1.11×10^{-15}. The posterior mode is not the finite-sample posterior mean or an asserted finite-sample Bayes optimum.

Table 9: **All cells of the non-Gaussian design check.** Normalized coefficient risk $N\mathbb{E}(\widehat{W} - W)^2$, estimated on 20,000 independent tasks per cell. The paired 95% Monte Carlo interval is for degree-four MAP minus ridge. The asymptotic reference coefficients are 1 and 0.72948. These intervals describe Monte Carlo error, not training variability.

Design	N	Degree-two ridge	Degree-four MAP	Paired 95% interval
Gaussian	128	1.0193	0.7559	$[-0.2768, -0.2500]$
Gaussian	512	1.0035	0.7425	$[-0.2745, -0.2476]$
Gaussian	2048	0.9962	0.7281	$[-0.2814, -0.2547]$
Gaussian mixture	128	1.0039	0.7546	$[-0.2626, -0.2360]$
Gaussian mixture	512	1.0144	0.7438	$[-0.2840, -0.2572]$
Gaussian mixture	2048	0.9966	0.7315	$[-0.2784, -0.2517]$
Laplace	128	1.0385	0.7931	$[-0.2592, -0.2317]$
Laplace	512	1.0079	0.7414	$[-0.2800, -0.2531]$
Laplace	2048	0.9940	0.7215	$[-0.2858, -0.2592]$

The three designs all show a degree-four benefit across the fixed tested lengths; their largest-length coefficients are compatible with the common Fisher reference. This checks the useful retained statistic under these laws, rather than claiming that any particular polynomial one-step approximation is stable at every finite sample size.

The protocol was fixed after an independent exploratory calculation motivated exact convex decoding. It uses wholly fresh banks and reports all nine cells. Seeds are $260919731 + 10000j + N$, with $j = 0, 1, 2$ in the listed design order. The implementation is `experiments/moment_map_check.py`; `experiments/moment_map_results/` contains all task-level losses, JSON summaries, and the script hash. The computation took 40.1 seconds on the evaluation CPU and used no trained models.

M ORIGINAL-LAW EQUAL-GRAM TRANSFORMER DIAGNOSTIC

This follow-up was specified after the earlier rotation results, using independent task banks and unchanged fitted checkpoints. The transition and its stationarity argument are given in Proposition 5.1.

Frozen protocol. The six cells use $p \in \{1, 2, 4\}$, $N \in \{32, 128\}$, $d = 2$, and 256 independent original tasks each. The task prior and covariates are standard Gaussian; noise has unit variance and density proportional to $\exp(-|e/\alpha_p|^p)$. A NumPy generator uses seed $2609177300 + 1000p + N$. The proposal angle is fixed at 0.25 and the direction threshold at 0.01 before observing the new banks. Each task receives exactly one Metropolis proposal; no burn-in, repeated attempts, angle tuning, or rejection filtering is used. Every raw proposal, acceptance log-ratio, acceptance uniform, accepted indicator, original coefficient, and both label endpoints are retained. Exact rejections reuse their original posterior result, so numerical recomputation cannot create a spurious change. The sampler alone uses the true coefficient; the Transformer and posterior estimator never receive it.

The unchanged final-step checkpoints have training seeds 26091611/12/13 under each noise law and 12,000 training updates. Inference runs in CPU float64, with the saved normalization epsilon, using batches of 16. This is distinct from the earlier H200 bfloat16 risk evaluation and requires no model retraining. Checkpoint and source hashes are frozen in `protocol.json`.

Numerical posterior verification. The initial integration uses tolerance 10^{-4} and whitened domain radii 8/12; the stricter pass uses 10^{-5} and radii 10/14, with refinement through 2,049 points per axis. Tolerances additionally tighten with the query's ℓ_1 norm. Both endpoint passes must converge and their prediction discrepancy must be at most 0.00025. Gaussian posterior means use analytic ridge, checked by the strict quadrature on the first eight tasks at each length. The complete augmented Gram is checked in float64, as is invariance of the learned Gram control.

All 1,536 original pairs are retained; there are 0 numerical failures. The maximum primary/strict prediction discrepancy is 9.04×10^{-5}, the maximum relative Gram error is 1.56×10^{-15}, and the maximum Gram-control prediction change is 4.44×10^{-15}.

Uncertainty. All primary squared-error contrasts include every original pair, including zero-change rejected transitions. Losses are first averaged over the three fitted models within a pair; normal Monte Carlo intervals then use 256 independent pairs, not 768 independent observations. Training-seed SD is reported separately. Direction intervals bootstrap the pair-average indicator on the threshold-eligible set, with 4,000 resamples. These intervals condition on the fitted models and are pointwise, not simultaneous or population-of-training-seeds guarantees.

Table 10: All original-law diagnostic cells. Error contrasts are in units of 10^{-4} and use all 256 pairs. Direction uses only the explicitly listed k threshold-eligible pairs. Gaussian direction is undefined.

N	Law	Accepted	k	Direction [95% CI]	Error contrast [95% CI]
32	Laplace	142	103	82.8% [76.4, 89.0]	−13.029 [−20.409, −5.650]
32	Gaussian	256	0	–	0.423 [0.329, 0.516]
32	Quartic	184	137	90.0% [86.1, 93.7]	−12.518 [−17.832, −7.204]
128	Laplace	72	49	79.6% [69.4, 89.1]	−3.420 [−5.878, −0.961]
128	Gaussian	256	0	–	0.108 [0.078, 0.138]
128	Quartic	116	69	87.4% [79.7, 94.2]	−2.461 [−3.465, −1.458]

Both endpoint risks estimate the same marginal risk by stationarity; their finite-bank difference is not a treatment benefit. The primary functional diagnostic instead compares the model's posterior

Table 11: Change magnitudes and fitted-seed variability. The model RMS is averaged over three fitted seeds. The contrast SD is across seed-specific mean contrasts; it is distinct from the conditional pair intervals above. The final two columns are mean normalized clean-target risks at the original and transitioned endpoint, on the same 256 dependent pairs.

N	Law	Posterior RMS	Model RMS	Contrast SD	Risk D	Risk D'
32	Laplace	0.04699	0.03842	0.000051	0.7352	0.7455
32	Gaussian	0.00000	0.00650	0.000002	1.3297	1.3346
32	Quartic	0.04222	0.03699	0.000063	1.3318	1.2570
128	Laplace	0.02735	0.01391	0.000011	0.6662	0.6845
128	Gaussian	0.00000	0.00328	0.000002	0.7509	0.7471
128	Quartic	0.01934	0.01464	0.000015	0.7060	0.7034

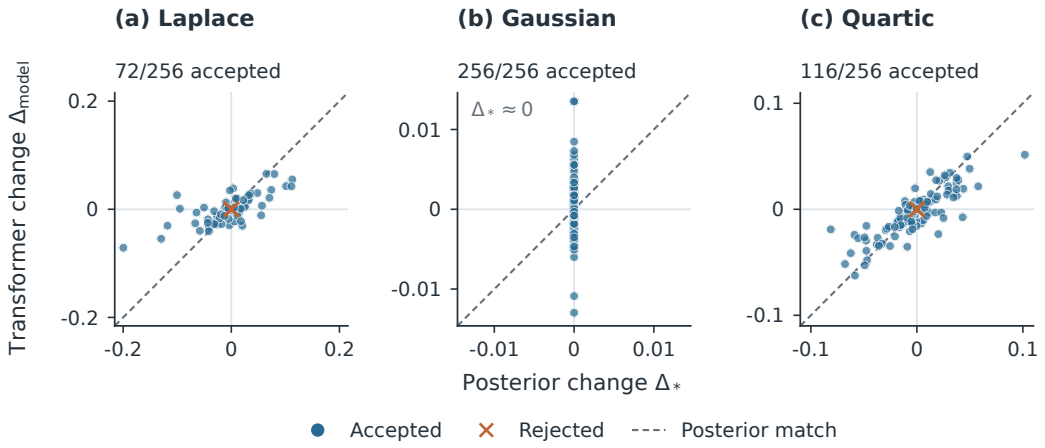

Figure 4: **Transformer responses at fixed Gram under the original task law.** All 256 pairs per law at $N = 128$. Blue points are accepted transitions; orange crosses at the origin are overlapping rejected transitions. Predictions are averaged across three fitted seeds for visualization; the discrepancy analysis averages squared errors across seeds. Dashed diagonals indicate exact posterior tracking. Panels use separate scales and equal horizontal and vertical aspect ratios.

change error with the exact zero-change restriction imposed by the Gram. The evidence directory contains `rerun_portable.py`, a portable export of the frozen numerical implementation, source snapshots, protocols, all proposal and prediction arrays, task-level integration histories, and offline analysis and plotting scripts. The portable wrapper accepts a checkpoint root and an output directory, verifies the recorded hashes, and can rerun the two lengths independently with one CPU thread per process. The export manifest records path/metadata changes and verifies that checkpoint inference tensors are bitwise unchanged.

N Finite-context information witnesses

N.1 Finite-sample witnesses from stationary summary fibers

Fix a sample size; none of the following statements takes an asymptotic limit. Write $U = (D, X_q)$, $S = (T_N(D), X_q)$, and $m(U) = \mathbb{E}[Z \mid U]$, where $\mathbb{E}Z^2 < \infty$. The optimal squared-loss risks over all measurable rules satisfy

$$B_N := R^*_{S,N} - R^*_{\text{full},N} = \mathbb{E}\{m(U) - \mathbb{E}[m(U) \mid S]\}^2.$$

Thus B_N is a decision-information loss, independently of the learned predictor used to probe it.

Theorem N.1 (A finite-sample all-decoder witness)**.** *Let (U, U') be any coupling with equal marginal distributions and $S(U') = S(U)$ almost surely. Put $\Delta = m(U') - m(U)$ and $Q = \mathbb{E}\Delta^2$. Then*

$$B_N \geq \frac{Q}{4}. \tag{103}$$

For any square-integrable pair statistic F, let $a = \mathbb{E}[F\Delta]$ and $b = \mathbb{E}F^2$. If $b > 0$, then

$$B_N \geq \frac{a^2}{4b} = \frac{1}{4}\sup_{\lambda\in\mathbb{R}}(2\lambda a - \lambda^2 b). \tag{104}$$

In particular, take $F = g(U') - g(U)$ for any fixed fitted predictor g. If its paired squared discrepancy improves on the zero-change reference,

$$\mathbb{E}(F - \Delta)^2 - \mathbb{E}\Delta^2 = b - 2a < 0,$$

then $B_N > b/16 > 0$. Negative a also yields the squared witness (104), although it does not indicate useful directional alignment. When $b = 0$, $a = 0$ and the model-indexed witness is defined as zero.

Proof. Let $r(U) = m(U) - \mathbb{E}[m(U) \mid S]$. Summary preservation gives $\Delta = r(U') - r(U)$, and equal marginals give $\mathbb{E}r(U')^2 = \mathbb{E}r(U)^2 = B_N$. Hence

$$Q \leq 2\mathbb{E}r(U')^2 + 2\mathbb{E}r(U)^2 = 4B_N.$$

Cauchy–Schwarz gives $a^2 \leq bQ$. Completing the square proves the variational equality; equivalently, $2\lambda F\Delta - \lambda^2 F^2 \leq \Delta^2$ pointwise. Finally, $b - 2a < 0$ implies $a > b/2$, and therefore $a^2/(4b) > b/16$. □

Interpretation of the lower bound. The conclusion lower-bounds the risk penalty of every summary decoder relative to the full-data Bayes rule at this actual sample size. It does not say that g beats the best summary decoder. For example, $g = m + C$ has exactly the same paired changes as m and thus achieves the best possible paired discrepancy, while its ordinary risk is $R^*_{\mathrm{full},N} + C^2$. Taking $C^2 > B_N$ makes it worse than the optimal summary rule. A comparison with a learned Gram control supplies a different empirical statement and must not be substituted for the all-decoder conclusion.

N.2 A stronger witness from conditional model drift

The simulation may expose latent variables to its transition rule without exposing them to the fitted predictor. Let Ω denote the full simulation state, including (W, D, X_q), with its original probability law μ. Let K be a reversible stationary Markov kernel on this state that preserves the summary $S = (T_N(D), X_q)$ pathwise. Write $m(\Omega) = \mathbb{E}[Z \mid D, X_q]$ and $g(\Omega) = g(D, X_q)$, with both in $L^2(\mu)$. Define

$$d_g = (I - K)g, \qquad c_g = \mathbb{E}d_g^2, \qquad a_g = \mathbb{E}[(g' - g)(m' - m)].$$

Although g is an observed-data function, d_g can depend on W because the diagnostic kernel can depend on W.

Theorem N.2 (Drift witness for every summary decoder)**.** *Under the preceding conditions, for every fixed $t \in \mathbb{R}$,*

$$B_N \geq ta_g - t^2 c_g. \tag{105}$$

If $c_g > 0$, its optimized form is

$$B_N \geq \frac{a_g^2}{4c_g} \geq \frac{a_g^2}{4\mathbb{E}(g' - g)^2}. \tag{106}$$

If $c_g = 0$, then $a_g = 0$ and the optimized witness is zero. For any square-integrable function r invariant under K, including any summary-measurable function,

$$a_g = 2\mathbb{E}[(m - r)d_g]. \tag{107}$$

In particular, the witness requires a posterior calculation only at the original endpoint. If K also preserves the true target Z pathwise, the optimized witness is at most $R^*_{\mathrm{full},N}$.

Proof. Reversibility is self-adjointness of K in $L^2(\mu)$. Expanding the paired product and using stationarity gives

$$a_g = 2\langle m, g\rangle - \langle m, Kg\rangle - \langle g, Km\rangle = 2\langle m, (I-K)g\rangle.$$

For any invariant r, self-adjointness gives $\langle r, (I-K)g\rangle = \langle (I-K)r, g\rangle = 0$. This proves (107). In particular, testing against all bounded summary-measurable functions gives $\mathbb{E}[d_g \mid S] = 0$. Set $m_S = \mathbb{E}[m \mid S]$. Then

$$0 \le \mathbb{E}(m - m_S - td_g)^2 = B_N - ta_g + t^2 c_g,$$

which proves the variational bound. Completing the square proves the optimized form. Conditional Jensen gives

$$c_g = \mathbb{E}\{\mathbb{E}[g - g' \mid \Omega]\}^2 \le \mathbb{E}(g - g')^2.$$

Finally, if $KZ = Z$, then $a_g/2 = \langle m - Z, d_g\rangle$. Cauchy–Schwarz implies $a_g^2/(4c_g) \le \mathbb{E}(m-Z)^2 = R^*_{\mathrm{full},N}$. □

Corollary N.3 (A sharper witness when the target is preserved)**.** *Under the assumptions of Theorem N.2, suppose additionally that $KZ = Z$. Write $R = R^*_{\mathrm{full},N}$ and $L_g = a_g^2/(4c_g)$ when $c_g > 0$, with $L_g = 0$ when $c_g = 0$. If $B_N + R > 0$, then*

$$L_g \le \frac{B_N R}{B_N + R}. \tag{108}$$

When $B_N + R = 0$, one has $L_g = 0$. Consequently, for any finite population risk upper bound $U \ge R$ and any $0 < \ell \le L_g$,

$$U > \ell, \qquad B_N \ge \frac{U\ell}{U - \ell}. \tag{109}$$

One may take $U = \mathbb{E}(g_0(D, X_q) - Z)^2$ for any square-integrable observed-data predictor g_0. A positive population fixed-gain witness $\ell = ta_g - t^2 c_g$ also satisfies the premise $\ell \le L_g$.

Proof. Let $m_S = \mathbb{E}[m \mid S]$, $e = m - m_S$, and $\eta = Z - m$. Conditional expectation gives $\mathbb{E}[e\eta] = 0$, $\mathbb{E}e^2 = B_N$, and $\mathbb{E}\eta^2 = R$. Since K is reversible and preserves both S and Z, the drift $d_g = (I-K)g$ is orthogonal to m_S and Z. Thus, for every $\lambda \in \mathbb{R}$,

$$\frac{a_g}{2} = \mathbb{E}\big[\{m - \lambda Z - (1-\lambda)m_S\}d_g\big] = \mathbb{E}\big[\{(1-\lambda)e - \lambda\eta\}d_g\big].$$

If $c_g > 0$, Cauchy–Schwarz gives

$$L_g \le (1-\lambda)^2 B_N + \lambda^2 R.$$

For $B_N + R > 0$, minimizing at $\lambda = B_N/(B_N + R)$ proves (108). If $B_N + R = 0$, the same inequality gives $L_g = 0$. The case $c_g = 0$ is immediate from Theorem N.2. A positive $\ell \le L_g$ forces $B_N, R > 0$ and $\ell \le B_N R/(B_N + R) < R \le U$. Rearranging yields $B_N \ge R\ell/(R - \ell)$. For fixed positive ℓ, the function $u \mapsto u\ell/(u - \ell)$ decreases on (ℓ, ∞), so replacing R by $U \ge R$ proves (109). Finally, the Bayes risk is bounded above by every observed-data predictor's risk, and completing the square gives $ta_g - t^2 c_g \le L_g$. □

The sharper bound can use the population risk of a separate predictor without evaluating the full-data Bayes risk. For finite-bank inference, one needs joint uncertainty control for a lower bound on the witness and an upper bound on that risk; the numerical experiments here report the original fixed-gain witnesses.

No claim that d_g is observed-data measurable is needed: it is a test function in the full simulation L^2 space, orthogonal to the summary space. This distinction is essential when the transition fixes a latent coefficient. The statistic gives a lower bound on a property of the observed-data experiment, rather than a predictor that uses the latent coefficient at deployment.

Independent branches estimate the squared conditional drift. For each independent original task Ω_i, simulate $k \geq 2$ conditionally independent branches $\Omega'_{ij} \sim K(\Omega_i, \cdot)$, always restarting at the same full state. Put

$$D_{ij} = g(\Omega_i) - g(\Omega'_{ij}), \qquad \overline{D}_i = k^{-1} \sum_{j=1}^{k} D_{ij},$$

and compute

$$A_i = 2\{m(\Omega_i) - r(\Omega_i)\}\overline{D}_i, \tag{110}$$

$$C_i = \frac{(\sum_j D_{ij})^2 - \sum_j D_{ij}^2}{k(k-1)}. \tag{111}$$

Conditional independence gives $\mathbb{E}[\overline{D}_i \mid \Omega_i] = d_g(\Omega_i)$ and $\mathbb{E}[C_i \mid \Omega_i] = d_g(\Omega_i)^2$. Consequently $\mathbb{E}A_i = a_g$ and $\mathbb{E}C_i = c_g$. For a fixed t, $W_i(t) = tA_i - t^2 C_i$ is an unbiased estimator of a population lower bound on B_N. The n original tasks are the independent statistical units, not the nk branches. Rejected proposals remain zero-displacement branches. A branch may itself contain a fixed number of transitions, but all branches must start independently from the same original state, using the same specified kernel.

Individual C_i can be negative even though $c_g \geq 0$. The ratio $\overline{A}^2/(4\overline{C})$ is therefore neither automatically stable nor a confidence bound. A useful protocol chooses t on an independent calibration bank and then estimates $\mathbb{E}W_i(t)$ on a fresh confirmation bank. Fixing t this way removes the need to estimate a ratio on the confirmation data; ordinary uncertainty estimates still require their own moment or boundedness conditions. Alternatively, a predetermined finite set of t values can use simultaneous confidence bounds with a stated multiplicity correction.

For example, $m, r, g \in L^4(\mu)$ is a sufficient condition for finite variances of A_i, C_i, hence for the iid central limit theorem applied to the fixed-t task witness. Cauchy–Schwarz bounds the product in A_i^2, while products of squared branch differences in C_i^2 are bounded using their fourth moments and equal endpoint marginals. Square integrability alone guarantees unbiasedness and integrability, not these variances. In the present Gaussian-task, exponential-power-noise experiments the target has all moments, $m \in L^4$ follows from conditional Jensen, and the ridge prediction obeys $|r| \leq \frac{1}{2}\|X_q\|\|Y\|$. The frozen final-RMSNorm readout is bounded, so the sufficient variance conditions hold. The resulting normal Monte Carlo intervals are asymptotic in the number of independent tasks; they are not nonasymptotic confidence bounds.

Numerical posterior error. Suppose the original endpoint has a certified numerical approximation $|\widetilde{m}_i - m_i| \leq \epsilon_i$. An observable conservative witness is

$$\widetilde{W}_i(t) = 2t(\widetilde{m}_i - r_i)\overline{D}_i - 2|t|\epsilon_i\, k^{-1} \sum_j |D_{ij}| - t^2 C_i. \tag{112}$$

On the certification event this is no greater than the corresponding exact-posterior $W_i(t)$, because $|\overline{D}_i| \leq k^{-1}\sum_j |D_{ij}|$. If certification holds as a function of the original task, then $\mathbb{E}\widetilde{W}_i(t) \leq ta_g - t^2 c_g \leq B_N$. For population L^2 error $\|\widetilde{m} - m\|_2 \leq \epsilon$, one can instead use

$$|\widetilde{a}_g - a_g| \leq 2\epsilon\sqrt{c_g}, \qquad B_N \geq \left(\frac{|\widetilde{a}_g|}{2\sqrt{c_g}} - \epsilon\right)_+^2,$$

where $\widetilde{a}_g = 2\mathbb{E}[(\widetilde{m} - r)d_g]$. These formulas do not replace an established quadrature error guarantee.

A bounded, fully finite-bank version. An explicit route to a finite-bank bound is available when the original posterior residual $H = m - r$ has a known second-moment bound $\mathbb{E}H^2 \leq V_H$, and the diagnostic predictor is bounded by $|g| \leq G$. For example, a frozen predictor may be clipped in advance, and taking $r = 0$ gives $V_H = \mathbb{E}Z^2$ by conditional Jensen. For a threshold $M > 0$ fixed in advance, replace H in (110) by $H_M = \max(-M, \min(H, M))$. Since $(|x| - M)_+ \leq x^2/(4M)$ and $|d_g| \leq 2G$,

$$|2t\mathbb{E}[(H - H_M)d_g]| \leq \frac{|t|GV_H}{M}.$$

The clipped task witness $W_i^M = 2tH_{M,i}\overline{D}_i - t^2C_i$ therefore satisfies

$$B_N \geq \mathbb{E}W_i^M - \frac{|t|GV_H}{M}.$$

Moreover $|W_i^M| \leq R := 4|t|MG + 4t^2G^2$ because $|D_{ij}| \leq 2G$ and $|C_i| \leq 4G^2$. For n independent original tasks, Hoeffding gives, with confidence $1-\delta$,

$$B_N \geq \left[\overline{W}^M - \frac{|t|GV_H}{M} - 2R\sqrt{\frac{\log(1/\delta)}{2n}}\right]_+. \tag{113}$$

If the numerical posterior is used, clip $\widetilde{m} - r$ to the same interval and subtract the penalty from (112); clipping is one-Lipschitz, so the pointwise comparison remains valid. Apply Hoeffding to the unobserved exact W_i^M and lower-bound its sample mean by the numerical conservative sample mean. A simultaneous numerical-certification failure probability δ_{num} adds to the statistical failure probability. This general bound may be conservative; it separates mathematical finite-sample coverage from ordinary Monte Carlo intervals rather than claiming that every small diagnostic bank yields a positive certificate.

O Independent finite-context Transformer drift confirmation

Target and frozen predictor. The target is $B_N = R^*_{\mathrm{Gram},N} - R^*_{\mathrm{full},N}$ at $N = 128, d = 2$, under the Gaussian task prior and covariates and unit-variance noise used in the standard-Transformer study. The primary diagnostic function g is the arithmetic mean of the predictions of the three unchanged final Transformer checkpoints, with training seeds 26091611/12/13. This differs from averaging their squared prediction losses in the earlier risk comparison. Each individual predictor is also evaluated secondarily, using the same law-specific calibration scalar. No checkpoint selection or model training occurs in this experiment. The known latent coefficient is used only by the Metropolis generator, never by the model or numerical posterior.

Calibration and fresh confirmation. We use the earlier, already inspected 256-task $N = 128$ bank for each of Laplace and quartic noise as explicitly post-hoc calibration data. New proposal generators have seed $2609189100 + 1000p$. Each task receives $k = 16$ conditionally independent branches of the single-step $\theta = 0.25$ kernel in Proposition 5.1; every branch restarts from the identical original (W, X, Y, X_q), with independent proposal and acceptance draws. These are not successive states of a Markov chain. Rejections reuse the original prediction exactly. With ridge as the Gram-measurable reference r, compute A_i, C_i by (110)–(111). Before observing these new calibration branches we fix the rule

$$t = \begin{cases} \mathrm{clip}_{[-100,100]}\left(\dfrac{\overline{A}}{2\max\{\overline{C} + 1.96\,\mathrm{SE}(C), 10^{-12}\}}\right), & \overline{C} > 0, \\ 0, & \overline{C} \leq 0. \end{cases}$$

This gives $t = 7.3071055428$ for Laplace and 7.7782532126 for quartic. The scalar may depend arbitrarily on the independent calibration bank; the population inequality then holds conditional on that bank.

The confirmation protocol is frozen before generating new tasks. Fresh task seeds are $2609189200 + 1000p$ and proposal seeds are $2609189300 + 1000p$. There are 1,024 original tasks for each non-Gaussian law and 256 for Gaussian noise, again with 16 restarted branches each. The Gaussian control uses a predetermined $t = 1$, rather than a scalar selected to vanish: its exact posterior equals ridge, so $A_i = 0$ and the population variational witness is $-c_g \leq 0$. All 2,304 original tasks and all 36,864 confirmation proposals are retained. Acceptance counts are 4,115/16,384 for Laplace, 6,904/16,384 for quartic, and 4,096/4,096 for Gaussian. No law, task, branch count, scalar or rejection is selected using confirmation outcomes.

Estimator and statistical uncertainty. For each original task we evaluate $W_i = tA_i - t^2C_i$. Its expectation, with the exact posterior, is the theoretical lower witness. The confirmation mean is not itself a deterministic bound on B_N. Negative C_i values are allowed by the distinct-branch U-statistic; replacing them by zero or squaring the branch mean would change the estimator. Independent

original tasks, not branches or trained seeds, are the statistical units. Primary intervals are $\overline{W} \pm 1.95996398454\, s_W/\sqrt{M}$, conditional on the calibration and fitted models. Multiplying by $N/d = 64$ gives the reported normalization. These are pointwise Monte Carlo approximations, without simultaneous coverage across laws or seeds. The moment conditions for this approximation are discussed in Section N.2; normal coverage is asymptotic in the number of evaluation tasks, even though the population inequality is exact at finite context length.

Table 12: All frozen-predictor confirmation results. Each law uses the same calibration scalar for the ensemble and all three individual seeds. Values and pointwise normal intervals are normalized by $N/d = 64$.

Law	Predictor	Witness estimate	95% interval
Laplace	Seed 26091611	0.067272	[0.036691, 0.097853]
Laplace	Seed 26091612	0.063306	[0.033414, 0.093197]
Laplace	Seed 26091613	0.061580	[0.030826, 0.092334]
Laplace	Mean predictor	0.066170	[0.036230, 0.096110]
Quartic	Seed 26091611	0.117968	[0.075167, 0.160769]
Quartic	Seed 26091612	0.104389	[0.068803, 0.139976]
Quartic	Seed 26091613	0.111058	[0.072091, 0.150026]
Quartic	Mean predictor	0.113776	[0.075213, 0.152339]
Gaussian	Seed 26091611	−0.000027	[−0.000051, −0.000004]
Gaussian	Seed 26091612	−0.000022	[−0.000045, 0.000001]
Gaussian	Seed 26091613	−0.000009	[−0.000020, 0.000002]
Gaussian	Mean predictor	−0.000013	[−0.000029, 0.000003]

The raw-scale ensemble estimates are 0.00103391 for Laplace, 0.00177775 for quartic, and -2.07016×10^{-7} for Gaussian. All three single-seed intervals are positive under both non-Gaussian laws. These statements concern the information loss of the summary; they do not assert that the corresponding trained predictor attains the full-data Bayes risk or beats every Gram decoder in total risk.

Numerical posterior and implementation checks. All inference uses CPU float64 and the original normalization epsilon. The original-point predictions on the calibration tasks match the six saved non-Gaussian model predictions bitwise. Only original-prompt posterior means are needed; branch posteriors are not computed. Gaussian means are analytic ridge. For other laws, primary integration uses tolerance 10^{-4}, radii 8/12, and grids through 1,025 points per axis; strict integration uses 10^{-5}, radii 10/14, and grids through 2,049 points. Tolerances tighten with query norm as in Section M. Both passes must converge and their prediction difference must not exceed 0.00025.

There are zero posterior-check failures. The largest primary/strict prediction difference is 3.7934×10^{-5} for Laplace and 1.7764×10^{-14} for quartic. The corresponding changes in the normalized ensemble witness means are 4.5401×10^{-7} and -6.11×10^{-16}. The maximum relative Gram discrepancy over confirmation proposals is 1.970×10^{-15}. Independent saved-array checks reproduce likelihood ratios, every acceptance decision, the angle relative to each original residual, Gram preservation, exact rejection reuse, distinct-branch products, the fixed ensemble, the frozen scalars, and the primary means and standard errors. Numerical source and checkpoint hashes remain fixed. Agreement between quadrature settings is a sensitivity check, not a rigorous posterior-error certificate; the numerical-error bounds in Section N.2 require an independently justified error bound.

Heavy tails and secondary sensitivity. As a secondary check we bootstrap original tasks 10,000 times, retaining all branches and predictors within each selected task. Confirmation bootstrap seeds are $2609189400 + 1000p$. The normalized ensemble percentile intervals are $[0.03596, 0.09603]$ for Laplace, $[0.07969, 0.15608]$ for quartic, and $[-0.00003144, 0.00000022]$ for Gaussian. These checks do not replace the prespecified normal intervals or establish exact coverage.

Task contributions are heavy-tailed. Laplace has empirical excess kurtosis 46.5; its largest positive and negative task contributions to the reported mean are 0.005253 and −0.005150. Its largest centered squared contribution accounts for 11.7% of the total. Quartic has skewness 14.1 and excess kurtosis 318.7; its largest task contributes 0.014776 to the mean and accounts for 55.6% of the centered sum

of squares. The Gaussian control also has heavy-tailed contributions (excess kurtosis 104.4), despite its exactly zero numerator. Thus the small Monte Carlo banks do not support a claim of precisely Gaussian sampling errors. All tasks remain in the primary calculation. As descriptive influence checks, deleting any single task leaves the Laplace mean within $[0.06098, 0.07139]$ and the quartic mean within $[0.09910, 0.11605]$; these ranges are not alternate estimators or confidence intervals. The complete quantiles and all seed-specific diagnostics are retained with the task arrays.

Reproduction and cost. The evidence archive contains the frozen protocols, executable simulator, source snapshot, shared inference-only checkpoints, all proposal and prediction arrays, task-level posterior histories, independent verifier, and analysis scripts. The two non-Gaussian laws run in separate CPU processes with one numerical thread each; the Gaussian control uses a freed process slot. No model retraining or GPU computation is required. Recorded wall times include periods with little accumulated CPU time, consistent with host suspension or scheduling pauses; they should not be interpreted as uninterrupted algorithm runtimes. The replay instructions separate unchanged execution data from anonymized host metadata and record every export transformation.

O.1 Independent replication under heavy tails

After inspecting the original $1{,}024$-task confirmations, we froze a separate $4{,}096$-task replication for each of $p \in \{1, 4\}$. The fixed three-model ensemble, $N = 128$, $d = 2$, $K = 16$ independently restarted MH branches, angle 0.25, numerical quadrature, and coefficients $t_1 = 7.307105542760461$ and $t_4 = 7.7782532125503545$ were unchanged. Task and proposal seeds are $26091815200 + 1000p$ and $26091816200 + 1000p$. All rejected branches and negative contributions are retained.

The primary statistic is the uncapped mean of $\widetilde{W}_i = (N/d)[2t(m_i - r_i)\widehat{d}_i - t^2\widehat{c}_i]$. Approximate normal 95% intervals use independent original tasks; $10{,}000$ task-bootstrap percentile intervals are secondary. Neither interval accounts for uncertified numerical integration error. The fixed models and old calibration are conditioned on.

Table 13: Independent replication, normalized by $N/d = 64$. Raw scale is obtained by division by 64. Capping is a prespecified secondary statistic.

Law	Statistic	Mean	Normal 95%	Bootstrap 95%
$p = 1$	Primary	0.0666	$[0.0529, 0.0802]$	$[0.0530, 0.0802]$
$p = 1$	Upper cap	0.0553	$[0.0436, 0.0670]$	$[0.0435, 0.0665]$
$p = 4$	Primary	0.0950	$[0.0825, 0.1076]$	$[0.0825, 0.1075]$
$p = 4$	Upper cap	0.0865	$[0.0755, 0.0974]$	$[0.0756, 0.0973]$

The upper caps $L_1 = 1.550484757998018$ and $L_4 = 1.615409970871054$ were fixed before replication outcomes from the nonnegative part of the 99th percentile of the original 256-task calibration contributions. Only positive upper tails are capped: $\min(\widetilde{W}_i, L_p) \le \widetilde{W}_i$, so the capped population expectation remains a lower witness. Every negative contribution is unchanged. The capped result does not replace the primary raw mean.

Table 14: Secondary fixed-model results on the independent replication. Same fixed t_p for all models; task normal intervals.

Law	Predictor	Raw mean	Normalized mean [normal 95%]	Bootstrap 95%
$p = 1$	Seed 26091611	0.0010242	0.0655 $[0.0511, 0.0800]$	$[0.0512, 0.0799]$
$p = 1$	Seed 26091612	0.0009633	0.0617 $[0.0478, 0.0755]$	$[0.0479, 0.0753]$
$p = 1$	Seed 26091613	0.0010460	0.0669 $[0.0531, 0.0808]$	$[0.0532, 0.0807]$
$p = 1$	ensemble	0.0010400	0.0666 $[0.0529, 0.0802]$	$[0.0530, 0.0802]$
$p = 4$	Seed 26091611	0.0014372	0.0920 $[0.0792, 0.1048]$	$[0.0794, 0.1047]$
$p = 4$	Seed 26091612	0.0014818	0.0948 $[0.0823, 0.1074]$	$[0.0824, 0.1074]$
$p = 4$	Seed 26091613	0.0014446	0.0925 $[0.0794, 0.1056]$	$[0.0793, 0.1052]$
$p = 4$	ensemble	0.0014849	0.0950 $[0.0825, 0.1076]$	$[0.0825, 0.1075]$

For $p = 1$, the acceptance rate was 24.84%; there were 0 numerical posterior failures. The maximum primary/strict posterior prediction difference was 4.91×10^{-5} and the maximum relative Gram

Table 15: Secondary upper-capped replication results for every fixed predictor. All intervals are normalized task intervals.

Law	Predictor	Capped tasks	Mean	Normal 95%	Bootstrap 95%
$p = 1$	Seed 26091611	53	0.0531	$[0.0407, 0.0654]$	$[0.0407, 0.0650]$
$p = 1$	Seed 26091612	48	0.0502	$[0.0383, 0.0620]$	$[0.0384, 0.0616]$
$p = 1$	Seed 26091613	47	0.0561	$[0.0441, 0.0681]$	$[0.0439, 0.0677]$
$p = 1$	ensemble	47	0.0553	$[0.0436, 0.0670]$	$[0.0435, 0.0665]$
$p = 4$	Seed 26091611	48	0.0832	$[0.0721, 0.0943]$	$[0.0720, 0.0941]$
$p = 4$	Seed 26091612	48	0.0863	$[0.0753, 0.0972]$	$[0.0754, 0.0973]$
$p = 4$	Seed 26091613	45	0.0835	$[0.0721, 0.0950]$	$[0.0719, 0.0948]$
$p = 4$	ensemble	46	0.0865	$[0.0755, 0.0974]$	$[0.0756, 0.0973]$

discrepancy was 2×10^{-15}. Uncapped task skewness/excess kurtosis were $0.27/64.53$, and the largest task accounted for 9.33% of centered sum of squares. Removing any one task gave normalized means in $[0.0651, 0.0687]$. All tasks were retained in the reported results.

Primary versus strict quadrature changes the normalized ensemble mean by 2.99×10^{-7}, with maximum normalized task-contribution change 0.000648. This comparison is not a certified numerical error bound.

For $p = 4$, the acceptance rate was 41.83%; there were 0 numerical posterior failures. The maximum primary/strict posterior prediction difference was 1.15×10^{-14} and the maximum relative Gram discrepancy was 1.98×10^{-15}. Uncapped task skewness/excess kurtosis were $2.39/34.71$, and the largest task accounted for 5.80% of centered sum of squares. Removing any one task gave normalized means in $[0.0935, 0.0961]$. All tasks were retained in the reported results.

Primary versus strict quadrature changes the normalized ensemble mean by 9.71×10^{-17}, with maximum normalized task-contribution change 1.45×10^{-13}. This comparison is not a certified numerical error bound.

Declared secondary pooling. The original $1{,}024$ confirmations and independent $4{,}096$ replication tasks are concatenated without reweighting, for $5{,}120$ per law. This analysis is secondary and does not replace the independent replication.

For $p = 1$, the pooled normalized uncapped mean is 0.0665 with normal 95% interval $[0.0540, 0.0790]$; the capped mean is 0.0547 with interval $[0.0441, 0.0653]$.

For $p = 4$, the pooled normalized uncapped mean is 0.0988 with normal 95% interval $[0.0861, 0.1114]$; the capped mean is 0.0869 with interval $[0.0772, 0.0967]$.

All per-seed statistics, raw scales, bootstrap draws, numerical histories, rejected branches, and independent arithmetic checks are included in the evidence archive. No extra Gaussian run was planned; the original analytic $a = 0$ negative control is preserved. These lower witnesses do not identify the exact Bayes gap or the fraction of a separate risk contrast explained by that gap.

Execution scheduling. A recorded resource-only change used up to four numerical processes with disjoint task shards. Completed records were preserved, and bitwise checks confirmed unchanged posterior calculations. Full timing, coverage and record hashes are archived; resumed aggregation time is not total computation time.