TASK GEOMETRY: RESPONSE CHANNELS PREDICT EXPLICIT RECOVERY AND ATTENTION LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning on fresh low-dimensional tasks requires attention to infer which input directions define task geometry. We study an explicit response-derived route to that geometry. A clipped second-moment response channel yields cross-fitted quadratic-softmax attention with log-free, dimension-explicit subspace recovery and the intrinsic point-query rate when the channel is open; when it is closed, that statistic is non-identifying even though its finite-sample geometry need not be Haar. In a controlled replication with frozen hypotheses on ten synthetic tasks at $d = 16$ and $N = 256$, the mean Haar-minus-explicit geometry advantage is $0.608\,[0.600, 0.617]$ larger on six open tasks than on three closed tasks, with one preregistered numerically weak task outside the contrast. On the three closed odd-link tasks, a small task-specific Transformer attains lower prediction risk than both the Haar control and the explicit route. This trained result is predictive rather than geometric: it uses task information beyond the selected second-moment statistic, while an open first-moment channel provides a classical alternative explanation. Together, response channels predict explicit geometry recovery and separate it from broader task-specific prediction.

1 INTRODUCTION

An attention head decides which context points count as neighbours of a query. When a task's response depends on only r of the d input directions, and those directions change from one context to the next, the head cannot know in advance where to localize. What is fresh task geometry worth, and what in the responses determines whether an explicit attention rule can recover it?

Theory of attention as an in-context learner mostly fixes the target class and asks whether a Transformer attains its rate (Kim et al., 2024; Ching et al., 2026); low-dimensional structure, when it enters, is usually shared across tasks and absorbed in pretraining (Oko et al., 2024). Geometric attention designs reshape the neighbourhood metric (Nielsen et al., 2024; Kuang et al., 2025), while classical sufficient dimension reduction shows that response moments can identify predictive directions and that particular moments have systematic blind spots (Li, 1991; 1992; Cook, 1998). Concurrent work supplies the complementary information-theoretic background: for fresh hidden-ridge tasks, task geometry is a statistical resource whose information budget controls the attainable localization volume of positive response-blind attention (Anonymous, 2026).

Background and current contributions. We use the geometry–risk frontier of Anonymous (2026) as attributed background and restate only the result needed here (Section 3); its proof remains with that concurrent work. The same work also introduced the earlier cross-fitted response-moment attention construction and its single radial-link comparison. This paper adds: a deterministic fixed clipping level with a derived eigengap and log-free, dimension-explicit recovery/risk analysis including the boundary $2s + r = 2$; a comparison-selected quadratic-softmax predictor that does not require an open second-moment channel; a zero-channel counterexample showing that non-identification need not yield Haar geometry; and the controlled ten-task explicit-versus-trained study below. We use the recovered-projector information count for conservative asymptotic accounting, since its constants and quantifiers differ from the finite experiment.

We build on that background with three contributions. **(1) Explicit response routes.** For bounded Hölder links with an identifying clipped second-moment channel, Theorem 2 gives fixed clipping,

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.

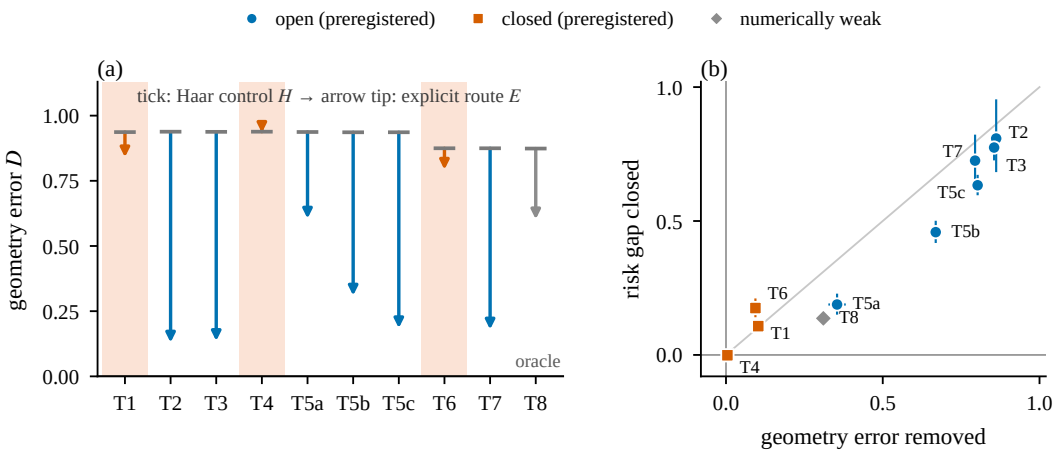


Figure 1: **Controlled replication of the explicit channel prediction.** (a) Mean geometry error of the matched Haar control (tick) and the explicit route (arrow tip), in frozen task order. (b) Share of Haar geometry error removed against share of the Haar-to-oracle risk gap closed. Labels and group contrasts were frozen before the current outcomes; because the same open/closed split had appeared in an earlier unpublished estimator benchmark, H1 is interpreted as a controlled replication with a matched Haar reference.

log-free projector recovery and intrinsic-rate point prediction. A complementary comparison-selected construction, proved in Appendix B, compares candidate smoothers rather than requiring projector consistency, so accurate prediction need not be tied to this one eigengap. Proposition 1 shows that a zero population channel can remain non-Haar in finite samples. **(2) Controlled channel test.** On ten synthetic task cases, the frozen channel map predicts where the explicit route recovers geometry against a matched Haar-projector control. The open-minus-closed contrast is $0.608\ [0.600, 0.617]$ in geometry and $0.277\ [0.258, 0.297]$ in risk; all six open tasks separate geometrically, including the genuine rank-2 interaction T7. Because several tasks fall outside the response theorem's assumptions, this experiment tests the channel mechanism at one operating point rather than its finite-sample rate. **(3) Trained prediction beyond the selected statistic.** On the three closed odd-link tasks, a small task-specific Transformer has lower mean risk than both controls, by $0.181\ [0.178, 0.184]$ versus Haar and $0.162\ [0.158, 0.165]$ versus the explicit route, with intervals over contexts for three fixed models. The trained result is predictive rather than geometric: no projector is observed, and the genuine rank-2 trained test T7 is negative. The same closed tasks have an open first-moment channel, providing a classical alternative explanation that this experiment does not distinguish.

H1/H2 and the rank-2 components were frozen before the current outcomes, with 512 contexts per task and simultaneous context-level intervals. Because the same task split had appeared in an earlier benchmark, H1 is a controlled replication. Post-hoc analyses are labelled throughout, and the post-outcome learning-rate correction is documented with the retained sensitivity run in Appendix D.5.

2 Attention as a localization rule

An attention head that answers a query x_0 from a context $\{(X_i, Y_i)\}_{i \le n}$ returns an offset a_0 plus a convex combination of the context values (a *positive normalized* row),

$$\widehat{f}(x_0) = a_0 + \sum_{i=1}^{n} a_i Y_i, \qquad a_i \ge 0, \qquad \sum_{i=1}^{n} a_i = 1. \tag{1}$$

Its canonical instance is the quadratic-softmax row

$$a_i(A) = \frac{\exp[-\frac{1}{2}(X_i - x_0)^\top A (X_i - x_0)]}{\sum_j \exp[-\frac{1}{2}(X_j - x_0)^\top A (X_j - x_0)]}, \qquad A \succeq 0, \tag{2}$$

which a dot-product softmax realizes exactly after explicit quadratic feature augmentation. The precision A is the head's localization geometry: its eigenvectors orient the neighbourhood of x_0 and

its eigenvalues set its scale in each direction. For standard Gaussian keys and $x_0 = 0$, the expected unnormalized kernel mass is $\mu(A) = \det(I_d + A)^{-1/2}$. The isotropic choice $A = h^{-2} I_d$ gives $\mu \asymp h^d$; the aligned choice $A = h^{-2}P + b^{-2}(I_d - P)$ for a rank-r projector P gives $\mu \asymp h^r b^{d-r}$, hence $\mu \asymp h^r$ when the transverse scale b is of order one (Figure 2a; the corresponding determinant criterion is given in Appendix B.5).

Fresh low-dimensional tasks. Each context is a fresh task. Its geometry is a projector $P = BB^\top$ drawn from the Haar (uniform) law on the Grassmannian $\mathrm{Gr}(d, r)$, with orthonormal frame $B \in \mathrm{St}(d, r)$ and $1 \le r < d$. Given the task,

$$X_i \overset{\text{iid}}{\sim} N(0, I_d), \qquad Y_i = g(B^\top X_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{3}$$

The link $g : \mathbb{R}^r \to \mathbb{R}$ lies in the class $\mathcal{H}_{s,r}(L, G)$ of functions bounded by G and s-Hölder with constant L, $0 < s \le 1$. For identifiability we use $\mathcal{G}_{s,r}(L, G, \gamma) = \{g \in \mathcal{H}_{s,r}(L, G) : s_r(\Gamma_g) \ge \gamma\}$, where $\Gamma_g = \mathbb{E}[g(Z)(ZZ^\top - I_r)]$, $Z \sim N(0, I_r)$, and s_r is the r-th singular value. Rotation-invariant keys carry no information about P, so geometry must come from side information or the responses.

Geometry as information. Before the current responses are revealed, the head may use a public variable U, independent of the task, and a geometry signal $\mathcal{S}$ drawn from a conditional law $\mathbb{P}_{\mathcal{S}|P,U}$, independently of the design, the noise and the query, with

$$I(P; \mathcal{S} \mid U) \le \mathcal{I} \quad \text{nats, evaluated under Haar } P. \tag{4}$$

A *response-blind* row chooses a_0 and $(a_i)_{i \ge 1}$ in (1) as measurable functions of $(x_0, X_{1:n}, U, \mathcal{S})$ only; $\mathcal{A}_+$ denotes these rows. At the point query $x_0 = 0$, the geometry-constrained minimax risk is

$$\mathfrak{R}_n(\mathcal{I}) = \inf_{\mathbb{P}_{\mathcal{S}|P,U} :\, I(P;\mathcal{S}|U) \le \mathcal{I}} \; \inf_{\widehat{f} \in \mathcal{A}_+} \; \sup_{P \in \mathrm{Gr}(d,r)} \; \sup_{g \in \mathcal{G}_{s,r}(L,G,\gamma)} \mathbb{E}\big[\{\widehat{f}(0) - g(0)\}^2\big]. \tag{5}$$

The signal law is chosen outside the minimax problem, the budget is averaged over Haar P, and the risk is worst case over (P, g).

We measure recovered geometry by $D(P, \widehat{P}) = \|P - \widehat{P}\|_F^2/(2r) = 1 - \mathrm{tr}(P\widehat{P})/r$ for rank-r projectors. An independent Haar projector has $\mathbb{E}D = 1 - r/d$.

3 Background: the geometry–risk frontier

We use the geometry–risk frontier of Anonymous (2026) as background. Localizing at scale h along the r predictive directions costs squared bias of order $L^2 h^{2s}$, and a normalized row has variance of order $\sigma^2/(n\mu)$, set by the neighbourhood mass μ. A head that does not know P must also be narrow in the $d - r$ irrelevant directions, so $\mu \asymp h^d$; a head that knows P can stay wide across them, so $\mu \asymp h^r$. Partial knowledge interpolates. If a geometry signal locates P within one of $K = e^{\mathcal{I}}$ quasi-uniform Grassmann cells, the transverse scale can grow to $b = hK^{1/(r(d-r))}$ while projected displacement remains $O(h)$, giving

$$V_{\text{eff}}(h, \mathcal{I}) \asymp \min\{h^r,\ h^d e^{\mathcal{I}/r}\}. \tag{6}$$

Balancing bias and variance yields a saturation information

$$\mathcal{I}_\star = \frac{r(d-r)}{2s + r} \log n. \tag{7}$$

Theorem 1 (Geometry–risk frontier; Anonymous, 2026)**.** *Fix finite $1 \le r < d$, $0 < s \le 1$ and $\sigma > 0$, and let $\gamma \le \kappa_{r,s} \min\{L, G\}$, where $\kappa_{r,s} I_r = \mathbb{E}[\rho(Z)(ZZ^\top - I_r)]$ for $\rho(z) = 1 \wedge \|z\|^s$; this condition keeps the standard radial hard family inside the model class. Consider (3) with the positive normalized point-query response-blind rows and pre-response geometry signal defined in Section 2, and worst-case risk over P and $g \in \mathcal{G}_{s,r}(L, G, \gamma)$ as in (5). There are n_0 and constants depending only on $(d, r, s, L, G, \gamma, \sigma)$ such that, for all $n \ge n_0$ and all $\mathcal{I} \ge 0$,*

$$\mathfrak{R}_n(\mathcal{I}) \asymp \max\Big\{ n^{-2s/(2s+r)},\ n^{-2s/(2s+d)} \exp\Big[-\frac{2s\mathcal{I}}{r(2s+d)} \Big] \Big\}. \tag{8}$$

The branches meet at $\mathcal{I}_\star$ in (7), which is sharp at rate level up to $O(1)$ nats. The proof is due to Anonymous (2026) and is not reproduced here.

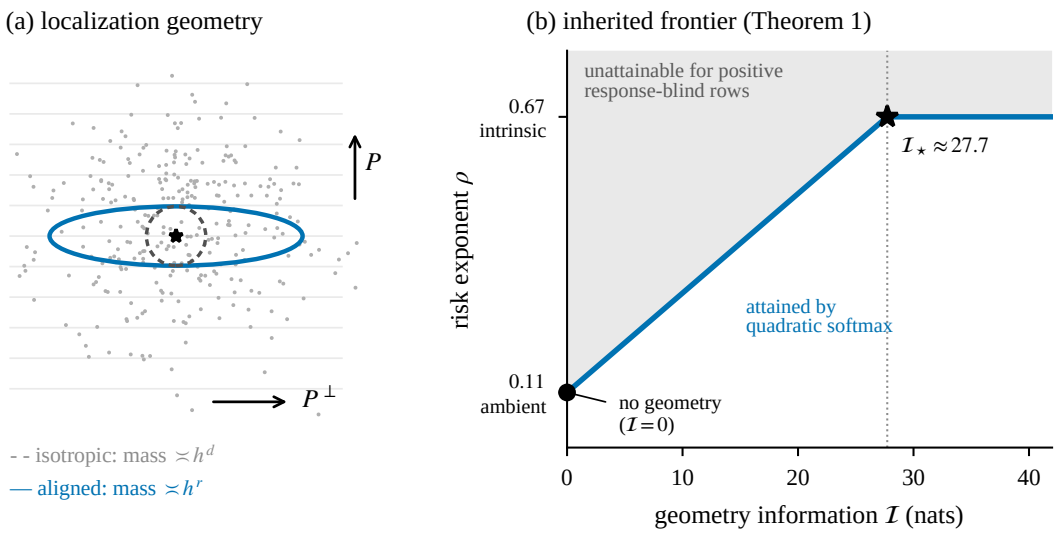


Figure 2: **Background resource picture.** (a) Conceptual isotropic versus subspace-aligned localization. (b) The inherited geometry–risk frontier of Anonymous (2026), shown as background at $d = 16$, $r = 1$, $s = 1$, $n = 256$; constants and $O(1)$ nats are suppressed.

The same work gives a quadratic-softmax row attaining the frontier: with a Grassmann code of $K \approx e^{\mathcal{I}}$ cells and transverse scale $b = hK^{1/(r(d-r))}$, its mass is $\min\{h^d e^{\mathcal{I}/r}, h^r\}$ up to constants. The result applies to positive normalized response-blind rows and to the signal law specified in its minimax formulation, pricing geometry available before row construction. The task-trained Transformer in Section 6 and the response-derived information count in Appendix B.1.4 have different quantifiers and are analyzed separately.

4 Where geometry comes from: response channels

Theorem 1 prices geometry but does not say where a fresh task's geometry comes from. Keys are rotation invariant, so any orientation information must be read from the responses; the rest of the paper asks which response statistics supply it. A classical statistic does this. By Stein's identity for Gaussian keys, every scalar map ϕ gives

$$\mathsf{H}_\phi := \mathbb{E}\big[(\phi(Y) - \mathbb{E}\phi(Y))(XX^\top - I_d)\big] = B\,\Gamma_\phi B^\top, \qquad \Gamma_\phi := \mathbb{E}\big[\phi(Y)(ZZ^\top - I_r)\big], \tag{9}$$

with $Z = B^\top X$; its range lies in the predictive subspace. We call H_ϕ the *explicit second-moment channel*; for $\phi(y) = y$ it is the principal-Hessian-directions (pHd) moment (Li, 1992; Cook, 1998).

Definition 1 (Channel status)**.** For a task and a map ϕ, the channel is *open* if $\operatorname{rank}\Gamma_\phi = r$, *partial* if $0 < \operatorname{rank}\Gamma_\phi < r$, and *closed* if $\Gamma_\phi = 0$. Its strength is $s_r(\Gamma_\phi)$. Separately, a channel is flagged *numerically weak* when a nonzero singular value of Γ_ϕ is too small to be estimated from one context; here it is a judgement recorded in the frozen channel map (Section 5).

The explicit route, and what an open channel buys. Clip the responses of a pilot sample, centre them by their sample mean (*pilot-centred*), form the moment $\widehat{\mathsf{H}}_m$ of (9), and estimate P by its r eigenvectors of largest absolute eigenvalue. Then attend with precision $h^{-2}\widehat{P}$, cross-fitting geometry and smoothing on disjoint data. Theorem 2 summarizes the fixed-clipping, log-free projector-recovery and cross-fitted risk results proved in Appendix B. It sharpens the fixed-fold route of Anonymous (2026), which already attains the intrinsic rate for $2s + r > 2$ with a clipping level growing like $\sqrt{\log N}$ and an assumed gap for the clipped moment: here the level is fixed, the clipped gap follows from γ, the constants are explicit, and no logarithmic factor is lost at $2s + r = 2$.

Theorem 2 (Explicit response route; point query)**.** *Assume (3) with $g \in \mathcal{G}_{s,r}(L, G, \gamma)$. Let $\phi = \operatorname{clip}(\cdot, -\tau_*, \tau_*)$, where the deterministic level τ_* derived in Appendix B.1 depends only on (G, σ, γ) and guarantees $s_r(\Gamma_\phi) \geq \gamma/2$. Let $\widehat{P}_m$ be the top-r absolute eigenprojector of the pilot-centred*

clipped moment computed from $m \geq 2$ samples, and let $v^2 = 4G^2 + \sigma^2$. Then: (i) $\mathbb{E}\|\widehat{P}_m - P\|_{\mathrm{op}}^{2s} \leq \min\{1, (C_M v^2 d/(m\gamma^2))^s\}$ for a universal constant C_M; (ii) with $k \geq 2$ folds of size n_{f}, context size $N = kn_{\mathrm{f}}$ and $m = (k-1)n_{\mathrm{f}}$, cross-fitted quadratic softmax with precision $h^{-2}\widehat{P}$ has, for all $0 < h \leq 1$, risk at most $2L^2 r^s h^{2s} + 2L^2 r_0^s \theta_m^s + 4\sigma^2/(n_{\mathrm{f}}\rho_r(h)) + (4G^2+\sigma^2)e^{-n_{\mathrm{f}}\rho_r(h)/8}$, where $\theta_m = \min\{1, C_M v^2 d/(m\gamma^2)\}$, $r_0 = \min(r, d-r)$ and $\rho_r(h) \asymp_r h^r$; (iii) if $2s + r \geq 2$ and $h \asymp N^{-1/(2s+r)}$, the risk is $O(N^{-2s/(2s+r)})$ for $N \geq N_0 < \infty$. (Proof: Appendix B.)

With an open channel, the route thus reaches the intrinsic rate. The geometry term costs $v^2 d/(m\gamma^2)$, so for a fixed noise level and link bound the channel strength γ controls it; Section 5 checks this ordering post hoc. For $2s + r < 2$ (rank 1 with $s < 1/2$) the route has a slower N^{-s} geometry term, and at $2s + r = 2$ only the onset of projector accuracy, not the rate, depends on the parameters (see Appendix B). When $\Gamma_\phi = 0$ it says nothing.

A closed channel limits this route, not what the responses carry. Our comparison selector searches a net of candidate subspaces but scores them by symmetric *prediction comparisons*; it never needs to prove that the selected orientation is close to P. The resulting quadratic-softmax predictor reaches the intrinsic polynomial exponent, up to a logarithmic selection factor, without assuming a second-moment eigengap and hence also on closed tasks, at a search cost exponential in dr_0; the full theorem and proof are in Appendix B.2; implementation details are in Appendix B.4. This construction belongs to the classical lineage of direction aggregation and adaptive index prediction (Gaïffas & Lecué, 2007; Lepski & Serdyukova, 2014; Conn & Li, 2019); its distinct ingredient is the symmetric pair-smoother comparison that keeps likelihood-ratio variance at intrinsic order h^{-r}.

Information accounting for an open channel. After cross-fitting, each fold's row is a positive normalized, response-blind row whose geometry signal is the pilot projector $\widehat{P}$, drawn independently of the fold. Classical Grassmann rate–distortion (Dai et al., 2008; Riegler et al., 2018) turns Theorem 2(i) into the conservative asymptotic lower bound $I(P; \widehat{P}) \geq \frac{r(d-r)}{2} \log \frac{m\gamma^2}{Cv^2 d}$, with $C = C_{d,r}$; the derivation is in Appendix B.1.4. At rate level this reaches the inherited frontier's saturation scale when projector error reaches the squared intrinsic bandwidth. We use the bound only as asymptotic accounting: its conservative constant and different minimax quantifiers preclude numerical calibration to the $m = 192$ experiment.

Channels close by symmetry. Odd links close the channel for every odd map, including clipping. For an odd link, $g(-z) = -g(z)$, with symmetric noise, the reflection $(Z, \varepsilon) \mapsto (-Z, -\varepsilon)$ flips $\phi(Y)$ and leaves $ZZ^\top - I_r$ unchanged. Hence $\Gamma_\phi = 0$: an odd-link task is closed for every odd ϕ, which is the parity blind spot of pHd-type moments. Even links monotone in z^2 keep the channel open. Evenness alone does not suffice: for the bounded even link $|z|e^{-z^2/2}$, $\mathbb{E}[Z^2 \mid g(Z)] = 1$, so $\Gamma_\phi = 0$ for every ϕ. The interaction $z_1 z_2$ is open, because its Γ_ϕ has eigenvalues $\pm c$ and is visible to absolute-eigenvalue selection. Channel status is thus a population property of (task, ϕ), computable before any data; an even link closed for every ϕ, like $|z|e^{-z^2/2}$, has generative exponent at least four (Damian et al., 2024).

A closed channel need not give chance geometry. When $\Gamma_\phi = 0$ one might expect a uniformly random estimated subspace, with $\mathbb{E}D = 1 - r/d$. That expectation is false in general: a zero population moment fails to identify P *through that moment*, but its empirical fluctuations can still carry task-aligned structure.

Proposition 1 (Zero channel, non-Haar geometry)**.** *Let $d = 2$, $r = 1$, $X = (Z, U)^\top \sim N(0, I_2)$, $P = e_1 e_1^\top$, and $Y = Z + \sigma\varepsilon$ with $\sigma \geq 0$ and $\varepsilon \sim N(0,1)$ independent of X. Let $\phi(y) = y$ and $\widehat{\mathsf{H}}_m = m^{-1}\sum_{i \leq m}(\phi(Y_i) - \overline{\phi(Y)})(X_i X_i^\top - I_2)$. Then $\mathsf{H}_\phi = 0$, but $\sqrt{m}\,\widehat{\mathsf{H}}_m$ converges in law to a centred Gaussian symmetric matrix with independent entries of variances $(10 + 2\sigma^2,\ 3 + \sigma^2,\ 2 + 2\sigma^2)$ for the $(1,1)$, $(1,2)$ and $(2,2)$ entries. The absolute-leading eigenprojector $\widehat{P}_m$ satisfies $\lim_{m\to\infty} \mathbb{E}D(P, \widehat{P}_m) < 1/2 = 1 - r/d$. The same conclusion holds for $\phi(y) = \mathrm{clip}(y, -2, 2)$ with $\sigma = 0$, with the corresponding clipped-limit variances. Hence, for all sufficiently large m, the estimated line is not Haar distributed relative to P. (Proof: Appendix C.)*

Section 5 therefore tests against a matched Haar control rather than against $1 - r/d$.

5 OPEN VERSUS CLOSED CHANNELS: A CONTROLLED TEST

Theorem 2 needs an open channel, but several tasks below violate its other hypotheses (T2 and T7 have unbounded links, T6 heteroscedastic noise, T5a has $2s + r < 2$). Accordingly, the experiment targets the channel mechanism at one operating point: does the rank of Γ_ϕ, computed from population quantities for the clipping map used, predict where the explicit route, with fixed hyperparameters, recovers geometry and improves prediction? Labels were frozen before any outcome was evaluated. On this suite every closed label is a parity label, and the same split had appeared with a related $\tau = 2$ clipped moment in our estimator benchmark, so H1 is a controlled replication with a matched Haar control; the non-parity closed link of Section 4 is not in the suite.

Tasks and frozen channel map. All tasks use the design of (3) with $d = 16$, Haar frames, base noise standard deviation 0.25 and $N = 256$ context points (full task formulas and frozen channel labels are in Appendix D.1). The frozen map labels the odd links $\tanh z$ (T1), $\sin 2z$ (T4) and $\tanh z_1$ with z_2-dependent noise (T6; rank 2, conditional mean of rank 1) *closed*; the even links $z^2 - 1$ (T2), $0.75\tanh(z^2 - 1)$ (T3) and $\min(|z|^q, 1)$, $q \in \{0.25, 0.5, 1\}$ (T5a–T5c, Hölder $s = q$), and the interaction $z_1 z_2$ (T7, eigenvalues $\pm c$) *open*; and $\tanh z_1 + 0.5\tanh(z_2^2 - 1)$ (T8) *numerically weak* ($\Gamma_{11} \approx -3 \times 10^{-6}$), excluded from both group contrasts.

Arms and controls. The explicit route E is a four-fold cross-fitted implementation motivated by Theorem 2. The confirmatory suite is deliberately broader than the theorem: among the open tasks, T2 and T7 are unbounded and T5a lies below the $2s + r \geq 2$ rate regime; the empirical clipping level and transverse precision also differ from the theorem. On the bounded open rank-1 tasks T3, T5b, and T5c, the qualitative assumptions are closest to the theorem. E otherwise follows the frozen historical construction. On each fold, $\widehat{P}$ comes from the pilot-centred moment of the other 192 points, clipped at $\tau = 1 + 0.25\sqrt{4\log 192} \approx 2.146$ (the $\sqrt{\log m}$ rule of Anonymous, 2026, not the τ_* of Theorem 2), and the fold's 64 points are the smoothing sample for quadratic softmax with precision $h^{-2}\widehat{P} + \alpha(I - \widehat{P})$, $\alpha = 0.25$ (Theorem 2 uses $\alpha = 0$). The hyperparameters were selected earlier on a separate single-task radial link (Anonymous, 2026) and were not tuned here; full arm definitions are in Appendix D.2. The Haar control H replaces $\widehat{P}$ by an independent Haar rank-r projector, and the oracle O uses the true P. All arms share contexts, folds, smoother and queries, so E versus H isolates the value of the estimated geometry. The theorem motivates this construction, while the experiment evaluates the frozen implementation with its fixed hyperparameters and $\tau \neq \tau_*$.

Outcomes and inference. Each task has 512 fresh contexts with 32 Gaussian queries each. A context's risk R is the query-averaged squared error to the noiseless conditional mean, and D is averaged over folds. The per-task coordinates are $C_D(t) = \mathbb{E}_c[D_H - D_E]$ and $C_{HE}(t) = \mathbb{E}_c[R_H - R_E]$. Hypothesis H1 states that the explicit route gains more over Haar on open tasks than on closed ones; it is tested by the contrasts $K_{\mathrm{H1},D} = \overline{C_D}_{\mathrm{open}} - \overline{C_D}_{\mathrm{closed}}$ and $K_{\mathrm{H1},R}$, defined in the same way from C_{HE}, and holds if both lower bounds exceed zero. Intervals are simultaneous 95% max-t intervals from a 9,999-replicate context bootstrap, simultaneous within each of three preregistered families: risk (34 quantities), geometry (11) and a secondary $R_H - R_O$ family (see Appendix D.4 for the full inferential family).

Result. H1 holds in both geometry and risk (Figure 1, Table 1): $K_{\mathrm{H1},D} = 0.608\ [0.600, 0.617]$ and $K_{\mathrm{H1},R} = 0.277\ [0.258, 0.297]$. Every open task separates from its Haar control in both measures (C_D from 0.33 to 0.81; smallest C_{HE}(T5a) $= 0.0032\ [0.0027, 0.0038]$): the route removes 35–86% of the control's geometry error and closes 19–81% of its risk gap to the oracle (Figure 1b). This includes the rank-2 interaction: on T7, E recovers both directions ($D_E = 0.180$ against 0.875 for Haar), with $C_D = 0.695\ [0.688, 0.702]$ and $C_{HE} = 0.388\ [0.361, 0.416]$. Closed tasks separate much less in geometry. T4 is indistinguishable from Haar in both measures ($C_D = 0.004\ [-0.004, 0.013]$). T1 and T6 lie modestly above Haar ($C_D = 0.096$ and 0.082; $C_{HE} = 0.025$ and 0.033), as Proposition 1 permits: a closed channel removes identification, not all task-aligned fluctuation. The map orders the tasks cleanly in geometry. In risk it holds as a group contrast only: T1 and T6 gain more than open T5a and T5b (0.003 and 0.020), and $K_{\mathrm{H1},R}$ is dominated by T2 ($C_{HE} = 1.18$). As a descriptive sensitivity check, removing T2 leaves the point contrast positive at about 0.100 (not a preregistered coordinate and no interval is attached). A scale-free secondary (post hoc): mean G_E is 0.60 on open

Table 1: **Headline confirmatory quantities** (simultaneous 95% intervals; 512 contexts per task). H1 is the controlled open-versus-closed replication; H2 compares the trained Transformer S with H and E on closed tasks. Intervals involving S cover context sampling conditional on the three fixed trained models. The genuine rank-2 trained test T7 is shown explicitly; the favorable preregistered H3 aggregate is reported for completeness because it is driven by the unattributable T8.

Quantity	Estimand	Versus	Estimate [95% simult.]
$K_{\mathrm{H1},D}$	$\overline{C_{D\mathrm{open}}} - \overline{C_{D\mathrm{closed}}}$	H	$0.608\ [0.600,\ 0.617]$
$K_{\mathrm{H1},R}$	$\overline{C_{HE\mathrm{open}}} - \overline{C_{HE\mathrm{closed}}}$	H	$0.277\ [0.258,\ 0.297]$
C_D(T7)	T7 only	H	$0.695\ [0.688,\ 0.702]$
C_{HE}(T7)	T7 only	H	$0.388\ [0.361,\ 0.416]$
$K_{\mathrm{H2},HS}$	$\mathbb{E}_{\mathrm{closed}}[R_H - R_S]$	H	$0.181\ [0.178,\ 0.184]$
$K_{\mathrm{H2},ES}$	$\mathbb{E}_{\mathrm{closed}}[R_E - R_S]$	E	$0.162\ [0.158,\ 0.165]$
C_{HS}(T7)	genuine rank-2 trained test	H	$-0.096\ [-0.108,\ -0.084]$
$K_{\mathrm{H3},HS}$	prereg. T7/T8 average	H	$0.068\ [0.060,\ 0.076]$ (not rank-2 evidence)

and 0.09 on closed tasks, but the smallest open value (T5a, 0.19) and the largest closed one (T6, 0.18) have overlapping intervals.

Post hoc, channel strength orders the rank-1 open tasks. Computing $\gamma_\tau = s_r(\Gamma_\phi)$ by quadrature for the clipping map used (Appendix F.1), the five rank-1 open tasks order identically by strength (1.26, 0.57, 0.31, 0.22, 0.13 for T2, T3, T5c, T5b, T5a), by geometry error ($D_E = 0.130 < 0.136 < 0.185 < 0.310 < 0.605$) and by gap closed; T7's off-diagonal strength is 0.81. Five tasks and a post hoc reading are thin evidence, but the ordering is consistent with the γ-dependence of the bound in Theorem 2.

The same logic on published estimators. In our estimator benchmark (Appendix H), estimators with a zero response moment (sliced inverse regression, SIR, on even links; pHd-type moments on odd links) did not recover geometry, and refined MAVE (Xia et al., 2002) was near-oracle on five tasks; the explicit route is a probe of its channel, not a competitive estimator.

6 A TASK-TRAINED TRANSFORMER PREDICTS BEYOND THE SELECTED SECOND-MOMENT STATISTIC

The closed label concerns the chosen clipped second-moment map, not the responses as a whole. On the three closed tasks, parity makes that moment vanish while a first response moment remains open: for the rank-1 odd links, direct Gaussian calculation gives $\mathbb{E}[YX] \parallel P$, the classical SIR direction (Li, 1991); T6 has a rank-1 conditional mean inside its rank-2 central subspace. Transformers can also implement linear-regression-like algorithms in context (Garg et al., 2022; Akyürek et al., 2023). These facts motivate a first-moment explanation, while our experiment asks the narrower question of whether a trained model obtains predictive information beyond the selected second-moment statistic.

Trained model and estimand. The trained arm S is a 2-layer, 4-head Transformer encoder regressor (Vaswani et al., 2017) with width 64 and 101,249 parameters. Each task has three preregistered seeds and uses the pre-existing radial-link recipe with no per-task tuning: Adam at learning rate 10^{-3}, early stopping and at most 3,000 steps (full architecture and training protocol in Appendix D.3). S sees up to 24,000 fresh training contexts per link with noiseless conditional-mean targets, whereas E uses only the current noisy context; we therefore interpret the comparison as a capability test rather than a data- or compute-matched comparison. The latent subspace remains fresh in every context. Risk is averaged over the three fixed trained models, so intervals cover context sampling rather than training randomness. The 32 unlabeled query tokens may attend to one another (never to query responses), making the empirical S estimand a query-batch predictor rather than the point-query row of the theory.

H2: predictive escape from this explicit statistic. Preregistered H2 requires the seed-averaged S risk to beat both H and E on the three closed tasks. It does: $K_{\mathrm{H2},HS} = 0.181\ [0.178, 0.184]$ and

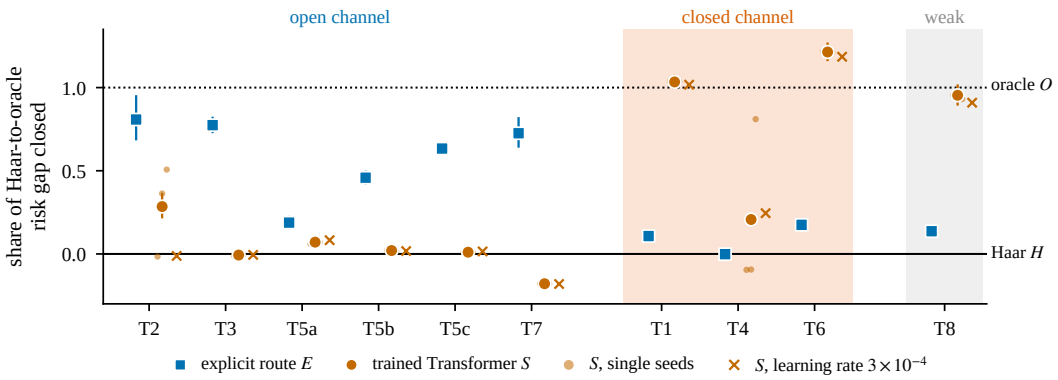


Figure 3: **Post hoc descriptive comparison of the explicit route and trained Transformer.** Share of the Haar-to-oracle smoothing gap closed, $G_A = (R_H - R_A)/(R_H - R_O)$, grouped by the preregistered explicit-channel label; points show the three trained seeds and the retained learning-rate sensitivity run. O uses the true projector but only 64 smoothing points per fold at the same fixed bandwidth, so it is a smoothing anchor rather than Bayes risk and $G_S > 1$ is possible.

Table 2: Closed tasks, primary run: controls, fixed trained-seed risks and seed robustness. No seed-level confidence interval is claimed.

Task	R_H	R_E	seed 1	seed 2	seed 3	below both
T1	0.2521	0.2270	0.0098	0.0113	0.0100	3/3
T4	0.4688	0.4693	0.5017	0.5012	0.1895	1/3
T6	0.2397	0.2065	0.0089	0.0097	0.0095	3/3

$K_{\mathrm{H2},ES} = 0.162$ $[0.158, 0.165]$ (Table 1). Seed outcomes are less uniform than these context-level intervals:

H2 is seed-robust on T1 and T6, while T4's seed-average advantage is carried by one of the three seeds. On T1 and T6, $R_S = 0.010$ and 0.009, far below both controls. These results establish use of current-context information beyond the selected second-moment statistic, combined with task-level information learned during training. The experiment measures prediction rather than an internal projector and does not isolate attention from MLP/readout contributions. The earlier SDR benchmark shows that first-moment methods exploit the same parity pattern (Appendix H), although it is not a matched first-moment attention arm in this study.

The genuine rank-2 trained test is negative. On T7, whose conditional mean depends on both directions and whose explicit route recovers both, all trained seeds at both learning rates stop near the constant-predictor plateau; $C_{HS}(\mathsf{T7}) = -0.096$ $[-0.108, -0.084]$. The preregistered aggregate $K_{\mathrm{H3},HS} = 0.068$ $[0.060, 0.076]$ is positive only because of T8, whose risk is compatible with using z_1 alone. We report the frozen aggregate for completeness but treat T7 as the genuine trained rank-2 result (Table 1; full tables in Appendix E). The negative conclusion is scoped to the frozen early-stopped recipe.

Post hoc pattern. On most explicit-open tasks S stays near Haar or the constant-predictor level while E succeeds; on the closed odd-link tasks the ordering reverses. The sign pattern coincides with the earlier benchmark's first-moment column, but those labels are nearly complementary to the selected second-moment map on this suite. We therefore treat the pattern as descriptive complementarity rather than mechanism identification (Appendix F).

7 Related Work

In-context learning and geometric attention. Softmax attention can be read as kernel smoothing (Tsai et al., 2019; Santos et al., 2026), and trained Transformers can implement regression and optimization-like procedures in context (Garg et al., 2022; Akyürek et al., 2023; von Oswald et al.,

2023; Bai et al., 2023). Existing theory studies nonparametric rates (Kim et al., 2024; Ching et al., 2026) or low-dimensional structure shared across pretraining tasks (Oko et al., 2024; Collins et al., 2024). Following Anonymous (2026), we instead redraw the predictive subspace in every context. Geometry-sensitive scores (Nielsen et al., 2024; Kuang et al., 2025) motivate the setting; our focus is when responses identify fresh task geometry and how that geometry affects localization.

Dimension reduction and response preprocessing. SIR uses a first inverse moment (Li, 1991); pHd, SAVE and directional regression use second moments (Li, 1992; Cook & Weisberg, 1991; Cook, 1998; Li & Wang, 2007); MAVE targets the central mean subspace (Xia et al., 2002); and slicing-free variants include WIRE/weighted PHD (Li et al., 2024; Lee & Shao, 2018). Response-preprocessed spectral estimators have sharp weak-recovery conditions (Mondelli & Montanari, 2019; Lu & Li, 2020; Luo et al., 2019), while modern multi-index work characterizes easy and trivial subspaces (Troiani et al., 2025). The response moment and its parity blind spots are therefore classical; our contribution is its fixed-clipping/log-free attention consequence, matched-Haar controlled replication, and zero-channel/non-Haar counterexample.

Prediction without first recovering a unique orientation also has classical antecedents. Direction aggregation and adaptation attain single-index rates (Gaïffas & Lecué, 2007; Lepski & Serdyukova, 2014); PSD bandwidth selection (Conn & Li, 2019) and estimated-index analysis (Klock et al., 2021) provide related routes. Our comparison selector adds a symmetric pair-smoother construction and quadratic-softmax realization. For the trained arm, link-dependent optimization mechanisms include the information exponent (Ben Arous et al., 2021; Oko et al., 2024), leap complexity (Abbe et al., 2023) and low-order learnable components (Damian et al., 2022); these motivate the post-hoc interpretation without determining the network's mechanism here.

8 LIMITATIONS

The confirmatory study uses synthetic Gaussian index tasks at $d = 16$ and $N = 256$, validating open/closed routing at one operating point rather than the theorem's m, d, γ, N scaling or information threshold. The selected second-moment channel is map-specific: all closed tasks here are parity blind spots, a first-moment channel remains open, and this study does not include a matched first-moment attention arm. The trained arm covers one small architecture and three fixed models per task; it has no projector readout, and its genuine rank-2 result on T7 is negative under the frozen early-stopped recipe. The oracle is a true-projector fixed-bandwidth smoother rather than Bayes risk, and the trained estimand allows interaction among the 32 unlabeled query covariates, whereas the theory is point-query. H1 is a controlled replication on a task family previously used in the estimator benchmark. The main response and comparison theorems are proved in Appendix B; the orientation lower bound remains a proof sketch and is not used in the main claims.

9 CONCLUSION

Response channels make fresh task geometry an explicit statistical object. For an identifying clipped second-moment channel, cross-fitted quadratic-softmax attention admits log-free subspace recovery and intrinsic-rate point prediction, and a controlled ten-task replication shows a large open-versus-closed geometry contrast against a matched Haar reference. A comparison-selected construction extends intrinsic-rate prediction beyond this eigengap condition. On the closed odd-link tasks, a task-trained Transformer also achieves low risk using information beyond the selected statistic, while the negative T7 result separates predictive success from demonstrated rank-2 geometry recovery. Together, the results position response channels as a map-specific diagnostic for explicit geometry identification and clarify the distinction between recovered geometry and broader task-specific prediction.

AI USE STATEMENT

Generative AI was used extensively: to formulate, prove and, in separate passes, audit the new response-channel and null-channel results; to help design the preregistered hypotheses and protocols; to generate the synthetic data, implement the estimators and the training and evaluation code, run and audit the experiments and analyses, including the documented learning-rate repair; to interpret

the results; and to draft the manuscript, figures and bibliography. The authors directed the research, made all scope and protocol decisions, reviewed all AI-assisted work, and take full responsibility for the content.

Reproducibility statement

The geometry–risk frontier is attributed to Anonymous (2026); proofs of the new theorem-level results are provided in the appendices, which also record the orientation lower-bound proof sketch, the full protocol, frozen task definitions and every preregistered empirical quantity.

References

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2552–2623. PMLR, 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Anonymous. Geometry-adaptive attention enables intrinsic-dimensional prediction, 2026. Concurrent submission, under review.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211, 2023. doi: 10.52202/075280-2498. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b2e63e36c57e153b9015fece2352a9f9-Abstract-Conference.html.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021. URL http://jmlr.org/papers/v22/20-1288.html.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax optimal in-context nonparametric regression with transformers. In *International Conference on Machine Learning*, 2026. doi: 10.17863/CAM.130396. URL https://www.repository.cam.ac.uk/handle/1810/403442. Accepted version.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function Lipschitzness. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92638–92696, 2024. doi: 10.52202/079017-2942. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a8633d27d782f66fe660c2fb4bae446e-Abstract-Conference.html.

Daniel J. Conn and Gang Li. An oracle property of the Nadaraya–Watson kernel estimator for high-dimensional nonparametric regression. *Scandinavian Journal of Statistics*, 46(3):735–764, 2019. doi: 10.1111/sjos.12370.

R. Dennis Cook. Principal Hessian directions revisited. *Journal of the American Statistical Association*, 93(441):84–94, 1998. doi: 10.1080/01621459.1998.10474090.

R. Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991. URL https://www.jstor.org/stable/2290564.

Wei Dai, Youjian Liu, and Brian Rider. Quantization bounds on Grassmann manifolds and applications to MIMO communications. *IEEE Transactions on Information Theory*, 54(3):1108–1123, 2008. doi: 10.1109/TIT.2007.915691.

Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(53):1647–1678, 2008. URL http://jmlr.org/papers/v9/dalalyan08a.html.

Alex Damian, Loucas Pillaud-Vivien, Jason Lee, and Joan Bruna. Computational-statistical gaps in Gaussian single-index models (extended abstract). In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1262–1262. PMLR, 2024. URL https://proceedings.mlr.press/v247/damian24a.html.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452. PMLR, 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Stéphane Gaïffas and Guillaume Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, 1:538–573, 2007. doi: 10.1214/07-EJS077. URL https://doi.org/10.1214/07-EJS077.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106667–106713, 2024. doi: 10.52202/079017-3387. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c11daad0a48ea5f3c5c6390c7b060720-Abstract-Conference.html.

Timo Klock, Alessandro Lanteri, and Stefano Vigogna. Estimating multi-index models with response-conditional least squares. *Electronic Journal of Statistics*, 15(1):589–629, 2021. doi: 10.1214/20-EJS1785.

Yilun Kuang, Noah Amsel, Sanae Lotfi, Shikai Qiu, Andres Potapczynski, and Andrew Gordon Wilson. Customizing the inductive biases of softmax attention using structured matrices. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31752–31776. PMLR, 2025. URL https://proceedings.mlr.press/v267/kuang25a.html.

Chung Eun Lee and Xiaofeng Shao. Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series. *Journal of the American Statistical Association*, 113(521):216–229, 2018. doi: 10.1080/01621459.2016.1240083.

Chung Eun Lee and Xiaofeng Shao. Volatility martingale difference divergence matrix and its application to dimension reduction for multivariate volatility. *Journal of Business & Economic Statistics*, 38(1):80–92, 2020. doi: 10.1080/07350015.2018.1458621.

Oleg Lepski and Nora Serdyukova. Adaptive estimation under single-index constraint in a regression model. *The Annals of Statistics*, 42(1):1–28, 2014. doi: 10.1214/13-AOS1152. URL https://doi.org/10.1214/13-AOS1152.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007. doi: 10.1198/016214507000000536.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. doi: 10.1080/01621459.1992.10476258.

Lu Li, Xiaofeng Shao, and Zhou Yu. A slicing-free perspective to sufficient dimension reduction: Selective review and recent developments. *International Statistical Review*, 92(3):355–382, 2024. doi: 10.1111/insr.12565.

Yue M. Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020. doi: 10.1093/imaiai/iaz020.

Wangyu Luo, Wael Alghamdi, and Yue M. Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9):2347–2356, 2019. doi: 10.1109/TSP.2019.2904918.

Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19(3):703–773, 2019. doi: 10.1007/s10208-018-9395-y.

Stefan K. Nielsen, Laziz U. Abdullaev, Rachel S. Y. Teo, and Tan M. Nguyen. Elliptical attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109748–109789, 2024. doi: 10.52202/079017-3482. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c63908a3e946af0e7978c23737229137-Abstract-Conference.html.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77316–77365, 2024. doi: 10.52202/079017-2459. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8d6c356c82252032a20c749dda36f2d8-Abstract-Conference.html.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Erwin Riegler, Günther Koliander, and Helmut Bölcskei. Rate-distortion theory for general sets and measures. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 101–105, 2018. doi: 10.1109/ISIT.2018.8437740.

Saul Santos, Nuno Gonçalves, Daniel C. McNamee, Marcos Treviso, and André F. T. Martins. Sparse attention as compact kernel regression. arXiv preprint arXiv:2601.22766, 2026. URL https://arxiv.org/abs/2601.22766.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. doi: 10.1214/aos/1176345969.

Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborova, Bruno Loureiro, and Florent Krzakala. Fundamental computational limits of weak learnability in high-dimensional multi-index models. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2467–2475. PMLR, 2025. URL https://proceedings.mlr.press/v258/troiani25a.html.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4344–4353, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1443. URL https://aclanthology.org/D19-1443/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 363–410, 2002. doi: 10.1111/1467-9868.03411.

APPENDIX SCOPE

These appendices contain proofs and detailed evidence for the results new to this submission, together with the protocol, full tables, robustness analyses, and extended related work. The geometry–risk frontier stated as Theorem 1 is inherited from Anonymous (2026); its proof and the historical radial-link results are not reproduced. Appendix I summarizes the relationship between that background and the results introduced here.

A NOTATION AND CONVENTIONS

Table A.1: Notation.

Symbol	Meaning
d, r	ambient dimension; rank of the predictive subspace ($1 \le r < d$)
$P = BB^\top$, $\mu_{d,r}$	task projector with frame $B \in \mathrm{St}(d,r)$; Haar measure on $\mathrm{Gr}(d,r)$
g, s, L, G, σ	link; Hölder exponent and constant; sup-norm bound; noise level
Γ_g, γ	$\mathbb{E}[g(Z)(ZZ^\top - I_r)]$; identifiability level $s_r(\Gamma_g) \ge \gamma$
n, N, m	sample/fold size as specified by each theorem; in Theorem 2, fold size n (n_{f} in the main text), context size N and pilot size m
K	number of candidate/code elements where used; number of cross-fitting folds in the response analysis (k in the main text)
A, h, b, $\mu(A)$	precision of the quadratic-softmax row; predictive and transverse scales; kernel mass $\det(I+A)^{-1/2}$
U, S, K, $\mathcal{I}$	public variable; geometry signal ($\mathcal{S}$ in the main text); signal kernel $\mathbb{P}_{S\mid P,U}$; information budget (nats)
V_{eff}, $\mathcal{I}_\star$	usable localization volume; saturation information $r(d-r)\log n/(2s+r)$
ϕ, H_ϕ, Γ_ϕ	response map; explicit second-moment channel $B\Gamma_\phi B^\top$; its reduced matrix
$D(P,\widehat{P})$	$\|P-\widehat{P}\|_F^2/(2r)$; independent-Haar expectation $1-r/d$
S, E, H, O (rows)	trained Transformer; explicit route; Haar-projector control; true-projector oracle
$C_D, C_{HE}, C_{HS}, C_{EO}$	per-task context-mean differences $D_H - D_E$, $R_H - R_E$, $R_H - R_S$, $R_E - R_O$
$K_{\mathrm{H1},\cdot}, K_{\mathrm{H2},\cdot}, K_{\mathrm{H3},HS}$	preregistered summary contrasts (main Sections 5–6; Supplement D.4)

The main paper defines the statistical model and the inherited geometry-information formulation. In these appendices, "response channel" always refers to the response statistic of Section 4; any information-theoretic signal law is called a geometry signal.

B THE EXPLICIT RESPONSE ROUTE AND PREDICTION WITHOUT EXPLICIT RECOVERY

This appendix gives the detailed statements combined in Theorem 2 (Section B.1), the prediction-consistent comparison construction (Section B.2), and complete proofs of the main results. The population moment, the fixed-fold cross-fitted construction and Proposition B.3 are from Anonymous (2026), which clips at a level growing like $\sqrt{\log N}$. New here are the fixed clipping level (Proposition B.1), the log-free concentration, recovery and risk bounds (Proposition B.2 and Theorems B.1 and B.2), and the comparison selector (Theorem B.3). A final orientation-recovery argument is retained as a proof sketch and is not used elsewhere in the main claims.

B.1 STATEMENTS: POPULATION CHANNEL, FIXED CLIPPING, LOG-FREE RECOVERY AND CROSS-FITTED RISK

The information frontier treats geometry as a pre-design resource. Current responses provide a different route: they can identify a task-specific predictive subspace and thereby define an anisotropic attention metric. Beyond population identifiability, this appendix asks when response-derived geometry is accurate enough relative to the prediction bandwidth.

B.1.1 POPULATION GEOMETRY AND FIXED CLIPPING

For $X \sim N(0, I_d)$, $Y = g(B^\top X) + \varepsilon$, and $P = BB^\top$,

$$H := \mathbb{E}[Y(XX^\top - I_d)] = B\Gamma_g B^\top. \tag{B.1}$$

Thus the response-weighted second-order moment has range equal to the predictive subspace whenever $s_r(\Gamma_g) > 0$.

For finite samples let $c_i = T_\tau(Y_i)$, $\bar{c} = m^{-1}\sum_i c_i$, and

$$\widetilde{H}_m = \frac{1}{m}\sum_{i=1}^{m}(c_i - \bar{c})(X_i X_i^\top - I_d). \tag{B.2}$$

Writing $\Psi_\tau(a) = \mathbb{E}_\varepsilon T_\tau(a + \varepsilon)$ and $\Gamma_{g,\tau} = \mathbb{E}[\Psi_\tau(g(Z))(ZZ^\top - I_r)]$, the exact finite-$m$ population object is

$$\mathbb{E}\widetilde{H}_m = \frac{m-1}{m} B\Gamma_{g,\tau} B^\top. \tag{B.3}$$

The next result removes the need to assume a clipped eigengap separately.

Proposition B.1 (Uniform fixed-clipping gap preservation)**.** *Let $\kappa = \mathbb{E}|U^2 - 1| = 4\phi(1)$ for $U \sim N(0,1)$ and suppose $s_r(\Gamma_g) \geq \gamma$. Define*

$$\tau_* = G + \sigma\sqrt{2\log_+\left(\frac{2\kappa\sigma}{\gamma\sqrt{2\pi}}\right)}, \qquad \log_+ x := \max\{\log x, 0\}. \tag{B.4}$$

Then, uniformly over the original identifying class,

$$\|\Gamma_{g,\tau_*} - \Gamma_g\|_{\mathrm{op}} \leq \gamma/2, \qquad s_r(\Gamma_{g,\tau_*}) \geq \gamma/2. \tag{B.5}$$

The level τ_* is deterministic and independent of context length.

The proof in Appendix B.3 uses the one-dimensional quadratic-form bound

$$\|\Gamma_{g,\tau} - \Gamma_g\|_{\mathrm{op}} \leq \frac{\kappa\sigma}{\sqrt{2\pi}} \exp\left[-\frac{(\tau - G)^2}{2\sigma^2}\right],$$

so no unnecessary factor of r appears.

B.1.2 LOG-FREE DIMENSION-EXPLICIT SUBSPACE RECOVERY

Set $v^2 = 4G^2 + \sigma^2$. A variance-sensitive analysis of the centred estimator yields the matrix deviation

$$\left\|\widetilde{H}_m - \frac{m-1}{m}H_{\tau_*}\right\|_{\mathrm{op}} \leq C\left[v\sqrt{\frac{d + \log(1/\eta)}{m}} + \tau_*\frac{d + \log(1/\eta)}{m}\right] \tag{B.6}$$

with probability at least $1 - \eta$. The random centering product $\bar{c}\,\bar{W}$ is controlled explicitly. Since $\gamma\tau_* \lesssim G^2 + \sigma^2$, the linear Bernstein term disappears after passing to the bounded projector loss.

Theorem B.1 (Log-free response-derived projector recovery)**.** *Assume $m \geq 2$ and let $\widehat{P}_m$ project onto the r eigenvectors of $\widetilde{H}_m$ with largest absolute eigenvalues, with measurable tie breaking. There are universal constants C_P, C_M such that*

$$\|\widehat{P}_m - P\|_{\mathrm{op}} \leq \min\left\{1, C_P\frac{v}{\gamma}\sqrt{\frac{d\log 9 + \log(6/\eta)}{m}}\right\} \tag{B.7}$$

with probability at least $1 - \eta$, and

$$\mathbb{E}\|\widehat{P}_m - P\|_{\mathrm{op}}^{2s} \leq \min\left\{1, \left(\frac{C_M v^2 d}{m\gamma^2}\right)^s\right\}. \tag{B.8}$$

The nonzero population eigenvalues may have mixed signs; the proof works with the absolute spectral cluster separated from zero.

B.1.3 From recovered geometry to intrinsic prediction

For fixed $K \geq 2$, write $N = Kn$ and $m = (K-1)n$. Estimate $\widehat{P}_k$ from the complement of fold k and use, on the held-out smoother fold,

$$A_{k,h} = h^{-2}\widehat{P}_k. \tag{B.9}$$

The choice of zero transverse precision is already contained in the earlier admissible family and is useful when dimension dependence is exposed: a fixed positive transverse precision preserves the h^r exponent for fixed d but carries an exponentially small factor in $d - r$.

Define

$$\rho_r(h) = \left[\frac{h\sqrt{h^2+2}}{1+h^2}\right]^r \asymp_r h^r, \qquad r_0 = \min(r, d-r), \tag{B.10}$$

and

$$\theta_m = \min\left\{1, \frac{C_M v^2 d}{m\gamma^2}\right\}.$$

Theorem B.2 (Log-free response-to-attention risk)**.** *Under the model above, for every $0 < h \leq 1$,*

$$\mathbb{E}\{\widehat{f}_{\mathrm{cf}}(0) - g(0)\}^2 \leq 2L^2 r^s h^{2s} + 2L^2 r_0^s \theta_m^s + \frac{4\sigma^2}{n\rho_r(h)} + (4G^2 + \sigma^2)\exp\{-n\rho_r(h)/8\}. \tag{B.11}$$

The bound is uniform over the original unclipped identifying class. Foldwise pilot/smoother independence suffices; the final fold predictors need not be independent.

Let $p = 2s + r$ and choose $h_* = bn^{-1/p}$ with

$$b = \left(\frac{\sigma^2 r^{1-s}}{s c_r L^2}\right)^{1/p}, \qquad c_r = (\sqrt{3}/2)^r. \tag{B.12}$$

For $p > 2$, a sufficient geometry-onset condition is

$$n \geq n_{\mathrm{geom}} := \left[\frac{C_M(4G^2+\sigma^2)d}{(K-1)\gamma^2 b^2}\right]^{p/(p-2)}. \tag{B.13}$$

Together with the explicit bandwidth and denominator conditions in Appendix B.3, this gives a finite N_0 such that

$$N \geq N_0 \quad \Longrightarrow \quad R_N^{\mathrm{cf}}(0) \lesssim N^{-2s/(2s+r)}. \tag{B.14}$$

At $p = 2$ (necessarily $r = 1, s = 1/2$), the same oracle polynomial exponent is obtained without a logarithmic clipping loss, directly from the risk bound of Theorem B.2; only $\|\widehat{P} - P\| \lesssim h_*$ becomes a parameter condition rather than an eventual-in-N threshold. At $p < 2$, the explicit orientation-then-smoothing construction has an N^{-s} geometry term; this is a limitation of that route, not a prediction impossibility theorem.

B.1.4 Channel nats: what an open channel supplies

Corollary B.1 (Information carried by the recovered projector)**.** *Let $P \sim \mu_{d,r}$ be Haar, let every task satisfy the assumptions of Theorem B.1 with the same (v, γ), and let $\widehat{P}_m$ be its projector. There is a constant C depending only on (d, r) such that*

$$I(P; \widehat{P}_m) \;\geq\; \frac{r(d-r)}{2}\log\frac{m\gamma^2}{Cv^2 d}.$$

Proof. Put $W = \mathrm{range}(I - \widehat{P}_m) \in \mathrm{Gr}(d, d-r)$, a function of $S = \widehat{P}_m$. For rank-r projectors, $\mathrm{tr}(PP_W) = \mathrm{tr}(P(I - \widehat{P}_m)) = \frac{1}{2}\|P - \widehat{P}_m\|_F^2 \leq r\|P - \widehat{P}_m\|_{\mathrm{op}}^2$, so (B.8) with exponent one (it integrates (B.7), which uses no Hölder continuity) gives $\mathbb{E}\,\mathrm{tr}(PP_W) \leq rC_M v^2 d/(m\gamma^2)$. The Grassmann small-ball/rate–distortion lower bound of Dai et al. (2008), equivalently the subregular-measure Shannon bound of Riegler et al. (2018), gives for $t = d - r$ and $a = 1$ that $\mathbb{E}\,\mathrm{tr}(PP_W) \geq c\,e^{-2I(P;S)/(r(d-r))}$ with $c = c_{d,r} > 0$. Combining the two bounds and solving for $I(P; S)$ proves the claim with $C = rC_M/c$. □

Comparison with the frontier (accounting only). At the fold size n_{f} the frontier saturates at $\mathcal{I}_\star = r(d-r)\log n_{\mathrm{f}}/(2s+r)$. The bound of Corollary B.1 exceeds $\mathcal{I}_\star$ when $Cv^2 d/(m\gamma^2) \le n_{\mathrm{f}}^{-2/(2s+r)}$, which is the geometry-onset condition (B.13) up to constants. This is not a corollary of Theorem 1: its converse requires a signal law $\mathbb{P}_{S|P,U}$ that does not depend on the link, whereas the law of $\widehat{P}_m$ does, and a row that knew the link could localize on its level sets rather than on neighbourhoods. Section F.1 reports a per-task proxy.

B.2 Statement: prediction without bandwidth-scale subspace recovery

The finite-context condition above answers when one explicit response-moment route recovers orientation finely enough. It does not imply that every accurate point predictor must first estimate the whole subspace to $O(h)$. A complementary construction selects an attention geometry by comparing the predictions induced by candidate Gaussian smoothers rather than by requiring projector consistency.

For a candidate rank-r projector Q and $0 < h \le 1/2$, let

$$\Sigma_Q = I - (1-h^2)Q. \tag{B.15}$$

For each pair P, Q, Appendix B.4 constructs a symmetric covariance $\Sigma_{P,Q} = F_h(P+Q)$ such that

$$P\Sigma_{P,Q}P = P\Sigma_Q P + h^2 P \tag{B.16}$$

and symmetrically with P, Q exchanged. The corresponding Gaussian likelihood-ratio statistics have second moments of order h^{-r} rather than h^{-d}.

Let $\mathcal{Q}_h$ be a deterministic operator-norm h-net of rank-r projectors with

$$|\mathcal{Q}_h| \le (1 + 4\sqrt{r_0}/h)^{dr_0}. \tag{B.17}$$

Using median-of-means estimates of the single- and pair-smoother expectations, choose a candidate whose symmetric comparison discrepancy is minimal, then use held-out normalized quadratic-softmax attention with precision $(h^{-2}-1)\widehat{Q}$.

Theorem B.3 (Comparison-selected quadratic-softmax prediction)**.** *Fix $K \ge 2$, write $N = Kn$, and allocate $m = (K-1)n$ selection samples and n independent smoothing samples as in Appendix B.4. For every $1 \le r < d$, $0 < s \le 1$, and bounded s-Hölder link, without assuming $s_r(\Gamma_g) \ge \gamma$, the comparison selector satisfies*

$$\begin{aligned} R_N^{\mathrm{cmp}}(0) \le{}& 400L^2 r^s h^{2s} + 4608(G^2+\sigma^2)\frac{J}{mh^r} + \frac{8(4G^2+\sigma^2)}{n\rho_r^{\mathrm{cmp}}(h)} \\ &+ 2(4G^2+\sigma^2)e^{-n\rho_r^{\mathrm{cmp}}(h)/8} + 8G^2\eta, \end{aligned} \tag{B.18}$$

where $\rho_r^{\mathrm{cmp}}(h) = h^r(2-h^2)^{r/2} \ge h^r$ and

$$J \lesssim dr_0 \log(1/h) + \log(1/\eta).$$

Consequently, with $\eta = N^{-2}$ and $h \asymp (H_N/N)^{1/(2s+r)}$ for $H_N \asymp dr_0 \log N$,

$$R_N^{\mathrm{cmp}}(0) \lesssim \left(\frac{dr_0 \log N}{N}\right)^{2s/(2s+r)} \tag{B.19}$$

for fixed model parameters and fixed K.

The comparison theorem itself makes no projector-consistency conclusion: it controls the selected smoothed value directly through symmetric comparisons. For context, Section B.3.7 records a standard Assouad-style route suggesting that uniform $O(h)$ orientation recovery on an explicit smooth radial subfamily requires order $\sigma^2[d+\log(1/\eta)]/(\gamma^2 h^2)$ samples. Because the final Assouad/confidence steps are not written out here, this remains a proof sketch and is not used to establish a formal recovery-versus-prediction separation.

B.3 LOG-FREE CENTRED RESPONSE MOMENT, FINITE-CONTEXT ONSET, AND ORIENTATION SKETCH

B.3.1 EXACT POPULATION OBJECT AND THE FIXED-CLIPPING GAP

Let $c_i = T_\tau(Y_i)$, $W_i = X_i X_i^\top - I_d$, $\bar{c} = m^{-1}\sum_i c_i$, and $\bar{W} = m^{-1}\sum_i W_i$. Then

$$\widetilde{H}_m = \frac{1}{m}\sum_i (c_i - \bar{c}) W_i = \frac{1}{m}\sum_i c_i W_i - \bar{c}\,\bar{W}. \tag{B.20}$$

Writing $H_\tau = \mathbb{E}[cW]$ and using $\mathbb{E}W = 0$ gives

$$\mathbb{E}\widetilde{H}_m = \left(1 - \frac{1}{m}\right) H_\tau. \tag{B.21}$$

If $X = BZ + B_\perp V$ with Z, V independent standard Gaussians, then

$$H_\tau = B\Gamma_{g,\tau}B^\top, \qquad \Gamma_{g,\tau} = \mathbb{E}[\Psi_\tau(g(Z))(ZZ^\top - I_r)]. \tag{B.22}$$

Clipping therefore changes the signal matrix inside the true subspace but creates no population direction outside it.

For $\tau \geq G$, define $Q(a) = \phi(a) - a\overline{\Phi}(a) = \mathbb{E}(U-a)_+$ for $U \sim N(0,1)$. Direct integration gives

$$\Psi_\tau(u) - u = \sigma\left[Q\left(\frac{\tau+u}{\sigma}\right) - Q\left(\frac{\tau-u}{\sigma}\right)\right].$$

Hence

$$\sup_{|u|\leq G} |\Psi_\tau(u) - u| \leq \frac{\sigma}{\sqrt{2\pi}} \exp\left[-\frac{(\tau-G)^2}{2\sigma^2}\right]. \tag{B.23}$$

For a unit $a \in \mathbb{R}^r$,

$$|a^\top(\Gamma_{g,\tau} - \Gamma_g)a| \leq \sup_{|u|\leq G} |\Psi_\tau(u) - u|\,\mathbb{E}|U^2 - 1|.$$

Thus, with $\kappa = 4\phi(1)$,

$$\boxed{\|\Gamma_{g,\tau} - \Gamma_g\|_{\mathrm{op}} \leq \frac{\kappa\sigma}{\sqrt{2\pi}} \exp\left[-\frac{(\tau-G)^2}{2\sigma^2}\right].} \tag{B.24}$$

The choice (B.4) proves Proposition B.1. In addition, nonemptiness implies $\gamma \leq \|\Gamma_g\|_{\mathrm{op}} \leq \kappa G$. The elementary maximization $\sup_{t>0} t\sqrt{2\log_+(c/t)} = c/\sqrt{e}$ yields the useful parameter inequality

$$\gamma\tau_* \leq G^2 + \sigma^2 \leq v^2, \qquad v^2 := 4G^2 + \sigma^2. \tag{B.25}$$

B.3.2 VARIANCE-SENSITIVE CENTRED CONCENTRATION

Let $\mu_c = \mathbb{E}c$. Since Ψ_τ is odd and 1-Lipschitz, $|\mu_c| \leq G$. Conditionally on X, clipping is a 1-Lipschitz function of Gaussian noise and therefore

$$\mathbb{E}[(c-\mu_c)^2 \mid X] \leq v^2, \qquad |c - \mu_c| \leq 2\tau. \tag{B.26}$$

Fix a unit vector a and write $w_a = (a^\top X)^2 - 1$. For every integer $k \geq 2$,

$$\mathbb{E}|(c-\mu_c)w_a|^k \leq (2\tau)^{k-2} v^2 \mathbb{E}|U^2 - 1|^k.$$

Using $\mathbb{E}|U^2-1|^k \leq 4^k k!$ and centering gives a Bernstein moment condition. A $1/4$-net of the sphere with at most 9^d points then yields, for

$$q = d\log 9 + \log(6/\eta),$$

$$\left\|\frac{1}{m}\sum_i (c_i - \mu_c)W_i - H_\tau\right\|_{\mathrm{op}} \leq 32v\sqrt{\frac{q}{m}} + 32\tau\frac{q}{m} \tag{B.27}$$

outside an event of probability $\eta/3$. The same net argument gives

$$\|\bar{W}\|_{\mathrm{op}} \leq 4\sqrt{q/m} + 4q/m.$$

Gaussian Lipschitz concentration conditionally on X, followed by Hoeffding's lemma for the conditional mean in $[-G, G]$, gives

$$|\bar{c} - \mu_c| \le v\sqrt{2q/m}$$

with another failure probability at most $\eta/3$. Applying these bounds to

$$\widetilde{H}_m = \frac{1}{m}\sum_i (c_i - \mu_c)W_i - (\bar{c} - \mu_c)\bar{W}$$

and using a deterministic 2τ bound when $q/m > 1$ proves

Proposition B.2 (Variance-sensitive centred concentration)**.** *For every $m \ge 1$ and $\eta \in (0,1)$,*

$$\left\| \widetilde{H}_m - \left(1 - \frac{1}{m}\right) H_\tau \right\|_{\mathrm{op}} \le 64 \left[v\sqrt{\frac{q}{m}} + \tau \frac{q}{m} \right] \tag{B.28}$$

with probability at least $1 - \eta$.

B.3.3 Absolute-spectrum perturbation and projector moments

Let M be symmetric of rank r with range projector P, all nonzero eigenvalues satisfying $|\lambda_j(M)| \ge \lambda$, and $\widehat{M} = M + E$. If $\widehat{P}$ is the top-r absolute-eigenvalue spectral projector of $\widehat{M}$, then

$$\|\widehat{P} - P\|_{\mathrm{op}} \le \min\left\{1, \frac{2\|E\|_{\mathrm{op}}}{\lambda}\right\}. \tag{B.29}$$

Indeed, on $\|E\|_{\mathrm{op}} < \lambda/2$, singular-value perturbation separates the selected absolute spectral cluster; if $\widehat{M}\widehat{U} = \widehat{U}\widehat{\Lambda}$, then $(I - P)\widehat{U}\widehat{\Lambda} = (I - P)E\widehat{U}$ and $\|\widehat{\Lambda}^{-1}\| \le (\lambda - \|E\|)^{-1}$. The complementary event is handled by the trivial projector bound.

For $M = (1 - 1/m)H_{\tau_*}$ and $m \ge 2$, Proposition B.1 gives $\lambda \ge \gamma/4$. Combining (B.25), (B.28) and (B.29) yields (B.7); integrating the tail gives (B.8). The numerical constants are intentionally loose; one fully explicit admissible choice is $C_P = 1024$, $C_M = 4C_P^2$.

B.3.4 Attention geometry and cross-fitted risk

Condition on an independent rank-r projector $\widehat{P}$ and set $A = h^{-2}\widehat{P}$ for $0 < h \le 1$. For $w(X) = e^{-X^\top AX/2}$,

$$\mu = \mathbb{E}w = (1 + h^{-2})^{-r/2}, \qquad \nu = \mathbb{E}w^2 = (1 + 2h^{-2})^{-r/2}.$$

The effective mass is

$$\rho = \frac{\mu^2}{\nu} = \left[\frac{h\sqrt{h^2 + 2}}{1 + h^2}\right]^r = \rho_r(h). \tag{B.30}$$

The tilted Gaussian covariance is $(I + A)^{-1} = \frac{h^2}{1+h^2}\widehat{P} + (I - \widehat{P})$. If

$$T = \mathrm{tr}\{P(I - \widehat{P})\} = \sum_j \sin^2\theta_j, \qquad \delta = \|P - \widehat{P}\|_{\mathrm{op}},$$

then $\delta^2 \le T \le r_0\delta^2$ and

$$D_A := \mathbb{E}_{\mathrm{tilt}}\|B^\top X\|^2 = \frac{rh^2 + T}{1 + h^2}. \tag{B.31}$$

Thus the squared Hölder bias is at most

$$L^2 D_A^s \le L^2 r^s h^{2s} + L^2 r_0^s \delta^{2s}.$$

For iid smoother weights w_i, $S = \sum_i w_i$, the elementary bound $e^{-tw} \le 1 - tw + t^2w^2/2$ and Chernoff's method with $t = \mu/(2\nu)$ give

$$\mathbb{P}(S < n\mu/2) \le e^{-n\rho/8}, \qquad \mathbb{E}\sum_i (w_i/S)^2 \le \frac{4}{n\rho} + e^{-n\rho/8}. \tag{B.32}$$

Weighted Jensen controls design bias and conditional mean-zero noise contributes $\sigma^2 \sum_i (w_i/S)^2$. This proves the conditional fold bound underlying Theorem B.2. For cross-fitting, each $\widehat{P}_k$ is measurable with respect to the complement of its own fold and hence independent of that smoother fold. The final fold predictions may be dependent, but

$$\left(K^{-1} \sum_k e_k \right)^2 \le K^{-1} \sum_k e_k^2$$

transfers the foldwise expected-risk bound without any independence assumption across final predictions.

B.3.5 Finite-context onset and boundary regimes

Put

$$p = 2s + r, \qquad a = \frac{2s}{p}, \qquad B_0 = 2L^2 r^s, \qquad D_0 = 2L^2 r_0^s,$$

$$V_0 = \frac{4\sigma^2}{c_r}, \qquad Q_0 = 4G^2 + \sigma^2, \qquad \zeta = \frac{C_M v^2 d}{\gamma^2}.$$

Using $\rho_r(h) \ge c_r h^r$ gives

$$R_N^{\mathrm{cf}} \le B_0 h^{2s} + \frac{V_0}{nh^r} + D_0 \min\{1, (\zeta/m)^s\} + Q_0 e^{-c_r n h^r/8}.$$

Minimizing the first two terms yields (B.12). Requiring the geometry moment to be at most the oracle bandwidth scale gives

$$\frac{\zeta}{(K-1)n} \le h_*^2 \iff n^{1-2/p} \ge \frac{\zeta}{(K-1)b^2}. \tag{B.33}$$

For $p > 2$ this gives (B.13). A slightly weaker expected-risk threshold replaces the right side by the same quantity multiplied by $\frac{r_0}{r}(r/p)^{1/s}$. An explicit denominator condition is

$$n \ge n_{\mathrm{den}} := \left[\frac{16}{c_r b^r} \log_+ \left(\frac{4s(4G^2+\sigma^2)}{p\sigma^2} \right) \right]^{p/(2s)}. \tag{B.34}$$

Write $O_0 n^{-2s/p} := B_0 h_*^{2s} + V_0/(nh_*^r)$ for the oracle part of the bound at h_*, so that $O_0 = B_0 b^{2s} + V_0 b^{-r}$, and let n_{risk} be the expected-risk threshold just described, that is, (B.13) with its right side multiplied by $\frac{r_0}{r}(r/p)^{1/s}$. Therefore one valid complete onset is

$$N_0 = K \left\lceil \max\left\{ 1, \frac{2}{K-1}, b^p, n_{\mathrm{risk}}, n_{\mathrm{den}} \right\} \right\rceil, \tag{B.35}$$

which implies $R_N^{\mathrm{cf}}(0) \le 3O_0 (N/K)^{-2s/p}$ for all divisible $N \ge N_0$. These constants are worst-case certificates, not calibrated predictions of the empirical onset.

At $p = 2$, $r = 1$, $s = 1/2$ and the geometry term has the same N^{-s} polynomial order as the oracle term, without a clipping-induced $\log N$ factor. At $p < 2$, necessarily $r = 1$, $s < 1/2$, the explicit moment route proves $R_N^{\mathrm{cf}} \lesssim N^{-2s/(2s+1)} + CN^{-s}$; no minimax impossibility claim follows from this upper-bound bottleneck.

Proof of Theorem 2. The deterministic level τ_* and the bound $s_r(\Gamma_{g,\tau_*}) \ge \gamma/2$ are Proposition B.1, proved in Section B.3. Part (i) is (B.8) of Theorem B.1, proved above from the variance-sensitive concentration and the absolute-spectrum perturbation bound. Part (ii) is Theorem B.2, proved above from foldwise pilot/smoother independence. Part (iii) is (B.14), obtained above from the onset (B.35), which combines the geometry threshold (B.13) with the bandwidth and denominator conditions, for $2s + r > 2$. At $2s + r = 2$ it follows directly from part (ii) with $h \asymp N^{-1/2}$, because the geometry term $\theta_m^s \asymp N^{-s}$ then has the order of the intrinsic rate. □

B.3.6 A LOWER BOUND FOR EXPLICIT ORIENTATION RECOVERY

The following lower bound is useful precisely because it is *not* a prediction lower bound. Let

$$A_* = \gamma r(1 + 1/r)^{r/2+1}$$

and, for this subsection only, assume

$$G \geq A_*, \qquad L \geq A_*(er)^{-s/2}. \tag{B.36}$$

The smooth radial link $g_*(z) = A_* e^{-\|z\|^2/(2r)}$ belongs to the identifying class and satisfies $\Gamma_{g_*} = -\gamma I_r$. For $f_P(x) = A_* e^{-x^\top Px/(2r)}$,

$$\mathbb{E}(f_P(X) - f_Q(X))^2 \leq \frac{A_*^2}{2r^2}\|P - Q\|_F^2,$$

so Gaussian response laws obey

$$\mathrm{KL}(\mathbb{P}_P^m, \mathbb{P}_Q^m) \leq C\frac{m\gamma^2}{\sigma^2}\|P - Q\|_F^2.$$

A local Grassmann hypercube with $M = \max(r, d - r) \geq d/2$ suggests the following standard Assouad completion.

B.3.7 ORIENTATION-RECOVERY LOWER-BOUND SKETCH (NOT USED)

Under (B.36), the preceding Kullback–Leibler bound and hypercube are the ingredients for a bound of the form

$$\sup_P \mathbb{E}_P\|\widehat{P} - P\|_{\mathrm{op}}^2 \gtrsim \min\left\{1, \frac{\sigma^2 d}{m\gamma^2}\right\}, \tag{B.37}$$

and, after the corresponding confidence-version argument, the heuristic requirement

$$m \gtrsim \frac{\sigma^2[d + \log(1/\eta)]}{\gamma^2 h^2} \tag{B.38}$$

for uniform probability-$1 - \eta$ orientation accuracy h. We have not written the Assouad decoding and confidence conversion in full, so these displays remain a proof sketch rather than a theorem. The main-text and comparison-theorem claims do not depend on them.

For this radial family $f_P(0) = A_*$ for every P. Hence the orientation lower bound cannot itself imply a point-prediction lower bound; this motivates the comparison construction below.

B.4 PREDICTION-CONSISTENT GAUSSIAN COMPARISON: FULL CONSTRUCTION

B.4.1 SYMMETRIC COMPARISON COVARIANCE

Fix $0 < h \leq 1/2$. For any rank-r projector Q define $\Sigma_Q = I - (1 - h^2)Q$. For rank-$r$ projectors P, Q, let $S = P + Q$ and define

$$F_h(t) = \begin{cases} 1, & t = 0, \\ 1 + h^2, & 0 < t \leq 1, \\ 1 + h^2 - \dfrac{2(1 - h^2)(t - 1)^2}{t}, & 1 < t \leq 2. \end{cases}$$

Set $\Sigma_{P,Q} = F_h(P + Q)$ by spectral calculus. Then

$$\Sigma_{P,Q} = \Sigma_{Q,P}, \qquad 2h^2 I \preceq \Sigma_{P,Q} \preceq (1 + h^2)I, \tag{B.39}$$

and

$$P\Sigma_{P,Q}P = P\Sigma_Q P + h^2 P, \qquad Q\Sigma_{P,Q}Q = Q\Sigma_P Q + h^2 Q. \tag{B.40}$$

To verify this, work in a principal-angle plane with $c = \cos\theta$. The eigenvalues of $P + Q$ are $1 + c$ and $1 - c$, with corresponding covariance eigenvalues

$$a_+ = 1 + h^2 - \frac{2(1 - h^2)c^2}{1 + c}, \qquad a_- = 1 + h^2.$$

A unit predictive direction in P has squared coordinates $(1+c)/2$ and $(1-c)/2$, so its variance is $1+h^2-(1-h^2)c^2$, exactly its variance under Σ_Q plus h^2.

For a positive covariance $\Sigma \prec 2I$, define the Gaussian likelihood ratio

$$W_\Sigma(x) = \frac{\phi_\Sigma(x)}{\phi_{I_d}(x)}.$$

Gaussian integration gives

$$\mathbb{E}W_\Sigma(X)^2 = \det(\Sigma)^{-1/2}\det(2I-\Sigma)^{-1/2}. \tag{B.41}$$

For Σ_Q, this equals $[h^2(2-h^2)]^{-r/2} \le h^{-r}$. For $\Sigma_{P,Q}$, at most r eigenvalues lie below one, all are at least $2h^2$, and the covariance differs from I on a space of dimension at most $2r$, yielding

$$\mathbb{E}W_{\Sigma_{P,Q}}(X)^2 \le h^{-r}. \tag{B.42}$$

The likelihood-ratio variance scales with the intrinsic dimension rather than d.

B.4.2 Finite candidate family and robust comparison statistics

Let $\mathcal{Q}_h$ be an operator-norm h-net of rank-r projectors. A frame-volume argument gives

$$M := |\mathcal{Q}_h| \le (1+4\sqrt{r_0}/h)^{dr_0}.$$

For $f(x) = g(B^\top x)$ define

$$\mu_Q = \mathbb{E}[W_{\Sigma_Q}(X)Y], \qquad \mu_{P,Q} = \mathbb{E}[W_{\Sigma_{P,Q}}(X)Y].$$

These are Gaussian-smoothed values and satisfy

$$\mathbb{E}[W_\Sigma(X)^2Y^2] \le (G^2+\sigma^2)h^{-r}$$

for every covariance used here. On a selection sample of size m, choose the smallest odd J with

$$J \ge \log_2((M+1)^2/\eta),$$

partition into J blocks of size $\ell = \lfloor m/J \rfloor$, and estimate every μ by the median of its block means. Chebyshev plus a median tail bound yields, with probability at least $1-\eta$,

$$\max_Q |\widehat{\mu}_Q - \mu_Q| \le t_h, \qquad \max_{P,Q} |\widehat{\mu}_{P,Q} - \mu_{P,Q}| \le t_h, \tag{B.43}$$

where

$$t_h \le 4\sqrt{\frac{2(G^2+\sigma^2)J}{mh^r}}.$$

No independence among comparison statistics is required.

B.4.3 Selection controls point prediction without controlling orientation

Define

$$\Delta(P) = \max_{Q\in\mathcal{Q}_h} |\widehat{\mu}_{P,Q} - \widehat{\mu}_Q|$$

and choose $\widehat{Q} \in \arg\min_{P\in\mathcal{Q}_h}\Delta(P)$ with deterministic tie breaking. Let $P_* \in \mathcal{Q}_h$ satisfy $\|P_* - P\|_{\mathrm{op}} \le h$. There exists a frame B_* spanning P_* such that $\|B - B_*\|_F^2 \le 2r_0h^2$. For $f_*(x) = g(B_*^\top x)$, (B.40) means that the projected law under $\Sigma_{P_*,Q}$ equals the projected law under Σ_Q plus independent $N(0,h^2I_r)$ noise. Hölder continuity therefore gives a uniform comparison discrepancy $O(Lr^{s/2}h^s)$; replacing f_* by f costs the same order.

On the uniform event (B.43), symmetry of the pair statistic provides the triangle

$$|\widehat{\mu}_{\widehat{Q}} - g(0)| \le |\widehat{\mu}_{\widehat{Q}} - \widehat{\mu}_{P_*,\widehat{Q}}| + |\widehat{\mu}_{\widehat{Q},P_*} - \widehat{\mu}_{P_*}| + |\widehat{\mu}_{P_*} - g(0)|,$$

which yields the explicit bound

$$|\mu_{\widehat{Q}} - g(0)| \le 10Lr^{s/2}h^s + 6t_h. \tag{B.44}$$

No conclusion about $\|\widehat{Q} - P\|_{\mathrm{op}}$ is needed.

B.4.4 HELD-OUT NORMALIZED ATTENTION AND RATE

On an independent smoother fold use the positive semidefinite precision

$$A = (h^{-2} - 1)\widehat{Q}.$$

Its tilted covariance is exactly $\Sigma_{\widehat{Q}}$ and its effective mass is

$$\rho_r^{\mathrm{cmp}}(h) = h^r(2 - h^2)^{r/2} \geq h^r.$$

The same denominator calculation as in Appendix B.3, now centred at $\mu_{\widehat{Q}}$, gives

$$\mathbb{E}[(\widehat{f} - \mu_{\widehat{Q}})^2 \mid \widehat{Q}] \leq \frac{4(4G^2 + \sigma^2)}{n\rho_r^{\mathrm{cmp}}(h)} + (4G^2 + \sigma^2)e^{-n\rho_r^{\mathrm{cmp}}(h)/8}.$$

Combining with (B.44) proves Theorem B.3. With

$$H_N = 5 + 2dr_0 \log_2(1 + 4\sqrt{r_0}N) + 2\log_2 N, \qquad h_N = (H_N/N)^{1/(2s+r)},$$

one obtains

$$R_N^{\mathrm{cmp}}(0) \lesssim (H_N/N)^{2s/(2s+r)}.$$

This risk bound is proved without establishing that $\widehat{Q}$ is close to P. The sketch in Section B.3.7 suggests a regime in which uniform bandwidth-scale orientation recovery is statistically harder; because its Assouad/confidence step is incomplete, we draw no formal separation or tightness conclusion from it.

B.5 QUADRATIC-ATTENTION DETERMINANT AND ALIGNMENT PROOF

Proposition B.3 (Determinant and alignment criterion)**.** *Let $D \sim N(0, \tau_0^2 I_d)$ and $A_h \succeq 0$, with eigenvalues $\lambda_{1,h}, \ldots, \lambda_{d,h}$. Then*

$$V_h(A_h) := \mathbb{E}e^{-D^\top A_h D/2} = \det(I + \tau_0^2 A_h)^{-1/2} = \prod_{j=1}^{d}(1 + \tau_0^2\lambda_{j,h})^{-1/2}.$$

Consequently $V_h(A_h) \asymp h^r$ if and only if

$$\sum_{j=1}^{d} \log(1 + \tau_0^2\lambda_{j,h}) = 2r\log(1/h) + O(1).$$

In particular, if exactly r eigenvalues are $\Theta(h^{-2})$, then $V_h(A_h) \asymp h^r$ exactly when all remaining eigenvalues are $O(1)$. For $A_h = h^{-2}M_h$, this requires the transverse eigenvalues of M_h to be $O(h^2)$. If, in addition, the remaining eigenvalues are $O(1)$ and E_h is the fast rank-r eigenspace, predictive displacement in a rank-r target projector P_B is $O(h^2)$ precisely when $\mathrm{tr}(P_B P_{E_h^\perp}) = O(h^2)$, up to fixed constants.

Proof. Diagonalize A_h and rotate the isotropic Gaussian displacement. One-dimensional Gaussian integration gives $\mathbb{E}e^{-\lambda D_j^2/2} = (1 + \tau_0^2\lambda)^{-1/2}$; independence proves the determinant identity. Taking logarithms proves the necessary-and-sufficient volume criterion. If exactly r eigenvalues are $\Theta(h^{-2})$, they contribute $2r\log(1/h) + O(1)$, while the remaining contribution is $O(1)$ exactly when every transverse eigenvalue is $O(1)$. Under the exponentially tilted displacement law the covariance is $(\tau_0^{-2}I + A_h)^{-1}$, equivalent up to fixed constants to $(I + A_h)^{-1}$. This is $\Theta(h^2)$ on E_h and $\Theta(1)$ on $E_h^\perp$, so

$$\mathrm{tr}\{P_B(I + A_h)^{-1}\} \asymp h^2\,\mathrm{tr}(P_B P_{E_h}) + \mathrm{tr}(P_B P_{E_h^\perp}) \asymp h^2 + \mathrm{tr}(P_B P_{E_h^\perp}).$$

□

C ZERO CHANNEL DOES NOT IMPLY HAAR GEOMETRY

This appendix proves the counterexample in Proposition 1. Earlier working materials also contained broader fixed-dimension null-channel claims and task-specific asymptotic corollaries. They are not needed for the results stated here and are omitted rather than carried as unproved claims.

C.1 Proof of Proposition 1

Proof of Proposition 1. Throughout, $\varepsilon \sim N(0,1)$ is independent of X, and $\widehat{\mathsf{H}}_m$ centres the transformed responses by their sample mean and does not centre the covariates. Write $s = \phi(Y)$ and $W = (s - \mathbb{E}s)(XX^\top - I_2)$. For $\phi(y) = y$, $\mathbb{E}s = 0$ and

$$W = \begin{pmatrix} (Z+\sigma\varepsilon)(Z^2-1) & (Z+\sigma\varepsilon)ZU \\ (Z+\sigma\varepsilon)ZU & (Z+\sigma\varepsilon)(U^2-1) \end{pmatrix}.$$

Each entry has mean zero, by oddness in (Z, ε) for the diagonal and by the centred independent factor U off the diagonal. Hence $\mathsf{H}_\phi = \mathbb{E}W = 0$. All moments are finite. The three distinct entries are pairwise uncorrelated: the pairs involving W_{12} are odd in U, and $\mathrm{Cov}(W_{11}, W_{22}) = \mathbb{E}[(Z+\sigma\varepsilon)^2(Z^2-1)]\,\mathbb{E}[U^2-1] = 0$. Their variances are

$$\mathbb{E}[(Z+\sigma\varepsilon)^2(Z^2-1)^2] = 10 + 2\sigma^2, \qquad \mathbb{E}[(Z+\sigma\varepsilon)^2 Z^2 U^2] = 3 + \sigma^2,$$
$$\mathbb{E}[(Z+\sigma\varepsilon)^2(U^2-1)^2] = 2 + 2\sigma^2,$$

using $\mathbb{E}Z^6 - 2\mathbb{E}Z^4 + \mathbb{E}Z^2 = 10$ and $\mathbb{E}Z^4 = 3$. By the multivariate central limit theorem, $\sqrt{m}\,\widehat{\mathsf{H}}_m = m^{-1/2}\sum_{i\le m} W_i \Rightarrow G = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ with independent $a \sim N(0, 10+2\sigma^2)$, $b \sim N(0, 3+\sigma^2)$ and $c \sim N(0, 2+2\sigma^2)$. With empirical centring of s, the difference is $-\bar{s}(S_m - I)$, where $S_m = m^{-1}\sum_i X_i X_i^\top$. Since $\bar{s} = O_p(m^{-1/2})$ and $S_m - I = O_p(m^{-1/2})$, the difference is $o_p(m^{-1/2})$, and Slutsky's lemma leaves the limit unchanged.

Put $t = (a+c)/2$, $x = (a-c)/2$ and $R = \sqrt{x^2 + b^2}$. Then $\mathrm{Var}\, t = \mathrm{Var}\, x = 3 + \sigma^2$, $\mathrm{Cov}(t, x) = (\mathrm{Var}\, a - \mathrm{Var}\, c)/4 = 2$, and b is independent of (t, x). The eigenvalues of G are $t \pm R$. Since $(t+R)^2 - (t-R)^2 = 4tR$, the eigenvalue of largest absolute value is $t + \mathrm{sign}(t)R$ almost surely, with eigenprojector

$$\Pi(G) = \tfrac{1}{2}\Big[I + \frac{\mathrm{sign}(t)}{R}\begin{pmatrix} x & b \\ b & -x \end{pmatrix}\Big], \qquad \mathrm{tr}(P\Pi(G)) = \tfrac{1}{2} + \tfrac{1}{2}\,\mathrm{sign}(t)\,\frac{x}{R}.$$

Gaussian regression gives $t \mid x \sim N(\beta x, v)$ with $\beta = 2/(3+\sigma^2) > 0$ and $v > 0$, so $\mathbb{E}[\mathrm{sign}(t) \mid x, b] = \mathbb{E}[\mathrm{sign}(t) \mid x] = 2\Phi(\beta x/\sqrt{v}) - 1$, which has the sign of x. Therefore

$$\mathbb{E}\,\mathrm{tr}(P\Pi(G)) = \tfrac{1}{2} + \tfrac{1}{2}\mathbb{E}\Big[\frac{x}{\sqrt{x^2+b^2}}\big\{2\Phi(\beta x/\sqrt{v}) - 1\big\}\Big] > \tfrac{1}{2},$$

because the integrand is bounded, nonnegative, and strictly positive when $x \neq 0$, an event of probability one. The absolute-leading projector is scale invariant and continuous at every symmetric matrix whose two eigenvalues have distinct absolute values. G has this property almost surely ($t \neq 0$ and $R \neq 0$). By the continuous-mapping theorem, $\widehat{P}_m = \Pi(\widehat{\mathsf{H}}_m) = \Pi(\sqrt{m}\widehat{\mathsf{H}}_m) \Rightarrow \Pi(G)$. Since $D = 1 - \mathrm{tr}(P\widehat{P}) \in [0,1]$ is bounded and continuous, $\mathbb{E}D(P, \widehat{P}_m) \to 1 - \mathbb{E}\,\mathrm{tr}(P\Pi(G)) < 1/2$. A Haar line has $\mathbb{E}D = 1/2$, so for all sufficiently large m the law of $\widehat{P}_m$ relative to P is not Haar. The argument gives no explicit threshold on m.

For $\phi(y) = \mathrm{clip}(y, -2, 2)$ with $\sigma = 0$, write $p = \varphi(2)$ and $q = 1 - \Phi(2)$, where φ is the standard normal density. The channel is again zero and the limit entries are again independent, by the same parity arguments. Splitting expectations at $|Z| = 2$ gives their variances $\gamma_0 = 10 - 72p - 4q$, $b_0 = 3 - 12p + 2q$ and $2a_0 = 2(1 - 4p + 6q)$. Then $\gamma_0 - 2a_0 = 8 - 64p - 16q > 0$, by $q < p/2$ and $p < 1/10$, so $\mathrm{Cov}(t, x) = (\gamma_0 - 2a_0)/4 > 0$, and the same sign-regression argument applies. □

Remark (selection rule matters). For $\phi(y) = y$ and every $\sigma \geq 0$, the *algebraically* largest eigenline of G is exactly Haar, because x and b are independent with equal variance $3+\sigma^2$. Only its absolute-value selection is biased toward P. The explicit route of Section 5 selects by absolute value.

D Experimental protocol

D.1 Tasks and the frozen channel map

All tasks use the Gaussian design of (3) with $d = 16$, a fresh Haar frame per context ($d \times r$ standard Gaussian matrix, thin QR, column signs canonicalized by $\mathrm{sign}\,\mathrm{diag}(R)$), and base noise standard

deviation 0.25; T6's noise scale depends on Z_2. Table D.1 lists the responses and the channel labels under the clipping map of the explicit route. The labels were frozen before any outcome of these runs was evaluated, and no outcome may alter a label, group, clipping level, rank or task membership. The same ten task cases were used earlier in our estimator benchmark (Section H), whose own population table (including the first-moment column used in Section F) had been frozen before that benchmark ran.

Table D.1: Frozen task suite and channel labels for the explicit route's map $\phi_\tau(y) = \mathrm{clip}(y, -\tau, \tau)$, $\tau \approx 2.146$. $e \sim N(0,1)$.

Task	r	Response Y	Label	Reason (population channel Γ_τ)
T1	1	$\tanh Z + 0.25e$	closed	joint sign symmetry: clipped response odd, $Z^2 - 1$ even
T2	1	$Z^2 - 1 + 0.25e$	open	conditional clipped response increasing in Z^2
T3	1	$0.75 \tanh(Z^2 - 1) + 0.25e$	open	monotone even channel
T4	1	$\sin 2Z + 0.25e$	closed	same symmetry as T1
T5a/b/c	1	$\min(\lvert Z\rvert^q, 1) + 0.25e$	open	$q = 0.25, 0.5, 1$; nonconstant monotone in Z^2
T6	2	$\tanh Z_1 + \frac{1}{4}\sqrt{1 + \frac{1}{2}\tanh(Z_2^2 - 1)}\, e$	closed	Z_2 symmetry kills off-diagonals; (Z_1, e) reflection kills diagonals
T7	2	$Z_1 Z_2 + 0.25e$	open	zero diagonal, nonzero off-diagonal: eigenvalues $\pm c$
T8	2	$\tanh Z_1 + 0.5 \tanh(Z_2^2 - 1) + 0.25e$	num. weak	$\Gamma_{11} \approx -2.92 \times 10^{-6}$, $\Gamma_{22} \approx 0.379$

The groups are open $= \{$T2, T3, T5a, T5b, T5c, T7$\}$, closed $= \{$T1, T4, T6$\}$ and numerically weak $= \{$T8$\}$. T8 is excluded from both H1 contrasts. T6's conditional mean $\tanh Z_1$ has rank one; its second direction enters only through the noise scale.

D.2 Arms

Explicit route E. Four deterministic contiguous folds of the $N = 256$ context points. For each fold, the pilot is the other 192 points and the smoother is the fold's 64 points. The pilot responses are clipped at $\tau = 1 + 0.25\sqrt{4 \log 192} = 2.14646$ and centred by their pilot mean, giving $c_1, \ldots, c_m$ with $\sum_i c_i = 0$; the moment $m^{-1} \sum_i c_i (X_i X_i^\top - I)$ is formed, and $\widehat{P}$ projects onto its r eigenvectors of largest absolute eigenvalue. The smoother fold predicts each query by the quadratic-softmax row with precision $h^{-2}\widehat{P} + \alpha(I - \widehat{P})$, $\alpha = 0.25$, $h = c_h\, 64^{-1/(2s_\mathrm{w}+r)}$ with $c_h = 2$ and working smoothness $s_\mathrm{w} = 1$ ($h = 0.5$ for $r = 1$, 0.7071 for $r = 2$). The four fold predictions are averaged. These hyperparameters were selected in the radial even-link study of Anonymous (2026), whose protocol provenance is summarized in Section G, and were not tuned here. *Haar control H.* Identical, except that each fold uses an independent Haar rank-r projector, drawn independently of the task frame, the data and the training randomness.[1] *Oracle O.* Identical, with the true P. It therefore remains a 64-point-per-fold fixed-bandwidth smoother, not the Bayes predictor; it is an anchor for the value of geometry under the common smoothing rule. *Trained Transformer S.* Described in Section D.3. The T6 secondary oracle smooths along the true mean direction z_1 alone (rank-1 projector; risk 0.017877).

[1] Haar projectors were drawn per context and fold and shared across tasks of the same rank, which departs from the preregistered seed derivation (a task-specific stream). Because task frames are independent Haar draws, the shared projector induces no cross-task dependence and leaves each task's H arm with its intended law. Over the 24 same-rank task pairs, 7 of the 48 same-context correlations of D_H and of R_H (mean $+0.0002$) have permutation $p < 0.05$, against 2.4 expected; the tests are dependent, and the smallest p, 0.002, is 0.09 after Bonferroni correction. The bootstrap resamples one context-index vector across tasks, so any residual dependence enters the intervals.

D.3 TRAINED TRANSFORMER

Support tokens are $(x_i, y_i, 1) \in \mathbb{R}^{d+2}$ and query tokens are $(x, 0, 0)$, with no positional encoding. There is no attention mask between the 32 query tokens: each query may therefore use the other query covariates, but no query response is present or revealed. This frozen architecture defines a 32-query-batch empirical predictor rather than the strict point-query row of the theory, and we do not use its risk as a direct test of the point-query theorem. The architecture has a linear token projection $\mathbb{R}^{d+2} \to \mathbb{R}^{64}$, a 2-layer TransformerEncoder with 4 heads, feed-forward width 256, dropout 0 and no attention mask over support and query tokens, and a linear readout $\mathbb{R}^{64} \to \mathbb{R}$. The model has 101,249 parameters. Training is task-specific: each step draws 8 fresh contexts (support 256, 32 queries, conditional-mean targets) from an online generator that implements the task formulas exactly, in float32 and with its own random-number stream; the development and evaluation banks come from the benchmark generator. The loss is query mean squared error. We use Adam ($\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay, no scheduler) at learning rate 10^{-3}. Development-bank MSE (64 contexts) is evaluated every 100 steps; training runs at least 800 and at most 3,000 steps with patience of 5 evaluations, and we keep the best development checkpoint. Confirmatory risk is never used for selection. Three training seeds per task were fixed in advance; the primary R_S averages the three seed *losses*, never their predictions. The architecture, input format and recipe follow Anonymous (2026); the 30 checkpoints here are new. There was no architecture search, no per-task tuning and no seed replacement.

D.4 INFERENCE

For each task we use 512 confirmatory contexts, 32 iid $N(0, I_{16})$ queries per context, and the same contexts, queries, folds and noise for all arms. The context is the inferential unit. Queries are averaged within a context, and seeds are never treated as replicates. Two separately seeded 9,999-replicate context-bootstrap index matrices, one for the risk family and one for the geometry family, are each shared across tasks and arms; the secondary $R_H - R_O$ family reuses the risk matrix. Intervals are fixed-studentized max-t with the higher empirical quantile, and are simultaneous over three separate families: the risk family (30 taskwise coordinates C_{HE}, C_{HS}, C_{EO} plus the four summaries $K_{\mathrm{H1},R}$, $K_{\mathrm{H2},HS}$, $K_{\mathrm{H2},ES}$, $K_{\mathrm{H3},HS}$; critical value 3.143443), the geometry family (10 C_D plus $K_{\mathrm{H1},D}$; 2.825593), and a secondary $R_H - R_O$ family (2.790503). The frozen readouts are: H1 is supported iff $\mathrm{LB}(K_{\mathrm{H1},R}) > 0$ and $\mathrm{LB}(K_{\mathrm{H1},D}) > 0$; H2 is supported iff $\mathrm{LB}(K_{\mathrm{H2},HS}) > 0$ and $\mathrm{LB}(K_{\mathrm{H2},ES}) > 0$; and the rank-2 components are $\mathrm{LB}(C_{HE}(\mathsf{T7})) > 0 \wedge \mathrm{LB}(C_D(\mathsf{T7})) > 0$ (explicit) and $\mathrm{LB}(K_{\mathrm{H3},HS}) > 0$ (trained). All ten tasks were complete, with zero incomplete cells. Evaluation ran on CPU. Explicit-route geometry and bootstrap computations take seconds per task. Transformer training used at most 90,000 optimizer steps in total, within a budget fixed in advance.

D.5 LEARNING-RATE CORRECTION (DISCLOSURE)

The Transformer learning rate was initially frozen at 3×10^{-4}, obtained by replaying the selection procedure of the radial-link comparison. After the first confirmatory run had been evaluated, checking a pre-existing archived training record showed that the recorded selection was 10^{-3}. All 30 models were then retrained at 10^{-3} with unchanged seeds and no selection, and the confirmatory evaluation was rerun. We report that run as primary and retain the original run as a sensitivity comparison. Every preregistered decision (H1, H2 and both rank-2 components) is the same under both learning rates. The E, H and O arms and the geometry family are bit-identical across the two runs. Means of risk quantities that do not involve S are unchanged, but their simultaneous bounds shift slightly, because the risk family's critical value is set jointly with the S coordinates (3.144175 at 3×10^{-4}, 3.143443 at 10^{-3}; Tables E.1 and E.2). The task whose qualitative status changes is $\mathsf{T2}$: at 3×10^{-4} all three seeds sit at the Haar level, while at 10^{-3} two of three seeds improve sharply.

D.6 CALIBRATION AND IMPLEMENTATION NOTES

The explicit route uses the same frozen construction as Anonymous (2026). On CPU, its expectation on that comparison's task is about 2.3% below the recorded CUDA value ($z \approx -2.2$), while the oracle arm's CPU value is within sampling error of its CUDA value ($z \approx -0.07$). The two evaluations use different random draws, because the original banks were generated on a GPU. The residual is

unresolved and has no bearing on the paired, same-device contrasts reported here. We make no claim of agreement tighter than this.

E FULL TEN-TASK RESULTS

Tables E.1 to E.4 report every preregistered coordinate and the secondary quantities. They are generated directly from the frozen analysis outputs by the script in the source package, so no number is transcribed by hand.

Table E.1: Raw context-mean risks (512 contexts $\times$ 32 queries) and mean projector error D per task. R_S: primary run (learning rate 10^{-3}), mean of three seed risks; R_S^{sens}: sensitivity run (3×10^{-4}). E, H, O and D do not depend on S and are identical in both runs. Haar references: $1 - r/d = 0.9375$ ($r = 1$), 0.875 ($r = 2$).

Task	channel	r	R_S	R_S^{sens}	R_E	R_H	R_O	D_E	D_H
T1	closed	1	0.01038	0.01429	0.22697	0.25210	0.01833	0.84038	0.93685
T2	open	1	1.52915	1.96353	0.76349	1.94609	0.48277	0.12970	0.93833
T3	open	1	0.21306	0.21287	0.07829	0.21191	0.03934	0.13580	0.93782
T4	closed	1	0.39744	0.38438	0.46929	0.46884	0.12407	0.93415	0.93838
T5a	open	1	0.02918	0.02897	0.02716	0.03039	0.01322	0.60517	0.93722
T5b	open	1	0.06557	0.06570	0.04683	0.06645	0.02366	0.30991	0.93616
T5c	open	1	0.11387	0.11348	0.06231	0.11474	0.03198	0.18478	0.93624
T6	closed	2	0.00938	0.01480	0.20646	0.23967	0.04996	0.79273	0.87484
T7	open	2	1.03451	1.03541	0.55038	0.93871	0.40344	0.17982	0.87481
T8	weak	2	0.10029	0.11108	0.29886	0.33205	0.08903	0.60245	0.87375

Table E.2: All 34 coordinates of the preregistered risk family with simultaneous 95% max-t intervals (9,999 context-bootstrap replicates, critical value 3.143443). $C_{HE} = \mathbb{E}_c[R_H - R_E]$, $C_{HS} = \mathbb{E}_c[R_H - R_S]$, $C_{EO} = \mathbb{E}_c[R_E - R_O]$. The last column gives C_{HS} in the sensitivity run (its own family). Summary rows: $K_{\text{H1},R}$, open-minus-closed mean of C_{HE}; $K_{\text{H2},HS}$ and $K_{\text{H2},ES}$, closed-task means of $R_H - R_S$ and $R_E - R_S$; $K_{\text{H3},HS}$, mean of C_{HS} over T7 and T8 (Section E). C_{HE} agrees on T6 and T8 to four decimals by coincidence: 0.03321 $[0.02730, 0.03912]$ and 0.03319 $[0.02731, 0.03908]$ (the raw means differ: $R_H - R_E = 0.03321$ and 0.03319).

Task	C_{HE}	C_{HS}	C_{EO}	C_{HS}^{sens}
T1	0.0251 [0.0199, 0.0304]	0.2417 [0.2365, 0.2470]	0.2086 [0.2027, 0.2146]	0.2378 [0.2326, 0.2430]
T2	1.1826 [1.0742, 1.2910]	0.4169 [0.3365, 0.4974]	0.2807 [0.2512, 0.3103]	−0.0174 [−0.0500, 0.0151]
T3	0.1336 [0.1287, 0.1385]	−0.0011 [−0.0042, 0.0019]	0.0390 [0.0364, 0.0415]	−0.0010 [−0.0040, 0.0021]
T4	−0.0004 [−0.0046, 0.0037]	0.0714 [0.0656, 0.0772]	0.3452 [0.3369, 0.3535]	0.0845 [0.0773, 0.0916]
T5a	0.0032 [0.0027, 0.0038]	0.0012 [0.0008, 0.0017]	0.0139 [0.0133, 0.0146]	0.0014 [0.0010, 0.0019]
T5b	0.0196 [0.0183, 0.0209]	0.0009 [−0.0000, 0.0018]	0.0232 [0.0220, 0.0243]	0.0007 [−0.0001, 0.0016]
T5c	0.0524 [0.0503, 0.0545]	0.0009 [−0.0006, 0.0023]	0.0303 [0.0288, 0.0319]	0.0013 [−0.0003, 0.0028]
T6	0.0332 [0.0273, 0.0391]	0.2303 [0.2247, 0.2358]	0.1565 [0.1510, 0.1620]	0.2249 [0.2193, 0.2305]
T7	0.3883 [0.3609, 0.4157]	−0.0958 [−0.1078, −0.0838]	0.1469 [0.1343, 0.1595]	−0.0967 [−0.1087, −0.0847]
T8	0.0332 [0.0273, 0.0391]	0.2318 [0.2226, 0.2409]	0.2098 [0.2027, 0.2169]	0.2210 [0.2121, 0.2299]
$K_{\text{H1},R}$	0.2773 [0.2578, 0.2969]			0.2773 [0.2578, 0.2969]
$K_{\text{H2},HS}$	0.1811 [0.1779, 0.1843]			0.1824 [0.1790, 0.1857]
$K_{\text{H2},ES}$	0.1618 [0.1583, 0.1654]			0.1631 [0.1594, 0.1668]
$K_{\text{H3},HS}$	0.0680 [0.0602, 0.0758]			0.0621 [0.0544, 0.0699]

Reading the trained rank-2 component. The preregistered trained rank-2 contrast is positive, $K_{\text{H3},HS} = (C_{HS}(\text{T7}) + C_{HS}(\text{T8}))/2 = 0.06799$ $[0.06016, 0.07582]$, but its positive lower bound comes entirely from T8. It averages a genuine rank-2 interaction on which S is worse than Haar with an interval excluding zero ($C_{HS}(\text{T7}) = -0.09580$ $[-0.10778, -0.08381]$) with a numerically weak mixed-parity task ($C_{HS}(\text{T8}) = 0.23177$ $[0.22259, 0.24095]$). We report the contrast because it was preregistered, but we do not use it as evidence of trained rank-2 use. The frozen metrics cannot attribute S's T8 risk to z_1 rather than z_2. For scale, the population-optimal z_1-only predictor on T8 has risk 0.0984 (deterministic quadrature), the same order as $R_S = 0.1003$. This neither establishes nor excludes that S uses z_1 only. T6's conditional mean has rank one, so S's success there is rank-1 mean prediction.

Table E.3: Geometry family (11 coordinates; critical value 2.825593), the secondary $R_H - R_O$ family (critical value 2.790503) and the oracle-normalized opportunity ratios $G_A = (R_H - R_A)/(R_H - R_O)$, reported because every $R_H - R_O$ lower bound is positive: point value and, in brackets, the preregistered conservative interval from the simultaneous bounds of numerator and denominator. $G_S > 1$ is possible because O is a fixed-bandwidth smoother given the true P, not the Bayes predictor, whereas S is trained on the task.

Task	$C_D = \mathbb{E}_c[D_H - D_E]$	$R_H - R_O$	G_E	G_S
T1	0.09647 [0.08370, 0.10924]	0.23377 [0.22939, 0.23814]	0.107 [0.084, 0.132]	1.034 [0.993, 1.077]
T2	0.80863 [0.80174, 0.81552]	1.46332 [1.35290, 1.57374]	0.808 [0.683, 0.954]	0.285 [0.214, 0.368]
T3	0.80202 [0.79509, 0.80895]	0.17257 [0.16794, 0.17721]	0.774 [0.726, 0.825]	−0.007 [−0.025, 0.011]
T4	0.00423 [−0.00437, 0.01282]	0.34477 [0.33737, 0.35217]	−0.001 [−0.014, 0.011]	0.207 [0.186, 0.229]
T5a	0.33204 [0.30599, 0.35809]	0.01717 [0.01659, 0.01775]	0.188 [0.150, 0.229]	0.071 [0.042, 0.101]
T5b	0.62626 [0.60718, 0.64534]	0.04279 [0.04171, 0.04387]	0.458 [0.418, 0.501]	0.020 [−0.000, 0.042]
T5c	0.75145 [0.74169, 0.76122]	0.08276 [0.08108, 0.08444]	0.634 [0.596, 0.672]	0.010 [−0.007, 0.029]
T6	0.08212 [0.07220, 0.09203]	0.18971 [0.18551, 0.19391]	0.175 [0.141, 0.211]	1.214 [1.159, 1.271]
T7	0.69499 [0.68828, 0.70171]	0.53528 [0.50550, 0.56505]	0.725 [0.639, 0.822]	−0.179 [−0.213, −0.148]
T8	0.27131 [0.26096, 0.28165]	0.24303 [0.23637, 0.24969]	0.137 [0.109, 0.165]	0.954 [0.891, 1.019]
$K_{\mathrm{H1},D}$	0.60830 [0.59978, 0.61681]			

Table E.4: Seed-specific context-mean risk of the trained Transformer (primary run), columns in increasing order of the preregistered seed values. Seeds are not inferential units; the primary R_S averages the three losses.

Task	seed 1	seed 2	seed 3
T1	0.00978	0.01135	0.01000
T2	1.96906	1.41439	1.20399
T3	0.21363	0.21273	0.21283
T4	0.50171	0.50115	0.18946
T5a	0.02937	0.02898	0.02917
T5b	0.06556	0.06540	0.06575
T5c	0.11408	0.11370	0.11385
T6	0.00893	0.00972	0.00948
T7	1.03456	1.03463	1.03434
T8	0.10037	0.09466	0.10583

T7 training trace. On T7, all three seeds early-stopped on a development-MSE plateau near 0.975 at their best steps (between 300 and 1,900), well before the 3,000-step cap, under both learning rates. The confirmatory risk (1.0345) matches the constant-zero predictor on the same contexts (1.0340; the population value is $\mathrm{Var}(Z_1 Z_2) = 1$). The negative result is therefore scoped to the frozen recipe, whose patience-based stopping is exactly where a longer budget could behave differently.

T2 and T4 seed dependence. On T2, S separates from Haar in the primary run through two of three seeds (1.414 and 1.204; the third stays at 1.969, near $R_H = 1.946$). In the sensitivity run it sits at the Haar level ($C_{HS} = -0.017$, interval containing 0). On T4, two primary seeds sit at the constant-predictor risk (≈ 0.50) and one reaches 0.189. The seed-averaged risk still beats both controls, so T4's contribution to H2 is real on the preregistered metric but rests on one seed. T1 and T6 are robust across seeds and learning rates.

F Post hoc analyses: two routes, two channels, channel strength and seeds

Everything in this appendix is descriptive and post hoc. The first-moment column comes from the task table frozen before our estimator benchmark (Section H), which records whether the sliced-inverse-regression channel $\mathrm{Cov}(\mathbb{E}[Z \mid Y])$ is nonzero. That column was *not* registered as a prediction for the trained Transformer S. On this suite its zeros are exactly the even-link tasks, where parity removes every first-order moment. Consequently the first-moment-open set equals the explicit-closed set plus T8, and the first-moment column carries almost no information beyond the explicit channel map.

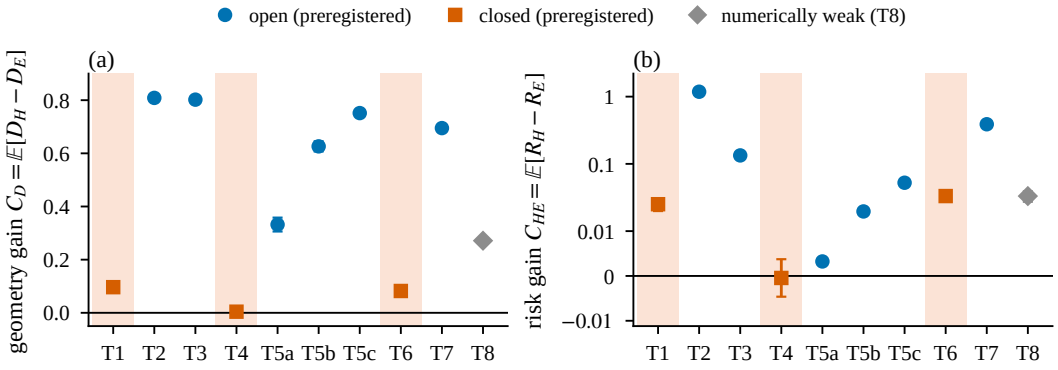


Figure E.1: Preregistered coordinates of the explicit route in frozen task order: (a) C_D; (b) C_{HE} on a symmetric-log axis, linear within ± 0.01. Bars: simultaneous 95% intervals over 512 contexts; closed tasks shaded.

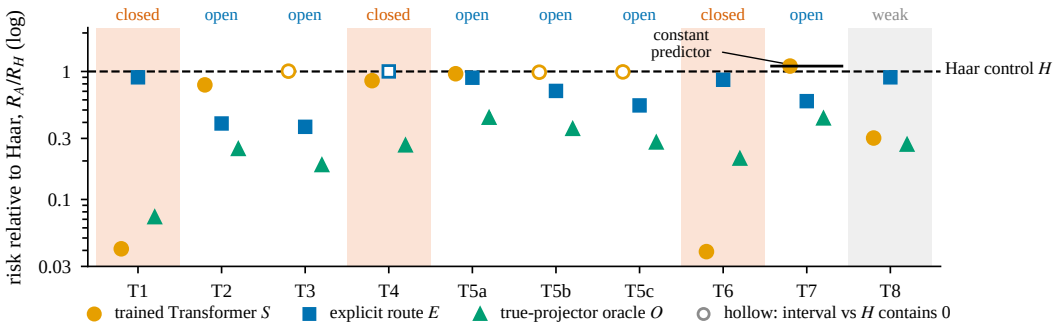


Figure E.2: Risk of S, E and O divided by the matched Haar-control risk (log scale; dashed: H), frozen task order. Filled markers: the paired difference from H has a simultaneous 95% interval excluding zero. Black bar on T7: constant predictor.

The 10/10 sign agreement below follows from the parity structure of the suite and is not independent confirmation. T8 is the only task where the two readings could come apart, because both channels are open there, and T8 cannot be attributed.

Table F.1: Per task, in frozen order: first-moment column of the task table, explicit-channel label, trained separation from Haar (C_{HS}, primary run, simultaneous 95% interval), and the point difference $R_E - R_S$ in both runs (arithmetic on reported means; not a preregistered quantity, no interval).

Task	first-moment col.	explicit channel	C_{HS} (primary)	$R_E - R_S$ (primary)	$R_E - R_S$ (sens.)
T1	1	closed	$0.2417\ [0.2365, 0.2470]$	$+0.2166$	$+0.2127$
T2	0	open	$0.4169\ [0.3365, 0.4974]$	-0.7657	-1.2000
T3	0	open	$-0.0011\ [-0.0042, 0.0019]$	-0.1348	-0.1346
T4	1	closed	$0.0714\ [0.0656, 0.0772]$	$+0.0718$	$+0.0849$
T5a	0	open	$0.0012\ [0.0008, 0.0017]$	-0.0020	-0.0018
T5b	0	open	$0.0009\ [-0.0000, 0.0018]$	-0.0187	-0.0189
T5c	0	open	$0.0009\ [-0.0006, 0.0023]$	-0.0516	-0.0512
T6	1 (mean)	closed	$0.2303\ [0.2247, 0.2358]$	$+0.1971$	$+0.1917$
T7	0	open	$-0.0958\ [-0.1078, -0.0838]$	-0.4841	-0.4850
T8	1 (z_1)	num. weak	$0.2318\ [0.2226, 0.2409]$	$+0.1986$	$+0.1878$

On the four first-moment-open tasks, S separates from Haar by a wide margin in both runs (on T4 in seed average only; Section F.2). On the six first-moment-zero tasks, three are null against Haar (T3, T5b, T5c), one shows a negligible positive separation (T5a), one is negative (T7), and one is partial and learning-rate dependent (T2). The descriptive pattern is complementary: by point estimate, S is worse than E on every open task and better than E on every closed task (T2 learning-rate dependent

in size, T5a negligible). A natural hypothesis links this pattern to gradient training on index models (Ben Arous et al., 2021; Abbe et al., 2023), in which links with a nonzero first Hermite coefficient are learned fastest, while links with information exponent at least 2 (including every even link) and interaction components that require a leap spend most of training in a search phase or plateau. This hypothesis remains untested here. Separating "reads a first-moment channel" from "succeeds where the explicit channel is closed" would require a new, preregistered task on which both channels are open with separately controlled strengths.

F.1 Channel strength, geometry error and channel nats

The strength $\gamma_\tau = s_r(\Gamma_\phi)$ is computed for the explicit route's map $\phi(y) = \mathrm{clip}(y, -\tau, \tau)$, $\tau = 2.14646$, with $\Gamma_\phi = \mathbb{E}[\phi(Y)(ZZ^\top - I_r)]$, for rank-1 tasks by adaptive quadrature with the clipped Gaussian mean in closed form (split at the kinks $z = 0, \pm 1$), and for rank-2 tasks by 160×160 Gauss–Hermite quadrature (T6's noise scale as in Table D.1); an independent Monte Carlo check agrees to three decimals. The quadrature reproduces the frozen T8 entry $\Gamma_{22} = 0.379$. D_E, D_H and G_E are the frozen outcomes. The last column, $\frac{r(d-r)}{2}\log(D_H/D_E)$, is a rate–distortion proxy for the nats carried by $\widehat{P}$: it calibrates the unknown additive constant of Corollary B.1 by the Haar value and is not an estimate of $I(P; \widehat{P})$.

Table F.2: Post hoc channel strength and frozen outcomes per task. Rank-1 open tasks: γ_τ, D_E and G_E are perfectly rank-ordered (Spearman ± 1, five tasks).

Task	label	r	γ_τ	D_E	D_H	G_E	proxy (nats)
T1	closed	1	0	0.840	0.937	0.107	0.8
T2	open	1	1.258	0.130	0.938	0.808	14.8
T3	open	1	0.569	0.136	0.938	0.774	14.5
T4	closed	1	0	0.934	0.938	−0.001	0.0
T5a	open	1	0.133	0.605	0.937	0.188	3.3
T5b	open	1	0.217	0.310	0.936	0.458	8.3
T5c	open	1	0.314	0.185	0.936	0.634	12.2
T6	closed	2	0	0.793	0.875	0.175	1.4
T7	open	2	0.814	0.180	0.875	0.725	22.1
T8	num. weak	2	3×10^{-6}	0.602	0.874	0.137	5.2

The proxy is at most 1.4 nats on the closed tasks, 3–15 nats on the rank-1 open tasks and 22 nats on T7. Its constant is uncontrolled, so only its ordering is informative. A test of the $1/m$ scaling in Theorem 2(i) needs geometry errors at several pilot sizes, which the confirmatory run does not have.

F.2 Per-seed outcomes for H2

H2's intervals condition on three fixed trained models. Table F.3 gives the seed-level picture on the closed tasks (primary run). With the median seed in place of the seed average, $\mathbb{E}_{\mathrm{closed}}[R_H - R_S] = 0.147$ and $\mathbb{E}_{\mathrm{closed}}[R_E - R_S] = 0.127$ (point estimates, no intervals). In the sensitivity run, two of three T4 seeds beat both controls and all seeds do on T1 and T6.

Table F.3: Closed tasks, primary run: control risks, the three seed risks of S (increasing seed value), and the number of seeds below both controls.

Task	R_H	R_E	seed 1	seed 2	seed 3	seeds below both
T1	0.2521	0.2270	0.0098	0.0113	0.0100	3/3
T4	0.4688	0.4693	0.5017	0.5012	0.1895	1/3
T6	0.2397	0.2065	0.0089	0.0097	0.0095	3/3

G Historical radial-link provenance

The explicit arm and Transformer training recipe were inherited from the single radial even-link study of Anonymous (2026), where the second-moment channel is open. We reuse that frozen

construction only as protocol provenance: the ten-task experiment in this paper uses new evaluation populations and 30 new task-specific Transformer checkpoints, and no current hypothesis is supported by re-counting the historical radial-link result. The near-matched T5c task provides the relevant within-study comparison in the current evidence tables. The CPU-versus-CUDA implementation check for the inherited explicit arm is recorded in Section D.6; it is a calibration note, not current confirmatory evidence.

H THE SAME CHANNEL LOGIC ON PUBLISHED ESTIMATORS

Before the main experiment, the same ten task cases were used in our benchmark of published sufficient-dimension-reduction estimators feeding a common prediction pipeline. The benchmark tests the channel logic of Section 4 on estimators other than the explicit route, and it records several negative results that bound this paper's claims. We summarize it here. All numbers come from the archived benchmark analysis tables.

Design. The setting is $d = 16$ with iid Gaussian design and Haar frames. Each estimator uses $m \in \{128, 512, 2048\}$ geometry samples and a prediction support of $n = 512$; there are 256 contexts and 128 conditional-mean queries per context. Each estimated projector feeds a common Gaussian local-constant (Nadaraya–Watson) predictor. The core estimators are SIR (Li, 1991), pHd (Li, 1992; Cook, 1998), directional regression (DR) (Li & Wang, 2007), the fixed clipped response moment (clipping at ± 2), and, at appendix level, a learned scalar response map (a globally shared scalar map trained through the downstream prediction loss and then frozen). Isotropic and true-projector smoothers are the anchors. The normalized regret is $\mathrm{NR} = (R_a - R_{\mathrm{oracle}})/(R_{\mathrm{iso}} - R_{\mathrm{oracle}})$: 0 is oracle-like, 1 is isotropic-like, and values above 1 are worse than using no geometry. Simultaneous 95% max-t bands were computed over a 680-coordinate risk family and a 150-coordinate geometry family, and all denominators were stable. A population table, frozen before any benchmark run, recorded for each task whether the first-moment (SIR) channel and the scalar second-moment channel of the clipped map are zero.

Channel predictions hold. Where an estimator's population channel was zero, its geometry was not recovered: its D was near, and on some tasks measurably below, the Haar reference value. The pattern is an observation on these tasks, not a consequence of the zero channel (Proposition 1). SIR fails on the even links (T2, T3, T5a–T5c) and succeeds on T1; it is also informative on the odd T4 in the rank-1 benchmark table. pHd and the fixed clipped moment fail on the odd links T1 and T4. This pre-existing SIR result is important for interpreting the current Transformer experiment: the opposite parity pattern is not Transformer-specific. It is not a matched first-moment attention control for the present experiment, because the benchmark uses a different geometry sample size and prediction pipeline. On six of the ten tasks no response is clipped at ± 2, and there the fixed clipped moment and raw-response pHd coincide as estimators, so their agreement is not independent evidence. Table H.1 expresses risk relative to a random-direction reference. The reference comes from a prediction-trained metric router that turned out to output a near-constant direction (see below), so this table is a post hoc, appendix-strength reading.

Table H.1: Benchmark normalized regret at $m = 512$. First column: the random-direction reference's own NR. Other columns: each estimator's NR divided by the reference's (1.00 = indistinguishable from a random direction). Rank-2 tasks are omitted because the reference outputs one direction. Post hoc; the reference was not preregistered as a control.

Task	random-dir. NR	SIR	pHd	DR	fixed clip	learned map
T1	2.024	0.05	0.88	0.06	0.88	0.54
T2	1.320	0.98	0.14	0.15	0.10	0.10
T3	1.046	0.98	0.14	0.21	0.14	0.13
T4	0.977	0.56	1.00	0.98	1.00	1.00
T5a	0.492	1.00	0.57	0.90	0.57	0.52
T5b	0.688	1.00	0.36	0.65	0.36	0.30
T5c	0.808	1.00	0.23	0.44	0.23	0.20

A corollary is methodological. On T5, a random direction attains $\mathrm{NR} = 0.49$–$0.81 < 1$: restricting localization to *any* one-dimensional subspace in $d = 16$ buys enough variance reduction to beat isotropic smoothing. Beating the isotropic reference is therefore not evidence of geometry recovery. The main experiment therefore uses a matched Haar-projector row.

The fixed clipped moment does not transfer. At $m = 512$, prediction with the fixed clipped moment is worse than isotropic smoothing on T1 ($\mathrm{NR} = 1.784$), T6 (1.702) and T8 (1.493), and at the random-direction level on T4 (0.980). This covers four of ten task cases and two of the three rank-2 cases, is stable across m, and on T1 and T6 it replicates under two alternative designs (correlated Gaussian and spherical t), the only failure tasks the robustness runs covered. The pHd parity blind spot, which the fixed map does not repair, predicted it in advance. Where its channel is open, the map is useful: $\mathrm{NR} = 0.131$–0.279 on T2, T3 and T5a–T5c, and 0.203 on T7. At rank 2 its geometry error is $D = 0.068$ on T7 (both directions recovered), 0.794 on T6 (Haar reference 0.875) and 0.491 on T8, the signature of recovering one of two directions. The learned scalar map reaches $\mathrm{NR} = 1.10$, 1.06 and 1.08 on T1, T6 and T8, and 0.39 on T7. The learned map is reported only as a descriptive appendix comparison.

Forward and slicing-free estimators. An extended panel ran on T1, T3, T6, T7 and T8 with 128 contexts and $m \in \{128, 512\}$. The kernel-based conditional-distribution method was dropped for runtime before any extended outcome was seen, so no statement here covers that class. The forward central-mean estimator RMAVE (Xia et al., 2002) attains $\mathrm{NR} = 0.011$, 0.011, 0.005, 0.007 and 0.018 at $m = 512$, one to two orders of magnitude below the fixed clipped moment on the same cells (1.725, 0.154, 1.728, 0.196, 1.498). In 45.5% of fits it hit its frozen 50-sweep optimizer cap. An under-converged fit should be worse, not better, so the cap does not discount this result. Its objective is aligned with conditional-mean prediction, a plausible but untested reason for much of the margin. The weighted PHD (Li et al., 2024) tracks the fixed map where the fixed map's second-moment channel is open (0.152 against 0.154 on T3) but partly escapes the parity blind spot where it is closed (0.321 on T1 and 0.308 on T6), as expected from aggregating over a family of response weights. Accordingly, the explicit route is used here as a mechanism probe rather than as a competitive geometry estimator.

A prediction-trained metric router does not find geometry. A small raw-context metric router trained end-to-end from prediction loss alone collapses to a near-constant read-out direction. Across 448 held-out contexts the pairwise $|\cos|$ is at least 0.9986, and scrambling the responses changes the direction by $\cos > 0.9999999$. Its alignment with the true subspace matches the uniform-direction law in $d = 16$ (mean $D = 0.9368$ against 0.9377 simulated) on the seven rank-1 task cases and on its own training distribution. Generic metric capacity did not identify task geometry in this architecture; the explicit response-moment construction did where its channel was open. The router result is negative for this architecture and training run; this benchmark contains no successful learned or task-adaptive geometry method. Anonymous (2026) (its App. M.4) report a matching null result for a related prediction-trained metric router.

Disclosed failure modes. (i) The parity blind spot of pHd-type moments on odd-mean links. (ii) Rank-2 non-generality of the fixed map. (iii) An empty-neighbourhood fallback of the compact-kernel isotropic reference on 30–42% of queries, which affects compact-kernel transfer ratios (not used here). (iv) The T6 estimand mismatch: the central subspace has rank 2 but the central mean has rank 1. (v) The chance floor $\mathrm{NR} < 1$. (vi) The identity of the fixed map and raw pHd on six tasks. (vii) The RMAVE optimizer cap. (viii) The absence of any conditional-distribution method.

I Extended related work

Minimax localization and dimension reduction. The ambient rate $n^{-2s/(2s+d)}$ and the intrinsic rate $n^{-2s/(2s+r)}$ are classical endpoints for nonparametric regression (Stone, 1982). The index space can be estimated at a near-parametric rate (Dalalyan et al., 2008, for structural dimension at most 4), smoothing on data projected onto a $\sqrt{n}$-consistent estimate attains index-dimension rates (Klock et al., 2021), and a kernel smoother that tunes a full bandwidth matrix by cross-validation reaches rates governed by the number of indices (Conn & Li, 2019). Direction-aggregation and adaptive-kernel procedures can also estimate the regression function while adapting to an unknown single-index

direction (Gaïffas & Lecué, 2007; Lepski & Serdyukova, 2014); these are direct antecedents of our comparison-selector viewpoint. Theorem 1, from Anonymous (2026), is complementary. It fixes the class of positive normalized response-blind rows, lets the geometry change with every task, and characterizes the whole path between the two endpoints as a function of the information available about that geometry. The information cost enters through a posterior Grassmann incidence inequality. The matching construction uses Grassmann codes (Dai et al., 2008) and connects to rate–distortion for general sets (Riegler et al., 2018).

Mechanistic and in-context learning studies. Controlled synthetic studies of trained transformers identify algorithms and learning dynamics in context (Akyürek et al., 2023; von Oswald et al., 2023; Reddy, 2024). Our experiment is of this kind in its use of controlled tasks. It differs in making a population prediction per task before training, with a matched random-geometry control, and in reporting where the trained route fails.

Response moments and their blind spots. The parity blind spot of second-moment methods on odd links, and of first-moment methods on even links, is classical in sufficient dimension reduction and motivates combined and forward methods (Li, 1991; 1992; Cook, 1998; Xia et al., 2002). The slicing-free weighted PHD aggregates the scalar-weighted second-moment field over response values (Li et al., 2024); the martingale difference divergence matrix (Lee & Shao, 2018) and its volatility version (Lee & Shao, 2020) are related distance-weighted conditional-second-moment constructions. We use this classical structure as a pre-data prediction for an attention route rather than as a new SDR estimator.

Table I.1: Positioning against the closest results (qualitative; each entry paraphrases the cited abstract).

Work	Geometry across tasks	Estimator / mechanism	Guarantee
Stone (1982)	none (full d-dimensional class)	any	ambient minimax rate
Conn & Li (2019)	fixed index space	Nadaraya–Watson with cross-validated bandwidth matrix	rate governed by the number of indices
Kim et al. (2024)	shared basis learned in pretraining	deep features + linear attention	in-context minimax optimality over Besov-type classes
Oko et al. (2024)	single index drawn from a shared r-dim. subspace	pretrained nonlinear transformer	prompt length depending on r, not d
Ching et al. (2026)	none (Hölder in d dims)	transformer emulating local polynomials	ambient in-context minimax rate
Theorem 1 (Anonymous, 2026)	fresh Haar subspace per context; $\mathcal{I}$ nats of side information	positive normalized response-blind rows; quadratic softmax attains	sharp ambient-to-intrinsic frontier in $\mathcal{I}$
Theorem 2 (ours; sharpens Anonymous, 2026)	fresh subspace read from responses	cross-fitted clipped moment + quadratic softmax	intrinsic rate at a point query when the channel is open and $2s+r \geq 2$, with a fixed clipping level and explicit constants
Theorem B.3 (ours)	fresh unknown subspace; no second-moment eigengap	net of candidate Gaussian smoothers + symmetric prediction comparisons	intrinsic polynomial exponent up to a logarithmic selection factor; no projector-consistency conclusion