SUBSPACE RECOVERY FROM REPEATED BINARY RULES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study recovery of an unknown subspace from binary classification rules whose parameters lie in that subspace. Each hidden rule is chosen uniformly and labels two independent Gaussian inputs; the learner observes the inputs and labels, but not the rule. With one observation per independently drawn rule, the subspace is unidentifiable. With two observations, we derive the exact likelihood and Fisher information. For every fixed ambient dimension d and known rank $3 \le k < d$, we determine the asymptotic minimax constant for normalized squared projector loss and give an estimator that attains it. The estimator uses spectral pilots, smoothly truncated likelihood corrections, and sensitivity calibration. The truncation is needed for our proof: at rank three, a raw score can have infinite variance at an arbitrarily nearby incorrect pilot. We also obtain the exact asymptotic risk of a simpler signed-moment estimator and a two-lag covariance formula for overlapping histories. In a separate synthetic transfer study with unmarked historical task boundaries, the temporal spectral memory improves early accuracy on an independently drawn target by 2.84 percentage points over erased history (nominal 99% interval $[2.21, 3.49]$; 64 fresh contexts). The transfer experiment uses supplied rank and long histories; it is not an empirical verification of the minimax theorem.

1 INTRODUCTION

Consider binary classification tasks with parameter vectors in a common, unknown low-dimensional subspace. A learner observes inputs and binary labels from earlier tasks and then receives examples from a new task. A representation of the common subspace can reduce the number of directions that must be learned from the new examples. We ask what information about that representation is created by reusing a hidden rule over time.

This question depends on the observation model. If a homogeneous linear rule is redrawn independently from an antipodally symmetric distribution after every example, its label is a fair sign conditional on the input. In the model considered here, the resulting law contains no information about the subspace. Reusing a rule for a second example changes the joint label law. We characterize that change and the best possible asymptotic accuracy of subspace recovery.

Previous work. Shared representations and transfer between tasks have a substantial literature. Tripuraneni et al. (2021) prove guarantees for learning linear representations across tasks. Closest to our moment construction, Gulluk et al. (2021) study nonlinear and binary tasks, estimate shared subspaces using cross-products of taskwise label–feature sums, and prove benefits for future tasks. For two observations per task, their cross-product construction already has the symmetrized pair form used here. Consequently, neither the use of cross-moments nor the general idea of binary-task subspace transfer is a contribution of this paper. Distribution-free few-sample meta-learning, including halfspace representations, is studied by Aliakbarpour et al. (2024); our Gaussian design and uniform homogeneous teachers impose different and stronger assumptions.

Other related observation models are also important. Chi and Fu (2017) recover principal subspaces from binary comparisons of aggregated energy projections; those observations are not labels from a reused hidden classifier. Gautier and Kitamura (2013) study nonparametric recovery of random coefficients in binary-choice models. Their inverse problem differs from recovering the support subspace of the symmetric paired-rule distribution considered here. One-bit covariance estimation and subspace estimation under coarse quantization also have rigorous recovery guarantees (Dirksen et al., 2022; 2025); those observations quantize signal coordinates rather than labels produced by

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.

one latent classifier reused on two covariates. Temporal covariance and factor-space estimation are established methods (Lam and Yao, 2012); we do not claim temporal spectral estimation itself as new. Work on task-unsegmented meta-learning (Harrison et al., 2020) also predates our experiment, whose target phase is explicitly marked.

Our statistical arguments use established tools. Quadratic-mean differentiability and information in transformation models are treated by Ruckdeschel (2010). We use the van Trees approach to minimax lower bounds (Gill and Levit, 1995; Gassiat and Stoltz, 2024), including prior work tailoring van Trees and group-equivariance arguments to principal-subspace estimation (Wahl, 2023), and standard one-step and Grassmann constructions (Luedtke and Chung, 2023; Bendokat et al., 2023). Our results concern the exact likelihood, information, regularity boundary, and attainable minimax constant for the specified paired-rule experiment, not a new general one-step or subspace lower-bound principle.

Our results. For every fixed $d > k \geq 3$, we compute the Fisher information $\mathcal{I}_k$ in one unit subspace-rotation direction and prove

$$\lim_{n\to\infty} nR_n^*(d,k) = \frac{d-k}{\mathcal{I}_k},$$

where n is the number of independent pairs and $R_n^*(d,k)$ is the minimax normalized projector risk. We construct an explicit regularized estimator attaining this constant uniformly over subspace orientations. We also show that a simpler dense moment estimator has risk coefficient $(d-k)/(2\lambda^2)$, where λ is its population eigengap. For $k \geq 4$, the ratio of these coefficients is at most $k/(k-3)$. These are fixed-dimensional asymptotic statements.

We then consider the actual overlapping statistic used in a temporal-memory assay. Conditional on the teacher schedule, its signal depends on lag one, whereas its cross-subspace noise also depends on lag two. Finally, we report a separately frozen transfer experiment with an independent future target and substantially lower typical historical-task resemblance than a rank-four independent-target study. The theorem concerns independent, clock-grouped pairs and subspace loss. The experiment concerns long histories and a particular classification head. We keep these two claims separate.

2 Model and preliminaries

Let $\mathcal{P}_{d,k}$ be the set of rank-k orthogonal projectors in $\mathbb{R}^d$. The unknown parameter is $P \in \mathcal{P}_{d,k}$, and $S = \mathrm{range}(P)$. For each $j = 1, \ldots, n$, independently draw a teacher u_j uniformly on the unit sphere of S, and draw $X_j, X_j' \sim N(0, I_d)$ independently. The observations are

$$(X_j, Y_j, X_j', Y_j'), \qquad Y_j = \mathrm{sgn}(u_j^\top X_j), \quad Y_j' = \mathrm{sgn}(u_j^\top X_j'). \tag{2.1}$$

The teachers are not observed. The rank and the grouping into pairs are supplied. Thus n pairs contain $2n$ labeled examples. Throughout the minimax analysis, d and k are fixed.

For $Q, P \in \mathcal{P}_{d,k}$, define

$$L(Q,P) = 1 - \frac{\mathrm{tr}(QP)}{k} = \frac{\|Q-P\|_{\mathrm{F}}^2}{2k}, \qquad R_n^*(d,k) = \inf_{\widehat{P}} \sup_{P\in\mathcal{P}_{d,k}} \mathbb{E}_P L(\widehat{P}, P). \tag{2.2}$$

The infimum includes all measurable randomized or deterministic rank-k projector estimates based on the full observations. It is not restricted to unbiased or regular estimators.

Proposition 2.1 (Likelihood and information-preserving normalization)**.** *Put* $K(r) = 2\arcsin(r)/\pi$ *and*

$$r_P(x,x') = \frac{x^\top P x'}{\sqrt{(x^\top P x)(x'^\top P x')}}.$$

Outside null sets, the conditional probability of the labels is

$$p_P(y,y' \mid x,x') = \tfrac{1}{4}\{1 + yy' K(r_P(x,x'))\}. \tag{2.3}$$

The signed unit vectors $z = YX/\|X\|$ *and* $z' = Y'X'/\|X'\|$ *are sufficient for* P*. Relative to product-uniform sphere measure, their density is* $g_P(z,z') = 1 + K(r_P(z,z'))$.

Proof sketch. A uniform random hyperplane separates two fixed directions with probability equal to their angle divided by π; equivalently, use the classical Gaussian sign identity (Cho and Saul, 2009). The two marginal labels are fair, giving (2.3). The transformation to signed directions removes the factor yy' from r_P. Conditional on these directions, the discarded lengths and residual sign bits have a law independent of P. Appendix A gives the change of variables. □

Set $q = \sqrt{d}\, z$, $q' = \sqrt{d}\, z'$ and

$$A = \mathrm{sym}(qq'^{\top}), \quad \mathrm{sym}(M) = \tfrac{1}{2}(M + M^{\top}), \quad b_d = \frac{\sqrt{d}\,\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d+1)/2)}, \quad \lambda = \frac{b_d^2}{k}. \tag{2.4}$$

Conditional on a teacher, $\mathbb{E}[q \mid u] = b_d u$ and $\mathbb{E}[qq^{\top} \mid u] = I_d$. Hence $\mathbb{E}_P A = \lambda P$, which identifies P. In contrast, with a fresh independent symmetric teacher for each single observation, Y is a fair sign conditional on X and the entire observed law is independent of P. This impossibility is specific to that teacher law.

3 Fisher information and minimax risk

We use canonical Grassmann coordinates: rotating one unit vector in S towards one in $S^{\perp}$ is a unit tangent direction. Its projector derivative has Frobenius norm $\sqrt{2}$. This convention fixes the factor of two in the risk.

Theorem 3.1 (Information and regularity)**.** *For fixed $d > k \geq 3$, the experiment* (2.1) *is differentiable in quadratic mean. Its Fisher information in an orthonormal tangent frame is $\mathcal{I}_k I_{k(d-k)}$, where*

$$\mathcal{I}_k = \frac{8F_k}{\pi^2 k(k-2)}, \qquad F_k = \mathbb{E}\left[\frac{1}{1 - \{2\arcsin(R)/\pi\}^2}\right], \qquad R^2 \sim \mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{k-1}{2}\right). \tag{3.1}$$

Here R is symmetric. The constant is finite and positive for $k \geq 3$ and does not depend on d.

Proof outline. At $P = \mathrm{diag}(I_k, 0)$, write the projected Gaussian inputs as as and bt, with s, t unit and $r = s^{\top} t$. For an inside coordinate i and outside coordinate α, differentiation gives

$$D_{\alpha i} r = X_\alpha (t_i - r s_i)/a + X'_\alpha (s_i - r t_i)/b.$$

Averaging over the four labels gives conditional score variance $(K'(r) D_{\alpha i} r)^2/(1 - K(r)^2)$. Radial independence and symmetry imply $\mathbb{E}[(D_{\alpha i} r)^2 \mid r] = 2(1 - r^2)/(k(k-2))$, yielding (3.1). To justify regularity, rather than just formal differentiation, Appendix A.3 proves an L^2 derivative of the square-root density along elementary rotations and then a full local chart. □

The rank restriction is essential. At $k = 2$ the formal angular information diverges. At $k = 1$, if two unoriented lines differ by angle ϑ, their squared Hellinger distance is $4p(1-p)$ with $p = \vartheta/\pi$; it is linear near zero. Neither case is covered by the regular theorem.

Theorem 3.2 (Exact fixed-dimensional asymptotic minimax risk)**.** *In* (2.1)*, for every fixed $d > k \geq 3$,*

$$\lim_{n \to \infty} n R_n^*(d, k) = \frac{d-k}{\mathcal{I}_k}. \tag{3.2}$$

The regularized cross-fitted estimator of Section 4 satisfies

$$\sup_{P \in \mathcal{P}_{d,k}} \left| n \mathbb{E}_P L(\widehat{P}_{\mathrm{eff}}, P) - \frac{d-k}{\mathcal{I}_k} \right| \longrightarrow 0. \tag{3.3}$$

In particular, its worst-case risk divided by $R_n^*(d, k)$ converges to one.

The lower bound. Use a local graph chart $P(T) = [I_k; T](I_k + T^{\top}T)^{-1}[I_k, T^{\top}]$ and the bounded target $\psi(T) = T(I_k + T^{\top}T)^{-1}$. Every projector estimate Q, including one outside the chart, obeys

$$L(Q, P(T)) \geq k^{-1} \left\| Q_{21} - \psi(T) \right\|_{\mathrm{F}}^2.$$

A compact product prior in a shrinking chart and a matrix van Trees inequality give the lower constant $(d-k)/\mathcal{I}_k$. No condition is imposed on estimator bias. Appendix B verifies the integration-by-parts hypotheses and also derives, for $0 < s < 1/\sqrt{3}$, the finite-sample bound

$$R_n^*(d,k) \geq \frac{(d-k)(1-3s^2)^2}{n\mathcal{I}_k + \pi^2 k(d-k)/s^2}. \tag{3.4}$$

The compact-prior method is standard (Gill and Levit, 1995; Gassiat and Stoltz, 2024); the information and projector-loss specialization are given here in full.

A simpler estimator. Let $\widehat{P}_{\mathrm{mom}} = \mathrm{Top}_k(n^{-1}\sum_j A_j)$, with any measurable top-k projector choice at a tie.

Proposition 3.3 (Moment risk)**.** *Uniformly over P at fixed $d > k \geq 3$,*

$$n\mathbb{E}_P L(\widehat{P}_{\mathrm{mom}}, P) \longrightarrow \frac{d-k}{2\lambda^2}. \tag{3.5}$$

For $k \geq 4$, the ratio of this coefficient to the minimax coefficient is between 1 and $k/(k-3)$.

The exact ratios are 1.3825 at $(d,k) = (32,8)$ and 1.1582 at $(64,16)$. Thus the simple moment estimator becomes asymptotically efficient in the ordered limit that first sends $n \to \infty$ at each fixed d, k, and then sends $k \to \infty$. This is not a simultaneous high-dimensional guarantee. Appendix D proves the coefficient and the ratio bound.

4 An asymptotically optimal estimator

A raw one-step likelihood correction cannot be justified here solely by finite on-model Fisher information. At rank three, a score evaluated at an arbitrarily nearby wrong pilot can have infinite second moment under the true law (Appendix A.5). We therefore regularize the correction explicitly. We make no efficiency claim for the unregularized construction.

For a projector Q, put $u = z^\top Q z$, $v = z'^\top Q z'$ and $r = r_Q(z,z')$. Define the tangent projection $\Pi_Q(H) = (I-Q)HQ + QH(I-Q)$. Where the expressions are nonsingular, the embedded score is

$$B_Q = \frac{zz'^\top + z'z^\top}{\sqrt{uv}} - r\left(\frac{zz^\top}{u} + \frac{z'z'^\top}{v}\right), \tag{4.1}$$
$$W_{Q,0} = \frac{K'(r)}{1+K(r)}\Pi_Q(B_Q).$$

For a tangent D, the log-likelihood derivative is $\mathrm{tr}(W_{Q,0}D)/2$.

Let χ be the nondecreasing C^3 cutoff equal to zero on $(-\infty,1]$, one on $[2,\infty)$, and to $35t^4 - 84t^5 + 70t^6 - 20t^7$ at $1+t$, $0 < t < 1$. Set

$$w_h = \chi(u/h)\chi(v/h)\chi((1-r^2)/h), \qquad W_{Q,h} = w_h W_{Q,0}. \tag{4.2}$$

The field is set to zero before evaluating a singular expression when a cutoff argument is at most one. Write $m = k(d-k)$ and define

$$j_h = \frac{\mathbb{E}_P[w_h \|W_{P,0}\|_{\mathrm{F}}^2]}{2m}, \qquad v_h = \frac{\mathbb{E}_P[w_h^2 \|W_{P,0}\|_{\mathrm{F}}^2]}{2m}. \tag{4.3}$$

These scalars depend on d, k, h, not on P. The correction uses the sensitivity j_h, not the variance v_h. Both converge to $\mathcal{I}_k$ as $h \downarrow 0$. Appendix C.1 gives one-dimensional integrals for their exact definition.

Write $Z_j = (z_j, z_j')$. For $n \geq 2$, split the pairs into sets A, B of sizes $a = \lfloor n/2 \rfloor$ and $b = n - a$. Compute moment pilots Q_A, Q_B on the respective halves. Use $h_n = [16 + \log(n+1)]^{-1}$ and form

$$D_A = \frac{1}{bj_{h_n}} \sum_{j \in B} W_{Q_A,h_n}(Z_j), \qquad C_A = \mathrm{Exp}_{Q_A}(D_A), \tag{4.4}$$
$$D_B = \frac{1}{aj_{h_n}} \sum_{j \in A} W_{Q_B,h_n}(Z_j), \qquad C_B = \mathrm{Exp}_{Q_B}(D_B),$$
$$\widehat{P}_{\mathrm{eff}} = \mathrm{Top}_k\big((b/n)C_A + (a/n)C_B\big).$$

Here $\mathrm{Exp}_Q(D) = e^{[D,Q]}Qe^{-[D,Q]}$. The weights are the *correction-half* sizes. They recover the full-data first-order sum even when n is odd. The two corrected projectors need not be independent.

Risk of the estimator. Appendix C proves the uniform expansion

$$\widehat{P}_{\text{eff}} - P = \frac{1}{nj_{h_n}} \sum_{j=1}^{n} W_{P,h_n}(Z_j) + R_n, \qquad \sup_P \mathbb{E}_P \|R_n\|_{\text{F}}^2 \le \frac{Ch_n^{-32}}{n^2}. \tag{4.5}$$

The proof uses global inequalities for the pilot, exponential map and final spectral projection, so no bad-pilot or tie event is discarded. Since h_n^{-32} is polylogarithmic, the squared L^2 remainder is $o(1/n)$. The leading term has squared Frobenius mean $2mv_{h_n}/(nj_{h_n}^2)$, which proves (3.3) directly in expected loss, rather than inferring it from a weak limit.

This construction establishes attainability. It is not a low-memory online learner and has not been shown superior at a practical feedback budget. The theorem uses exact sensitivity calibration. Vanishing deterministic calibration error suffices, but a fixed numerical tolerance alone is not an arbitrary-n certificate.

5 Subspace estimation from overlapping observations

The previous theorem is deliberately a minimal experiment. The transfer assay instead accumulates the overlapping statistic

$$C_N = \frac{1}{N-1} \sum_{t=1}^{N-1} \mathrm{sym}(q_t q_{t+1}^\top). \tag{5.1}$$

No historical episode boundaries are supplied to this update. Adjacent summands share an observation, and cannot be treated as independent pairs.

Let B be an orthonormal basis of the true k-dimensional span. Conditional on an exogenous teacher schedule, assume that q_t are independent, $\mathbb{E}q_t = B\mu_t$, $\mathbb{E}q_t q_t^\top = I_d$, and their norms are uniformly bounded. Define

$$H_N = \frac{1}{N-1} \sum_{t=1}^{N-1} \mathrm{sym}(\mu_t \mu_{t+1}^\top), \qquad J_N = \frac{1}{N-2} \sum_{t=1}^{N-2} \mathrm{sym}(\mu_t \mu_{t+2}^\top).$$

Proposition 5.1 (Overlapping-lag risk)**.** *If $H_N \to H \succ 0$ and $J_N \to J$ at fixed d, k, then the leading-k projector of* (5.1) *satisfies*

$$\lim_{N\to\infty} N\mathbb{E}\left[L(\mathrm{Top}_k(C_N), BB^\top) \mid \textit{schedule}\right] = \frac{d-k}{2k} \operatorname{tr}\left[H^{-1}(I_k + J)H^{-1}\right]. \tag{5.2}$$

The exact finite-N covariance in Appendix D.3 gives the formula. Lag one determines the signal; lag two supplies an additional covariance term. This conditional asymptotic law describes subspace estimation, not the finite-budget target-head accuracy below. The smooth likelihood correction in (4.4) is not used in the long-history experiment.

6 Experiments

We test whether the overlapping signed statistic can improve early target learning. This assay does not run the efficient estimator in (4.4), does not test a minimax coefficient, and does not claim to replace gradient-based representation learning.

Experimental setup. The primary setting has $d = 64$, supplied latent rank $k = 16$, and 64 independent historical teachers. Each teacher labels a block of independent Gaussian inputs. Block lengths are independently sampled from $\{2047, 4096, 8191\}$. The memory learner receives the resulting uninterrupted input–label stream, without historical task identifiers or boundary markers. The target is then drawn independently and uniformly within the same latent subspace. There is no target-dependent rejection. A common dense orthogonal rotation is applied to inputs and task

Table 1: Related spectral history minus each prespecified causal control, in 64 fresh confirmation contexts. "pp" denotes percentage points. The query, not just the history, is shared within each comparison.

Comparator	Early difference (pp)	Nominal 99% interval (pp)
Erased history	$+2.844$	$[+2.211, +3.487]$
Unrelated history	$+4.967$	$[+4.099, +5.818]$
Shuffled labels	$+4.099$	$[+3.169, +4.998]$
Input-only history	$+3.833$	$[+2.919, +4.747]$
Scrambled directions	$+3.823$	$[+3.079, +4.595]$

directions; this is the primary coordinate system. The identity rotation is a paired diagnostic, not another independent replicate.

The memory produces a flattened leading spectral filter $\widehat{\Pi}_k$. Its target metric is $F = I_d/2 + d\widehat{\Pi}_k/(2k)$, with an unchanged bias coordinate. The head is reset at the explicitly marked target phase and receives 32 labeled support examples. All arms use the same retention $\rho = 0.5$, normalized metric-preconditioned tanh-residual update, supports, and 2048 target queries. Appendix E gives the full update, tie rule and controls.

One setting was frozen before a 24-context development study. Only after its declared joint gate passed were 64 seed-disjoint confirmation contexts activated. The primary endpoint averages query accuracy at shots 1, 2 and 4 within each context. All contrasts use paired contexts. Nominal two-sided 99% percentile intervals use 20,000 whole-context bootstrap resamples. They are individual intervals, not a simultaneous-coverage guarantee. The setting was chosen after earlier research; seed-disjoint confirmation does not make the entire project a nonadaptive search.

Results. Table 1 gives the primary comparisons. Absolute early accuracy is 57.02% with related spectral history, 54.18% with erased history, and 52.06% with unrelated history. All five lower interval endpoints are positive. Mean span overlap is 85.98% and mean target-direction coverage is 86.14%, but mean worst-direction coverage is only 24.54%. Mean recovery therefore does not certify recovery of every direction. The true-subspace oracle exceeds the related learner by 1.30 early percentage points. Inherited diagonal and matrix-softmax comparators are worse by 2.82 and 2.87 points, respectively; they were not retuned for this setting.

Comparison with retrieval. For H independent uniform teachers and an independent target in k dimensions, the maximum absolute teacher–target cosine M satisfies

$$\mathbb{P}(M \le t) = I_{t^2}\big(1/2, (k-1)/2\big)^H,$$

where I_x is the regularized incomplete beta function. The present geometry predicts $\mathbb{E}M = 0.60483$; confirmation gives 0.60808. In 60 of 64 contexts, $M \le 0.75$; the other four are retained. This is lower typical resemblance, not a deterministic no-near-copy guarantee.

A prespecified secondary diagnostic receives the true historical teacher dictionary and selects one signed teacher using target-support agreement, then signed margin. It cannot adapt or combine teachers. The spectral learner exceeds this restricted, privileged diagnostic by 1.48 early points (nominal 99% interval $[0.36, 2.57]$), but the ordering reverses later (Figure 1). At shot 32, the accuracies are 59.86% and 65.93%, respectively. The population ceiling obtained by choosing the best signed teacher with knowledge of the true target is 70.92%. We do not exclude stronger retrieval, mixtures, or adaptive selection.

Computational cost and additional experiments. Confirmation used 19,379,539 historical examples, a mean of 302,805 per context. The history state is dense $O(d^2)$; the bulk evaluation additionally materializes histories. This is not a sample-efficiency or measured runtime improvement. A previous independent-target study at $(d, k, H) = (64, 4, 128)$ yielded a 10.74-point early advantage over erased history, but mean nearest-task cosine was 0.96675. A separate target-first study enforced a cosine cap by rejecting historical teachers. These establish different properties and must not be pooled or represented as matched-budget replications (Appendix E.6).

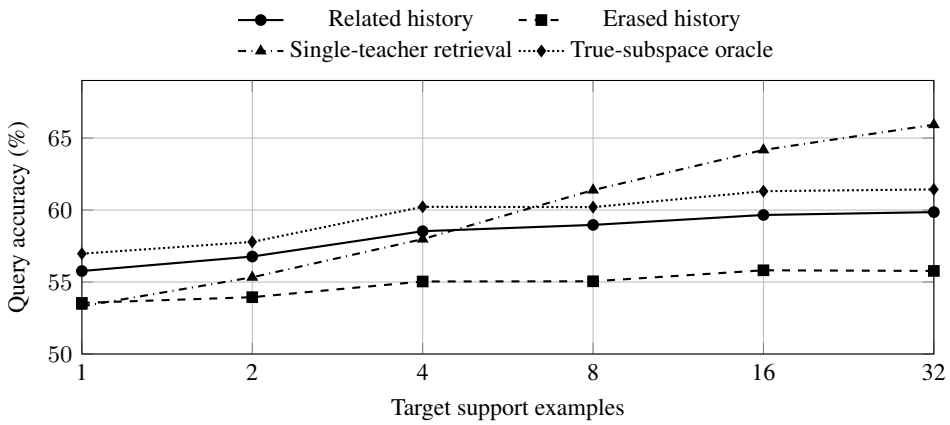

Figure 1: Confirmation mean learning curves. Retrieval is worse at the declared early endpoint but better later. Curves are descriptive means; no pointwise uncertainty band or shot-32 significance claim is inferred from the early-endpoint intervals.

7 Concluding remarks

The minimax result fixes known $d, k \geq 3$, Gaussian inputs, noiseless homogeneous labels, uniform teachers, independent clock-grouped pairs, and projector loss. Its attainable constant is an asymptotic statement with dimension-dependent remainder constants. The efficient construction uses dense pilots and exact mathematical sensitivity calibration. We have not established a useful finite-budget gain for it, the nonregular rates at ranks one and two, or efficiency when dimensions grow with the sample size.

The confirmed transfer study uses a different temporal law and a fixed head. Historical boundaries are hidden, but target onset is marked and the head is reset. Rank is supplied; histories are long; the inputs are synthetic. The controls do not include a newly tuned modern meta-learning system. Four prospective low-memory streaming variants failed their declared rank-eight overlap gate. They remain negative results, not successful implementations of the theorem (Appendix E.7). We claim neither whole-life dominance nor superiority to backpropagation.

We have determined the information created by two uses of a hidden symmetric binary rule, the exact fixed-dimensional asymptotic minimax projector risk, and an estimator attaining that risk. An overlapping-history calculation and a separate confirmed transfer assay show how related temporal information can be used outside the pair model. The unresolved question is whether that statistical information can be used with substantially smaller live state and feedback budgets, without supplying the latent rank.

AI USE STATEMENT

ChatGPT was used extensively throughout this work, including substantive assistance with the theoretical model, mathematical claims and proofs, identification of proof gaps, hypotheses, experimental design, synthetic-data generation and processing, implementation, deterministic checks, analysis and interpretation, literature searches, and manuscript drafting and editing. The assistance was not limited to language polishing. No translation or analysis of human-participant data was involved.

Verification included executable algebra and numerical checks, source and seed checksums, AI-assisted proof challenges, and comparisons with saved experimental analyses. The supplementary material distinguishes preserved source files from reconstructed reproduction scripts and regenerated experimental records. These checks do not constitute formal theorem verification or independent human peer review. The earlier unregularized score argument is not used to claim attainability; the regularized construction and its stated assumptions are the basis of the theorem. Responsibility for the final text, mathematical claims, experimental results, references, and supplementary artifacts rests with the human authors.

REPRODUCIBILITY STATEMENT

Appendices A–D give the model assumptions, loss normalization, likelihood, information calculation, estimator, and proofs. Appendix E specifies the experimental input law, update, controls, and statistical unit. The accompanying anonymous code supplement contains the frozen low-redundancy generator and learner, checksum-verified seed ledgers, the saved confirmation analysis, numerical proof checks, and scripts for reproducing the reported statistics. Its provenance record distinguishes original preserved files, reconstructed wrappers, and regenerated per-context outputs; the full original raw-data archive is not included. The theory checks are computational checks of selected identities, not formal proof certificates.

ETHICS STATEMENT

The experiments use synthetic inputs and do not involve human participants or personal data. Claims are restricted to the stated statistical and synthetic experimental settings. The main research-integrity issues are correct attribution, disclosure of AI assistance, separation of development from confirmation, and accurate statements of the theorem's scope.

REFERENCES

Maryam Aliakbarpour, Konstantina Bairaktari, Gavin Brown, Adam Smith, Nathan Srebro, and Jonathan Ullman. Metalearning with very few samples per task. *Proceedings of the 37th Conference on Learning Theory, PMLR 247:46–93*, 2024. https://proceedings.mlr.press/v247/aliakbarpour24a.html.

Thomas Bendokat, Ralf Zimmermann, and P.-A. Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *arXiv:2011.13699, version 3*, 2023. https://arxiv.org/abs/2011.13699v3.

Yuejie Chi and Haoyu Fu. Subspace learning from bits. *arXiv:1407.6288, version 3*, 2017. https://arxiv.org/abs/1407.6288v3.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. *Advances in Neural Information Processing Systems 22*, 2009. Proceedings version.

Saptati Datta, Nicolas W. Hengartner, Yulia Pimonova, Natalie E. Klein, and Nicholas Lubbers. Effects of structural allocation of geometric task diversity in linear meta-learning models. *arXiv:2509.18349, version 3*, 2026. https://arxiv.org/abs/2509.18349v3.

Sjoerd Dirksen, Johannes Maly, and Holger Rauhut. Covariance estimation under one-bit quantization. *The Annals of Statistics, 50(6):3538–3562*, 2022. https://doi.org/10.1214/22-AOS2239.

Sjoerd Dirksen, Weilin Li, and Johannes Maly. Subspace and DOA estimation under coarse quantization. *IEEE Transactions on Information Theory, 71(10):8149–8168*, 2025. https://doi.org/10.1109/TIT.2025.3598702.

Elisabeth Gassiat and Gilles Stoltz. The van Trees inequality in the spirit of Hájek and Le Cam. *Statistical Science, 39(4):644–653*, 2024. https://arxiv.org/abs/2402.06431.

Eric Gautier and Yuichi Kitamura. Nonparametric estimation in random coefficients binary choice models. *Econometrica, 81(2):581–607*, 2013. https://doi.org/10.3982/ECTA8675.

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

Halil Ibrahim Gulluk, Yue Sun, Samet Oymak, and Maryam Fazel. Sample efficient subspace-based representations for nonlinear meta-learning. *arXiv:2102.07206, version 2*, 2021. https://arxiv.org/abs/2102.07206v2.

James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. *Advances in Neural Information Processing Systems 33*, 2020. https://arxiv.org/abs/1912.08866v2.

De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-PCA: Efficient guarantees for Oja's algorithm, beyond rank-one updates. *Proceedings of the 34th Conference on Learning Theory, PMLR 134:2463–2498*, 2021. https://proceedings.mlr.press/v134/huang21a.html.

Vladimir Koltchinskii and Karim Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *arXiv:1504.07333*, 2015. https://arxiv.org/abs/1504.07333.

Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics, 40(2):694–726*, 2012. https://doi.org/10.1214/12-AOS970.

Alex Luedtke and Incheoul Chung. One-step estimation of differentiable Hilbert-valued parameters. *arXiv:2303.16711, version 3*, 2023. https://arxiv.org/abs/2303.16711v3.

Peter Ruckdeschel. Fisher information in group-type models. *arXiv:1005.1027*, 2010. https://arxiv.org/abs/1005.1027.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. *Proceedings of the 38th International Conference on Machine Learning, PMLR 139:10434–10443*, 2021. https://proceedings.mlr.press/v139/tripuraneni21a.html.

Martin Wahl. Van Trees inequality, group equivariance, and estimation of principal subspaces. In *Foundations of Modern Statistics*, Springer Proceedings in Mathematics & Statistics 425, pages 301–321. Springer, Cham, 2023. https://doi.org/10.1007/978-3-031-30114-8_8.

A LIKELIHOOD AND REGULARITY

This appendix supplies the details of Proposition 2.1 and Theorem 3.1. All observations here are independent pairs as in (2.1); the overlapping-history experiment is not used in these proofs.

A.1 LIKELIHOOD OF A PAIR

Fix P and inputs x, x' whose projections are nonzero. Their directions in S make angle $\theta = \arccos r_P(x, x')$. A uniform teacher direction has the same signs against the two projected inputs with probability $1-\theta/\pi$. This follows by restricting to their two-dimensional span: the angular wedge on which the signs differ has total measure 2θ in a circle of measure 2π. By antipodal symmetry both labels are fair. Thus

$$\mathbb{P}(YY' = 1 \mid x, x') = 1 - \theta/\pi, \qquad \mathbb{E}[YY' \mid x, x'] = 1 - 2\theta/\pi = K(r_P),$$

which gives (2.3). Projected zero vectors or collinear directions form null sets for the regular-rank model. Values there may be assigned measurably without affecting the experiment.

Write $x = ra$, $x' = r'a'$ with a, a' uniform on the sphere. Gaussian radii r, r' are independent of these directions. Change variables to $z = ya$, $z' = y'a'$. Surface measure is unchanged and $r_P(a, a') = yy' r_P(z, z')$, so the joint density of (z, z', y, y') is

$$\tfrac{1}{4}\{1 + K(r_P(z, z'))\}.$$

Summing over y, y' gives $g_P = 1 + K(r_P)$; the two remaining bits are conditionally uniform. The radii retain their independent chi distributions. Their joint conditional law given (z, z') is independent of P, proving sufficiency. This does not say that the subsequent empirical mean of A is sufficient.

For a unit teacher u, rotational symmetry gives $\mathbb{E}[q \mid u] = b_d u$. Its scalar coefficient is $\sqrt{d}\,\mathbb{E}|a_1| = b_d$. Since a sign disappears from a second outer product, $\mathbb{E}[qq^\top \mid u] = I_d$. Conditional independence within a pair yields

$$\mathbb{E}[A \mid u] = b_d^2 uu^\top, \qquad \mathbb{E}_P A = (b_d^2/k)P.$$

This moment identifies P. With one freshly drawn symmetric teacher per observation, however, $\mathbb{P}(Y = 1 \mid X = x) = 1/2$ for almost every x. Independence of the draws makes every finite sequence's law independent of P.

A.2 FISHER INFORMATION

Take $P_0 = \operatorname{diag}(I_k, 0)$, and write $P_0 x = as$, $P_0 x' = bt$ in inside coordinates, with $a, b > 0$ and s, t unit. Let $r = s^\top t$. Rotating the inside basis vector i towards the outside basis vector α gives

$$D_{\alpha i} r = X_\alpha (t_i - r s_i)/a + X'_\alpha (s_i - r t_i)/b. \tag{A.1}$$

The label score in this direction is

$$S_{\alpha i} = \frac{yy' K'(r) D_{\alpha i} r}{1 + yy' K(r)}.$$

Conditioning on the inputs and summing over the four label probabilities gives

$$\mathbb{E}[S_{\alpha i}^2 \mid x, x'] = \frac{(K'(r) D_{\alpha i} r)^2}{1 - K(r)^2}. \tag{A.2}$$

The outside coordinates are independent centered standard Gaussians. Conditional on r, rotational symmetry gives

$$\mathbb{E}[(t_i - r s_i)^2 \mid r] = \mathbb{E}[(s_i - r t_i)^2 \mid r] = \frac{1 - r^2}{k}.$$

The lengths are independent chi-k variables, with $\mathbb{E}a^{-2} = \mathbb{E}b^{-2} = 1/(k-2)$ for $k > 2$. Combining with (A.1) proves

$$\mathbb{E}[(D_{\alpha i} r)^2 \mid r] = \frac{2(1 - r^2)}{k(k-2)}.$$

Insert $K'(r) = 2/(\pi\sqrt{1-r^2})$ in (A.2) to obtain (3.1). Independent inside and outside reflections annihilate mixed information entries; permutations equate the diagonal entries. Equivalently, the information tensor is invariant under $O(k) \times O(d-k)$ and equals $\mathcal{I}_k$ times the canonical tangent metric.

The angle $\Theta = \arccos R$ has density proportional to $\sin^{k-2}\theta$. An equivalent integral is

$$F_k = \frac{2}{\mathrm{B}(1/2, (k-1)/2)} \int_0^{\pi/2} \frac{\pi^2 \sin^{k-2}\theta}{4\theta(\pi-\theta)}\, d\theta. \tag{A.3}$$

The integrand is $O(\theta^{k-3})$ near zero, so it is integrable for $k > 2$. The absence of d from $\mathcal{I}_k$ reflects the fact that one chosen outside coordinate has unit Gaussian variance; the number $k(d-k)$ of tangent coordinates still depends on d.

A.3 Quadratic-mean differentiability

Finite formal Fisher information alone does not imply differentiability in quadratic mean. We verify that property directly. Let μ be product Gaussian input measure times counting measure on the four labels, and let $f_P = \sqrt{p_P(y, y' \mid x, x')} \in L^2(\mu)$. It has squared norm one. Orthogonal input transformations act by

$$(\mathcal{U}_R f)(x, x', y, y') = f(R^\top x, R^\top x', y, y').$$

This is a unitary representation because the Gaussian measure is invariant. It is strongly continuous: first prove continuity for continuous compactly supported functions, and then use their density in L^2 and the isometry. Moreover, $f_{RP_0R^\top} = \mathcal{U}_R f_{P_0}$.

Consider one elementary rotation $R(t)$ mixing an inside coordinate i with an outside coordinate α. For almost every input pair, the remaining $k-1$ inside coordinate vectors are noncollinear. This uses $k \geq 3$. They remain fixed along the entire rotation. Neither full projected input can therefore vanish, and the full projected vectors cannot become collinear, anywhere on this path. Consequently $t \mapsto \mathcal{U}_{R(t)} f_{P_0}$ is pointwise continuously differentiable almost everywhere. At zero its derivative is

$$D_{\alpha i} f_{P_0} = \tfrac{1}{2} S_{\alpha i} f_{P_0}, \qquad \|D_{\alpha i} f_{P_0}\|^2_{L^2(\mu)} = \mathcal{I}_k/4.$$

The pointwise fundamental theorem of calculus gives

$$\mathcal{U}_{R(t)} f_{P_0} - f_{P_0} = \int_0^t \mathcal{U}_{R(s)} D_{\alpha i} f_{P_0}\, ds. \tag{A.4}$$

The right side is also a Bochner integral in L^2: its integrand is strongly continuous and has constant finite norm. Fubini and Cauchy–Schwarz identify the pointwise and L^2 integrals. Dividing by t in (A.4) and using strong continuity proves the L^2 derivative. No singular log-likelihood Hessian is interchanged with an integral.

List all $m = k(d-k)$ elementary inside–outside rotations and put $R(\theta) = R_1(\theta_1) \cdots R_m(\theta_m)$. The map $\theta \mapsto R(\theta) P_0 R(\theta)^\top$ is a smooth local chart, since its derivative at zero spans the tangent space independently. For $U_j = \mathcal{U}_{R_j(\theta_j)}$,

$$U_1 \cdots U_m f - f = \sum_{j=1}^m U_1 \cdots U_{j-1}(U_j - I) f.$$

Apply the elementary L^2 derivative and strong continuity to each of the finitely many summands. This yields

$$\left\| f_{P(\theta)} - f_{P_0} - \frac{1}{2} \sum_{j=1}^m \theta_j S_j f_{P_0} \right\|_{L^2(\mu)} = o(\|\theta\|).$$

This is quadratic-mean differentiability. Rotations transport it to all P. In smooth coordinates the information matrix is continuous; Appendix B gives its exact graph-chart metric. The corresponding iid local asymptotic normality statement follows from the standard DQM theorem. It is not needed to bypass the direct lower-bound argument below.

For completeness, DQM also gives the required L^1 derivative of the density: expand $(f+\Delta f)^2 - f^2$ and use Cauchy–Schwarz on the quadratic remainder. Continuity of the density derivative follows from the rotation representation and smooth changes of coordinates. Expectations of bounded statistics are therefore continuously differentiable locally. This is the regularity used in the van Trees integration by parts (Gassiat and Stoltz, 2024).

A.4 Ranks one and two

At $k = 2$, even after restricting both projected input lengths to a compact interval away from zero, the angular term in (A.2) has a nonintegrable logarithmic singularity. A finite-information quadratic-mean expansion cannot hold. We do not determine the rank-two estimation rate.

At $k = 1$, let the two lines make acute angle ϑ, and put $p = \vartheta/\pi$. For one Gaussian input, the oriented line predictions differ with probability p. The two conditional pair-label laws have disjoint supports precisely when exactly one of the two predictions changes. That event has probability $2p(1-p)$. Squared Hellinger distance between disjoint probability distributions is two under the convention $\int(\sqrt{f} - \sqrt{g})^2$. Thus the unconditional squared Hellinger distance is $4p(1-p)$, linear rather than quadratic near zero.

A.5 Variance of the unregularized score

The following counterexample concerns the score averaged under a fixed true distribution, not its own-model Fisher information. Take $d = 4, k = 3$ and

$$Q = \operatorname{span}(e_1, e_2, e_3), \qquad P_\beta = \operatorname{span}(e_1, e_2, \cos\beta\, e_3 + \sin\beta\, e_4),$$

where $\beta > 0$ is arbitrarily small. For signed unit observations

$$z = (e_1 + e_4)/\sqrt{2}, \qquad z' = (-\cos\varepsilon\, e_1 + \sin\varepsilon\, e_2 + e_4)/\sqrt{2},$$

the canonical score at Q has its nonzero row equal to

$$\left(\frac{\sin\varepsilon}{\varepsilon}, \frac{1+\cos\varepsilon}{\varepsilon}, 0\right), \qquad \|S_Q\|^2 = \frac{4\cos^2(\varepsilon/2)}{\varepsilon^2}.$$

The pilot density is $g_Q = 2\varepsilon/\pi$, whereas

$$g_{P_\beta}(z, z') \longrightarrow \frac{4}{\pi}\arctan(\sin\beta) > 0.$$

This is not only a calculation on a null curve. Restrict the projected radii to small intervals around $1/\sqrt{2}$, the outside signs to be positive, and the first projected direction to a small neighborhood of e_1. On a sufficiently small fixed neighborhood, the true density remains bounded below. Parameterize the second projected direction by angle $\pi - \varepsilon$ from the first. The leading score norm remains bounded below by a positive multiple of $1/\varepsilon$. In a three-dimensional inside space the angular measure is a positive multiple of $\sin\varepsilon\, d\varepsilon$, hence of $\varepsilon\, d\varepsilon$ near zero. The second-moment integral is bounded below by a positive multiple of $\int_0^{\varepsilon_0} d\varepsilon/\varepsilon = \infty$.

Therefore $\mathbb{E}_{P_\beta}\|S_Q\|^2 = \infty$ for arbitrarily nearby wrong pilots. This disproves a raw-score fixed-true-law L^2 continuity argument at rank three. It does not disprove DQM, finite on-model information, or eventual efficiency of the raw final projector by some other proof. We make no such raw-estimator claim.

B Proof of the lower bounds

We give a direct matrix van Trees proof, including an explicit coordinate metric. This supplies the lower bound in Theorem 3.2 without assuming unbiasedness or regularity of the estimator. The statistical inequality is standard (Gill and Levit, 1995; Gassiat and Stoltz, 2024); we verify its hypotheses and constants for this experiment.

B.1 A LOCAL PARAMETRIZATION

At $P_0 = \mathrm{diag}(I_k, 0)$, write $p = d - k$, $m = pk$, and for $T \in \mathbb{R}^{p\times k}$ put

$$J(T) = \begin{bmatrix} I_k \\ T \end{bmatrix}, \quad G(T) = (I_k + T^\top T)^{-1}, \quad P(T) = J(T)G(T)J(T)^\top. \tag{B.1}$$

Vectorization uses column-major order. For a displacement V, differentiating gives

$$DP(T)[V] = \begin{bmatrix} 0 \\ V \end{bmatrix} GJ^\top + JG \begin{bmatrix} 0 & V^\top \end{bmatrix} - JG(V^\top T + T^\top V)GJ^\top. \tag{B.2}$$

Let $C(T) = (I_p + TT^\top)^{-1}$. The canonical squared tangent norm is

$$\tfrac{1}{2}\,\|DP(T)[V]\|_{\mathrm{F}}^2 = \mathrm{tr}\{G(T)V^\top C(T)V\}. \tag{B.3}$$

For example, this follows by projecting the derivative of the orthonormal frame $J(T)G(T)^{1/2}$ onto its orthogonal complement; the within-frame derivative does not change the projector. The identity also follows directly from (B.2). The coordinate Fisher matrix is therefore

$$\mathcal{I}(T) = \mathcal{I}_k\{G(T) \otimes C(T)\}. \tag{B.4}$$

It is continuous, equals $\mathcal{I}_k I_m$ at zero, and satisfies

$$(1 + \|T\|_{\mathrm{op}}^2)^{-2}\mathcal{I}_k I_m \preceq \mathcal{I}(T) \preceq \mathcal{I}_k I_m.$$

Quadratic-mean differentiability in this chart follows from Appendix A.3 by a smooth change of local coordinates. Thus bounded-statistic expectations have continuous coordinate derivatives on compact chart subsets.

B.2 REDUCTION TO A VECTOR ESTIMATION PROBLEM

Define

$$\psi(T) = \mathrm{vec}\{T(I_k + T^\top T)^{-1}\}.$$

It is the lower-left block of $P(T)$ in vectorized form. Its derivative at zero is I_m. For an arbitrary rank-k projector Q, let $Z_Q = \mathrm{vec}(Q_{21})$. The two off-diagonal blocks of the symmetric difference $Q - P(T)$ give

$$L(Q, P(T)) \geq k^{-1}\,\|Z_Q - \psi(T)\|^2\,. \tag{B.5}$$

This holds even if Q lies far outside the chart; its cross-block is bounded. Randomized estimates may be treated by adjoining an independent randomization variable to the data, which does not change the likelihood information.

For later use, differentiation before vectorization yields

$$D(TG)[V] = VG - TG(V^\top T + T^\top V)G.$$

Since $\|G\|_{\mathrm{op}} \leq 1$ and $\|G - I_k\|_{\mathrm{op}} \leq \|T\|_{\mathrm{op}}^2$, it follows that

$$\|D\psi(T) - I_m\|_{\mathrm{op}} \leq 3\,\|T\|_{\mathrm{op}}^2\,. \tag{B.6}$$

B.3 THE ASYMPTOTIC LOWER BOUND

The one-dimensional density $q_0(t) = \cos^2(\pi t/2)$ on $[-1, 1]$, zero elsewhere, integrates to one, vanishes at its endpoints, and has information $\int (q_0')^2/q_0 = \pi^2$. Take its m-fold product and scale it to a small chart cube. Let $\theta = \mathrm{vec}\,T$, let $q(\theta)$ be this prior, and let V now denote the joint score of the n observations and the prior. Coordinatewise integration by parts gives

$$\begin{aligned} \mathbb{E}[(Z_Q - \psi(\theta))V^\top] &= A_q := \mathbb{E}_q D\psi(\theta), \\ \mathbb{E}[VV^\top] &= n\mathbb{E}_q\mathcal{I}(\theta) + \mathcal{I}_q. \end{aligned} \tag{B.7}$$

The boundedness of Z_Q, L^1 differentiability of the density, continuity of the information on a compact cube, and the vanishing prior boundary terms justify these identities. One may first integrate

expectations of bounded statistics over the data, then integrate over the parameter, as in Gassiat and Stoltz (2024). Every product is integrable by Cauchy–Schwarz. The likelihood score has mean zero, so the cross term with the prior score vanishes.

For any random vectors U, V with finite second moments and positive definite $\mathbb{E}VV^\top$, positivity of the residual covariance gives

$$\mathbb{E}UU^\top \succeq \mathbb{E}UV^\top(\mathbb{E}VV^\top)^{-1}\mathbb{E}VU^\top.$$

Apply this to $U = Z_Q - \psi(\theta)$ and combine with (B.5):

$$\mathbb{E}_q\mathbb{E}_{P(\theta)}L(Q, P(\theta)) \geq \frac{1}{k}\operatorname{tr}\left[A_q\{n\mathbb{E}_q\mathcal{I}(\theta) + \mathcal{I}_q\}^{-1}A_q^\top\right]. \tag{B.8}$$

There is no estimator-dependent term on the right.

For a fixed $a > 0$, take $\theta = (a/\sqrt{n})t$ with independent coordinates of density q_0. Then $\mathcal{I}_q = n\pi^2a^{-2}I_m$, $A_q \to I_m$, and $\mathbb{E}_q\mathcal{I}(\theta) \to \mathcal{I}_kI_m$. Worst-case risk is at least this prior average. Taking the infimum over all estimates in (B.8), then $n \to \infty$, gives

$$\liminf_{n\to\infty} nR_n^*(d, k) \geq \frac{d-k}{\mathcal{I}_k + \pi^2/a^2}.$$

Finally let $a \to \infty$. This yields

$$\liminf_{n\to\infty} nR_n^*(d, k) \geq (d-k)/\mathcal{I}_k. \tag{B.9}$$

The limits have not been interchanged: the local scale a is fixed while n grows. The prior remains inside a valid local chart for all sufficiently large n at each fixed a.

B.4 A FINITE-SAMPLE LOWER BOUND

Instead scale the cube by $s/\sqrt{m}$, where $0 < s < 1/\sqrt{3}$. On its support, $\|T\|_{\mathrm{F}} \leq s$. The prior information is $\pi^2ms^{-2}I_m$. By (B.4), the joint information is bounded above by $(n\mathcal{I}_k + \pi^2m/s^2)I_m$. By (B.6), every singular value of A_q is at least $1 - 3s^2$. Substitution into (B.8) proves (3.4) for every n.

The bound may be conservative. Letting $s = s_n \downarrow 0$ with $ns_n^2 \to \infty$ recovers (B.9). This provides a second route to the asymptotic lower constant without invoking efficiency of any particular estimate.

C PROOF OF THE UPPER BOUND

We prove (3.3) for the *regularized* estimator (4.4). No claim for a raw or fixed-correlation-clipped score follows. Constants in this appendix may depend on fixed d, k, the cutoff function, λ, $\mathcal{I}_k$ and a positive calibration lower bound, but not on the unknown orientation, n, or the choice within a tied eigenspace.

C.1 SCORE REGULARIZATION AND CALIBRATION

For an orthonormal frame U of P, the canonical horizontal score is a $d \times k$ matrix Ψ with $U^\top\Psi = 0$. Its symmetric embedding is $W = U\Psi^\top + \Psi U^\top$, so

$$\|W\|_{\mathrm{F}}^2 = 2\|\Psi\|_{\mathrm{F}}^2.$$

The likelihood derivative in direction $D = UH^\top + HU^\top$ is $\operatorname{tr}(WD)/2 = \operatorname{tr}(\Psi^\top H)$. Differentiation of the projected correlation gives precisely (4.1). The factor of two is therefore fixed by the canonical tangent normalization, not chosen in the risk calculation.

Let $W_{P,h} = w_hW_{P,0}$ as in (4.2). Reflecting an outside coordinate reverses its corresponding score row while leaving the density and cutoff unchanged. Inside/outside sign and permutation symmetries imply, in canonical coordinates,

$$\mathbb{E}_P(w_h\Psi_P) = 0, \qquad \mathbb{E}_P\{\operatorname{vec}(w_h\Psi_P)\operatorname{vec}(\Psi_P)^\top\} = j_hI_m, \qquad \mathbb{E}_P\{\operatorname{vec}(w_h\Psi_P)\operatorname{vec}(w_h\Psi_P)^\top\} = v_hI_m. \tag{C.1}$$

Consequently $\mathbb{E}_P W_{P,h} = 0$ and $\mathbb{E}_P \|W_{P,h}\|_{\mathrm{F}}^2 = 2mv_h$. Since $0 \le w_h \le 1$ and $w_h \to 1$ almost surely as $h \downarrow 0$, the true-score second moment and dominated convergence give

$$j_h \to \mathcal{I}_k, \quad v_h \to \mathcal{I}_k, \quad \|w_h \Psi_P - \Psi_P\|_{L^2(P)} \to 0. \tag{C.2}$$

These are on-model statements, not continuity at wrong pilots. Also $j_h^2 \le \mathcal{I}_k v_h$ by Cauchy–Schwarz.

The following deterministic integrals specify the calibration completely. Let $T \sim \mathrm{Beta}(k/2, (d-k)/2)$ and let Θ have normalized density proportional to $\sin^{k-2}\theta$ on $(0,\pi)$. For $\ell = 1, 2$, put

$$a_\ell(h) = \mathbb{E}\left[\chi(T/h)^\ell \frac{1-T}{T}\right], \qquad b_\ell(h) = \mathbb{E}[\chi(T/h)^\ell],$$
$$c_\ell(h) = \mathbb{E}\left[\frac{\chi(\sin^2\Theta/h)^\ell}{1-(1-2\Theta/\pi)^2}\right].$$

Then

$$j_h = \frac{8a_1(h)b_1(h)c_1(h)}{\pi^2 k(d-k)}, \qquad v_h = \frac{8a_2(h)b_2(h)c_2(h)}{\pi^2 k(d-k)}. \tag{C.3}$$

To verify this, condition on the two projected squared radii u, v and their angle under product-uniform reference measure. Complementary directions are independent and symmetric, so the cross term in the squared horizontal score vanishes. Multiplication by the signed-pair density leaves

$$\frac{4}{\pi^2}\left(\frac{1-u}{u} + \frac{1-v}{v}\right)\frac{1}{1+K(r)}.$$

The weight is a product of radial factors and an even angular factor. Averaging the last expression over r and $-r$ replaces $(1+K(r))^{-1}$ by $(1-K(r)^2)^{-1}$. This proves (C.3) after division by m.

For $0 < h \le 1/16$, monotonicity of χ implies $j_h \ge j_{1/16} > 0$. Positivity follows from a positive-measure set where both projected squared radii exceed $1/8$, the angle stays away from zero and π, and the score is nonzero. This is a fixed-dimensional positivity assertion; it is not a useful lower numerical bound uniformly over d, k.

C.2 Bounds on the score field

The matrix formula (4.1) is basis-free. Along a geodesic $Q(t) = e^{tK_0}Q(0)e^{-tK_0}$, with K_0 skew-symmetric and $\kappa = \|K_0\|_{\mathrm{F}}$, all nonzero cutoff terms and their first two derivatives have $u, v, 1-r^2 \ge h$. Singular regions are masked before evaluation. At transition boundaries the first three derivatives of χ match zero, so the bounds extend by continuity.

For clarity, the following are deliberately loose derivative envelopes; $c(r) = K'(r)/(1+K(r))$.

Table 2: Derivative envelopes for the smooth score field. Constants may depend on the fixed dimensions.

Quantity	Value	First path derivative	Second path derivative
r	1	$C\kappa h^{-1}$	$C\kappa^2 h^{-2}$
$\Pi_Q B_Q$	Ch^{-1}	$C\kappa h^{-2}$	$C\kappa^2 h^{-3}$
$c(r)$	Ch^{-1}	$C\kappa h^{-3}$	$C\kappa^2 h^{-5}$
w_h	1	$C\kappa h^{-2}$	$C\kappa^2 h^{-4}$
$W_{Q,h}$	Ch^{-2}	$C\kappa h^{-4}$	$C\kappa^2 h^{-6}$

Indeed $u, v, z^\top Q z'$ have first derivatives bounded by 2κ and second derivatives by $4\kappa^2$. Differentiating $r = (z^\top Q z')/\sqrt{uv}$ uses $u, v \ge h$ and $|r| \le 1$. The first two ordinary derivatives of c are bounded by Ch^{-2} and Ch^{-3}: its strongest singularity, at $r = -1$, is a simple pole. The product rule gives the table. We may therefore take one common envelope

$$\Lambda_h = Ch^{-8} \ge 1. \tag{C.4}$$

Any $P, Q \in \mathcal{P}_{d,k}$ are joined by a minimizing geodesic whose initial symmetric tangent V satisfies

$$\|V\|_{\mathrm{F}}^2 = 2\sum_i \theta_i^2 \le (\pi^2/4)\|Q-P\|_{\mathrm{F}}^2,$$

where the principal angles lie in $[0, \pi/2]$. Integrating the derivative bound gives the global fixed-input inequality

$$\|W_{Q,h}(Z) - W_{P,h}(Z)\|_{\mathrm{F}} \le \Lambda_h \|Q - P\|_{\mathrm{F}}. \tag{C.5}$$

A minimizing geodesic need not be unique; any one suffices. No global smooth eigenbasis is selected.

Let $m_P(Q) = \mathbb{E}_P W_{Q,h}$. Differentiate the identity $\mathbb{E}_Q W_{Q,h} = 0$ at $Q = P$. The field and its derivative are bounded for fixed h, and Appendix A.3 supplies the density's L^1 derivative. By (C.1), the likelihood term is $j_h D$ in tangent direction D. Hence

$$Dm_P(P)[D] = -j_h D. \tag{C.6}$$

The global exponential remainder proved next and a second-order expansion along a minimizing geodesic give

$$\begin{aligned} \|m_P(Q) + j_h(Q - P)\|_{\mathrm{F}} &\le \Lambda_h \|Q - P\|_{\mathrm{F}}^2, \\ \|m_P(Q)\|_{\mathrm{F}} &\le \Lambda_h \|Q - P\|_{\mathrm{F}}. \end{aligned} \tag{C.7}$$

The normal part of $Q - P$ is quadratic. This statement does not identify tangent spaces at distinct projectors or invoke an unregularized off-pilot score moment.

C.3 Geometric inequalities

Lemma C.1 (Three global inequalities)**.** *For rank-k projectors and Frobenius norm, the following hold.*

1. *If D is symmetric tangent at Q, then*

$$\left\|\mathrm{Exp}_Q(D) - Q\right\|_{\mathrm{F}} \le \|D\|_{\mathrm{F}}, \qquad \left\|\mathrm{Exp}_Q(D) - Q - D\right\|_{\mathrm{F}} \le \|D\|_{\mathrm{F}}^2.$$

2. *For every symmetric E and any rank-k top projector Q of $P + E$,*

$$\|Q - P - \Pi_P(E)\|_{\mathrm{F}} \le C_k \|E\|_{\mathrm{F}}^2, \qquad C_k = 16\sqrt{2k} + 4.$$

3. *For $\lambda > 0$, any rank-k top projector Q of $\lambda P + E$ obeys $\|Q - P\|_{\mathrm{F}} \le 2\|E\|_{\mathrm{F}}/\lambda$.*

All statements include large errors and arbitrary valid tie choices.

Proof. For the first statement put $K_0 = [D, Q]$. Then K_0 is skew-symmetric, $[K_0, Q] = D$ and $\|K_0\|_{\mathrm{F}} = \|D\|_{\mathrm{F}}$. For $f(t) = e^{tK_0} Q e^{-tK_0}$, orthogonal conjugation gives $\|f'(t)\|_{\mathrm{F}} = \|D\|_{\mathrm{F}}$ and $\|f''(t)\|_{\mathrm{F}} \le 2\|D\|_{\mathrm{F}}^2$. Integrate the first derivative and the second-order integral remainder. This is the usual Grassmann exponential in projector coordinates (Bendokat et al., 2023).

For the second statement, first suppose $\|E\|_{\mathrm{op}} \le 1/4$. On the complex contour $|z - 1| = 1/2$, the resolvents of P and $P + E$ have operator norms at most two and four. The identity

$$R = R_0 + R_0 E R_0 + R_0 E R_0 E R$$

gives a first-order contour integral $\Pi_P(E)$ and a Frobenius remainder at most $8\|E\|_{\mathrm{F}}^2$. The separated cluster contains the top k eigenvalues. Such local projector expansions are standard (Koltchinskii and Lounici, 2015). If $\|E\|_{\mathrm{op}} > 1/4$, let $x = \|E\|_{\mathrm{F}} > 1/4$ and use

$$\|Q - P - \Pi_P(E)\|_{\mathrm{F}} \le \sqrt{2k} + x \le (16\sqrt{2k} + 4)x^2.$$

This bounds the error at a tie, not continuity of the top-projector map.

Finally the variational property of a top projector gives $\mathrm{tr}[(\lambda P + E)Q] \ge \mathrm{tr}[(\lambda P + E)P]$. Hence

$$\frac{\lambda}{2}\|Q - P\|_{\mathrm{F}}^2 \le \mathrm{tr}[E(Q - P)] \le \|E\|_{\mathrm{F}} \|Q - P\|_{\mathrm{F}},$$

which proves the third statement. □

C.4 Pilot estimators and sample splitting

The pair matrices are bounded by d in operator norm and have mean λP. In fixed dimensions, centered independent mean matrices satisfy second and fourth moment bounds $O(a^{-1})$ and $O(a^{-2})$. These follow by expanding the sums: terms with a singleton centered index vanish, leaving only paired indices and the all-equal term. Apply Lemma C.1(3) to the pilot. For $H_A = Q_A - P$,

$$\sup_P \mathbb{E}\,\|H_A\|_{\mathrm{F}}^2 \le C/a, \qquad \sup_P \mathbb{E}\,\|H_A\|_{\mathrm{F}}^4 \le C/a^2, \tag{C.8}$$

and similarly for B. These inequalities do not condition on a good-pilot event.

Conditional on Q_A, its opposite-half data are independent of it. Set

$$\Delta_A = \frac{1}{b}\sum_{j\in B}\{W_{Q_A,h}(Z_j) - W_{P,h}(Z_j)\} - m_P(Q_A).$$

Equation (C.5) and centering imply

$$\mathbb{E}[\|\Delta_A\|_{\mathrm{F}}^2 \mid Q_A] \le \frac{C\Lambda_h^2}{b}\,\|H_A\|_{\mathrm{F}}^2\,.$$

Its unconditional L^2 norm is $O(\Lambda_h/n)$. The drift remainder in (C.7), together with the fourth pilot moment, has the same order. Since $j_h \ge j_{1/16} > 0$, the correction in (4.4) satisfies

$$H_A + D_A = \frac{1}{bj_h}\sum_{j\in B} W_{P,h}(Z_j) + R_A^{\mathrm{lin}}, \qquad \left\|R_A^{\mathrm{lin}}\right\|_{L^2} \le C\Lambda_h/n. \tag{C.9}$$

The corresponding statement holds for the other half. This is precisely the step that would require unjustified off-pilot control for the raw score; the cutoff provides it explicitly.

C.5 Bounds on the remainders

Conditional fourth-moment bounds for a mean of centered bounded independent fields, and (C.7)–(C.8), imply

$$\mathbb{E}\,\|D_A\|_{\mathrm{F}}^4 \le C\Lambda_h^4/n^2. \tag{C.10}$$

Lemma C.1(1) therefore bounds the exponential remainder in L^2 by $C\Lambda_h^2/n$. Combining with (C.9) yields

$$C_A - P = \frac{1}{bj_h}\sum_{j\in B} W_{P,h}(Z_j) + R_A, \qquad \|R_A\|_{L^2} \le C\Lambda_h^2/n.$$

The displacement bound also gives $\mathbb{E}\,\|C_A - P\|_{\mathrm{F}}^4 \le C\Lambda_h^4/n^2$. The same holds for C_B. Jensen's inequality gives the same fourth-moment order for $M - P = (b/n)(C_A - P) + (a/n)(C_B - P)$; no independence between the corrected halves is used.

Their leading terms recombine exactly to $Z_n = (nj_h)^{-1}\sum_{j=1}^n W_{P,h}(Z_j)$, which is tangent at P. Apply Lemma C.1(2) to $M = P + (M - P)$. Since $\Pi_P(Z_n) = Z_n$ and Π_P is a contraction,

$$\widehat{P}_{\mathrm{eff}} - P = Z_n + R_n, \qquad \|R_n\|_{L^2} \le C\Lambda_h^2/n \le Ch^{-16}/n.$$

This proves (4.5). Large steps, bad pilots and final ties all satisfy the same global inequalities; no omitted event requires a separate uniform-integrability assertion.

C.6 Expected risk

At the true parameter the fields are iid and centered. Therefore

$$\mathbb{E}_P\,\|Z_n\|_{\mathrm{F}}^2 = \frac{2mv_h}{nj_h^2}.$$

Use the exact projector loss (2.2) and Cauchy–Schwarz for the cross term with R_n. Uniformly in P,

$$\left|n\mathbb{E}_P L(\widehat{P}_h, P) - (d-k)\frac{v_h}{j_h^2}\right| \le C\left(\frac{h^{-16}}{\sqrt{n}} + \frac{h^{-32}}{n}\right). \tag{C.11}$$

For $h = h_n = [16 + \log(n+1)]^{-1}$ the right side tends to zero. Equation (C.2) gives (3.3). Together with (B.9), it proves Theorem 3.2. This argument establishes expected loss directly; a central limit theorem is not substituted for moment control.

For a deterministic approximate calibration $\widetilde{j}_n$ with $|\widetilde{j}_n - j_{h_n}| \le \varepsilon_n \to 0$ and an eventual positive lower bound, the uncancelled pilot term has L^2 norm $O(\varepsilon_n/\sqrt{n})$. Repeating the argument adds that quantity to the remainder bound. This statement leaves all other numerical operations exact. Fixed floating-point tolerances for all operations are not covered by an unlimited-n theorem.

D MOMENT ESTIMATORS

D.1 SECOND MOMENTS OF THE PAIR STATISTIC

Conditional on the hidden teacher, q and q' are independent and each has second moment I_d. Expanding the four terms of $A_{ij}A_{\ell m}$ therefore gives

$$\mathbb{E}_P[A_{ij}A_{\ell m}] = \tfrac{1}{2}(\delta_{i\ell}\delta_{jm} + \delta_{im}\delta_{j\ell}). \tag{D.1}$$

In particular $\mathbb{E}_P A^2 = (d+1)I_d/2$. This calculation does not assume that q and q' are unconditionally independent. It uses only their conditional independence and conditional second moments. In coordinates with $P = \operatorname{diag}(I_k, 0)$, the cross-subspace entries have mean zero and covariance $\frac{1}{2}I_{k(d-k)}$.

Let $E_n = n^{-1}\sum_j A_j - \lambda P$ and let $G_n = (E_n)_{21}$. Bounded independent observations give $\mathbb{E}\,\|E_n\|_{\mathrm{F}}^2 = O(n^{-1})$ and $\mathbb{E}\,\|E_n\|_{\mathrm{F}}^4 = O(n^{-2})$, with constants independent of orientation. Lemma C.1(2), applied after division by λ, yields

$$\widehat{P}_{\mathrm{mom}} - P = \lambda^{-1}\Pi_P(E_n) + R_n^{\mathrm{mom}}, \qquad \|R_n^{\mathrm{mom}}\|_{L^2} = O(n^{-1}).$$

This formula is global and includes ties. The leading term has squared Frobenius mean $k(d-k)/(n\lambda^2)$ by (D.1). The exact loss and Cauchy–Schwarz give (3.5) uniformly over P.

D.2 COMPARISON WITH THE MINIMAX RISK

The ratio of the moment coefficient to the minimax coefficient is

$$\frac{\mathcal{I}_k}{2\lambda^2} = \frac{4kF_k}{\pi^2 b_d^4(k-2)}. \tag{D.2}$$

The identity $\mathbb{E}[A_{\alpha i}S_{\alpha i}] = \lambda$ follows by differentiating the bounded-statistic mean $\mathbb{E}_P A = \lambda P$. Since $\operatorname{Var}(A_{\alpha i}) = 1/2$, Cauchy–Schwarz gives $\mathcal{I}_k \ge 2\lambda^2$, proving the lower ratio bound one.

For $k > 3$, convexity of $\arcsin$ on $[0,1]$ gives $|2\arcsin r/\pi| \le |r|$, and hence

$$1 \le F_k \le \mathbb{E}(1-R^2)^{-1} = \frac{k-2}{k-3}.$$

If $G \sim N(0, I_d)$, radial independence and Jensen's inequality give

$$b_d = \frac{\sqrt{d}\sqrt{2/\pi}}{\mathbb{E}\,\|G\|} \ge \sqrt{2/\pi}.$$

Substitution in (D.2) proves $\mathcal{I}_k/(2\lambda^2) \le k/(k-3)$. In conjunction with Theorem 3.2, it proves the ordered-limit statement of Proposition 3.3. No uniform remainder has been supplied for a simultaneous growth of n, d, k.

D.3 COVARIANCE OF THE LAG-ONE STATISTIC

We prove Proposition 5.1. Condition on the exogenous teacher schedule. Write $a_t = B^\top q_t$ and $\xi_{t,\alpha} = b_{\perp,\alpha}^\top q_t$ for an outside unit vector. The assumptions imply

$$\mathbb{E}a_t = \mu_t, \quad \mathbb{E}\xi_{t,\alpha} = 0, \quad \mathbb{E}[a_t a_t^\top] = I_k, \quad \mathbb{E}[\xi_{t,\alpha}a_t] = 0, \quad \mathbb{E}[\xi_{t,\alpha}\xi_{t,\beta}] = \delta_{\alpha\beta}.$$

The αth row of the cross-subspace block of the summand is

$$G_{t,\alpha} = \tfrac{1}{2}(\xi_{t,\alpha} a_{t+1}^\top + \xi_{t+1,\alpha} a_t^\top).$$

Its mean is zero. Conditional independence over time gives

$$\mathbb{E}[G_{t,\alpha}^\top G_{t,\beta}] = \tfrac{1}{2}\delta_{\alpha\beta} I_k.$$

For adjacent summands, the only surviving cross term shares the outside component at $t+1$:

$$\mathbb{E}[G_{t,\alpha}^\top G_{t+1,\beta}] = \tfrac{1}{4}\delta_{\alpha\beta} \mu_t \mu_{t+2}^\top.$$

More distant cross covariances vanish. Therefore different outside rows of $G_N = B_\perp^\top C_N B$ are uncorrelated, and each row has exact covariance

$$W_N = \frac{I_k}{2(N-1)} + \frac{N-2}{2(N-1)^2} J_N. \tag{D.3}$$

This accounts for overlap; it does not replace overlapping products with iid pairs.

The conditional mean of C_N is $B H_N B^\top$. For sufficiently large N, its k positive eigenvalues are separated from zero because $H_N \to H \succ 0$. The first-order cross-block of its leading projector is $G_N H_N^{-1}$. Thus the first-order loss is

$$k^{-1} \left\| G_N H_N^{-1} \right\|_{\mathrm{F}}^2,$$

whose conditional expectation is $(d-k)\,\mathrm{tr}(H_N^{-1} W_N H_N^{-1})/k$.

To control the remainder, split the centered lag-one summands into odd and even indices. Within each subset the summands are independent conditional on the schedule and uniformly bounded. Fourth moments of their normalized means are $O(N^{-2})$; bounded concentration also controls the fixed-gap complement. A fixed-gap projector expansion and bounded loss then make the expected remainder $o(N^{-1})$, uniformly for the deterministic schedule's bounded observations. Multiplying the first-order expression by N and using (D.3) proves (5.2).

The convergence assumptions do not impose a rate on $H_N - H$: the conditional population eigenspace is exactly $\mathrm{range}(B)$ for every sufficiently large N. The finite-N first-order term uses H_N^{-1} before taking the limit. This point prevents an unjustified $\sqrt{N}$-rate assumption on the schedule.

D.4 Limitations of the risk formulas

In the pair model, isotropic task diversity makes the population operator λP. In the long-history result, directional diversity is encoded in the eigenvalues of H and persistence also affects J. Weak useful directions are penalized through H^{-1}. This is a quantitative statement about the specified statistic. Autocovariance-based factor identification and the importance of task diversity are not new general principles (Lam and Yao, 2012; Tripuraneni et al., 2021; Datta et al., 2026). Neither formula is a minimax theorem for arbitrary unknown episode laws or a risk guarantee for the target-phase classifier.

E Details of the experiments

The mathematical estimator and the experimental spectral memory are distinct. This appendix specifies the latter. Numerical results below are transcribed from frozen analyses; no new outcome is generated for the manuscript.

E.1 Data generation

In each rank-sixteen context, choose a 16-dimensional coordinate subspace of $\mathbb{R}^{64}$ and then apply a common independent dense orthogonal rotation. Historical teacher directions are independent normalized Gaussian vectors in that subspace. Each of the 64 block lengths is independently and uniformly drawn from $\{2047, 4096, 8191\}$. Inputs are standard Gaussian vectors and labels are noiseless signs. A target direction is generated independently within the shared subspace, without

accepting or rejecting it based on any historical teacher. It receives 32 independent support inputs and 2048 independent query inputs. A paired identity-coordinate run is retained only as a diagnostic.

The representation learner receives inputs and feedback, but no historical teacher vectors, task identifiers or historical block boundaries. It uses all overlapping lag-one products, including products crossing blocks. The evaluator knows the latent subspace and teachers in order to construct controls and compute coverage. The target-phase transition is explicitly supplied: the historical metric is frozen, and the classifier is reset. Thus this is not a wholly boundary-free lifelong-learning pipeline.

Separate deterministic seed namespaces generate the shared geometry, history teachers, durations, history inputs, coordinate rotation, controls, target direction, target supports and queries. Target-seed interventions were specified to leave history-derived quantities unchanged, and history-seed interventions to leave the independently sampled target unchanged in the latent coordinates. A shared coordinate transform is subsequently applied to both histories and target data.

E.2 The learning algorithm

Given C_N in (5.1), let its eigenvalues be ordered and let γ be the rank-k cutoff eigenvalue. At a simple cutoff, $\widehat{\Pi}_k$ is the ordinary leading-k projector. The implementation uses tolerance

$$\delta = 10^{-12} \max\{1, \max_i |\lambda_i(C_N)|\}.$$

Eigenvectors above $\gamma + \delta$ receive weight one. Those within δ of γ share the remaining spectral mass equally; all others receive zero. This filter has trace k and spectrum in $[0, 1]$, and is unambiguous within the tied eigenspace. It need not be an idempotent projector at a numerical tie. Theoretical top-projector estimators in the minimax theorem instead choose an arbitrary valid rank-k projector; their risk bounds include ties. These conventions must not be conflated.

For the experimental filter, define

$$F = \tfrac{1}{2} I_d + \frac{d}{2k} \widehat{\Pi}_k, \qquad \mathcal{F} = \mathrm{diag}(F, 1), \qquad f(x) = \frac{(x, 1)}{\sqrt{d+1}}.$$

The classifier begins at $w_0 = 0$. For a support (x_t, y_t), it updates by

$$w_{t+1} = \rho w_t + \{y_t - \tanh(w_t^\top f(x_t))\} \frac{\mathcal{F} f(x_t)}{\|\mathcal{F} f(x_t)\|} \frac{\|f(x_t)\|}{1 + \|f(x_t)\|^2}, \qquad \rho = 0.5. \tag{E.1}$$

The prediction is $+1$ when $w_t^\top f(x) \geq 0$, and -1 otherwise. The same head, supports and queries are used in every metric arm. No result here establishes that (E.1) is an optimal readout. The normalization $q = \sqrt{d}\, yx / \|x\|$ is used for history, whereas $f(x)$ is used for the head; they are different feature maps.

E.3 Controls and comparators

Erased history uses $\mathcal{F} = I_{d+1}$. Unrelated history uses the same inputs and durations but labels generated by teachers in an orthogonal k-dimensional subspace. Shuffled feedback globally permutes the related history labels. Input-only history replaces all labels by $+1$. Scrambled directions orthogonally conjugate the related spectral metric, preserving its spectrum but disrupting its alignment with the true span. These are mechanism controls, not a comprehensive set of state-of-the-art learning baselines.

Two additional inherited comparators use an exponential moving average of the same pair statistic, with step $\eta = 1/(128d)$ and zero initial matrix. The first pending predecessor is zero. One keeps only its diagonal; the other uses the full matrix. For eigenvalues or diagonal entries c_i, their feature-coordinate metric eigenvalues are

$$\tfrac{1}{2} + \tfrac{d}{2} \frac{\exp(\beta c_i)}{\sum_j \exp(\beta c_j)}.$$

The inherited parameters are $\beta = 16$ for the diagonal comparator and $\beta = 4$ for the full matrix. They were not retuned for rank sixteen. The soft true-subspace oracle uses the same $I_d/2 + dP/(2k)$ metric with the actual P. The hard oracle uses P on feature coordinates and retains the bias coordinate.

E.4 Experimental protocol and statistical analysis

The frozen design specified one geometry and one duration distribution, 24 development contexts, then 64 disjoint confirmation contexts. Its joint gate required all five individual lower 99% confidence endpoints to be positive, mean early gains of at least two percentage points over both erased and unrelated histories, mean nearest absolute teacher–target cosine at most 0.7, and at least 90% of targets with nearest cosine at most 0.75. Every sampled context had to be retained. The development study passed before the confirmation activation receipt was created; the setting and endpoints were unchanged.

For each context and arm, query accuracy is measured at shots $0, 1, 2, 4, 8, 16, 32$. The primary statistic is the mean of shots $1, 2, 4$. The bootstrap resamples entire independent contexts with replacement 20,000 times and uses the 0.005 and 0.995 quantiles of the paired mean difference. Two rotations of one context are not treated as independent replicates. The intervals are nominal percentile-bootstrap intervals and do not claim exact finite-sample or simultaneous coverage.

Confirmation mean span overlap is $\mathrm{tr}(B^\top \widehat{\Pi}_k B)/k$. Target-direction coverage is $v^\top \widehat{\Pi}_k v$, and worst-direction coverage is the smallest eigenvalue of $B^\top \widehat{\Pi}_k B$. The filter convention above also defines these quantities at ties. The total number of history examples in confirmation is 19,379,539, not a fixed 64 times one common history length. The 24 development and 64 confirmation contexts together use 26,650,978 history examples.

E.5 Comparison with retrieval

For a fixed independent unit target and a uniform historical teacher in k dimensions, squared cosine has distribution $\mathrm{Beta}(1/2, (k-1)/2)$. Independence of H teachers gives the maximum-cosine formula in Section 6. At $k = 16, H = 64$, its expected maximum is 0.604829, and the probability that at least one absolute cosine exceeds 0.75 is 0.033096. The observed mean is 0.608079; 4 of 64 contexts exceed 0.75 and remain included; the largest cosine is 0.791913.

The prespecified retrieval diagnostic receives all true historical teachers and their negatives. It chooses the teacher with the most target-support label agreements, breaking ties by the largest signed support margin and then the first index. Its output must be this single teacher: it may not adapt the vector, combine teachers, or make query-dependent selections. Its early contrast is a secondary endpoint, not a gate. The population accuracy of the best signed teacher when the true target is known is $1/2 + \arcsin(M)/\pi$. The resulting mean ceiling, 70.9157%, is privileged and not deployable. These restrictions limit the interpretation of the early advantage over support-based retrieval.

E.6 Additional experiments

The low-redundancy study should not be pooled with the earlier ones. The following comparison is descriptive, not a common-budget test.

Table 3: Complementary independent-target studies. The history budgets differ, so this is a descriptive comparison.

Study	d, k, H	Early gain vs erased	Sampling qualification
Independent target, rank 4	$64, 4, 128$	10.735 pp	Target independent; mean nearest absolute cosine 0.96675.
Independent target, rank 16	$64, 16, 64$	2.844 pp	Much longer histories; mean nearest cosine 0.60808; no rejection.

The rank-four independent-target confirmation interval against erased history is $[8.874, 12.615]$ pp. Its block lengths are sampled from $\{31, 64, 127\}$, rather than the much longer rank-sixteen lengths. An earlier target-first branch imposed $|u^\top v| \le 0.75$ by rejecting historical teachers after choosing the target. It therefore enforces an exclusion property but does not have target-independent history. We do not use it to claim both properties simultaneously. The minimax theorem concerns neither of these long-history sampling laws.

E.7 Experiments with streaming algorithms

Four prospective low-memory attempts were tested in a separate pair-persistence sanity family at $d = 32$, including supplied rank eight, using 32,768 labeled observations per context. Each was developed on its own fresh contexts; the rows below are not paired with each other. Each rank-eight condition had eight contexts and a prespecified mean-overlap threshold of 80%.

Table 4: Negative streaming development results. Each variant was evaluated on its own fresh contexts; the gate was a mean overlap of at least 80%.

Prospective variant	Rank-eight mean overlap	Gate
Signed tangent update with polar retraction	72.09%	Failed
Slow-gain update with tail averaging	75.18%	Failed
Trace-calibrated Oja, last iterate	76.06%	Failed
Capacity-first acquisition and fresh compression	73.75%	Failed

The wider acquisition space in the last variant did not change the scored rank-eight output. These failures were preserved and did not trigger confirmation of those learners. They do not prove that an efficient streaming solution is impossible, and they do not contradict a minimax theorem about a dense, regularized estimator. General symmetric-update Oja theory already exists (Huang et al., 2021); these are failures of particular implementations and schedules, not a negative result for all streaming PCA.

E.8 Code and reproducibility

The anonymous supplement includes the low-redundancy protocol, development and confirmation seed ledgers, preserved aggregate results, and the frozen generator and learner. The generator and learner files and both seed ledgers match the checksums recorded in the prospective protocol. The public runner and analysis scripts are reconstructed wrappers; they are not represented as the original frozen runner and analyzer.

The original audit record reports deterministic replay of all 88 development and confirmation contexts, independent query aggregation, bootstrap reconstruction, and source-hash checks. The full original raw-data archive is not included in this supplement. Instead, per-context confirmation outputs are regenerated using the checksum-verified core and frozen seeds; these are identified as regenerated records and checked against the preserved aggregate analysis. Regeneration uses the original contexts and does not supply new scientific evidence. The supplement documents the environment, checks performed, and any differences from the archived records.

Numerical proof checks and their source files accompany the written proofs. They test selected algebraic identities and finite-dimensional calculations; they do not establish the theorem by numerical experiment or replace verification of its assumptions and proof.