LEARNING TO CORRECT OVERLAPPING FORECASTS: ACCURACY, QUERIES, AND LOCAL ERRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

When several models predict the same outcome from partly overlapping inputs, combining their outputs can count shared evidence repeatedly or lose useful information. Residual correction offers an alternative: each model predicts the error remaining after earlier predictions, using only its own inputs, and its output is added as a correction. Learning these corrections from finite data introduces errors that affect later prediction targets. We develop a theory of the accuracy attainable with imperfect corrections and the conditions under which learning them reduces rather than amplifies error. In an independent additive-signal model, we identify which local errors later experts can remove and which survive, and show that residual queries can outperform ordinary forecasts even with imperfect responses. When response error scales with the predictable residual, we determine the optimal worst-case risk for unequal expert precisions and every repeated-query budget. An optimal allocation and ordering rule attains it without knowing the information overlap. We then analyze corrections learned by local ridge regression on fresh residual labels. For iid symmetric features with finite variance, we give a necessary and sufficient stability condition for fixed covering sweeps with common training parameters. Gaussian least-squares correction requires more samples per fit than the total feature dimension plus one to converge, even with correlated features. Regularization can stabilize smaller batches. Matched-data and fixed-budget experiments test how regularization and sample allocation turn these stability conditions into useful correction.

1 INTRODUCTION

Consider several models that predict the same outcome from different inputs. For example, two models forecasting electricity demand might both use weather data, while one also uses past consumption and the other uses industrial activity. Their inputs contain both shared information and information available to only one model. We study how to combine their predictions when the underlying inputs are not pooled. In our theory, each such predictor is called an *expert*: it can use only its own information. Simply adding predictions can count shared evidence twice, while averaging can dilute information available to only one expert. Moreover, each prediction compresses its inputs, so some information can remain unrecoverable even to an optimal aggregator that knows the prior (Neyman, 2024). Our goal is to understand how changing what these models predict can improve information recovery.

Residual queries change what each expert predicts (Frongillo et al., 2025). Each expert predicts what the current combined predictor still gets wrong, and its response is added as a correction. The expert still uses only its own evidence: earlier prediction functions change the target, not the expert's observed information. In the independent additive-signal model, exact responses recover all information in one pass. Imperfect responses change later targets, and learning each correction from finite data can introduce further errors. This raises our central question: *when do imperfect corrections recover information, and when do they amplify error?*

The background is prediction from distributed evidence, including forecasting tournaments (Mellers et al., 2014), economic and operational forecast combination (Wang et al., 2023), and collaborative or networked machine learning (Collina et al., 2026; Kearns et al., 2025). Partial-information models distinguish overlap from measurement noise (Satopää et al., 2015), while work on shared information studies how it affects aggregated judgments (Palley & Soll, 2019). Robust aggregation addresses unknown information structures (Arieli et al., 2018; Neyman & Roughgarden, 2022), and higher-order

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.

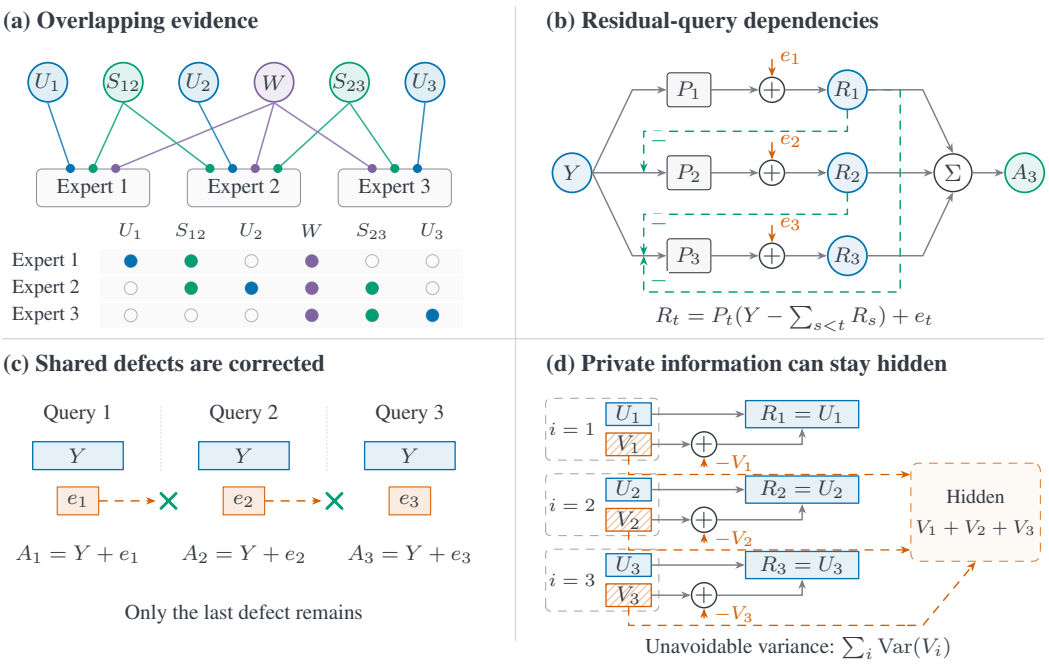

Figure 1: Information overlap enables correction but leaves a private-error floor. (a) An ownership graph and its incidence matrix. (b) Residual queries: dashed edges enter prediction targets, not experts' observed features. (c) A later fully informed expert corrects an earlier shared defect. (d) Local errors $e_i = -V_i$ hide private information. The separation theorem combines shared and private components in one instance; the learning analysis measures which fitted errors survive correction.

reports elicit evidence absent from ordinary forecasts (Prelec et al., 2017; Chen et al., 2021). Our focus is the accuracy and learnability of residual responses when exact answers are unavailable.

We develop a theory linking the value of imperfect queries to the conditions for learning their responses. We first determine attainable accuracy from response precision and query budgets, then study when corrections learned by local ridge regression are stable. Errors are measured against the *actual* residual target. After training, a predictor using only its own features has a local error, even when training was random or shared. Such errors need not be linear, unbiased, or independent. Fresh randomness added at prediction time is different: it may be invisible to later predictors and need not satisfy this locality condition.

The central distinction is whether approximation consumes an absolute error budget or a fraction of the predictable residual. Absolute tolerances create an irreducible private-error floor. Relative tolerances permit successive corrections to become more accurate as the residual shrinks. Learning these corrections introduces a second boundary: repeated local regression can amplify estimation error even though exact residual queries recover the target. Our contributions characterize the accuracy attainable from imperfect queries and the training conditions under which local correction converges.

1. An error decomposition and a noisy-query separation. We characterize error exactly as uncovered information plus orthogonal surviving local defects, yielding the heterogeneous absolute-error minimax frontier. An overlapping Gaussian construction then separates residual queries from ordinary forecasts with *any* measurable decoder throughout the nontrivial absolute-error regime.

2. Optimal precision allocation and query order. When error scales with the predictable actual target, we determine the exact minimax risk for heterogeneous expert precisions and every repeated-query budget. A joint allocation and ordering rule attains it over arbitrary linear target graphs and measurable decoders. Equal counts need not prevent exponential risk gaps in residual summation.

3. An exact stability boundary for learned correction. For iid symmetric features, we derive the finite-sample energy dynamics of local ridge regression and an exact convergence criterion for covering sweeps with common training parameters. The Gaussian sample threshold holds under

arbitrary positive definite feature correlation. Matched-data and fixed-budget experiments isolate how regularization and sample allocation govern correction.

Market scoring rules reward changes in a common forecast's score (Neyman & Roughgarden, 2023), and dynamic-market convergence studies strategic information revelation (Ostrovsky, 2012). Our fixed local information and report-dependent targets follow the DAG mechanism of Frongillo et al. (2025); earlier report realizations are not additional observed features. This also distinguishes the learning construction from networked predictors that observe parent predictions (Kearns et al., 2025). Appendix H develops the scoring connection and its information requirements.

2 Information and imperfect residual predictions

There are n experts and independent centered primitive signals $X_1, \dots, X_m$ with finite variances. Signal ℓ has a nonempty owner set $T_\ell \subseteq [n]$. Multiple signals can have the same owners, and zero-variance signals are allowed. Define

$$\mathcal{F}_i = \sigma(X_\ell : i \in T_\ell), \qquad P_i Z = \mathbb{E}[Z \mid \mathcal{F}_i], \qquad Y = \sum_\ell X_\ell, \qquad \mathbb{E}Y^2 = 1. \tag{1}$$

The aggregator knows this model family but not the ownership sets or prior. For the oracle results, experts evaluate conditional means under the common prior. Section 5 replaces this oracle with predictors fitted from samples. All norms are population L^2 norms. Risk is mean squared error relative to Y, the target recoverable from the pooled primitive signals.

Queries and order of choices. An independent public seed Z fixes an ordered list of $k \le B \le n$ distinct experts, real coefficients, and a measurable decoder g. Reports are local functions satisfying

$$H_t = a_t Y + \sum_{s<t} b_{ts} R_s, \qquad R_t = P_{i_t} H_t + e_t, \qquad P_{i_t} e_t = e_t. \tag{2}$$

The prior is chosen before the seed; error functions may depend on the seed, and every oracle norm bound holds conditional on it. Selection and coefficients cannot depend on realized signals or the unknown prior. The graph describes dependencies among report functions, not conditioning on an observed transcript. All reports and targets are square-integrable. The budget includes every report node, even when its value is absent from the final decoder.

Absolute and relative precision. An absolute tolerance requires $\|e_t\| \le \eta_{i_t}$. For this model, the class $\mathcal{P}_B$ also requires the counterfactual reports Q_t, obtained by setting all errors to zero, to satisfy

$$\|Q_t\| \le 1 \quad \text{for every admissible prior, seed, and node.} \tag{3}$$

This fixes the units of each clean answer. Our relative model instead requires $\|e_t\| \le \eta_{i_t} \|P_{i_t} H_t\|$ and needs no coefficient normalization. Section 4 contrasts predictable and full-target tolerances.

Lemma 1 (One-pass direct gain)**.** *Condition* (3) *implies* $|a_t| \le 1$ *for every distinct-expert query.*

Proof. Place unit target variance in a signal private to the current expert. All preceding clean reports vanish; normalization of the current report $a_t Y$ forces the direct-gain bound. □

The lemma uses normalization over the full prior family in (1). Restricting the family, for example by ruling out private information, can weaken this implication. The protocol class and its worst-case value must then be specified for that restricted family.

Residual queries. The difference-query mechanism of Frongillo et al. (2025) sets $a_t = 1, b_{ts} = -1$, and sums the reports:

$$A_0 = 0, \qquad R_t = P_{i_t}(Y - A_{t-1}) + e_t, \qquad A_t = A_{t-1} + R_t. \tag{4}$$

Each clean report contains the additive coordinates first encountered at its expert and satisfies (3). Later experts also predict the visible part of earlier errors. With experts observing (U, W) and (V, W), the reports are $U + W + e_1$ and $V - P_2 e_1 + e_2$, yielding

$$A_2 - Y = (I - P_2)e_1 + e_2. \tag{5}$$

The shared part of the first error disappears. The surviving part is orthogonal to all functions of the second expert's information, including its own response error.

A three-expert calculation. Let independent W, U_1, U_2, U_3 each have variance $1/4$, let $Y = W + U_1 + U_2 + U_3$, and let expert i see (W, U_i). Take $e_1 = 0.2W + 0.1U_1$, $e_2 = -0.1W + 0.1U_2$, and $e_3 = 0.1W$. The successive reports are $1.2W + 1.1U_1$, $U_2 - 0.3W + 0.1U_2$, and $U_3 + 0.2W$. Thus $A_3 - Y = 0.1(U_1 + U_2 + W)$ has risk 0.0075: only private errors and the final shared error survive. The input errors have total energy 0.0200.

3 Surviving errors and the advantage of richer queries

Write $P_t = P_{i_t}$, $C_t = I - P_t$, and define the clean residual and surviving-error operators

$$F = C_k \cdots C_1, \qquad D_t = C_k \cdots C_{t+1} P_t. \tag{6}$$

Independent-coordinate conditional expectations commute on the full space L^2, including nonlinear error functions. Orthogonality is consequently produced by the correction operators and holds even for correlated input errors.

Theorem 2 (Exact surviving-error identity)**.** *For commuting orthogonal projections and local errors, $F, D_1, \ldots, D_k$ are mutually orthogonal projections summing to I, and*

$$Y - A_k = FY - \sum_t D_t e_t, \tag{7}$$

$$\|Y - A_k\|^2 = \|FY\|^2 + \sum_t \|D_t e_t\|^2 \le \|FY\|^2 + \sum_t \|e_t\|^2 . \tag{8}$$

Proof. The residual obeys $r_t = C_t r_{t-1} - e_t$. Expand this recursion and use $P_t e_t = e_t$. Products of commuting projections are projections. For $t < u$, $D_t D_u$ contains $C_u P_u = 0$, and $F D_t = 0$ similarly. Telescoping gives $F + \sum_t D_t = I$, so Pythagoras proves the identity. □

For a selected set S, additivity gives the uncovered-information term

$$\|FY\|^2 = \sum_{\ell : T_\ell \cap S = \varnothing} \mathrm{Var}(X_\ell). \tag{9}$$

Clean information is assigned to its first selected owner, whereas an error survives only after its last correcting owner. Constants, biased errors, and shared nonlinear functions are covered by the same identity. For noncommuting projections, the error-only energy bound still holds, but a cross term with clean bias may remain; Appendix E gives the exact operator statement.

3.1 The absolute-error minimax frontier

Let $\mathcal{E}(\boldsymbol{\eta}) = \sum_i \eta_i^2$ and define

$$\mathcal{R}_B^*(\boldsymbol{\eta}) = \inf_{\Pi \in \mathcal{P}_B} \sup_{\mathcal{D}, e} \mathbb{E}[(g(R) - Y)^2], \tag{10}$$

where the supremum uses (1) and local absolute tolerances. Expectation includes the public seed. The zero predictor always has risk one.

Theorem 3 (Budget and heterogeneous absolute precision)**.** *For $0 \le B \le n$ and $\boldsymbol{\eta} \ge 0$,*

$$\mathcal{R}_B^*(\boldsymbol{\eta}) = 1 - \frac{B}{n}\big(1 - \mathcal{E}(\boldsymbol{\eta})\big)_+. \tag{11}$$

Uniformly sampling B experts and using residual queries attains this value when $\mathcal{E}(\boldsymbol{\eta}) \le 1$; otherwise return zero. The lower bound allows every measurable decoder.

A uniform subset misses any nonempty owner set with probability at most $1 - B/n$ and selects each error budget with probability B/n. Equation (8) gives the upper bound. The matching lower bound hides independent private components of variances η_i^2 inside local errors and places the remaining variance at a least-queried expert. Lemma 1 makes this erasure valid for every normalized linear target graph. Appendix B proves both bounds and characterizes optimal balanced coverage. With homogeneous precision, even full participation guarantees an improvement over zero only when $n\eta^2 < 1$, exposing the population dependence of an absolute tolerance.

3.2 A SEPARATION WITH OVERLAPPING INFORMATION

The minimax frontier alone does not quantify the value of residual questions. Let $\mathcal{R}^*_{\mathrm{ord}}(\boldsymbol{\eta})$ be the minimax risk when every expert instead reports $P_iY + e_i$ once, with local absolute tolerances, and the aggregator applies any measurable decoder. The next theorem quantifies this loss.

Theorem 4 (Ordinary forecasts leave an additional information loss)**.** *For $n \geq 2$ and $\mathcal{E} = \mathcal{E}(\boldsymbol{\eta}) < 1$, put $c_n = ((n-1)/(n+1))^2$. Then*

$$\mathcal{R}^*_{\mathrm{ord}}(\boldsymbol{\eta}) \geq \mathcal{E} + (1-\mathcal{E})c_n, \qquad \mathcal{R}^*_{\mathrm{ord}}(\boldsymbol{\eta}) - \mathcal{R}^*_n(\boldsymbol{\eta}) \geq (1-\mathcal{E})c_n > 0. \tag{12}$$

The ordinary-report lower bound holds on an overlapping Gaussian instance even when the decoder knows its prior.

Proof. Take independent centered Gaussian $W, U_1, \ldots, U_n, V_1, \ldots, V_n$ with $\mathrm{Var}(W) = \mathrm{Var}(U_i) = (1-\mathcal{E})/(n+1)$ and $\mathrm{Var}(V_i) = \eta_i^2$. Expert i sees (W, U_i, V_i) and $Y = W + \sum_i (U_i + V_i)$. Errors $e_i = -V_i$ leave reports $W + U_i$. Their optimal conditional-mean decoder is $\frac{2}{n+1}\sum_i (W + U_i)$, whose residual variance is $(1-\mathcal{E})c_n + \sum_i \mathrm{Var}(V_i)$. Independence of the hidden components proves the claim for every decoder. Theorem 3 supplies the residual upper bound. □

The noiseless Gaussian obstruction c_n is established in Neyman (2024) and discussed by Frongillo et al. (2025). Adding private erasure preserves a strict gap throughout $\mathcal{E} < 1$, even against a decoder that knows the prior. The residual value is exact, while the ordinary value is bounded below by (12).

4 ALLOCATING AND ORDERING IMPERFECT QUERIES

An absolute error floor need not describe accuracy on progressively smaller targets. Give expert i a known *projected-relative* tolerance η_i:

$$\|e_t\| \leq \eta_{i_t} \|P_{i_t} H_t\|, \qquad P_{i_t} e_t = e_t. \tag{13}$$

Precision now scales with the ideal response to the actual target. Query coefficients may be arbitrary finite real numbers, without clean normalization or bounded direct gain. The homogeneous model uses a common precision $\eta_i = \eta$ for every expert and every query.

An incentive meaning for relative precision. The quadratic improvement score $H_t^2 - (R_t - H_t)^2$ gives abstention zero payoff and expected reward $\|P_t H_t\|^2 - \|e_t\|^2$. Thus $\eta_i \leq 1$ preserves a fraction $1 - \eta_i^2$ of the best expected improvement. This score-difference structure also appears in market scoring rules (Neyman & Roughgarden, 2023); our experts retain their fixed local information when predicting report-dependent targets.

Theorem 5 (Relative sweep contraction)**.** *For homogeneous $0 \leq \eta \leq 1$, commuting projections and any initial residual r_0, residual queries satisfy*

$$\|r_k\|^2 \leq \|F r_0\|^2 + \eta^2 \|(I - F) r_0\|^2. \tag{14}$$

The proof sums the one-step decrease $(1-\eta^2)\|P_t r_{t-1}\|^2$ and bounds it below by the newly covered components of r_0 (Appendix J). With unequal precisions, the proof must also control transfers of local error between shared and private components.

Theorem 6 (Optimal precision allocation and query order)**.** *Set $\rho_i = \min\{\eta_i^2, 1\}$, and let $g_i(x)$ linearly interpolate the points (k, ρ_i^k) for integers $k \geq 0$, with $0^0 = 1$. Over (1), arbitrary repeated-expert linear query graphs fixed by an independent public seed, and measurable decoders, the exact minimax risk at budget L is*

$$V_L(\boldsymbol{\eta}) = \min_{\substack{x_i \geq 0 \\ \sum_i x_i \leq L}} \max_i g_i(x_i). \tag{15}$$

If all $\rho_i < 1$ and $V_L > 0$, choose $x_i = g_i^{-1}(V_L)$, jointly round to adjacent integer counts N_i with $\mathbb{E}N_i = x_i$ and $\sum_i N_i = L$, and order tokens (i,k), $1 \leq k \leq N_i$, by increasing ρ_i^k. Residual queries in that order attain the value without knowing ownership. An expert with $\rho_i = 1$ forces value one at every budget. If all experts are perfect, the value is $(1 - L/n)_+$.

Proof idea. Private Gaussian erasures give every protocol lower risk $\max_i \mathbb{E}\rho_i^{N_i}$. Convexity and dependent rounding identify its minimum with (15). For the upper bound, a backward quadratic potential maintains a separate ceiling on each expert's entire local function space. Sorting the levels preserves all ceilings despite error transfer through shared information. The resulting risk is at most $\sum_\ell \mathrm{Var}(X_\ell) \min_{i \in T_\ell} \rho_i^{N_i}$. Averaging proves the matching bound uniformly over unknown ownership and all admissible local error functions (Appendix V). □

Corollary 7 (Homogeneous precision at every budget)**.** *Let $L = qn + r$, $0 \le r < n$, and $\rho = \min\{\eta^2, 1\}$. Then*

$$\mathcal{R}^*_{L,\mathrm{rel},\mathrm{rep}}(\eta) = \rho^q \left[1 - \frac{r}{n}(1-\rho)\right]. \tag{16}$$

Here $x_i = L/n$. For $\eta \le 1$, a uniform r-subset followed by q full residual sweeps attains the value; for $\eta > 1$, use zero prediction. For $L = B \le n$ this gives the one-pass frontier

$$\mathcal{R}^*_{B,\mathrm{rel}}(\eta) = 1 - \frac{B}{n}(1-\eta^2)_+. \tag{17}$$

The partial sweep comes first, applying its coverage guarantee before seed-dependent errors change the residual. Appendix O also gives a direct proof and the full-information value $\min\{1, \eta^{2L}\}$.

For example, squared precisions $(0.04, 0.25, 0.64)$ give optimal six-query risk 0.18577, whereas two visits per expert have lower bound 0.4096 (Figure 7). The policy revisits less precise experts more often because the adversary can place target information in any expert's private component.

Equal counts can conceal ineffective correction. For $0 < \rho < 1$ and two experts queried $K > 2$ times each, residual summation over full sweeps guarantees risk ρ^K. Yet the order $1^{K-1}, 2, 1, 2^{K-1}$ permits risk arbitrarily close to ρ^2: errors move from shared information into private information after its final owner visit. The ratio grows as ρ^{2-K} (Appendix V). Thus the guarantee depends both on how often each expert is queried and on which information remains available for later correction.

Full-target tolerance $\|e_t\| \le \eta \|H_t\|$ is different: unpredictable variation enlarges the error budget. Even two private signals can then exceed the sweep bound η^2 (Appendix J). Absolute-error amplification and its relative-precision limit are derived in Appendices C and J.

5 When learned residual correction converges

Residual fitting is the squared-loss boosting update (Murphy, 2022, Section 18.5.2). Here each report uses a fixed local feature space. Node t fits labels $H_t = Y - \sum_{s<t} \widehat{f}_s$, holding earlier fitted functions fixed. Conditional on training data $\mathcal{T}$, its error satisfies

$$e_t = \widehat{f}_t - P_t H_t, \qquad \|e_t\|^2 = \left\|\widehat{f}_t - H_t\right\|^2 - \|P_t H_t - H_t\|^2. \tag{18}$$

Thus excess prediction risk is exactly local error energy. Theorem 2 applies conditional on any fitted collection, including predictors trained on shared data. Locality alone, however, does not make repeated learning converge: a regression can inject more error from unobserved coordinates than it removes from observed ones. The distribution of fitted errors determines which effect dominates when experts repeatedly learn corrections from fresh residual labels.

A statistical model with unknown responses. Let $Z_1, \ldots, Z_m$ be independent, identically distributed, symmetric, unit-variance features, and let $Y = \theta^\top Z$ for an unknown fixed θ with $\|\theta\|_2 = 1$. Features are population-standardized; their law need not be Gaussian or have a fourth moment. Expert t observes coordinates J_t. Its report is a zero-intercept ridge regression on N_t fresh labeled examples, with objective $\|y - Gb\|_2^2 + \lambda_t \|b\|_2^2$. The training coordinator forms actual residual labels by evaluating preceding reports on these examples. At prediction time only local features are used. Neither θ nor the feature distribution is supplied to the learner. The coordinator accesses outcomes and black-box evaluations of local predictors on matched example IDs, while local feature matrices stay at the experts. Each update fits the current residual locally. This is a specified training workflow, not a lower bound over all distributed learning procedures; if raw features can instead be pooled, noiseless Gaussian OLS recovers θ from a full-rank design and permits assigning the recovered coefficients directly to their owners for local prediction.

Theorem 8 (Exact energy dynamics beyond Gaussian features)**.** *Fix the ownership sequence and training parameters before sampling, with $\lambda_t > 0$. At a node with $d = |J| \geq 1$, N examples, and penalty λ, let G be its $N \times d$ iid design and $A = (G^\top G + \lambda I)^{-1}$. Define*

$$\alpha = \lambda^2 \mathbb{E} A_{11}^2, \qquad \gamma = \lambda^2 \mathbb{E} A_{12}^2, \qquad \beta = \mathbb{E}(AG^\top GA)_{11}, \tag{19}$$

with $\gamma = 0$ when $d = 1$. All three moments are finite. Starting from $x_j = \theta_j^2$, the expected squared residual coefficients update exactly as

$$x_i' = \begin{cases} x_i, & i \notin J, \\ \alpha x_i + \gamma \sum_{j \in J \setminus \{i\}} x_j + \beta \sum_{j \notin J} x_j, & i \in J. \end{cases} \tag{20}$$

The training-averaged population risk is $\sum_i x_i$ after every node.

Conditional on earlier fits, outside coordinates provide regression noise independent of the fresh local design. Sign symmetry eliminates mixed coefficient terms; exchangeability gives the three moments in (19). Bounded regularized inverses make them integrable with only second moments. Appendix Q proves the recurrence without assuming independence among earlier fitted coefficients.

The recurrence is a nonnegative linear map M_J on coordinate energies. It separates retained ridge bias within the local block from estimation error induced by coordinates the expert cannot observe. With common d, N, λ, every updated row has the same sum

$$c = \alpha + (d-1)\gamma + (m-d)\beta. \tag{21}$$

This scalar determines whether repeated training removes error or amplifies it, for every fixed ownership pattern whose queries collectively cover all features.

Theorem 9 (Exact stability boundary for repeated learning)**.** *Suppose every node uses the same d, N, λ and a fixed sweep covers all m coordinates. Let T be the ordered product of its maps (20). Repeating the sweep with fresh samples is mean-square consistent for every fixed unit θ if and only if $c < 1$, equivalently $\rho(T) < 1$. After K sweeps,*

$$c < 1 \quad \Longrightarrow \quad \mathbb{E}_{\mathcal{T}} \left\| Y - \textstyle\sum_t \widehat{f}_t \right\|^2 \leq m c^K \max_j \theta_j^2. \tag{22}$$

For $\theta_j = 1/\sqrt{m}$, risk is exactly one when $c = 1$ and at least c^K when $c > 1$. The condition $2\lambda \geq m - d$ guarantees $c < 1$ for every admissible feature law and every $N \geq 1$.

Proof. If $c < 1$, every node preserves the coordinate bound $x \leq \mathbf{1}$ and sets each visited coordinate to at most c. Full coverage gives $T\mathbf{1} \leq c\mathbf{1}$. If $c > 1$, the inequalities reverse, giving $T\mathbf{1} \geq c\mathbf{1}$; at $c = 1$, every map fixes $\mathbf{1}$. Nonnegativity proves all three risk statements and the spectral criterion. For the distribution-free condition, put $S = G^\top G$ and $\tau = d^{-1}\mathbb{E}\,\mathrm{tr}[S^2(S + \lambda I)^{-2}] > 0$. The trace identities give $1 - c = \tau + [2\lambda - (m-d)]\beta$, which is positive under the stated condition. □

Two explicit design thresholds. For Gaussian OLS, $N > d + 1$ gives $\alpha = \gamma = 0$ and $\beta = 1/(N - d - 1)$, recovering the Gaussian ownership recurrence in Appendix K. Within this finite-moment regime,

$$c = \frac{m-d}{N-d-1} < 1 \quad \Longleftrightarrow \quad N > m + 1. \tag{23}$$

A local regression can be well determined yet make the collective learning process unstable: the required batch size depends on the *total* feature dimension. The Rademacher singleton case in Appendix Q gives the exact condition $N + 2\lambda > m - 1$. Thus regularization can stabilize small batches, although excessive regularization slows correction. The next result shows that the Gaussian OLS threshold also survives correlation between the features.

Theorem 10 (Correlated Gaussian learning)**.** *Let $X \sim N(0, \Sigma)$ with arbitrary $\Sigma \succ 0$, and let a fixed sequence of size-$d < m$ coordinate sets cover all m features. Use $N > d + 1$ fresh samples per residual OLS fit. In globally whitened coordinates, let P_J project onto the span of the observed features and $C_J = I - P_J$. The second-moment matrix H of the residual coefficients obeys*

$$H' = C_J H C_J + \frac{\mathrm{tr}(C_J H)}{N-d-1} P_J. \tag{24}$$

Repeated covering sweeps are mean-square consistent for every fixed unit-variance linear target if and only if $N > m + 1$. With $F = C_{J_n} \cdots C_{J_1}$ and $c = (m-d)/(N-d-1) < 1$, risk after K sweeps is at most $m\kappa^K$, where $\kappa = c + (1-c)\,\|F\|_{\mathrm{op}}^2 < 1$.

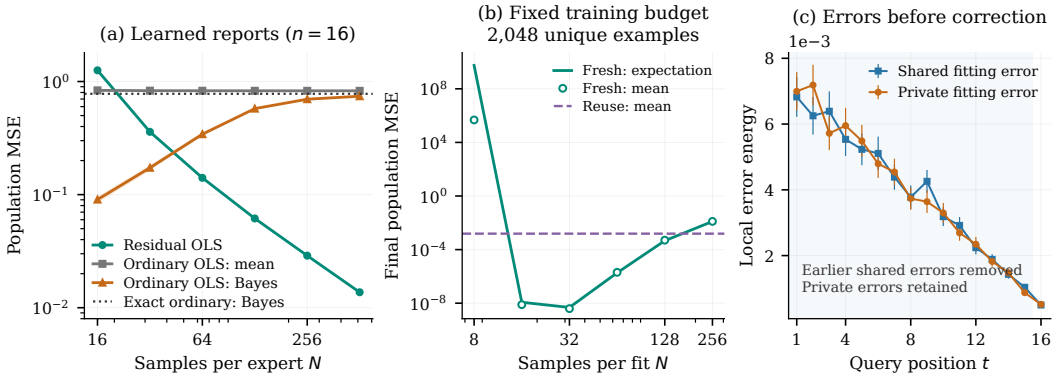

Figure 2: Information recovery depends on training precision. (a) Learned forecast comparison at $n = 16$. (b) The same 2,048 unique Gaussian examples are allocated to fresh sweeps with $8NK = 2{,}048$, or reused in one sweep; reuse processes each example at all eight nodes. Lines give exact fresh-data expectations and markers give sampled means. (c) Shared and private fitting errors at $n = 16, N = 128$. All empirical summaries use 1,000 runs; bands and error bars in (a,c) indicate two standard errors across independent training trajectories.

For the positive sweep map $\mathcal{T}$, each node maps I to $I+(c-1)P_J$. Telescoping and positivity compare $\mathcal{T}(I) - I$ with $(c-1)(I - FF^\top)$. Full coverage makes $\|F\|_{\mathrm{op}} < 1$, proving the threshold and rate bound; Appendix S gives the exact covariance calculation and necessity argument. Correlation changes the rate through the projection geometry, but not this sample-size threshold.

A sample–query consequence. For iid Gaussian features with $d < m$, $N = 2m - d + 1$ gives $c = 1/2$. Consequently $K = \lceil \log_2(m/\delta) \rceil$ sweeps suffice for expected risk δ, using nK report fits and $nNK = O(nm \log(m/\delta))$ fresh labeled examples. Increasing the query budget permits geometric improvement with a fixed batch size. The relative oracle with $\eta^2 = 1/2$ has risk 2^{-K} at the same full-sweep budget, but the error classes differ: a fitted regression can inject spurious coefficients when its predictable residual is zero. The common geometric rate describes two distinct guarantees, rather than a pathwise oracle property of the fitted responses.

Relation to adaptive filtering. Selective partial-update LMS has recent transient analyses (Siqueira et al., 2024); selective affine-projection methods have established mean-square analyses (Abadi & Palangi, 2010). Our residual fits use the latter structure. Gaussian sketch-and-solve gives the local covariance (Bartan & Pilanci, 2023); composing these maps yields the exact threshold over arbitrary covering subspaces. Appendix S provides the full covariance calculation and covering-product proof.

6 Experiments with trained local predictors

All experiments fit actual residual labels and evaluate population risk from the fitted coefficients. Each configuration uses 1,000 independently seeded trajectories on CPU. The simulator holds complete vectors to generate data and evaluate risk; each regression receives only its owned columns and residual labels. Exact expectations are computed separately from the proved moment maps. Finite averages can miss rare high-risk trajectories. Related stochastic-optimization analyses link multiplicative noise to heavy tails (Hodgkinson & Mahoney, 2021); here exact moment maps provide the comparison.

Information recovery and sample cost. In the Gaussian shared/private problem, expert i observes (W, U_i), where $W, U_1, \ldots, U_n$ are independent $N(0, 1/(n+1))$ and $Y = W + \sum_i U_i$. We train ordinary and residual OLS on the same stagewise examples for $n \in \{4, 8, 16, 32\}$ and $N \in \{16, 32, 64, 128, 256, 512\}$. Ordinary baselines include averaging and a Gaussian Bayes decoder given fitted coefficients and true covariance. That decoder can beat the exact-report information floor because estimation changes its report subspace. At $n = 32, N = 16$, residual risk is 6.474 versus 0.0367 for learned ordinary Bayes decoding; at $N = 512$, the risks are 0.03088 and 0.7672.

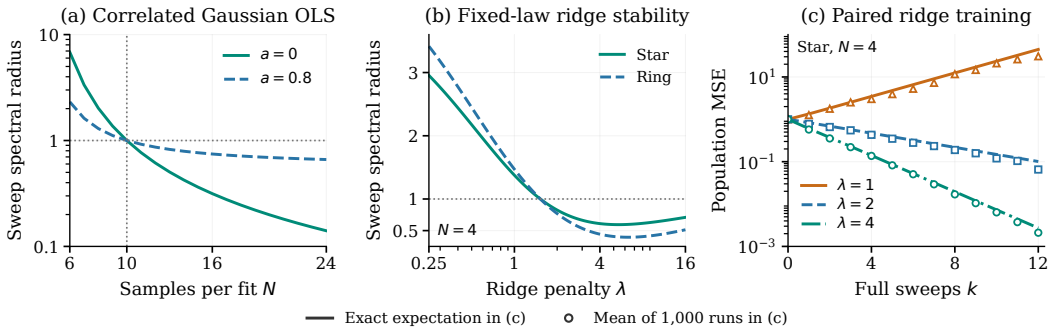

Figure 3: Correlation changes convergence speed; regularization can restore stability. (a) Exact Gaussian sweep radii with $\Sigma_{ij} = a^{|i-j|}$ at $a = 0, 0.8$, sharing the threshold $N = 10$. (b) Exact Rademacher sweep radii at fixed $N = 4$ cross one as the penalty changes. (c) Star-pattern fits use identical batches across penalties: lines are exact expected risks and markers are means of actual training runs. Appendix T reports all ten configurations and their full trajectory summaries.

The same protocol loses at the smaller sample size and wins at the larger one: its information advantage becomes useful only when the local fits are sufficiently accurate.

Sample allocation at fixed total data. With eight shared/private experts and $2{,}048$ unique labeled examples, we compare $N \in \{8, 16, 32, 64, 128, 256\}$ and $K = 256/N$ fresh sweeps, using the same examples across schedules. Exact expected risk is 0.01259 for one $N = 256$ sweep, but 4.83×10^{-9} for eight $N = 32$ sweeps. At $N = 8$, repeated fitting is unstable and expected risk reaches 5.60×10^{9}. This fixed-data comparison isolates the stability boundary: allocating more passes improves correction above the threshold and amplifies error below it. Reusing all $2{,}048$ examples in one sweep gives measured risk 0.00155, with eight times as many example–node visits during local fitting. Appendix U reports every schedule, sample cost, and trajectory summary.

Regularization with the data held fixed. We isolate regularization on $m = 9$ Rademacher features with $N = 4$ and 12 sweeps. An eight-node star owns $\{1, j\}$; a nine-node cycle owns adjacent pairs. For each pattern, every penalty uses the identical training batches. At $\lambda = 1$, the exact updated-row sum is $c = 1.178 > 1$; at $\lambda = 2$, it is $0.894 < 1$. For the star, final expected risks are 44.63 and 0.1009, with measured means 30.77 and 0.0654. Increasing λ to four lowers expectation to 0.00272, while increasing it to sixteen raises expectation to 0.01504. The paired batches isolate stabilization by regularization and show that stronger penalties do not always produce faster correction.

Expected risk and observed trajectories. At the critical Gaussian threshold, exact expected risk is one but final sample means are 0.237 and 0.134 for the star and cycle. The fixed-budget unstable configuration is even more concentrated: its measured mean is 4.75×10^{5} and median 2.22, far below the analytic expectation. These discrepancies show why finite-run averages alone cannot certify mean-square consistency. Appendices N and R give the full comparisons for shared-error correction, ordinary forecasts, and data reuse across the tested sample sizes.

7 Conclusion

Residual questions retain a provable advantage over ordinary forecasts under local approximation, including on genuinely overlapping information. The attainable accuracy depends on how approximation is measured: absolute precision creates a private-error floor, while projected-relative precision supports geometric correction. Exact learning dynamics identify when correction dominates error injected by unobserved features; regularization can restore convergence with small fresh batches. The Gaussian learning threshold survives feature correlation even when the expert projections no longer commute, while the rate depends on their geometry. With unequal precisions, optimal allocation and query order jointly control how errors move between shared and private information. Complementary target interactions and communication that enlarges the observed feature spaces are excluded.

AI USE STATEMENT

We used generative AI tools for language polishing and supplementary checks of theoretical arguments. The theoretical results and proofs were developed and verified by the authors, with AI serving only as an additional checking aid. All AI-assisted edits and suggestions incorporated into the paper were reviewed by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

To the best of our knowledge, this work does not raise specific ethical concerns requiring additional discussion under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Code provides finite-space identity checks, OLS and ridge experiments, matched-data and fixed-budget comparisons, and figure generation. Appendices I, N, R, T, and U document numerical procedures, data generation, training protocols, random seeds, evaluation, and uncertainty reporting. All experiments run on CPU. Analytic proofs are provided in the paper and appendices; the Lean 4 certificates described in Appendix I.4 verify selected projection-energy lemmas and their finite-recursion consequence. The reproducibility package, including experimental code and Lean 4 certificates, is available in an anonymous Git repository at https://anonymous.4open.science/r/forecast-correction-764A.

REFERENCES

Mohammad Shams Esfand Abadi and Hamid Palangi. Mean-square performance analysis of the family of selective partial update and selective regressor affine projection algorithms. *Signal Processing*, 90(1):197–206, 2010.

Itai Arieli, Yakov Babichenko, and Rann Smorodinsky. Robust forecast aggregation. *Proceedings of the National Academy of Sciences*, 115(52):E12135–E12143, 2018.

Burak Bartan and Mert Pilanci. Distributed sketching for randomized optimization: Exact characterization, concentration, and lower bounds. *IEEE Transactions on Information Theory*, 69(6): 3850–3879, 2023.

Heinz H Bauschke and Patrick L Combettes. Correction to: convex analysis and monotone operator theory in hilbert spaces. In *Convex analysis and monotone operator theory in Hilbert spaces*, pp. C1–C4. Springer, 2020.

Yi-Chun Chen, Manuel Mueller-Frank, and Mallesh M Pai. The wisdom of the crowd and higher-order beliefs, 2021. arXiv preprint arXiv:2102.02666.

Natalie Collina, Surbhi Goel, Varun Gupta, et al. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1532–1543, 2025.

Natalie Collina, Ira Globus-Harris, Surbhi Goel, et al. Collaborative prediction: Tractable information aggregation via agreement. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4712–4798. SIAM, 2026.

Dmitriy Drusvyatskiy and Adrian S. Lewis. Inexact alternating projections on nonconvex sets, 2018. arXiv preprint arXiv:1811.01298.

Rafael Frongillo, Mary Monroe, Eric Neyman, et al. Robust forecast aggregation via additional queries, 2025. arXiv preprint arXiv:2512.05271.

Yongkang Guo, Jason D. Hartline, Zhihuan Huang, et al. Algorithmic robust forecast aggregation. In *Proceedings of the 26th ACM Conference on Economics and Computation*, EC '25, pp. 1110–1129, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400719431. doi: 10.1145/3736252.3742674.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4262–4274. PMLR, 2021.

Michael Kearns, Aaron Roth, and Emily Ryu. Networked information aggregation via machine learning, 2025. arXiv preprint arXiv:2507.09683.

Tao Lin and Yiling Chen. Sample complexity of forecast aggregation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Barbara Mellers, Lyle Ungar, Jonathan Baron, et al. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, 25(5):1106–1115, 2014.

Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Eric Neyman. *Algorithmic bayesian epistemology*. Columbia University, 2024.

Eric Neyman and Tim Roughgarden. Are you smarter than a random expert? the robust aggregation of substitutable signals. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 990–1012, 2022.

Eric Neyman and Tim Roughgarden. From proper scoring rules to max-min optimal forecast aggregation. *Operations Research*, 71(6):2175–2195, 2023.

Michael Ostrovsky. Information aggregation in dynamic markets with strategic traders. *Econometrica*, 80(6):2595–2647, 2012.

Asa B Palley and Jack B Soll. Extracting the wisdom of crowds when information is shared. *Management Science*, 65(5):2291–2309, 2019.

Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

Ville A. Satopää, Shane T. Jensen, Robin Pemantle, et al. Partial information framework: Model-based aggregation of estimates from diverse information sources, 2015. arXiv preprint arXiv:1505.06472.

Newton N Siqueira, Leonardo C Resende, Fabio AA Andrade, et al. Transient analysis of a selective partial-update lms algorithm. *Applied Sciences*, 14(7):2775, 2024.

Jingrong Wang and Ben Liang. Gradient and projection free distributed online min-max resource optimization. In *Learning for Dynamics and Control Conference*, pp. 391–403. PMLR, 2022.

Xiaoqian Wang, Rob J Hyndman, Feng Li, et al. Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547, 2023.

Appendix Contents

A Probability model and projection details **15**

A.1 Commutation on the full function space 15

A.2 First-owner and last-owner assignments 15

A.3 Normalization applies to the clean graph 16

B Full minimax argument and boundary cases **16**

B.1 Achievability by uniform coverage 16

B.2 A private-information erasure lower bound 16

B.3 Consequences for resource allocation 17

B.4 Order of choices and randomization 17

B.5 Measurable decoders and hidden conditional variance 17

B.6 A single scalar private signal suffices 18

B.7 Why the hard response functions are admissible 18

B.8 Endpoints and deterministic selection 18

B.9 Balanced coverage 19

B.10 From the risk formula to accuracy requirements 19

C Repeated queries and the meaning of normalization **19**

D Sharp examples and repeated-agent queries **20**

D.1 Frozen and actual targets 20

D.2 The repeated-query erasure induction 20

D.3 Amplification as an infimum 20

E What survives beyond independent signals **21**

F General projection stability and repeated sweeps **22**

F.1 The synthesis operator 22

F.2 Why the general bound need not separate squared bias and error 22

F.3 Correlated binary information 22

F.4 A general repeated-sweep estimate 23

F.5 Two-step separation and approximate locality 23

G Approximate responses to quadratic scores **24**

H Quadratic scoring and approximation semantics **24**

H.1 Allocating score weight 25

I Numerical procedures and vector figures **26**

I.1 Finite-space conditional expectations 26

I.2 Independent linear programs 27
I.3 Noncommuting projections 27
I.4 Formalized energy lemmas 27

J Relative precision: full proofs **28**
J.1 The relative one-pass lower bound 28
J.2 First-exposure orthogonality and repeated sweeps 28
J.3 A fixed hard prior for repeated full-information queries 28
J.4 Amplification under relative precision 29
J.5 Why full-target relative error is different 29

K From regression samples to aggregate risk **29**
K.1 Conditional regression and the inverse-Gram expectation 29
K.2 Arbitrary Gaussian ownership 30
K.3 The exact recurrence and finite-sample comparison 30
K.4 The information retained by ordinary reports 31
K.5 Training randomness and invisible channel noise 31

L A sharp joint bound for approximate locality **31**

M Quadratic improvement payments and market information **31**

N Learned-predictor experiments **32**

O Exact relative precision at every report budget **33**
O.1 A partial sweep followed by full sweeps 33
O.2 One symmetric private prior for every graph 36
O.3 Balancing the integer visit counts 37
O.4 Exact responses, complete erasure, and full information 37

P One-pass learning formulas and local corruption **37**

Q Residual ridge regression and stability **39**
Q.1 Statistical model and ridge coefficients 39
Q.2 Proof of the exact ridge energy recurrence 40
Q.3 Proof of the coverage-independent stability threshold 41
Q.4 A distribution-free choice of ridge penalty 41
Q.5 Gaussian least squares and Rademacher thresholds 42
Q.6 Connection with partial-update affine projection 42

R Repeated learning: configurations and trajectory diagnostics **42**
R.1 Expected risks and training costs 43

R.2 Concentration across fitted trajectories 43
R.3 Reproduction 44

S Correlated Gaussian residual learning **44**
S.1 The exact second-moment update 44
S.2 The covering-sweep threshold 45
S.3 A sufficient training cost 46

T A paired ridge sweep under a fixed feature law **47**

U Allocating a fixed label budget **48**

V Heterogeneous relative precision and query order **49**
V.1 A private-information lower bound for every graph 49
V.2 An attaining order on overlapping information 50
V.3 Exact allocation by piecewise-linear interpolation 51
V.4 Why equal visit counts do not determine residual risk 52

A PROBABILITY MODEL AND PROJECTION DETAILS

This appendix supplies the measure-theoretic and algebraic details behind the main results. The primitive family is finite, but each primitive can have an arbitrary real distribution with finite second moment. Degenerate distributions are allowed. The underlying probability space can be taken to be the product of the primitive spaces; this ensures that locally measurable errors are functions of precisely the information assigned to the corresponding expert. All lower-bound errors are functions of these locally observed primitives, without auxiliary random inputs.

A.1 COMMUTATION ON THE FULL FUNCTION SPACE

For a coordinate subset $J \subseteq [m]$, write P_J for conditional expectation onto $\sigma(X_j : j \in J)$. We show

$$P_J P_L = P_L P_J = P_{J\cap L} \quad \text{on } L^2. \tag{25}$$

First consider a bounded product function $h(X) = \prod_{j=1}^{m} h_j(X_j)$. Independence gives

$$P_J h = \prod_{j\in J} h_j(X_j) \prod_{j\notin J} \mathbb{E}h_j(X_j).$$

Applying P_L integrates the coordinates in $J \setminus L$ and therefore leaves precisely the coordinates in $J \cap L$. The same expression results in either order. Finite linear combinations of bounded product functions are dense in the L^2 space of the product sigma-field: indicator functions of measurable rectangles generate a dense subspace, by the monotone class theorem and approximation of square-integrable functions by bounded simple functions. Continuity of the conditional-expectation contractions extends this identity from the dense subspace to every square-integrable function.

Expert projections correspond to $J_i = \{\ell : i \in T_\ell\}$, so (25) applies to arbitrary square-integrable local errors, including nonlinear functions of multiple observed coordinates. For a centered primitive X_ℓ, the simpler identity

$$P_i X_\ell = \begin{cases} X_\ell, & i \in T_\ell, \\ 0, & i \notin T_\ell \end{cases} \tag{26}$$

also follows from independence. Applying all complement projections to (1) yields (9) because the surviving independent centered coordinates have additive variances.

A.2 FIRST-OWNER AND LAST-OWNER ASSIGNMENTS

Let the selected list be $i_1, \dots, i_k$. In the error-free recursion, induction gives

$$Y - A_{t-1}^{(0)} = C_{t-1} \cdots C_1 Y, \tag{27}$$

$$Q_t = P_t C_{t-1} \cdots C_1 Y = \sum_{\ell:\, i_t \in T_\ell,\, T_\ell \cap \{i_1,\dots,i_{t-1}\} = \varnothing} X_\ell. \tag{28}$$

Thus a coordinate is assigned to its first selected owner. In contrast, the surviving-error operator $D_t = P_t \prod_{u>t}(I - P_u)$ retains the directions visible at node t and invisible to every later node. If an error itself is a linear combination of primitives, this assigns each of its coordinates to its last selected owner. The projection identity also applies to arbitrary nonlinear local errors.

For a direct proof of the identity partition, telescope in reverse order:

$$I - C_k \cdots C_1 = (I - C_k) + C_k(I - C_{k-1}) + \cdots + C_k \cdots C_2(I - C_1) \tag{29}$$

$$= \sum_{t=1}^{k} C_k \cdots C_{t+1} P_t. \tag{30}$$

The telescoping identity holds without commutation. To obtain orthogonal projections with mutually orthogonal ranges, use commutation to put the factors $C_u P_u$ next to each other in $D_t D_u$ for $t < u$. Their product vanishes. Likewise $F D_t = 0$ because F contains C_t and D_t contains P_t. Pythagoras therefore gives the exact error decomposition in Theorem 2.

A.3 Normalization applies to the clean graph

Every clean residual query has norm at most one even for noncommuting projections: it is a product of orthogonal projections and their complements applied to unit-norm Y. This statement does not bound the actual target after errors have been introduced.

A concrete example demonstrates the difference. Let U, V be independent, with variances $1-\varepsilon^2$ and ε^2, and let $Y = U + V$. Expert 1 observes V and expert 2 observes both coordinates. Choose $e_1 = \eta V/\varepsilon$, for $\varepsilon > 0$. Then the actual second query target is

$$P_2(Y - R_1) = U - \eta V/\varepsilon, \qquad \|P_2(Y - R_1)\|^2 = 1 - \varepsilon^2 + \eta^2.$$

It exceeds one when $\eta > \varepsilon$, although the clean queries V, U both have norm at most one. Theorems 3 and 14 use the normalization specified in (3), not a bound on this actual quantity.

B Full minimax argument and boundary cases

B.1 Achievability by uniform coverage

For a signal with $q \geq 1$ owners, a uniformly random B-subset misses it with probability

$$\frac{\binom{n-q}{B}}{\binom{n}{B}} \leq 1 - \frac{B}{n}, \tag{31}$$

with a zero numerator when $n - q < B$. Averaging (9) and using $\sum_\ell \mathrm{Var}(X_\ell) = 1$ gives $\mathbb{E}_Z \|FY\|^2 \leq 1 - B/n$. Each expert is selected with probability B/n, so (8) yields

$$\mathbb{E}_{Z,\mathcal{D}}(A_B - Y)^2 \leq 1 - \frac{B}{n} + \frac{B}{n}\mathcal{E}(\boldsymbol{\eta}). \tag{32}$$

Every clean residual query is a product of contractions applied to Y, and hence satisfies (3). Comparing (32) with the zero predictor proves the upper bound.

B.2 A private-information erasure lower bound

Fix any protocol, including its distribution over selected lists. Let p_i be the probability it queries expert i. Since $\sum_i p_i \leq B$, some i_* has $p_{i_*} \leq B/n$. Fix this prior before drawing the seed.

Suppose first $\mathcal{E}(\boldsymbol{\eta}) < 1$. Give each expert independent private Gaussian coordinates U_i, V_i with

$$\mathrm{Var}(V_i) = \eta_i^2, \qquad \mathrm{Var}(U_i) = \begin{cases} 1 - \mathcal{E}(\boldsymbol{\eta}), & i = i_*, \\ 0, & i \neq i_*. \end{cases} \qquad Y = \sum_i (U_i + V_i). \tag{33}$$

All coordinates are centered and mutually independent. Each expert sees both its coordinates, so these are legitimate signals, not extra channel randomness. At a node querying i_t, let

$$R_t = a_t U_{i_t}, \qquad e_t = -a_t V_{i_t}. \tag{34}$$

Earlier reports concern distinct private experts and have conditional mean zero given $\mathcal{F}_{i_t}$. Thus (34) satisfies the actual report equation (2), including arbitrary coefficients b_{ts}. The direct-gain bound in Lemma 1 gives $\|e_t\| = |a_t|\eta_{i_t} \leq \eta_{i_t}$, so every local defect obeys its budget.

The seed and transcript reveal no V_i. When i_* is unqueried, they also reveal no U_{i_*}. By independence, for every measurable decoder,

$$\begin{aligned} \mathbb{E}[(g_Z(R) - Y)^2] &\geq \mathcal{E}(\boldsymbol{\eta}) + (1 - \mathcal{E}(\boldsymbol{\eta}))\Pr(i_* \text{ unqueried}) \\ &= 1 - p_{i_*}(1 - \mathcal{E}(\boldsymbol{\eta})) \geq 1 - \frac{B}{n}(1 - \mathcal{E}(\boldsymbol{\eta})). \end{aligned} \tag{35}$$

This is a conditional-variance bound, rather than a restriction to linear estimation. The prior depends on the protocol's marginal selection probabilities, but not on its realized seed. Appendix B.6 realizes the same hidden-variance construction with one scalar, finite-support private signal per expert.

If $\mathcal{E}(\boldsymbol{\eta}) \geq 1$, instead give expert i a private Gaussian V_i with variance $\eta_i^2/\mathcal{E}(\boldsymbol{\eta})$ and set $U_i = 0$. Choosing $R_t = 0$ and $e_t = -a_t V_{i_t}$ erases every report within its tolerance. Since $\mathrm{Var}(Y) = 1$, every decoder has error at least one. This completes the proof of Theorem 3.

B.3 Consequences for resource allocation

Uniform precision. If all experts have tolerance η, then $\mathcal{E}(\boldsymbol{\eta}) = n\eta^2$. With all experts queried, the risk is $\min\{1, n\eta^2\}$. A nontrivial guarantee requires $\eta < n^{-1/2}$, and vanishing error along growing populations requires $n\eta^2 \to 0$. This requirement follows from a hard private-information instance; it holds even for arbitrary nonlinear measurable decoders.

Target accuracy. For desired mean squared error $\delta < 1$, feasibility requires $\mathcal{E}(\boldsymbol{\eta}) \le \delta$. When this holds,

$$B \ge \left\lceil \frac{n(1-\delta)}{1-\mathcal{E}(\boldsymbol{\eta})} \right\rceil \tag{36}$$

is necessary and sufficient. Each additional query reduces the minimax risk by $(1 - \mathcal{E}(\boldsymbol{\eta}))/n$ below the threshold $\mathcal{E}(\boldsymbol{\eta}) = 1$. Thus additional queries become less valuable as total reporting tolerance approaches $\mathcal{E}(\boldsymbol{\eta}) = 1$, where even full coverage cannot improve the worst-case guarantee.

Corollary 11 (Balanced coverage is necessary)**.** *Suppose $\mathcal{E}(\boldsymbol{\eta}) < 1$. Every minimax-optimal one-pass protocol selects exactly B experts almost surely and has marginal selection probabilities $p_i = B/n$ for all i. Conversely, every distribution over B-subsets with these marginals attains (11) when used with residual queries. If $p_{\min} = \min_i p_i$, any protocol with those marginals has excess worst-case risk at least*

$$\left(\frac{B}{n} - p_{\min}\right)(1 - \mathcal{E}(\boldsymbol{\eta})). \tag{37}$$

Appendix B.9 proves the coverage characterization directly from the erasure lower bound.

Thus unequal precision does not identify expendable experts. An expert with a large tolerance may own indispensable information, and the adversary can put the remaining visible variance at whichever expert is least likely to be queried. Preferential selection needs further information about ownership or signal value. Balanced marginals are forced by the unknown-information minimax problem, although the full distribution over subsets need not be uniform.

B.4 Order of choices and randomization

The minimax order in (10) first lets the aggregator select a distribution over graphs, coefficients, and decoders, after which the adversary chooses a single prior. A public seed is then drawn independently of the signals. The allowed reports at each seed are local functions obeying the expert-specific budgets, and the risk averages over both sources of randomness. Both expectations use this same prior, fixed before the independent public randomization occurs.

The normalization assumption holds at every seed over the entire prior family. Lemma 1 uses different test priors to infer properties of the coefficients; those test priors are not the prior used in the lower-bound risk evaluation. Once the coefficient bounds have been inferred, (33) defines one fixed lower-bound prior. This distinction justifies the use of a particular node's private-coordinate test in the normalization lemma without violating the minimax order.

B.5 Measurable decoders and hidden conditional variance

For any sigma-field $\mathcal{G}$ and any $\mathcal{G}$-measurable square-integrable estimator h,

$$\mathbb{E}[(h - Y)^2] = \mathbb{E}[(h - \mathbb{E}[Y \mid \mathcal{G}])^2] + \mathbb{E}[\mathrm{Var}(Y \mid \mathcal{G})]. \tag{38}$$

If the left side is infinite, the lower bound by the second term still holds. In the one-pass hard construction, take $\mathcal{G} = \sigma(Z, R_1, \ldots, R_k)$. For every fixed seed, the transcript consists of deterministic scalar multiples of the queried U_i coordinates. All hidden V_i and unqueried U_i remain independent of the transcript conditional on every realization of the public seed.

When $\mathcal{E}(\boldsymbol{\eta}) < 1$, the independent hidden sum $\sum_i V_i$ always contributes variance $\mathcal{E}(\boldsymbol{\eta})$. On the event that i_* is unqueried, U_{i_*} contributes an additional $1 - \mathcal{E}(\boldsymbol{\eta})$. Conditioning on the seed makes that event deterministic, so the lower bound is

$$\mathbb{E}[\mathrm{Var}(Y \mid \mathcal{G})] \ge \mathcal{E}(\boldsymbol{\eta}) + (1 - \mathcal{E}(\boldsymbol{\eta}))(1 - p_{i_*}).$$

Even if a queried coefficient a_t vanishes, the bound remains true because additional visible variance may also be hidden. No Gaussian conditioning formula is necessary; independence and second moments suffice. Gaussian coordinates are a convenient admissible choice.

For homogeneous tolerances, there is also a fully symmetric hard prior. Let $v = \min\{\eta^2, 1/n\}$ and give each expert independent U_i, V_i of variances $1/n - v, v$. Under (34), any list of k experts leaves conditional variance at least

$$nv + (n-k)(1/n - v) = 1 - k/n + kv.$$

For $k \le B$ this is at least $1 - B/n + Bv$, yielding (11). Unlike the heterogeneous construction, this symmetric prior does not depend on the protocol's marginal selection probabilities.

B.6 A SINGLE SCALAR PRIVATE SIGNAL SUFFICES

The lower bound does not require each expert to receive a separately labeled pair of primitives. It also holds with one scalar independent signal per private ownership class. Suppose $\mathcal{E}(\boldsymbol{\eta}) < 1$ and choose the same least-selected expert i_*. Let V_i take values $\pm\eta_i$ with equal probability, independently across experts. Set $U_i = 0$ for $i \neq i_*$. Independently choose

$$\Pr(U_{i_*} = M) = \Pr(U_{i_*} = -M) = \frac{1 - \mathcal{E}(\boldsymbol{\eta})}{2M^2}, \qquad \Pr(U_{i_*} = 0) = 1 - \frac{1 - \mathcal{E}(\boldsymbol{\eta})}{M^2},$$

where $M > 2\eta_{i_*}$ and $M^2 \ge 1 - \mathcal{E}(\boldsymbol{\eta})$. Define the single observed signal of expert i to be $X_i = U_i + V_i$. These signals are centered, independent across experts, and satisfy $\sum_i \mathrm{Var}(X_i) = 1$.

At expert i_* the possible values lie in the three disjoint intervals centered at $-M, 0, M$ with radius η_{i_*}. Consequently U_{i_*} is determined by which interval contains X_{i_*}, and $V_{i_*} = X_{i_*} - U_{i_*}$ is also locally measurable. The other experts have $X_i = V_i$. Thus the erasure defects $-a_t V_{i_t}$ and reports based on U_{i_t} remain local functions of a single scalar observation. All independence and conditional-variance arguments are unchanged. For $\mathcal{E}(\boldsymbol{\eta}) \ge 1$, use $X_i = V_i = \pm\eta_i / \sqrt{\mathcal{E}(\boldsymbol{\eta})}$ and erase all reports. This construction has finite support; no Gaussian decomposition is assumed recoverable from its sum.

B.7 WHY THE HARD RESPONSE FUNCTIONS ARE ADMISSIBLE

In the one-pass construction, all previously selected experts are different from i_t. Their reports have the form $a_s U_{i_s}$ and are centered and independent of $\mathcal{F}_{i_t}$. Hence

$$P_{i_t}\left(a_t Y + \sum_{s<t} b_{ts} R_s\right) = a_t (U_{i_t} + V_{i_t}), \tag{39}$$

$$R_t - P_{i_t}\left(a_t Y + \sum_{s<t} b_{ts} R_s\right) = -a_t V_{i_t}. \tag{40}$$

This defect is measurable with respect to the reporting expert's information, has norm at most its tolerance, and may be correlated with the target. The lower bound therefore uses a permitted local misreport, rather than an adversarial channel that accesses another expert's signals.

When $\mathcal{E}(\boldsymbol{\eta}) \ge 1$, choosing $\mathrm{Var}(V_i) = \eta_i^2 / \mathcal{E}(\boldsymbol{\eta})$ gives a unit-variance target and $\|a_t V_{i_t}\| \le \eta_{i_t}$. The all-zero report profile is then admissible. Since neither the independent public seed nor the zero reports reveal Y, its conditional variance remains one in (38).

B.8 ENDPOINTS AND DETERMINISTIC SELECTION

The theorem includes $B = 0$, in which case risk is one. At $B = n$ the risk is $\min\{1, \mathcal{E}(\boldsymbol{\eta})\}$. At $\boldsymbol{\eta} = 0$ the risk is $1 - B/n$, recovering the noiseless budget law. At $\mathcal{E}(\boldsymbol{\eta}) = 1$ the residual protocol and the zero predictor both have the optimal risk bound one.

Randomization is necessary for a nontrivial worst-case guarantee when $B < n$ in the one-pass class. A deterministic list omits some expert. Put all target variance in a private coordinate of that expert and choose zero defects. Every clean report is zero, so any decoder has mean squared error at least one

under a centered, unit-variance prior. Randomization protects against this concentration by giving every expert the same inclusion probability under the randomized query design.

Uniform sampling is one minimax design, not necessarily the unique design or unique optimal decoder. The optimal value and an attaining design are established, while other distributions with balanced marginals attain the same worst-case risk bound.

B.9 Balanced coverage

Proof of Corollary 11. Subtract (11) from the lower bound $1 - p_{\min}(1 - \mathcal{E}(\boldsymbol{\eta}))$. Equality at the optimum requires $p_{\min} \geq B/n$. Since $\sum_i p_i \leq B$, all marginals equal B/n and the size budget is saturated almost surely. Conversely, fix any owner of each signal. A design with these marginals misses that signal with probability at most $1 - B/n$, and its expected selected tolerance is $(B/n)\mathcal{E}(\boldsymbol{\eta})$. The surviving-error identity (8) then gives the required upper bound for every such balanced design. □

B.10 From the risk formula to accuracy requirements

When $\mathcal{E}(\boldsymbol{\eta}) < 1$, the condition $\mathcal{R}_B^* \leq \delta$ is equivalent to $B \geq n(1-\delta)/(1-\mathcal{E}(\boldsymbol{\eta}))$. Since B is an integer not exceeding n, this is feasible precisely when $\mathcal{E}(\boldsymbol{\eta}) \leq \delta$ for $\delta < 1$. The rounding in (36) follows. If $\eta_i = \eta$, full-query risk vanishes along a population sequence exactly when $n\eta^2 \to 0$.

These conditions concern worst-case information structures. On a fixed instance with substantial overlap, (8) can be much smaller than the minimax value because most error components are corrected. The sharp common-information example illustrates this distinction.

C Repeated queries and the meaning of normalization

The distinct-expert condition in Section 2 makes clean-query normalization control each direct target coefficient. Repeated agents break that implication. This matters because errors are measured in the units of each scalar answer.

Proposition 12 (Clean normalization permits unlimited amplification)**.** *For a fully informed expert and finite tolerance η, two queries with unit-norm clean reports have minimax infimum zero.*

Proof. For any $M > 0$, use

$$R_1 = Y + e_1, \qquad R_2 = MY - (M-1)R_1 + e_2. \tag{41}$$

Both clean reports equal Y and have norm one. Yet

$$\widehat{Y} = \frac{R_2 + (M-1)R_1}{M} = Y + \frac{e_2}{M} \tag{42}$$

has worst-case error at most η^2/M^2. Let $M \to \infty$. □

The first error is canceled algebraically, while the second is divided by an unrestricted gain. This example does not violate the stated normalization; it shows that normalization of clean answers alone is too weak for arbitrary repeated-agent graphs. It also explains why an error term cannot be transferred between different query conventions without rechecking scale.

Definition 13 (Bounded direct gain)**.** *A repeated-agent graph has bounded direct gain if every query in (2) satisfies $|a_t| \leq 1$, conditional on every seed. Each visit to expert i has a local defect of norm at most η_i. The graph and coefficients are fixed before signals are realized.*

No restriction is placed on the number of visits to an expert or on the finite coefficients b_{ts}. Clean-query normalization may additionally be imposed; the attaining protocol satisfies it. Write $\mathcal{R}^*_{L,\mathrm{gain}}$ for the infimum over these graphs with at most L total report nodes, where L may exceed n.

Theorem 14 (Repeated-agent graphs)**.** *For bounded-direct-gain graphs and arbitrary measurable decoders,*

$$\mathcal{R}^*_{L,\mathrm{gain}}(\boldsymbol{\eta}) = 1 - \frac{\min\{L,n\}}{n}(1 - \mathcal{E}(\boldsymbol{\eta}))_+. \tag{43}$$

Proof. For the lower bound, let p_i be the probability that expert i is ever queried. Then $\sum_i p_i \le \min\{L, n\}$. Use the private prior (33) at an expert with minimal p_i. Inject $e_t = -a_t V_{i_t}$ at every visit. Inductively, every report has the form $R_t = c_t U_{i_t}$, where

$$c_t = a_t + \sum_{s<t:\, i_s = i_t} b_{ts} c_s. \tag{44}$$

Reports from other experts have zero conditional mean. Hence all V_i remain hidden regardless of repeated visits, and the argument (35) applies. For $\mathcal{E}(\boldsymbol{\eta}) \ge 1$ all reports can again be erased. The upper bound uses Theorem 3 with $\min\{L, n\}$ distinct experts. □

Repeated visits cannot reveal the private component hidden by the lower-bound defects, so averaging does not improve (43). Gains from repetition require favorable instance structure or further restrictions on the errors, such as fresh independent errors at every visit.

D Sharp examples and repeated-agent queries

D.1 Frozen and actual targets

For k experts all observing a centered unit-variance Y, the clean residual reports are $Q_1 = Y$ and $Q_t = 0$ for $t > 1$. A frozen-target report is $Q_t + e_t$. With the sum decoder and $e_t = \eta Y$, all defects are local and have norm η, but the aggregate error is $k\eta Y$, with squared norm $k^2\eta^2$.

For actual residual targets, $P_t Y = Y$ and $P_t e_s = e_s$ on this shared-information instance. The first aggregate is $Y + e_1$. At every later step,

$$R_t = Y - (Y + e_{t-1}) + e_t = -e_{t-1} + e_t, \qquad A_t = Y + e_t.$$

Its final error is exactly e_k, with squared norm at most η^2. The example compares the same residual construction under two error semantics. It is not a lower bound for all frozen-target protocols: for this particular instance, a protocol that queries only one expert would also have error at most η^2.

For a different sharp example, let expert i privately observe X_i of variance $1/n$, and let $e_i = \eta\sqrt{n} X_i$. With all experts selected, the actual residual reports are $X_i + e_i$, since all other private coordinates have conditional mean zero. The aggregate error has squared norm $n\eta^2$. This attains the actual residual protocol's full-query worst-case bound. For $n\eta^2 > 1$, the zero predictor is minimax optimal.

D.2 The repeated-query erasure induction

For completeness, consider an arbitrary seed and an arbitrary repeated-agent graph with coefficients $|a_t| \le 1$. Its hard prior is the private U_i, V_i construction in Section 3.1. Define c_t by (44) and $R_t = c_t U_{i_t}$. Because the graph is finite and triangular, every c_t is a finite real number. At step t, independence gives

$$P_{i_t} R_s = \begin{cases} c_s U_{i_t}, & i_s = i_t, \\ 0, & i_s \ne i_t. \end{cases}$$

Consequently the actual target conditional mean equals

$$a_t(U_{i_t} + V_{i_t}) + \sum_{s<t:\, i_s = i_t} b_{ts} c_s U_{i_t} = c_t U_{i_t} + a_t V_{i_t}.$$

The defect $-a_t V_{i_t}$ produces exactly the stipulated report and satisfies the local budget. Thus repeated queries can change the coefficients of visible coordinates but cannot expose the hidden V_i. The joint transcript remains independent of all V_i. The conditional-variance lower bound therefore applies unchanged and proves Theorem 14 for arbitrary repeated visits.

D.3 Amplification as an infimum

Proposition 12 asserts an infimum of zero. For a fixed finite M and positive η, its worst-case error is exactly η^2/M^2: the upper bound follows from the algebraic cancellation, and a local defect $e_2 = \eta Y$ attains it. Taking arbitrarily large M gives the zero infimum without requiring an infinite coefficient in any particular protocol. If $\eta = 0$, exact recovery is already possible with one query.

This counterexample uses two visits to the same expert and an unbounded direct coefficient on Y. It does not contradict Lemma 1, whose proof requires distinct experts, or Theorem 14, which explicitly bounds the direct target gain and therefore excludes this amplification.

E What survives beyond independent signals

Commutation gives the exact separation in (8), but error energy remains controlled without it. This distinction separates stability of the correction mechanism from recoverability of the target. Hilbert-space projection methods have standard modern treatments (Bauschke & Combettes, 2020); inexact projection methods likewise study how local computational errors affect iterative procedures (Drusvyatskiy & Lewis, 2018).

Theorem 15 (Propagation under arbitrary information projections)**.** *Let $P_1, \dots, P_k$ be arbitrary orthogonal projections on a real Hilbert space. For local errors $P_t e_t = e_t$, define F as in (6). Then*

$$\|(Y - A_k) - FY\|^2 \le \sum_{t=1}^{k} \|e_t\|^2, \qquad \|Y - A_k\| \le \|FY\| + \sqrt{\sum_t \|e_t\|^2}. \tag{45}$$

More precisely, the operator $K(e_1, \dots, e_k) = \sum_t D_t e_t$ satisfies

$$KK^* = I - FF^*. \tag{46}$$

Proof. Set $z_0 = 0$ and $z_t = C_t z_{t-1} + e_t$. Since $C_t z_{t-1} \perp e_t$,

$$\|z_t\|^2 = \|C_t z_{t-1}\|^2 + \|e_t\|^2 \le \|z_{t-1}\|^2 + \|e_t\|^2.$$

Induction and the residual expansion prove (45). For (46), let $S_t = C_k \cdots C_{t+1}$. Since $D_t = S_t P_t$, summing $D_t D_t^* = S_t S_t^* - S_t C_t S_t^*$ telescopes to $I - FF^*$. □

Proposition 16 (When squared bias and error separate)**.** *Two orthogonal projections P, Q commute if and only if, for every Y and every $e \in \mathrm{ran}(P)$,*

$$\|(I - Q)(I - P)Y - (I - Q)e\|^2 = \|(I - Q)(I - P)Y\|^2 + \|(I - Q)e\|^2. \tag{47}$$

Expanding the cross term reduces the identity to $(I - P)QP = 0$; taking adjoints then gives $PQ = QP$. Appendix F.5 gives the complete argument in both directions.

Commutation is exactly the condition for universal two-step squared bias–error separation. Locality controls propagation even without commutation; commuting information additionally makes clean bias orthogonal to surviving defects. For general information structures, the interaction between clean bias and propagated noise must therefore be retained.

Correlation affects recovery. For balanced binary U, V with correlation $\rho \in (0, 1)$, let experts 1 and 2 see U and V, respectively, and take $Y = V$. One noiseless sweep leaves $\rho^2 V - \rho U$, of squared norm $\rho^2(1 - \rho^2)$. Despite local observability, correlation leaves a clean bias that repeated sweeps reduce geometrically, with explicit rates derived in Appendix F.

Nonlocal channel corruption. The locality condition is essential. If every expert observes Y but a channel adds $e_t = \eta W$ for an independent, unobserved unit-variance W, then $P_t e_t = 0$. Residual correction cannot predict these errors; the final error is $k\eta W$. The quadratic growth $k^2\eta^2$ reappears. Theorems 2 and 15 require the error to be measurable from the reporting expert's information; the independent channel variable violates that condition.

Approximate locality. The change from local response defects to unrestricted channel errors can be quantified. Write $e_t = \ell_t + u_t$, where $\ell_t = P_t e_t$ and $u_t = (I - P_t)e_t$, with $\|\ell_t\| \le \eta_{i_t}$ and $\|u_t\| \le \zeta_t$. Superposition, Theorem 15, and contraction of the suffix operators give

$$\|(Y - A_k) - FY\| \le \sqrt{\sum_t \eta_{i_t}^2} + \sum_t \zeta_t. \tag{48}$$

For commuting projections the stronger bound is

$$\|Y - A_k\| \le \sqrt{\|FY\|^2 + \sum_t \eta_{i_t}^2} + \sum_t \zeta_t. \tag{49}$$

Private local errors and shared nonlocal errors attain the square-root and linear orders separately. In a full homogeneous sweep, the local and nonlocal bounds coincide at tolerance $\zeta = \eta/\sqrt{n}$.

F General projection stability and repeated sweeps

F.1 The synthesis operator

Let H be a real Hilbert space and let $P_1, \dots, P_k$ be orthogonal projections, with no commutation assumption. On H^k define

$$K(x_1, \dots, x_k) = \sum_{t=1}^{k} S_t P_t x_t, \qquad S_t = C_k \cdots C_{t+1}.$$

Its adjoint is $K^* x = (P_t S_t^* x)_{t=1}^k$. Since $P_t^2 = P_t = P_t^*$,

$$KK^* = \sum_t S_t P_t S_t^* \tag{50}$$

$$= \sum_t \big(S_t S_t^* - S_t C_t S_t^*\big) \tag{51}$$

$$= I - (C_k \cdots C_1)(C_1 \cdots C_k) = I - FF^*. \tag{52}$$

The penultimate equality uses $C_t^2 = C_t$: $S_t C_t S_t^* = (S_t C_t)(S_t C_t)^*$. It is exactly the preceding suffix term in the telescoping sum. Therefore $KK^* \preceq I$, and $\|K\|_{\mathrm{op}} \le 1$. Applying this contraction to the local vector $(e_1, \dots, e_k)$ gives (45), because $P_t e_t = e_t$ at every node.

The same proof can be stated on the direct sum of the closed ranges $\mathrm{ran}(P_t)$. Restricting the domain to that direct sum does not change KK^*, because the blocks already contain P_t. This formulation makes the local information restriction part of the operator domain.

When the projections commute, F is self-adjoint and idempotent. The operator identity then reduces to $KK^* = I - F$. The stronger orthogonal decomposition in Theorem 2 additionally resolves the individual blocks and eliminates the cross term with FY.

F.2 Why the general bound need not separate squared bias and error

Without commutation, FY and the propagated local errors may have a nonzero inner product. Thus the valid universal statement is (45), rather than the stronger claim $\|Y - A_k\|^2 \le \|FY\|^2 + \sum_t \|e_t\|^2$.

For an explicit counterexample on $\mathbb{R}^2$, let P_1 project onto the first coordinate axis, let P_2 project onto $(1,1)/\sqrt{2}$, and let $Y = (0,1)$. Set $e_1 = (1,0)$ and $e_2 = 0$. Then $FY = (-1/2, 1/2)$ and $C_2 e_1 = (1/2, -1/2) = -FY$. Therefore

$$\|FY - C_2 e_1\|^2 = 2 > \|FY\|^2 + \|e_1\|^2 = \frac{3}{2}.$$

The Hilbert-space counterexample concerns bias–error separation; the propagated error still satisfies $\|C_2 e_1\|^2 = 1/2 \le 1$ as required by the local-energy bound.

F.3 Correlated binary information

Let $(U, V) \in \{-1, 1\}^2$ have distribution

$$\Pr(U = u, V = v) = \frac{1 + \rho uv}{4}, \qquad 0 < \rho < 1.$$

Then both marginals are balanced, $\mathbb{E}UV = \rho$, and

$$\mathbb{E}[V \mid U] = \rho U, \qquad \mathbb{E}[U \mid V] = \rho V.$$

For target $Y = V$, a noiseless sweep first reports ρU, then $(1-\rho^2)V$. The remaining residual is

$$FV = \rho^2 V - \rho U, \qquad \|FV\|^2 = \rho^2(1-\rho^2).$$

The span $H_0 = \mathrm{span}\{U, V\}$ is invariant under both projections. In the orthonormal basis $U, (V - \rho U)/\sqrt{1-\rho^2}$, the projections have matrices

$$P_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad P_2 = \begin{pmatrix} \rho^2 & \rho\sqrt{1-\rho^2} \\ \rho\sqrt{1-\rho^2} & 1-\rho^2 \end{pmatrix}.$$

The product $F = (I - P_2)(I - P_1)$ obeys $F^2 = \rho^2 F$ and $\|F\|_{\mathrm{op}} = \rho$ on H_0. Consequently

$$\left\|F^K V\right\|^2 = \rho^{4K-2}(1-\rho^2), \qquad \sup_{Y \in H_0 : \|Y\|=1} \left\|F^K Y\right\|^2 = \rho^{4K-2}. \tag{53}$$

The second expression is the diagnostic plotted in Figure 5. The maximizing unit target is taken over the two-dimensional linear span; it is not fixed to V.

F.4 A GENERAL REPEATED-SWEEP ESTIMATE

Suppose a sweep has complement product F and each sweep introduces local defects with total squared norm at most E. Let H_0 be a closed subspace invariant under every projection, containing the target and every injected defect. Assume $\|F|_{H_0}\|_{\mathrm{op}} \le \rho < 1$. Let $r^{(K)}$ denote the residual after K sweeps. Theorem 15 bounds each sweep's propagated error w_j by $\sqrt{E}$, so

$$r^{(K)} = F^K Y - \sum_{j=1}^{K} F^{K-j} w_j.$$

Applying the contraction on H_0 gives

$$\left\|r^{(K)}\right\| \le \rho^K \|Y\| + \sqrt{E}\,\frac{1-\rho^K}{1-\rho}. \tag{54}$$

Alternatively, apply the all-step energy bound to the entire concatenated sequence of K sweeps to obtain

$$\left\|r^{(K)} - F^K Y\right\| \le \sqrt{KE}.$$

Taking the smaller error bound yields

$$\left\|r^{(K)}\right\| \le \rho^K \|Y\| + \sqrt{E}\,\min\left\{\sqrt{K}, \frac{1-\rho^K}{1-\rho}\right\}.$$

The invariant-subspace and contraction hypotheses are needed for (54). A joint kernel of all projections is unchanged by every complement product, so a contraction factor below one cannot hold on a space containing such a direction. The general energy bound itself requires no contraction hypothesis.

F.5 TWO-STEP SEPARATION AND APPROXIMATE LOCALITY

To detail Proposition 16, expanding the square shows that its identity is equivalent to

$$\langle (I-Q)(I-P)Y, (I-Q)Pv\rangle = 0 \quad \text{for all } Y, v \in H.$$

Self-adjointness and idempotence of $I - Q$ make this equivalent to $(I-P)(I-Q)P = 0$, or $(I-P)QP = 0$. The latter gives $QP = PQP$, while its adjoint gives $PQ = PQP$. Hence the projections commute. Conversely their commutation makes $(I-P)(I-Q)P = 0$. Necessity concerns all targets and local errors simultaneously. A particular cross term may still vanish.

For (48), split each arbitrary defect as $\ell_t = P_t e_t$ and $u_t = (I - P_t)e_t$. Expanding the residual recursion gives

$$Y - A_k = FY - \sum_t S_t \ell_t - \sum_t S_t u_t.$$

Theorem 15 bounds the first error sum by $\sqrt{\sum_t \eta_{i_t}^2}$, while contraction of each suffix operator bounds the second sum by $\sum_t \zeta_t$. If the projections commute, Theorem 2 applies to $FY - \sum_t S_t \ell_t$ before adding the nonlocal term, yielding (49). The separate sharp examples identify each contribution's order, without establishing simultaneous equality for both terms on one instance.

G Approximate responses to quadratic scores

The local error model has a direct incentive interpretation in the existing DAG reporting mechanism of Frongillo et al. (2025). Suppose Y becomes verifiable after reports are collected. Conditional on the public list of distinct experts, each expert simultaneously submits one square-integrable function of its private information. Expert i_t receives score

$$s_t = -\left(R_t - H_t\right)^2, \qquad H_t = a_t Y + \sum_{s<t} b_{ts} R_s. \tag{55}$$

The ordering describes score dependencies; it does not reveal previous report realizations to the experts. The conditional mean minimizes expected squared loss (Murphy, 2022, Section 5.1.5.1), so a negative squared-error score elicits that mean. Scoring-based aggregation also studies how loss geometry shapes pooling rules (Neyman & Roughgarden, 2023); here the score determines which response defects the correction protocol must tolerate.

Proposition 17 (Regret equals local error energy)**.** *With the other experts' report functions fixed, the maximum expected improvement expert i_t can obtain by replacing R_t with a local square-integrable report is*

$$\mathrm{Reg}_t = \left\|R_t - P_{i_t} H_t\right\|^2. \tag{56}$$

Thus the local tolerance $\|e_t\| \le \eta_i$ in (2) holds exactly when the report is an η_i^2-approximate best response to the quadratic score.

Orthogonality separates the squared loss into response defect and unpredictable target variation. Minimizing the former proves the identity; Appendix H gives the full calculation.

Applying Theorem 2 to residual scores therefore bounds aggregate error by uncovered variance plus the sum of the selected experts' regrets. This is an expected, prior-averaged guarantee, conditional on the public seed. It requires neither pointwise regret bounds for every private type nor independent reporting mistakes. The repeated-agent theorem is an oracle result; a player controlling several linked reports would require a separate incentive analysis.

Allocating score weight. Suppose expert i has expected utility regret at most $\varepsilon_i > 0$ under score $-\lambda_i(R_i - H_i)^2$, giving $\eta_i^2 = \varepsilon_i/\lambda_i$. With positive weights totaling W, the optimal allocation is

$$\lambda_i^* = \frac{W\sqrt{\varepsilon_i}}{\sum_j \sqrt{\varepsilon_j}}, \qquad \mathcal{E}(\boldsymbol{\eta})^{\min} = \frac{\left(\sum_i \sqrt{\varepsilon_i}\right)^2}{W}. \tag{57}$$

Substitution into Theorem 3 gives the optimized risk. Stronger score weights reduce response defects, while balanced selection protects against unknown private information. Here W budgets score coefficients, not expected payments; the regret bounds must remain valid at the chosen weights. Appendix H.1 derives the allocation and the resulting optimized risk.

H Quadratic scoring and approximation semantics

For a fixed seed and selected distinct experts, regard report functions as pure strategies in a Bayesian game. The realized score of expert i_t is (55), and the expectation is over the common prior. The target Y is verified after reporting. All earlier reports entering H_t are chosen by different experts. Under a unilateral deviation by expert i_t, those report functions are held fixed; their realizations do not respond to the deviation. The regret calculation compares this player's local strategies against fixed report functions for every other player in the game.

For any $\mathcal{F}_{i_t}$-measurable h, orthogonality of $H_t - P_{i_t} H_t$ to local functions gives

$$\mathbb{E}[(h - H_t)^2] - \mathbb{E}[(P_{i_t} H_t - H_t)^2] = \left\|h - P_{i_t} H_t\right\|^2. \tag{58}$$

Taking $h = R_t$ proves Proposition 17. In particular, every pure-strategy profile has a well-defined local response defect; the squared norm of that defect is exactly its expected regret. A profile with regret at most $\eta_{i_t}^2$ at every selected node is a heterogeneous approximate Bayes–Nash equilibrium in expected utility. The claim is conditional on the public seed. An interim bound conditional on each private signal would be a stronger and different requirement.

For residual scores, Theorem 2 gives the more precise profile-specific formula

$$\mathbb{E}[(Y - A_k)^2] = \|FY\|^2 + \sum_t \|D_t(R_t - P_t H_t)\|^2 \le \|FY\|^2 + \sum_t \mathrm{Reg}_t \,.$$

Thus total response regret is an upper bound on surviving error energy, while defects predictable by later experts do not contribute to the final error. In the lower-bound construction every hidden private component contributes to an admissible response defect, so this regret guarantee cannot yield a uniformly smaller risk over the full information family.

This interpretation presumes that experts evaluate their score targets under the common prior and other fixed report functions. Section 5 supplies a learning construction under a Gaussian sampling model. The scoring interpretation does not supply an incentive argument for players who control multiple dependent nodes or a transcript-conditioned conversational protocol. These alternatives change the available information or unilateral deviations and require their own approximation models.

Interpreting a measured error. A reporting system should measure approximation against its actual score target. Raw target-prediction loss contains both response defect and the target variation unpredictable from an expert's information, as Proposition 17 shows. Only the former enters $\mathcal{E}(\boldsymbol{\eta})$. Errors measured against a counterfactual ideal report quantify a different object and cannot be inserted into (11) without an additional argument. Likewise, errors from a communication channel belong to the nonlocal term when they depend on information unavailable to the sender. These distinctions make the theorem useful as a criterion for a reporting model: it specifies which approximation guarantee a learning or incentive mechanism must provide before the aggregation frontier applies.

Predictions as additional observed features. Networked information aggregation through learned predictors lets each agent train on its local features and the predictions of its parents in a DAG (Kearns et al., 2025). The resulting depth guarantees concern information recovery through enlarged observation spaces. In (2), earlier reports enter a prediction target, while the conditioning sigma-field remains the original $\mathcal{F}_i$. Treating those reports as observed features would change the projection operators and the admissible local error functions. Consequently Theorem 14 is a statement about fixed-information conditional-expectation queries, and does not cover protocols that enlarge local observation spaces through learned communication.

Learning reports versus learning an aggregator. Sample-complexity analyses of forecast aggregation learn a rule from historical forecasts and outcomes (Lin & Chen, 2023); algorithmic robust aggregation optimizes over specified information structures (Guo et al., 2025). Our Gaussian construction instead learns the local report functions, while retaining their sum as the decoder. Tractable agreement replaces exact Bayesian computation with calibration conditions (Collina et al., 2025), and collaborative prediction develops reductions using local learning procedures (Collina et al., 2026). The exact ownership recurrence here applies to fresh-sample Gaussian least squares in fixed local feature spaces. Its guarantee follows from local excess risks and the actual residual labels, rather than a bound for these other learning and communication protocols.

H.1 Allocating score weight

The frontier also gives an explicit allocation rule when approximation is measured in utility units. Suppose expert i has expected utility regret at most $\varepsilon_i > 0$ and receives the weighted score $-\lambda_i(R_i - H_i)^2$. Its squared response tolerance is then $\eta_i^2 = \varepsilon_i/\lambda_i$. If the designer fixes positive weights with total $\sum_i \lambda_i = W$, the smallest total squared tolerance is

$$\min_{\lambda_i>0:\,\sum_i \lambda_i = W} \sum_i \frac{\varepsilon_i}{\lambda_i} = \frac{(\sum_i \sqrt{\varepsilon_i})^2}{W}, \qquad \lambda_i^* = \frac{W\sqrt{\varepsilon_i}}{\sum_j \sqrt{\varepsilon_j}}. \tag{59}$$

Cauchy–Schwarz gives $(\sum_i \sqrt{\varepsilon_i})^2 \le (\sum_i \varepsilon_i/\lambda_i)(\sum_i \lambda_i)$, and the stated weights give equality. Substituting (57) into Theorem 3 yields the exact optimized risk

$$1 - \frac{B}{n}\left(1 - \frac{(\sum_i \sqrt{\varepsilon_i})^2}{W}\right)_+ . \tag{60}$$

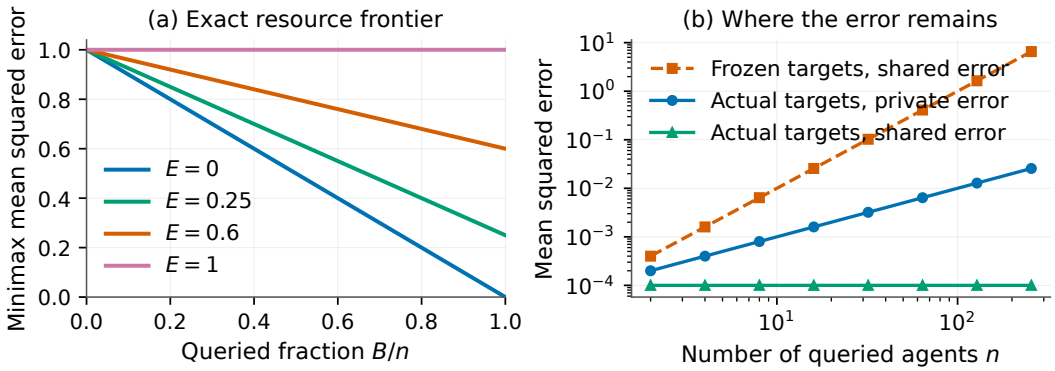

Figure 4: Absolute-error consequences. Left: the minimax budget frontier. Right: frozen and actual target examples with different ownership and error configurations. These are analytic curves, separate from the learned-predictor comparisons in Figure 2.

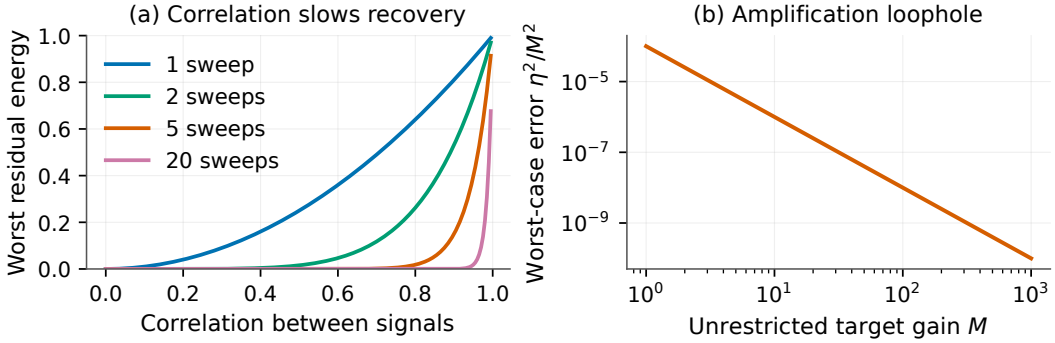

Figure 5: Analytic boundary diagnostics. Left: worst residual energy over unit targets in the linear span of two correlated binary signals after K noiseless sweeps, equal to ρ^{4K-2}. Right: the two-query construction of Proposition 12 has worst error η^2/M^2 at $\eta = 0.01$, despite both clean reports having norm one. Unbounded direct gain explains the vanishing error.

The quantity W is a budget on score coefficients, not expected monetary expenditure. Under this budget, score weight and query coverage solve different allocation problems: stronger score weights reduce response defects, while balanced selection protects against unknown private information. The derivation assumes the utility-regret bounds remain valid at the chosen weights; it does not predict how a human or learned expert's regret changes when incentives change.

I Numerical procedures and vector figures

The identity checks in this appendix use CPU linear algebra and deterministic finite-space expectations. No model, training data, or GPU computation is used. The script `experiments/verify_and_plot.py` sets random seed 20260912, exports its diagnostics as JSON, and produces vector PDF plots. Random generation selects the verification instances; expectations are computed exactly by enumerating every state in each finite probability space.

I.1 Finite-space conditional expectations

For each $n = 2, \ldots, 8$, we generate 24 configurations. Eight independent uniform Rademacher coordinates give a probability space of $2^8 = 256$ states. Each coordinate receives a uniformly sampled nonempty owner set. Random Gaussian target coefficients are normalized to Euclidean norm one, so the additive target has variance one. We sample a nonempty subset and its order.

For expert i, conditional expectation is computed by grouping states with identical locally visible bits and averaging within each group. A random function on the full state space is projected onto

that local information and rescaled to a tolerance drawn uniformly from $[0.01, 0.2]$. These defects are generally nonlinear, can have nonzero means, and can be correlated through shared coordinates. We evaluate the residual recursion, explicitly propagate every defect through later complement projections, and compare the two sides of (8). Pairwise inner products check that the uncovered target and all propagated defects are mutually orthogonal.

I.2 Independent linear programs

For each $n = 2, \dots, 8$, each $B = 1, \dots, n$, and each total error energy $E \in \{0.05, 0.4, 0.95, 1.2\}$, draw $\epsilon_i = \eta_i^2$ from a Dirichlet allocation of total E. This gives 140 configurations. Enumerate all subsets S of size at most B, including the empty subset. Write x_S for their probabilities. We solve the residual-protocol upper-bound design problem

$$\begin{aligned} &\text{minimize } t \\ &\text{subject to } \sum_S x_S = 1, \quad x_S \geq 0, \\ &\qquad 1 - \sum_{S: i \in S} x_S + \sum_S x_S \sum_{j \in S} \epsilon_j \leq t \quad (i = 1, \dots, n). \end{aligned} \tag{61}$$

The empty subset supplies the zero predictor. This is a finite optimization of a proved upper bound, rather than an enumeration of all legal query graphs. Its optimum agrees with (11).

For an independently drawn vector of inclusion probabilities $p \in [0, 1]^n$, we solve the private-erasure allocation problem

$$\begin{aligned} &\text{minimize } \sum_i p_i z_i \\ &\text{subject to } \sum_i v_i = 1, \quad v_i, z_i \geq 0, \quad z_i \geq v_i - \epsilon_i. \end{aligned} \tag{62}$$

Here v_i is expert i's private target variance and z_i its part that remains visible after erasure. The optimum is $\min_i p_i(1 - \sum_i \epsilon_i)_+$, matching the lower-bound construction. This LP independently checks that all tolerances can first be used to hide private variance, while allocating the remaining visible variance to an expert with minimal inclusion probability.

I.3 Noncommuting projections

For 300 configurations in dimension 12, draw between one and twelve projections. A QR decomposition of a Gaussian matrix produces an orthonormal basis, and a random rank from one to eleven selects each projection's range. Local errors are random vectors projected onto those ranges. We explicitly form the block synthesis matrix K and the complement product F, then check (46) in operator norm and the cumulative energy inequality. Table 1 reports the largest discrepancies.

Table 1: Numerical identity checks. These are floating-point discrepancies from exact analytic equalities or inequalities, not statistical prediction errors.

Quantity	Configurations	Largest discrepancy
Commuting energy identity	168	2.23×10^{-16}
Pairwise surviving-error orthogonality	168	4.97×10^{-17}
Heterogeneous primal/erasure LP values	140	1.67×10^{-14}
Noncommuting operator identity	300	3.27×10^{-15}
Local-energy upper-bound violation	300	0

I.4 Formalized energy lemmas

The accompanying Lean source checks the local Pythagorean identity, the one-step energy inequality, and their finite-recursion consequence for varying orthogonal projections in a real inner-product

space. In particular it checks

$$z_{t+1} = z_t - P_t z_t + e_t, \quad e_t \in \operatorname{ran}(P_t) \quad \Longrightarrow \quad \|z_k\|^2 \le \|z_0\|^2 + \sum_{t<k} \|e_t\|^2 .$$

The file uses Lean 4.34.0-rc2 and the corresponding mathlib release. It contains no proof holes or additional axioms beyond the standard library's logical axioms. This formalization covers the energy argument in Theorem 15; the conditional-expectation construction, commuting decomposition, minimax theorem, and game interpretation are proved mathematically in the preceding sections.

J Relative precision: full proofs

J.1 The relative one-pass lower bound

For $0 \le \eta < 1$, fix any one-pass protocol, with arbitrary finite target coefficients and any measurable decoder. Some expert j has inclusion probability $p_j \le B/n$. Before drawing the public seed, choose independent Gaussian primitives U, V, owned only by j, with variances $1-\eta^2$ and η^2, and put $Y = U + V$. All other experts have trivial information. Their reports can be zero. At the unique node querying j, take $R_t = a_t U$ and $e_t = -a_t V$. Earlier reports vanish, so $P_j H_t = a_t Y$ and $\|e_t\| = \eta \|P_j H_t\|$. Reports after this node again have conditional mean zero for every other expert. The construction therefore satisfies the actual query equations at every node.

If j is unqueried, the transcript reveals no target information. If it is queried, V remains independent of the transcript and the seed. Every decoder has risk at least $1 - p_j(1-\eta^2) \ge 1 - (B/n)(1-\eta^2)$. This argument imposes no clean normalization or direct-gain bound. For $\eta \ge 1$, zero reports with $e_t = -a_t Y$ at the informed expert erase the target entirely. For $B = 0$, the zero predictor attains unit risk, completing the boundary cases of the frontier in (17).

J.2 First-exposure orthogonality and repeated sweeps

Let $F_t = C_t \cdots C_1$ and $F_0 = I$. Expanding the actual residual recursion gives $r_{t-1} = F_{t-1} r_0 - v_{t-1}$, where v_{t-1} is a sum of earlier local errors propagated through commuting complements. Each summand remains in the range of its original projection. Set $z_t = P_t F_{t-1} r_0$. For every $s < t$, commutation gives $P_s z_t = 0$, so $z_t \perp v_{t-1}$ and $z_t \perp P_t v_{t-1}$. Consequently

$$\|P_t r_{t-1}\|^2 = \|z_t\|^2 + \|P_t v_{t-1}\|^2 \ge \|z_t\|^2 .$$

The projections $P_t F_{t-1}$ are mutually orthogonal and sum to $I - F_k$. Thus $\sum_t \|z_t\|^2 = \|(I - F_k) r_0\|^2$. Summing the one-step energy loss in the proof of Theorem 5 gives the sweep bound for arbitrary local errors, including errors with nonzero means and dependence across nodes.

Let F_{all} be the product of all expert complements. In the additive model it kills Y, and it kills every local error because its product contains the complement of that error's projection. All complements commute with F_{all}, so the residual at every full-sweep boundary belongs to $\ker F_{\mathrm{all}}$. Each subsequent full sweep contracts squared norm by at most η^2. This proves the η^{2K} bound for K sweeps, including nonlinear local defects and arbitrary dependence across visits. The bound holds over all ownership structures, but does not specify the exact repeated-query frontier for arbitrary report budgets.

J.3 A fixed hard prior for repeated full-information queries

For $0 < \eta < 1$, put $\rho = \eta^2$. For a budget of L report nodes choose independent centered Gaussian primitives $Z_1, \dots, Z_L, V$ with

$$\operatorname{Var}(Z_j) = (1-\rho)\rho^{j-1}, \qquad \operatorname{Var}(V) = \rho^L, \qquad Y = \sum_{j=1}^{L} Z_j + V.$$

Every expert observes every primitive. Their variances sum to one, and this single prior is fixed before the public seed. Write $W_t = \sum_{j>t} Z_j + V$, so $W_0 = Y$ and $\operatorname{Var}(W_t) = \rho^t$.

At actual node t choose $e_t = -a_t W_t$. Inductively, earlier reports are linear functions of $Z_1, \dots, Z_{t-1}$. The actual target therefore has the form $H_t = a_t W_{t-1} + L_t$, where L_t depends linearly on those

earlier components. Independence gives

$$\|H_t\|^2 \geq a_t^2 \rho^{t-1}, \qquad \|e_t\|^2 = a_t^2 \rho^t \leq \eta^2 \|H_t\|^2 .$$

Full information makes $P_t H_t = H_t$, so both relative conditions hold. Moreover, $R_t = H_t - a_t W_t$ is linear in $Z_1, \ldots, Z_t$, completing the induction. If the realized protocol uses $k \leq L$ nodes, its transcript is independent of W_k. Every measurable decoder has risk at least $\rho^k \geq \rho^L$. The same prior works for every public seed, including seeds that select different graphs and coefficients.

For $\eta = 0$, the upper bound is zero after one query. For $\eta \geq 1$, the error choice $e_t = -a_t Y$ with all reports zero is admissible and leaves unit conditional variance. Together with residual contraction, these constructions prove Theorem 18. The full-information family permits multiple primitives with the same owner set; restricting observations to $\sigma(Y)$ alone would be a different family.

J.4 Amplification under relative precision

For the two-query graph (41), take $M \geq 1$ and use its stated decoder. The second actual target is $Y - (M-1)e_1$. Under either relative condition with full information, the exact worst-case error of this decoder is

$$\sup_{e_1, e_2} \left\| \widehat{Y} - Y \right\|^2 = \left[\frac{\eta}{M} \big(1 + (M-1)\eta\big) \right]^2 .$$

The triangle inequality proves the upper bound, and $e_1 = -\eta Y$, $e_2 = \eta[1 + (M-1)\eta]Y$ attains it. For $0 < \eta < 1$ the expression tends to η^4, rather than zero. Two ordinary residual queries attain η^4 with finite coefficients; Theorem 18 shows their minimax optimality.

J.5 Why full-target relative error is different

Take independent private X_1, X_2 of variances v and $1 - v$, with $0 < v < 1$, and let $Y = X_1 + X_2$. Under full-target tolerance, set

$$e_1 = \eta \frac{X_1}{\sqrt{v}}, \qquad e_2 = \eta \sqrt{1 - v + \eta^2} \frac{X_2}{\sqrt{1-v}}.$$

The first residual is $r_1 = X_2 - e_1$, so $\|r_1\|^2 = 1 - v + \eta^2$ and both errors saturate their actual full-target tolerances. The final residual is $-e_1 - e_2$ and has risk $\eta^2 + \eta^2(1 - v + \eta^2)$. This counterexample concerns the residual sweep guarantee; it does not determine the arbitrary-decoder minimax value for partially informed full-target relative queries.

K From regression samples to aggregate risk

K.1 Conditional regression and the inverse-Gram expectation

Write $s = 1/(n+1)$. Conditional on the training history before stage t, let earlier fitted reports be $\widehat{f}_j = a_j W + b_j U_j$. Put $c_t = 1 - \sum_{j<t} a_j$ and $d_j = b_j - 1$. The actual residual target is

$$r_{t-1} = c_t W + U_t - \sum_{j<t} d_j U_j + \sum_{j>t} U_j.$$

The conditional mean given (W, U_t) is $c_t W + U_t$. Its independent Gaussian residual noise has variance

$$v_t = s \left(n - t + \sum_{j<t} d_j^2 \right). \tag{63}$$

Fresh data at stage t are independent of the training history. Conditional also on its $N \times 2$ design matrix X_t, the least-squares coefficient error is centered Gaussian with covariance $v_t (X_t^\top X_t)^{-1}$. This conditioning is why the finite-sample theorem uses fresh examples at each stage.

For completeness, let G be an $N \times d$ standard Gaussian matrix with $N > d + 1$. For column g_j, let P_{-j} project onto the other $d - 1$ columns. The Schur complement gives

$$[(G^\top G)^{-1}]_{jj} = \frac{1}{g_j^\top (I - P_{-j}) g_j}.$$

Conditional on the other columns, the denominator is χ^2_{N-d+1}. Integrating its gamma density yields $\mathbb{E}[1/\chi^2_m] = 1/(m-2)$ for $m > 2$. Sign symmetry makes expected off-diagonal entries zero; their integrability follows from positive semidefiniteness and finite diagonal expectations. Therefore

$$\mathbb{E}[(G^\top G)^{-1}] = \frac{I_d}{N-d-1}, \qquad \mathbb{E}[(X_t^\top X_t)^{-1}] = \frac{I_2}{s(N-3)}.$$

Writing $h_t = a_t - c_t$ for the shared coefficient error gives

$$\mathbb{E}[sh_t^2 \mid \text{history}] = \mathbb{E}[sd_t^2 \mid \text{history}] = \beta v_t. \tag{64}$$

Training expectations do not bound each fit's absolute or projected-relative error.

K.2 Arbitrary Gaussian ownership

To prove Theorem 19, condition on all earlier fits and write $r_{t-1} = \sum_j a_j X_j$. Its conditional mean at the current expert is $\sum_{j \in J_t} a_j X_j$, and the independent regression noise has variance $v_{\text{out}} = \sum_{j \notin J_t} v_j a_j^2$. Conditional on the local design, OLS coefficient-error covariance is $v_{\text{out}}(X^\top X)^{-1}$. Whitening the independent Gaussian columns in the inverse-Gram calculation gives

$$\mathbb{E}[v_j(\widehat{a}_j - a_j)^2 \mid \text{history}] = \frac{v_{\text{out}}}{N_t - d_t - 1} \qquad (j \in J_t).$$

After subtracting the fitted report, owned residual coefficients are minus these estimation errors. Unowned coefficients remain a_j. Taking training expectations proves (74). Independence and centering of the primitive coordinates make their energies additive for every fitted residual, giving final expected risk $\sum_j x_{kj}$. An empty feature set leaves the residual unchanged; positive variances ensure the nonempty Gaussian designs have full column rank almost surely.

The recurrence is a product of nonnegative linear maps applied to the initial variance vector. For a fixed ownership order and sample allocation, its column sums therefore determine the largest expected risk over unit total variance. This evaluation requires the ownership structure; it does not assert that a prior-agnostic aggregator can infer the best order without additional information. In the two-expert example, after the first query the energies are $(\beta v_3, \beta v_3, v_3)$, and after the second they are $(\beta v_3, \beta^2 v_3, \beta^2 v_3)$. Reversing the order exchanges v_1 and v_3, proving the claimed ordering effect.

K.3 The exact recurrence and finite-sample comparison

Subtracting the fitted report leaves

$$r_t = -h_t W - \sum_{j \le t} d_j U_j + \sum_{j > t} U_j.$$

Only the current shared error remains, whereas every earlier private error persists. Set $q_t = s\sum_{j \le t} d_j^2$ and $Q_t = \mathbb{E}q_t$. Equations (63)–(64) give

$$\mathbb{E}[q_t \mid \text{history}] = q_{t-1} + \beta[(n-t)s + q_{t-1}],$$

which proves (77). At the last node $v_n = q_{n-1}$, and final expected risk is $Q_{n-1} + 2\beta Q_{n-1}$. Expanding the recurrence establishes the exact sum in (76). The calculation conditions on each training history and does not assume that fitted errors at different stages are independent.

For $n \ge 2$ define $J = n(n-1)\beta/[2(n+1)]$. The exact sum gives

$$J \le \mathbb{E}_{\mathcal{T}} \|r_n\|^2 \le J(1+2\beta)(1+\beta)^{n-2} \le Je^{n\beta}.$$

Consequently the difference from J is at most $J(e^{n\beta} - 1)$. In any regime $n\beta \le C < \infty$, it is bounded by $(e^C/2)n^2\beta^2$. This establishes the stated asymptotic in the regime $n\beta \to 0$; for fixed small N, the exact risk can instead grow exponentially with population size n.

If $N \ge 3 + \lceil en/(2\delta) \rceil$ for $\delta \in (0,1)$, then $n\beta \le 2\delta/e < 1$ and $J \le n\beta/2 \le \delta/e$. Hence $Je^{n\beta} \le \delta$. A direct comparison also shows that $N \ge 4n + 3$ suffices to beat the exact ordinary-report Bayes risk c_n: the ratio is at most

$$\frac{n(n+1)\beta e^{n\beta}}{2(n-1)} \le \frac{n+1}{8(n-1)} e^{1/4} \le \frac{3}{8} e^{1/4} < 1.$$

This comparison concerns the fixed exact reports $W + U_i$. Learned ordinary reports have different coefficients and can retain a different amount of information.

K.4 THE INFORMATION RETAINED BY ORDINARY REPORTS

For $X_i = W + U_i$, $\mathrm{Cov}(X) = s(I + \mathbf{1}\mathbf{1}^\top)$ and $\mathrm{Cov}(Y, X) = 2s\mathbf{1}^\top$. Gaussian conditioning yields the optimal decoder $2\sum_i X_i/(n+1)$ and residual variance $1 - 4n/(n+1)^2 = c_n$. This recovers the known ordinary-forecast obstruction (Neyman, 2024; Frongillo et al., 2025). Adding independent hidden V_i in Theorem 4 contributes exactly $\mathcal{E}$ and scales the remaining conditional variance by $1 - \mathcal{E}$.

For fitted ordinary reports $a_i W + b_i U_i$, form the report matrix M whose ith row has entries a_i in the shared column and b_i in private column i. Given these coefficients, the Bayes risk is

$$s \left\| (I - M^\dagger M)\mathbf{1} \right\|_2^2,$$

where $M^\dagger$ is the Moore–Penrose inverse. Our experiments evaluate this expression through the one-dimensional nullspace, and check it against dense least-squares projection. Estimation can improve this benchmark by changing the nullspace's alignment with $\mathbf{1}$, so larger coefficient errors need not imply that the resulting ordinary reports retain less information about the target.

K.5 TRAINING RANDOMNESS AND INVISIBLE CHANNEL NOISE

Condition on all trained report functions and let u_t be centered prediction-time channel corruptions, independent of training, signals, and one another. Suppose these are added to fixed local reports. Their sum is orthogonal to the signal-dependent prediction error, so final risk equals the local-report risk plus $\sum_t \mathbb{E}u_t^2$. The same conclusion holds when actual residual targets are evaluated by conditional expectation and the fixed local response defects do not depend on the channel draws: each channel term has zero conditional mean for every expert. Shared channel noise instead adds coherently, as in (49). If local report functions adapt to realized channel corruption, the resulting dependence must be included when calculating cross terms in the prediction risk.

L A SHARP JOINT BOUND FOR APPROXIMATE LOCALITY

Let $z_0 = 0$ and $z_t = C_t z_{t-1} + \ell_t + u_t$, with $P_t \ell_t = \ell_t$ and $P_t u_t = 0$. Since the first two terms inside $C_t z_{t-1} + u_t$ lie in $\ker P_t$, Pythagoras and contraction give

$$\|z_t\|^2 = \|C_t z_{t-1} + u_t\|^2 + \|\ell_t\|^2 \le (\|z_{t-1}\| + \zeta_t)^2 + \eta_t^2.$$

Induction proves $\|z_k\| \le s_k$. The actual residual expansion gives $(Y - A_k) - FY = -z_k$, proving the upper bound in Proposition 21. No commutation is needed for this inequality.

To attain every step on a single instance, take k independent $N(0, 1/k)$ signals X_t, private to their respective experts, with target $Y = \sum_t X_t$. Enlarge the probability space by an independent unit-variance Gaussian channel variable W observed by no expert. Set $\ell_t = \eta_t \sqrt{k} X_t$. If $\|z_{t-1}\| > 0$, set $u_t = \zeta_t z_{t-1} / \|z_{t-1}\|$; otherwise set $u_t = \zeta_t W$. By induction, z_{t-1} is centered and depends only on $W, X_1, \ldots, X_{t-1}$, so $P_t z_{t-1} = P_t u_t = 0$. The local and nonlocal budgets are both saturated and $\|z_t\| = s_t$ at every step. Full coverage gives $FY = 0$, hence final residual risk is exactly s_k^2.

For $\zeta_t = 0$, the recursion reduces to $s_k^2 = \sum_t \eta_t^2$. For $\eta_t = 0$, it reduces to $s_k = \sum_t \zeta_t$. Mixed budgets obey the sharper sequential interpolation in (79); for example, at two equal-budget nodes $s_2^2 = (\sqrt{\eta^2 + \zeta^2} + \zeta)^2 + \eta^2$. This is generally smaller than $(\sqrt{2}\eta + 2\zeta)^2$. The attaining construction aligns each nonlocal corruption with the previously surviving error. Fresh independent prediction-time noise need not produce this alignment or attain the same worst-case risk.

M QUADRATIC IMPROVEMENT PAYMENTS AND MARKET INFORMATION

With other report functions fixed, replace the score $-(R_t - H_t)^2$ by $H_t^2 - (R_t - H_t)^2$. The added term does not depend on the reporting player's action, so it leaves unilateral incentives unchanged. Abstention $R_t = 0$ has zero realized payment. Orthogonal projection yields expected payment $\|P_t H_t\|^2 - \|R_t - P_t H_t\|^2$. Projected-relative precision with $\eta \le 1$ therefore guarantees nonnegative ex-ante expected payment, conditional on the public seed, and at least $(1 - \eta^2)\|P_t H_t\|^2$.

For $H_t = Y - A_{t-1}$ this payment equals

$$-(Y - A_{t-1} - R_t)^2 + (Y - A_{t-1})^2,$$

which is the algebraic quadratic market-score increment. In market scoring rules, a trader revises a publicly observed forecast (Neyman & Roughgarden, 2023, discussion of prediction markets); dynamic-market convergence studies the associated strategic revelation and learning from trading history (Ostrovsky, 2012). In our DAG mechanism, experts submit local functions and condition on their original information. Algebraic equality of score increments does not equate these information structures or their equilibrium analyses.

The nonnegative guarantee is ex ante over the signal prior, not conditional on every private type. Gaussian targets and unbounded report functions also permit unbounded realized transfers, so this construction does not establish bounded liability. Restricting target and report ranges would control transfers but would change the admissible prior and response classes used by the minimax theorems.

N LEARNED-PREDICTOR EXPERIMENTS

The study uses NumPy least squares on the complete Gaussian linear model in Section 5. For each of the 24 population/sample configurations, we draw 1,000 training repetitions with fresh stagewise data and 1,000 with one reused dataset. A master seed of 20260912 generates independent configuration seeds, recorded with each result. Each method in a configuration uses the same local designs and outcome labels; residual labels subtract the already fitted residual reports on the current examples. No model selection, early stopping, or evaluation-based switching is used.

Methods and evaluation. Actual residual OLS fits the two local coefficients without an intercept and sums the resulting reports. Ordinary OLS fits the original target at every expert. We evaluate its arithmetic mean, the fixed calibration $2/(n+1)$ times its sum, and its optimal Gaussian decoder given the fitted coefficients and true covariance. The last is an informational benchmark with additional knowledge, not a fitted aggregator. These are full least-squares procedures; all linear systems are solved directly. Population error is the squared norm of the target coefficient minus the aggregate coefficient, multiplied by $1/(n+1)$. Appendix K gives the orthogonal projection formula used to evaluate the Bayes risk for each collection of fitted ordinary reports.

The simulated design coordinates have unit variance; scaling every coordinate and response by $1/\sqrt{n+1}$ leaves the fitted coefficients unchanged. Evaluation explicitly includes this normalization. Standard errors are the sample standard deviations across independent training repetitions divided by $\sqrt{1000}$. Paired method differences are formed within each repetition before calculating their standard errors. The ± 2 standard-error bands describe the Monte Carlo precision of each configuration; they do not provide simultaneous confidence statements across the full experiment.

Training reuse. Fresh training uses nN distinct examples in total; reuse uses N distinct examples. Both give each node N training examples and use the same local feature spaces. This is a sampling-scheme comparison at equal per-node sample size, not equal total distinct examples. The pathwise error identity holds in either case. The inverse-Gram calculation conditional on earlier fits applies to fresh data, whereas reused-data labels depend on the same observations that generated the previous coefficients. Figure 6 measures the resulting difference directly.

Numerical and implementation checks. The 24 fresh-training residual means lie within 2.584 Monte Carlo standard errors of (76). A separate dense least-squares projection reproduces the nullspace evaluation of ordinary Bayes risk to within 4.45×10^{-16}. The archived per-node measurements include the actual target's shared coefficient and unpredictable variance, both fitted coefficients, local error energies, and surviving private error. The full experiment required about 38 seconds on CPU and used no GPU. The finite-space identity and optimization checks in Appendix I test the projection and minimax identities separately from these Gaussian training experiments.

General ownership and query order. We also fit actual residual OLS for 5,000 repetitions in a six-coordinate, four-expert overlapping instance and in both orders of the two-expert example with variances $(0.8, 0.1, 0.1)$ and $N = 32$. The general-instance expected risk is 0.0303482,

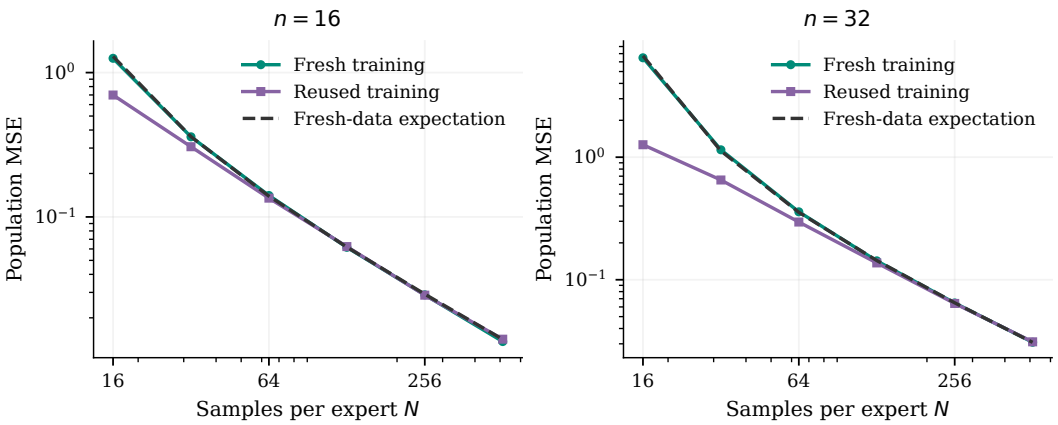

Figure 6: Fresh and reused training data for actual residual OLS. Points average 1,000 independent runs, and shading shows ± 2 standard errors. The dashed expectation applies only to fresh stagewise data. Each node receives N examples; the total count is nN for fresh training and N for reuse.

compared with empirical 0.0302681 (standard error 0.0006183). Learning from the larger unique signal first gives 0.0036435 empirically, versus 0.0286947 in reverse order; the exact expectations are 0.0036861 and 0.0294887. Across all recorded coordinate energies and stages, discrepancies from (74) are at most 1.379 standard errors. These additional results are produced by `experiments/verify_general_learning.py`.

O Exact relative precision at every report budget

This section proves Corollary 7 over the full additive ownership family. The budget counts every report node, including repeated visits and reports omitted by the decoder. Conditional on the independent public seed, the graph, expert indices, and finite coefficients are fixed. Error functions may depend on that seed, but the adversary chooses a single prior before it is drawn. No clean normalization or bound on direct target gain is imposed anywhere in this argument.

O.1 A partial sweep followed by full sweeps

First suppose $0 < \eta < 1$, and write $\rho = \eta^2$ and $L = qn + r$, with $0 \le r < n$. Choose a uniform r-subset S of experts and query them once in a fixed order. Then perform q full residual sweeps, each in a fixed order through all n experts. This defines an admissible ex ante graph with exactly L reports and a decoder that adds their values to form the final forecast.

Let F_S denote the product of the complements for the initial subset. Theorem 5, applied to the fixed initial target Y, gives for every seed and admissible error sequence

$$\|r_r\|^2 \le \rho + (1-\rho)\,\|F_S Y\|^2 . \tag{65}$$

The uncovered term is the sum of the variances of primitives with no owner in S. Each primitive has at least one owner, so a uniform r-subset misses it with probability at most $1 - r/n$. Since the primitive variances sum to one,

$$\mathbb{E}_Z \|r_r\|^2 \le 1 - \frac{r}{n}(1-\rho).$$

This averaging remains valid when the errors depend on the entire public seed: the right side of (65) involves only the fixed prior and the selected subset, not the resulting error functions.

Put $F_{\mathrm{all}} = \prod_{i=1}^{n}(I - P_i)$. Every target primitive has an owner, so $F_{\mathrm{all}} Y = 0$. If e_t is local to expert i_t, then $F_{\mathrm{all}} e_t = 0$, because the product includes $I - P_{i_t}$. All complement projections commute with F_{all}. Applying it to the residual recursion therefore shows that $F_{\mathrm{all}} r_t = 0$ at every stage, including the end of the initial partial sweep. Theorem 5 consequently contracts squared residual norm by at most ρ at each subsequent full sweep. Thus

$$\mathbb{E}_Z \|r_L\|^2 \le \rho^q \mathbb{E}_Z \|r_r\|^2 \le \rho^q \left[1 - \frac{r}{n}(1-\rho)\right]. \tag{66}$$

Table 2: Fresh training: population risk averaged over $1{,}000$ runs per configuration. Parentheses give one standard error. The Bayes column is evaluated using every fitted ordinary-report coefficient together with the known Gaussian population covariance.

n	N	Residual OLS	Ordinary mean	Ordinary Bayes	Exact expectation
4	16	0.1189 (0.0040)	0.4731 (0.0023)	0.2848 (0.0036)	0.1177
4	32	0.0450 (0.0014)	0.4627 (0.0015)	0.3396 (0.0024)	0.0463
4	64	0.0211 (0.0006)	0.4549 (0.0011)	0.3487 (0.0017)	0.0208
4	128	0.0097 (0.0003)	0.4523 (0.0007)	0.3565 (0.0013)	0.0099
4	256	0.0048 (0.0001)	0.4509 (0.0005)	0.3568 (0.0009)	0.0048
4	512	0.0023 (0.0001)	0.4508 (0.0004)	0.3594 (0.0007)	0.0024
8	16	0.3626 (0.0094)	0.6941 (0.0016)	0.2196 (0.0050)	0.3744
8	32	0.1359 (0.0032)	0.6869 (0.0011)	0.3575 (0.0054)	0.1316
8	64	0.0561 (0.0011)	0.6835 (0.0007)	0.5002 (0.0031)	0.0562
8	128	0.0255 (0.0005)	0.6820 (0.0005)	0.5611 (0.0014)	0.0261
8	256	0.0127 (0.0002)	0.6817 (0.0004)	0.5848 (0.0009)	0.0126
8	512	0.0063 (0.0001)	0.6812 (0.0003)	0.5954 (0.0006)	0.0062
16	16	1.2560 (0.0301)	0.8353 (0.0009)	0.0904 (0.0029)	1.2949
16	32	0.3590 (0.0063)	0.8301 (0.0007)	0.1727 (0.0050)	0.3597
16	64	0.1408 (0.0021)	0.8289 (0.0004)	0.3421 (0.0066)	0.1393
16	128	0.0616 (0.0009)	0.8279 (0.0003)	0.5761 (0.0047)	0.0618
16	256	0.0289 (0.0004)	0.8279 (0.0002)	0.6976 (0.0016)	0.0292
16	512	0.0137 (0.0002)	0.8274 (0.0002)	0.7408 (0.0006)	0.0142
32	16	6.4743 (0.1507)	0.9136 (0.0005)	0.0367 (0.0012)	6.7154
32	32	1.1489 (0.0176)	0.9127 (0.0004)	0.0521 (0.0019)	1.1247
32	64	0.3583 (0.0043)	0.9109 (0.0002)	0.1078 (0.0037)	0.3548
32	128	0.1426 (0.0015)	0.9107 (0.0002)	0.2743 (0.0066)	0.1435
32	256	0.0646 (0.0006)	0.9103 (0.0001)	0.5789 (0.0063)	0.0648
32	512	0.0309 (0.0003)	0.9100 (0.0001)	0.7672 (0.0022)	0.0308

Table 3: Reused training: population risk averaged over $1{,}000$ runs per configuration. Parentheses give one standard error. The Bayes column is evaluated using every fitted ordinary-report coefficient together with the known Gaussian population covariance.

n	N	Residual OLS	Ordinary mean	Ordinary Bayes	Fixed calibration
4	16	0.1188 (0.0036)	0.4980 (0.0036)	0.2912 (0.0049)	0.4713 (0.0074)
4	32	0.0458 (0.0015)	0.4724 (0.0023)	0.3377 (0.0038)	0.4158 (0.0049)
4	64	0.0214 (0.0006)	0.4588 (0.0016)	0.3497 (0.0029)	0.3864 (0.0033)
4	128	0.0098 (0.0003)	0.4537 (0.0011)	0.3541 (0.0021)	0.3702 (0.0022)
4	256	0.0049 (0.0001)	0.4528 (0.0007)	0.3571 (0.0015)	0.3652 (0.0015)
4	512	0.0024 (0.0001)	0.4516 (0.0005)	0.3596 (0.0010)	0.3635 (0.0011)
8	16	0.3154 (0.0066)	0.7349 (0.0036)	0.2351 (0.0059)	0.7788 (0.0120)
8	32	0.1302 (0.0027)	0.7064 (0.0018)	0.3802 (0.0063)	0.6892 (0.0060)
8	64	0.0575 (0.0011)	0.6918 (0.0011)	0.5171 (0.0042)	0.6482 (0.0038)
8	128	0.0266 (0.0005)	0.6873 (0.0008)	0.5711 (0.0026)	0.6265 (0.0026)
8	256	0.0124 (0.0002)	0.6839 (0.0005)	0.5881 (0.0019)	0.6156 (0.0019)
8	512	0.0063 (0.0001)	0.6821 (0.0004)	0.5967 (0.0012)	0.6086 (0.0012)
16	16	0.6991 (0.0100)	0.8928 (0.0033)	0.0993 (0.0035)	1.0119 (0.0130)
16	32	0.3065 (0.0043)	0.8570 (0.0016)	0.1792 (0.0056)	0.8917 (0.0068)
16	64	0.1350 (0.0019)	0.8418 (0.0009)	0.3500 (0.0075)	0.8271 (0.0036)
16	128	0.0623 (0.0009)	0.8341 (0.0005)	0.5951 (0.0058)	0.8049 (0.0026)
16	256	0.0287 (0.0004)	0.8310 (0.0003)	0.7246 (0.0020)	0.7941 (0.0017)
16	512	0.0142 (0.0002)	0.8288 (0.0002)	0.7535 (0.0013)	0.7842 (0.0011)
32	16	1.2640 (0.0123)	0.9774 (0.0034)	0.0413 (0.0016)	1.1263 (0.0136)
32	32	0.6516 (0.0064)	0.9390 (0.0014)	0.0600 (0.0023)	0.9949 (0.0062)
32	64	0.2957 (0.0028)	0.9259 (0.0008)	0.1145 (0.0042)	0.9419 (0.0037)
32	128	0.1372 (0.0014)	0.9171 (0.0004)	0.2820 (0.0073)	0.9098 (0.0021)
32	256	0.0640 (0.0006)	0.9141 (0.0002)	0.6111 (0.0069)	0.8989 (0.0014)
32	512	0.0311 (0.0003)	0.9119 (0.0001)	0.8015 (0.0021)	0.8892 (0.0009)

The partial sweep is placed first so that coverage is averaged against Y. A residual produced before a publicly chosen final partial sweep could depend on that choice through the earlier error functions. This ordering makes the coverage calculation independent of errors introduced by later queries.

O.2 One symmetric private prior for every graph

Fix any admissible protocol. For each expert i, choose independent centered Gaussian primitives $Z_{i,1}, \dots, Z_{i,L}, V_i$, with

$$\mathrm{Var}(Z_{i,h}) = \frac{(1-\rho)\rho^{h-1}}{n}, \qquad \mathrm{Var}(V_i) = \frac{\rho^L}{n}. \tag{67}$$

All these primitives are mutually independent, and each is owned only by its corresponding expert. Set

$$X_i = \sum_{h=1}^{L} Z_{i,h} + V_i, \qquad Y = \sum_{i=1}^{n} X_i.$$

Then $\mathrm{Var}(X_i) = 1/n$ and $\mathrm{Var}(Y) = 1$. Expert i observes its separate primitives, as permitted by (1). This symmetric prior is fixed before the public seed and applies to every admissible graph and all its realized visit counts, with unrestricted query coefficients.

For $0 \le h \le L$, define the tail

$$W_{i,h} = \sum_{j=h+1}^{L} Z_{i,j} + V_i, \qquad W_{i,0} = X_i, \qquad \mathrm{Var}(W_{i,h}) = \frac{\rho^h}{n}.$$

Condition on a seed. At node t querying expert $i = i_t$, let h be this expert's number of visits up to and including that node. Choose the local error

$$e_t = -a_t W_{i,h}. \tag{68}$$

We verify inductively that every report made by expert i through its hth visit is a centered linear function of $Z_{i,1}, \dots, Z_{i,h}$. Before the current visit, its earlier reports involve only its first $h-1$ increments. Reports from other experts are centered functions of independent private coordinates and have conditional mean zero given $\mathcal{F}_i$. Hence the actual conditional-mean target is

$$\begin{aligned} P_i H_t &= a_t X_i + \sum_{s<t:\, i_s = i} b_{ts} R_s \\ &= a_t W_{i,h-1} + M_t, \end{aligned} \tag{69}$$

where M_t is linear in $Z_{i,1}, \dots, Z_{i,h-1}$. Independence of this prefix and the remaining tail gives

$$\|P_i H_t\|^2 \ge \frac{a_t^2 \rho^{h-1}}{n}, \qquad \|e_t\|^2 = \frac{a_t^2 \rho^h}{n} \le \rho \|P_i H_t\|^2. \tag{70}$$

Thus the actual projected-relative constraint holds. The resulting report is

$$R_t = P_i H_t - a_t W_{i,h} = M_t + a_t Z_{i,h},$$

which completes the induction. Arbitrarily large or signed coefficients cause no difficulty. When $a_t = 0$, the chosen error is zero, and measurability holds with the preceding prefix alone.

Let N_i be the total number of visits to expert i for this seed. The transcript depends only on the prefixes $Z_{i,1}, \dots, Z_{i,N_i}$. The centered sum of tails $\sum_i W_{i,N_i}$ is independent of these prefixes and has variance $n^{-1} \sum_i \rho^{N_i}$. Splitting Y into its prefix sum and this independent hidden sum shows that every measurable decoder satisfies

$$\mathbb{E}[(g_Z(R) - Y)^2 \mid Z] \ge \frac{1}{n} \sum_{i=1}^{n} \rho^{N_i}. \tag{71}$$

An infinite-risk decoder also obeys this inequality. Although the visit counts depend on the seed, the primitive prior in (67) does not; the admissible errors and exposed prefixes vary with the graph, while the underlying Gaussian distribution remains fixed.

O.3 Balancing the integer visit counts

The counts satisfy $N_i \geq 0$ and $\sum_i N_i \leq L$. Since $0 < \rho < 1$, increasing any count strictly decreases $\sum_i \rho^{N_i}$, so its minimum uses the full budget. If two counts obey $a \geq b + 2$, transferring one visit from the larger to the smaller decreases their contribution, because

$$\rho^a + \rho^b - \rho^{a-1} - \rho^{b+1} = (1-\rho)(\rho^b - \rho^{a-1}) > 0.$$

At a minimum all counts differ by at most one. The budget identity $L = qn + r$ therefore forces $n - r$ counts equal to q and r counts equal to $q + 1$. Consequently

$$\frac{1}{n}\sum_i \rho^{N_i} \geq \frac{(n-r)\rho^q + r\rho^{q+1}}{n} = \rho^q\left[1 - \frac{r}{n}(1-\rho)\right]. \tag{72}$$

Combining (71) and (72) proves the lower bound for each seed. Averaging preserves it under the same prior, and (66) supplies the matching upper bound for the claimed minimax value.

The argument also gives necessary visit-count conditions for optimality when $0 < \eta < 1$. An optimal protocol uses all L nodes and visits each expert q or $q + 1$ times almost surely. To determine the marginals of the extra visits, assign arbitrary private total variances $v_i \geq 0$ summing to one and scale the expert-i variances in (67) by nv_i. The same proof gives lower risk $\sum_i v_i \mathbb{E}_Z \rho^{N_i}$, whose maximum over (v_i) is $\max_i \mathbb{E}_Z \rho^{N_i}$. With $p_i = \Pr(N_i = q + 1)$ and $\sum_i p_i = r$, optimality forces every $p_i = r/n$. These necessary conditions concern visit counts; the residual schedule with a partial sweep first additionally specifies targets and a decoder that attain the bound.

O.4 Exact responses, complete erasure, and full information

If $L = 0$, every report-free decoder has squared error at least one on a centered unit-variance target, and the zero predictor attains one. For $\eta = 0$, residual queries to a uniform $\min\{L, n\}$-subset attain $(1 - L/n)_+$. For the matching lower bound, give each expert a single independent centered Gaussian private coordinate X_i of variance $1/n$. With zero errors, every report is a centered scalar multiple of its expert's coordinate. At most L distinct coordinates can be queried, leaving conditional variance at least $(n - \min\{L, n\})/n$ for every seed and decoder.

For $\eta \geq 1$, use the same private prior and set every report to zero. Then $P_i H_t = a_t X_i$, and the error $e_t = -a_t X_i$ is local and obeys the projected-relative constraint. Unit target variance remains hidden at every budget. The zero predictor attains the matching risk one. These cases establish the endpoint conventions in (16), including $0^0 = 1$ when $\eta = 0$ and $L < n$.

Theorem 18 (Full-information special case)**.** *If every expert observes all primitive signals, the minimax risk with at most $L \geq 1$ reports is $\min\{1, \eta^{2L}\}$ under either (13) or $\|e_t\| \leq \eta \|H_t\|$.*

If every expert observes all primitives, homogeneous expert identities do not change the oracle class: each projection is the identity on every target and report. The problem reduces to $n = 1$, for which (16) gives $\min\{1, \eta^{2L}\}$ for $L \geq 1$. Projected-relative and full-target relative tolerances then coincide. This proves Theorem 18 as a special case; under partial information, the full-target counterexample in Appendix J rules out the analogous sweep contraction for general partially informed experts.

P One-pass learning formulas and local corruption

Residual fitting is the squared-loss boosting update (Murphy, 2022, Section 18.5.2). Here each learner uses a fixed local feature space, producing the report functions required by the theory. Train node t on labels $H_t = Y - \sum_{s<t} \widehat{f}_s$, holding earlier fitted functions fixed, and let $\widehat{f}_t$ depend only on its local features. Conditional on training data $\mathcal{T}$, define

$$e_t = \widehat{f}_t - P_t H_t, \qquad \epsilon_t(\mathcal{T}) = \|e_t\|^2 = \left\|\widehat{f}_t - H_t\right\|^2 - \|P_t H_t - H_t\|^2. \tag{73}$$

The excess prediction risk is precisely the local error energy. Theorem 2 therefore holds for every fitted collection, including dependent or shared training data. Fresh stagewise samples are needed for the finite-sample calculation below, not for this deterministic reduction. These learning guarantees average over training; they do not require or establish a pathwise projected-relative tolerance for every fitted predictor, even though each of its errors remains local.

Theorem 19 (Exact learning dynamics for Gaussian ownership)**.** *Let $X_j \sim N(0, v_j)$ be independent, with $v_j > 0$ and $\sum_j v_j = 1$. Fix the query order and sample sizes before training. At node t, zero-intercept least squares fits the actual residual on the d_t owned coordinates using $N_t > d_t + 1$ fresh examples. Let J_t be its owned-coordinate set. The expected residual energy in each coordinate obeys*

$$x_{0j} = v_j, \qquad x_{tj} = \begin{cases} x_{t-1,j}, & j \notin J_t, \\ \dfrac{\sum_{\ell \notin J_t} x_{t-1,\ell}}{N_t - d_t - 1}, & j \in J_t. \end{cases} \tag{74}$$

The expected population risk after node k is $\sum_j x_{kj}$.

Conditional on earlier fits, the unexplained coordinates provide Gaussian regression noise independent of the local design. Each fitted coordinate receives expected error energy equal to that noise variance divided by $N_t - d_t - 1$; unobserved coordinates retain their current errors. Averaging over earlier fits proves the recurrence. Appendix K gives its coefficient derivation. The resulting law tracks local feature dimension and sample size together with the unexplained variance inherited from every preceding fitted report as the queries successively change the target.

The recurrence also quantifies an ordering effect hidden by the unrestricted minimax frontier. With two experts owning $\{X_1, X_2\}$ and $\{X_2, X_3\}$ and equal $N > 3$, let $\beta = 1/(N-3)$. Querying them in this order gives expected risk $\beta v_3(1 + 2\beta)$; reversing them gives $\beta v_1(1 + 2\beta)$. Learning first from the expert with the larger unique-signal variance is therefore better. The two orders cover identical information and share X_2. Their learning risks differ because the first regression has different amounts of unpredictable private variation at that stage.

For known ownership and fixed sample sizes, write T_t for the nonnegative linear map in (74). The worst expected risk over positive variance vectors with unit sum is

$$\sup_{\boldsymbol{v} > 0:\, \sum_j v_j = 1} \mathbf{1}^\top T_k \cdots T_1 \boldsymbol{v} = \max_j \sum_\ell (T_k \cdots T_1)_{\ell j}. \tag{75}$$

Thus a proposed order and sample allocation can be evaluated without fitting predictors or knowing the actual variances. This is an exact design criterion for the specified OLS learner and known ownership, rather than an optimum over all learning algorithms.

An overlapping Gaussian learning problem. Let $W, U_1, \ldots, U_n$ be independent $N(0, 1/(n+1))$ variables, let $Y = W + \sum_i U_i$, and let expert i observe (W, U_i). In order $1, \ldots, n$, fit a zero-intercept ordinary least-squares predictor on $N > 3$ fresh labeled examples of its actual residual. Training provides the residual labels; at prediction time each fitted report uses only its two local features. Neither the report functions nor their error tolerances are supplied by an oracle. At each stage the training coordinator evaluates earlier fitted reports on the fresh examples, subtracts their sum from the observed outcome, and sends these scalar residual labels to the next local regression. Earlier predictions affect supervision but are never included as prediction-time features. This distinction preserves the original information projections while making the residual targets available from labeled data.

Theorem 20 (Exact finite-sample aggregation risk)**.** *For this learning procedure, put $\beta = 1/(N-3)$. Its expected population risk is*

$$\mathbb{E}_{\mathcal{T}} \left\| Y - \textstyle\sum_i \widehat{f}_i \right\|^2 = \frac{\beta(1+2\beta)}{n+1} \sum_{j=1}^{n-1} j(1+\beta)^{j-1} \le \frac{n(n-1)}{2(n+1)(N-3)} \exp\left(\frac{n}{N-3} \right). \tag{76}$$

For $n = 1$ the risk is zero. For $n \ge 2$ and $n/(N-3) \to 0$, the risk is asymptotic to $n(n-1)/(2(n+1)(N-3))$, whereas exact ordinary reports retain Bayes risk c_n.

Proof. Let Q_t be the expected sum of private fitting-error energies through node t. Before fitting node t, the unpredictable residual has conditional variance $(n-t)/(n+1)$ plus the realized earlier private-error energies. Gaussian least squares contributes β times this variance to each of the shared and private coefficient errors. Thus

$$Q_0 = 0, \qquad Q_t = (1+\beta) Q_{t-1} + \frac{\beta(n-t)}{n+1}. \tag{77}$$

Table 4: Population risk at $N = 128$ fresh examples per expert. Means and one standard error over 1,000 runs; the ordinary Bayes decoder knows the fitted coefficients and covariance.

n	Residual OLS	Exact expectation	Ordinary mean	Ordinary Bayes
4	0.0097 (0.0003)	0.0099	0.4523 (0.0007)	0.3565 (0.0013)
8	0.0255 (0.0005)	0.0261	0.6820 (0.0005)	0.5611 (0.0014)
16	0.0616 (0.0009)	0.0618	0.8279 (0.0003)	0.5761 (0.0047)
32	0.1426 (0.0015)	0.1435	0.9107 (0.0002)	0.2743 (0.0066)

Later nodes remove each shared error except the last, so final expected risk is $(1 + 2\beta)Q_{n-1}$. Expanding the recurrence gives (76); Appendix K derives the regression expectation and bound. □

For desired expected risk $\delta \in (0, 1)$, $N \geq 3 + \lceil en/(2\delta) \rceil$ is sufficient. The procedure then uses $nN = O(n^2/\delta)$ labeled examples across stages. This is a sufficient sample budget for the stated learner, not a sample-complexity lower bound. It exposes a cost omitted by fixed tolerances: local regression error depends on the unexplained variance of its actual target, so the accuracy of an earlier fit changes the statistical task faced by later experts.

When richer queries are harder. Different query difficulties can also be compared: residual tolerance energy $\mathcal{E}_R$ beats the ordinary worst-case guarantee whenever $\mathcal{E}_R < \mathcal{E}_P + (1 - \mathcal{E}_P)c_n$, where $\mathcal{E}_P < 1$ is the ordinary tolerance energy. Richer queries may therefore be less precise and still have a provable advantage despite the extra error introduced by answering them.

Locality after training and at prediction time. Random training does not violate locality after the fitted functions are fixed. A corruption u_t invisible to the reporting expert does: writing $e_t = \ell_t + u_t$, with $\ell_t = P_t e_t$, $\|\ell_t\| \leq \eta_t$, and $\|u_t\| \leq \zeta_t$, gives

$$\|Y - A_k\| \leq \sqrt{\|FY\|^2 + \sum_t \eta_t^2} + \sum_t \zeta_t. \tag{78}$$

A shared invisible corruption contributes coherently, whereas independent centered channel noises contribute their variances when they are independent of training, signals, and one another. The coherent term matches the homogeneous local bound at $\zeta = \eta/\sqrt{n}$. These describe different error mechanisms, not a universal threshold for all random prediction errors.

Proposition 21 (Sharp joint propagation radius)**.** *For local and nonlocal budgets (η_t, ζ_t), define $s_0 = 0$ and*

$$s_t^2 = (s_{t-1} + \zeta_t)^2 + \eta_t^2. \tag{79}$$

Under arbitrary orthogonal projections, the propagated error satisfies $\|(Y - A_k) - FY\| \leq s_k$. The bound is jointly attainable on independent private Gaussian information with full coverage and an additional unobserved channel variable.

To see the improvement over (78), the new local error is orthogonal both to the previous error after projection and to the new nonlocal component. Only the latter can align with previous surviving error. The recursion retains this temporal distinction instead of adding both budgets at the end. Appendix L proves the bound and joint attainment. The result is sharp for residual summation; optimizing over all decoders defines a more general minimax problem.

Q Residual ridge regression and stability

Q.1 Statistical model and ridge coefficients

Let $Z = (Z_1, \ldots, Z_m)$ have independent coordinates with a common distribution μ symmetric about zero and satisfying $\mathbb{E}Z_j^2 = 1$. The target is $Y = \theta^\top Z$ for a fixed unknown vector θ with $\|\theta\|_2^2 = 1$. An expert with coordinate set J predicts using Z_J. A fixed sequence of such sets determines the training order. Each node receives N fresh labeled examples, independent of all earlier training data, and fits the actual residual after subtracting the previously fitted reports. Throughout the ridge analysis, $N \geq 1$, $\lambda > 0$, and each local set has size $1 \leq d \leq m$.

For a local design matrix $G \in \mathbb{R}^{N\times d}$ and residual labels $y \in \mathbb{R}^N$, the fitted coefficient vector is

$$\widehat{b} = \arg\min_{a\in\mathbb{R}^d} \left\{ \|y - Ga\|_2^2 + \lambda \|a\|_2^2 \right\} = (G^\top G + \lambda I_d)^{-1} G^\top y. \tag{80}$$

Thus λ penalizes the sum of squared errors. Define $S = G^\top G$ and $A = (S + \lambda I_d)^{-1}$, and set

$$\alpha = \lambda^2 \mathbb{E}[A_{11}^2], \qquad \gamma = \lambda^2 \mathbb{E}[A_{12}^2], \qquad \beta = \mathbb{E}[(ASA)_{11}], \tag{81}$$

with $\gamma = 0$ when $d = 1$. These constants depend on (μ, N, d, λ) and describe the expected propagation of coefficient energy. The learner computes (80) from its local design and residual labels; it does not need the constants in (81) or the target coefficients θ.

All three expectations are finite under the stated second-moment assumption. Indeed, $0 \preceq A \preceq \lambda^{-1} I_d$, and the eigenvalues of ASA have the form $s/(s+\lambda)^2 \le 1/(4\lambda)$ for $s \ge 0$. Ridge therefore remains well-defined for discrete or singular local designs.

Q.2 Proof of the exact ridge energy recurrence

We prove Theorem 8 by conditioning on the complete training history before the current node. The fitted residual is then a fixed linear function $r = b^\top Z$. For the current set J, write the fresh residual labels as

$$y = Gb_J + \varepsilon, \qquad \sigma_{\text{out}}^2 = \sum_{j\notin J} b_j^2.$$

Independence and centering of the coordinates imply

$$\mathbb{E}[\varepsilon \mid G, \text{history}] = 0, \qquad \mathbb{E}[\varepsilon\varepsilon^\top \mid G, \text{history}] = \sigma_{\text{out}}^2 I_N.$$

The unobserved coordinates supply regression noise independent of the local design. Fresh examples make this statement valid after conditioning on all preceding fits.

Subtracting the new fitted report changes the owned coefficients to

$$b'_J = b_J - AG^\top y = \lambda A b_J - AG^\top \varepsilon, \tag{82}$$

and leaves $b'_j = b_j$ for $j \notin J$. Conditional expectation first over ε gives, for $i \in J$,

$$\mathbb{E}[(b'_i)^2 \mid G, \text{history}] = \lambda^2 (Ab_J)_i^2 + \sigma_{\text{out}}^2 (ASA)_{ii}. \tag{83}$$

This calculation uses the conditional covariance of ε, without a Gaussian assumption.

Column sign symmetry removes every mixed coefficient term in (83). For any diagonal sign matrix D, the designs G and GD have the same distribution, and their inverse matrices satisfy $A \mapsto DAD$. Hence

$$\mathbb{E}[A_{ij}A_{i\ell}] = 0 \qquad (j \neq \ell),$$

because changing the sign of column j changes the sign of this product. Permutation symmetry of the columns makes all diagonal square moments equal and all off-diagonal square moments equal. Averaging (83) over G consequently yields

$$\mathbb{E}[(b'_i)^2 \mid \text{history}] = \alpha b_i^2 + \gamma \sum_{j\in J\setminus\{i\}} b_j^2 + \beta \sum_{j\notin J} b_j^2, \qquad i \in J. \tag{84}$$

The required moments are finite by the bounds following (81), which justifies averaging the sign and permutation identities over the current random design matrix.

Let $b^{(t)}$ denote the residual coefficient vector after node t, and define $x_{tj} = \mathbb{E}[(b_j^{(t)})^2]$, with expectation over training. Applying the tower property to (84) proves (20), with $x_{0j} = \theta_j^2$ and unchanged unowned coordinates. For an independent prediction instance,

$$\left\| (b^{(t)})^\top Z \right\|_{L^2}^2 = \sum_j (b_j^{(t)})^2.$$

Thus the expected population risk equals $\sum_j x_{tj}$. Dependence among coefficients fitted at different nodes is retained through the conditioning argument.

Q.3 Proof of the coverage-independent stability threshold

Fix a full sweep $J_1, \dots, J_q$ whose union is $[m]$, using the same (d, N, λ) at every node. Let M_J be the nonnegative matrix implementing (20), and let

$$T = M_{J_q} \cdots M_{J_1}, \qquad c = \alpha + (d-1)\gamma + (m-d)\beta.$$

Each updated row of M_J sums to c, while each unowned row is an identity row. This common row sum determines the stability criterion in Theorem 9.

Suppose first that $c < 1$. Start a sweep at the all-ones vector $\mathbf{1}$. Every intermediate coordinate stays at most one. An updated coordinate becomes at most c; if it is updated again, the same upper bound holds. Full coverage therefore gives

$$T\mathbf{1} \le c\mathbf{1}, \qquad \|T\|_\infty \le c, \qquad \rho(T) \le c < 1. \tag{85}$$

Repeating the sweep K times and using nonnegativity yields

$$T^K x_0 \le c^K \|x_0\|_\infty \mathbf{1}, \qquad \mathbb{E}\left\|r^{(K)}\right\|_{L^2}^2 \le m c^K \max_j \theta_j^2 \le m c^K. \tag{86}$$

This proves consistency uniformly over unit-norm target coefficients.

When $c = 1$, every M_J is row-stochastic. Hence $T\mathbf{1} = \mathbf{1}$ and $\rho(T) = 1$. Choosing the fixed target $\theta_j = 1/\sqrt{m}$ gives $x_0 = \mathbf{1}/m$ and $T^K x_0 = x_0$ for every K, so the expected risk stays exactly one.

When $c > 1$, starting from $\mathbf{1}$ keeps all intermediate coordinates at least one. Every updated coordinate becomes at least c, and full coverage gives $T\mathbf{1} \ge c\mathbf{1}$. Induction gives $T^K\mathbf{1} \ge c^K\mathbf{1}$, which implies $\rho(T) \ge c > 1$. More generally, for any strictly positive initial energy vector,

$$\mathbf{1}^\top T^K x_0 \ge m c^K \min_j x_{0j} \longrightarrow \infty. \tag{87}$$

In particular, the same fixed target $\theta_j = 1/\sqrt{m}$ has expected risk at least c^K.

These three cases prove (21). The threshold $c < 1$ is exact for every fixed covering pattern with the stated uniform node parameters. The spectral radius $\rho(T)$ and the precise convergence rate can still depend on the overlap pattern and order; (86) gives a uniform upper bound.

Q.4 A distribution-free choice of ridge penalty

A penalty based on the number of unobserved coordinates suffices to place every admissible feature distribution in the stable regime.

Proposition 22 (A sufficient ridge penalty)**.** *Under the assumptions of Theorem 8, the penalty condition $2\lambda \ge m - d$ guarantees stability for every covering sweep.*

Proof. Permutation symmetry and $A = A^\top$ give

$$c_{\text{in}} := \alpha + (d-1)\gamma = \frac{\lambda^2}{d}\mathbb{E}\,\mathrm{tr}(A^2), \qquad \beta = \frac{1}{d}\mathbb{E}\,\mathrm{tr}(SA^2).$$

Since A is a function of S, the matrices commute, and

$$I_d - \lambda^2 A^2 = (S^2 + 2\lambda S)A^2.$$

Taking normalized expected traces shows

$$1 - c = \tau + \big(2\lambda - (m-d)\big)\beta, \qquad \tau = \frac{1}{d}\mathbb{E}\,\mathrm{tr}(S^2A^2). \tag{88}$$

Each eigenvalue contributing to τ is $s^2/(s+\lambda)^2 \in [0,1]$. Moreover, $\mathbb{E}\,\mathrm{tr}(S) = Nd > 0$, so S has a positive eigenvalue with positive probability. Consequently $\tau > 0$, establishing the strict inequality $1 - c_{\text{in}} > 2\lambda\beta$. If $2\lambda \ge m - d$, both terms on the right of (88) are nonnegative with a strictly positive first term, proving stability for the complete covering sweep. □

This choice requires no distributional moments beyond standardization. Increasing λ sufficiently guarantees stability, while the contraction rate remains determined by c and the sweep matrix.

Q.5 Gaussian least squares and Rademacher thresholds

For Gaussian zero-intercept least squares, take $\lambda = 0$ and $N > d + 1$. The local design has full column rank almost surely, and

$$\mathbb{E}[(G^\top G)^{-1}] = \frac{I_d}{N - d - 1}.$$

One direct derivation uses the Schur complement: the jth diagonal entry of $(G^\top G)^{-1}$ is the reciprocal squared length of column j after projection away from the other $d - 1$ columns. Conditional on those columns, that squared length is χ^2_{N-d+1}, whose reciprocal expectation is $1/(N - d - 1)$. Sign symmetry makes the expected off-diagonal entries zero. The unpenalized residual update has no coefficient bias, so

$$\alpha = \gamma = 0, \qquad \beta = \frac{1}{N - d - 1}, \qquad c = \frac{m - d}{N - d - 1}. \tag{89}$$

Within the domain $N > d + 1$, stability is therefore equivalent to $N > m + 1$. For $m > d$, equality $N = m + 1$ gives the critical regime. If $m = d$, the target is locally determined and the fitted residual vanishes after one full-information least-squares step; (89) gives $c = 0$ on its stated sample domain.

For a single Rademacher coordinate per node, $d = 1$ and $G^\top G = N$ deterministically. Hence

$$\alpha = \frac{\lambda^2}{(N + \lambda)^2}, \qquad \beta = \frac{N}{(N + \lambda)^2}, \qquad c = \frac{\lambda^2 + (m - 1)N}{(N + \lambda)^2}. \tag{90}$$

Because $N \geq 1$, comparison with one gives the exact threshold

$$c < 1 \quad \Longleftrightarrow \quad N + 2\lambda > m - 1.$$

Equality produces unit spectral radius, and the reverse strict inequality produces divergence for the fixed uniform-energy target. A fixed ridge penalty can thus make repeated correction consistent without increasing the number of fresh examples per query.

Q.6 Connection with partial-update affine projection

The identity

$$(G^\top G + \lambda I_d)^{-1} G^\top = G^\top (G G^\top + \lambda I_N)^{-1}$$

expresses the local ridge correction in the dual affine-projection form. Recent work analyzes the transient behavior of selective partial-update LMS (Siqueira et al., 2024); selective affine-projection algorithms also have established mean-square analyses (Abadi & Palangi, 2010). The procedure analyzed here has the latter update structure with regularization. The derivation specializes to independent fresh batches and iid symmetric coordinates: design symmetry removes all coefficient cross moments, leaving the three scalar coefficients in (81). Common updated-row sums then give an exact criterion for each fixed covering sweep and the explicit batch-size and penalty thresholds. These assumptions permit a closed finite-sample calculation; Theorem 10 separately handles correlated Gaussian OLS through a matrix recurrence, while general nonstationary regressors remain outside the present learning model.

R Repeated learning: configurations and trajectory diagnostics

We repeat complete ownership sweeps to examine expected-risk stability and the distribution of fitted trajectories. The experiment uses $m = 9$ independent standardized primitives with either a standard Gaussian or Rademacher law and target $Y = \frac{1}{3}\sum_{j=1}^{9} Z_j$. The star visits $\{1, j\}$ for $j = 2, \ldots, 9$, giving eight local fits per sweep. The ring visits $\{j, j + 1\}$ for $j = 1, \ldots, 8$ and then $\{9, 1\}$, giving nine fits per sweep. Both patterns cover every coordinate with two-feature local learners.

Each fit uses N fresh complete examples, independent of all earlier training. Its response is the actual residual after evaluating every previously fitted report on those examples. Gaussian learners use zero-intercept OLS with $N \in \{8, 10, 16\}$. Rademacher learners use zero-intercept ridge with $N = 8$ and $\lambda \in \{0.25, 1, 4, 16\}$, minimizing $\|y - Gb\|_2^2 + \lambda\|b\|_2^2$. We run 12 sweeps for each of 1,000 independent training trajectories. At every node, population MSE is the squared Euclidean norm of the residual coefficient vector. The reported variability therefore comes from training.

Table 5: Population risk after 12 sweeps, with 1,000 independent training trajectories per configuration. The expectation comes from the exact energy recurrence. Mean and median summarize observed trajectories. The last column is the maximum individual risk as a percentage of the total population risk observed across all sampled training trajectories.

Law	Pattern	N	λ	Expectation	Mean	Median	Largest share
Gaussian	Star	8	0	4.31×10^{3}	2.53×10^{3}	1.16	46.9%
Gaussian	Star	10	0	1	0.237	0.00106	13.5%
Gaussian	Star	16	0	1.19×10^{-6}	5.40×10^{-7}	8.64×10^{-9}	16.9%
Rademacher	Star	8	0.25	0.267	0.119	0.00154	21.2%
Rademacher	Star	8	1	0.00249	0.0035	3.81×10^{-5}	60.8%
Rademacher	Star	8	4	7.58×10^{-6}	8.95×10^{-6}	8.70×10^{-7}	22.1%
Rademacher	Star	8	16	4.33×10^{-4}	4.30×10^{-4}	3.07×10^{-4}	1.1%
Gaussian	Ring	8	0	1.24×10^{4}	4.79×10^{3}	1.27	71.5%
Gaussian	Ring	10	0	1	0.134	7.00×10^{-4}	11.2%
Gaussian	Ring	16	0	2.11×10^{-7}	8.99×10^{-8}	6.71×10^{-10}	24.7%
Rademacher	Ring	8	0.25	0.225	0.0963	8.16×10^{-4}	17.6%
Rademacher	Ring	8	1	9.49×10^{-4}	7.60×10^{-4}	6.01×10^{-6}	22.9%
Rademacher	Ring	8	4	2.77×10^{-8}	3.13×10^{-8}	8.87×10^{-10}	12.0%
Rademacher	Ring	8	16	4.12×10^{-7}	3.91×10^{-7}	1.70×10^{-7}	2.7%

R.1 Expected risks and training costs

Expected risks are computed from products of the coordinate-energy transition matrices. Gaussian OLS uses $\alpha = \gamma = 0$ and $\beta = 1/(N-3)$. For two-feature Rademacher designs, the Gram matrix has diagonal N and off-diagonal $S = 2K - N$, where $K \sim \mathrm{Binomial}(N, 1/2)$. We evaluate the ridge coefficients by summing over all $N+1$ possible values of S with their binomial probabilities. We evaluate the sums in rational arithmetic and the matrix products and spectral radii in floating-point arithmetic. A separate enumeration of all 512 complete Rademacher datasets at $m = N = 3$ and $\lambda = 0.25$ agrees with the predicted coordinate energies to within 5.7×10^{-15}.

The common full owned-row sum is $c = \alpha + \gamma + 7\beta$. Gaussian OLS is unstable at $N = 8$, critical at $N = 10$, and stable at $N = 16$. All four Rademacher ridge settings are stable: c is 0.9453, 0.7659, 0.5197, and 0.5570 in increasing order of λ. The two ownership patterns share this threshold, but the ring contracts faster per sweep at $\lambda = 16$: its spectral radius is 0.2910, compared with 0.5268 for the star. The star uses $8N$ examples per sweep; the ring uses $9N$, reflecting its additional local fit.

R.2 Concentration across fitted trajectories

Table 5 reports every configuration after 12 sweeps. At the critical Gaussian setting, the exact expected risk remains one from the uniform initial energy vector, whereas the observed means are 0.237 for the star and 0.134 for the ring. The corresponding sample standard errors are 0.0487 and 0.0289. These discrepancies show that the empirical means and their standard errors do not reliably capture the expectation at this trajectory budget. Concentration is also pronounced in the unstable setting: one run contributes 71.5% of the ring's total observed risk at $N = 8$. The table retains the median and the largest-run share to expose this variation.

All first-sweep means are within 1.20 sample standard errors of their exact expectations, but several late-sweep means depart substantially, including some stable configurations. The $\lambda = 16$ Rademacher runs remain less concentrated: the largest trajectories contribute 1.1% and 2.7% of total observed risk for the star and ring, respectively. Their mean risks are 4.30×10^{-4} and 3.91×10^{-7}, compared with expectations 4.33×10^{-4} and 4.12×10^{-7}. The expectation curves characterize the proved stability criterion, while the trajectory summaries describe the finite training experiment.

Table 6: Configuration seeds and fresh training examples per trajectory over 12 sweeps. Each configuration contains 1,000 independently fitted trajectories. The last two columns give the total numbers of fresh examples used by the star and ring patterns.

Law	N	λ	Star seed	Ring seed	Star samples	Ring samples
Gaussian	8	0	2186910931	683571107	768	864
Gaussian	10	0	1528497176	2619801859	960	1080
Gaussian	16	0	676216922	604998489	1536	1728
Rademacher	8	0.25	3981068538	1125504861	768	864
Rademacher	8	1	2095245375	3429162737	768	864
Rademacher	8	4	1886044348	4197942453	768	864
Rademacher	8	16	301669058	1838440900	768	864

R.3 Reproduction

The experiment script is `experiments/repeated_ridge.py`. It uses NumPy's default random-number generator with base seed 20260914. Table 6 records the derived configuration seeds and the number of fresh examples used by each trajectory. The saved data include per-node population risks, sweep means, sample standard errors, quantiles, transition coefficients, and sweep matrices. The plotting script `experiments/plot_stability.py` reads these saved summaries. Increasing the number of sweeps increases training cost and can leave the estimated mean increasingly dominated by a small number of sampled training trajectories.

S Correlated Gaussian residual learning

Correlation changes the subspaces visible to the experts. The proof of Theorem 10 tracks the full second-moment matrix of the residual coefficients, including the cross moments induced by feature correlation and noncommuting information projections.

S.1 The exact second-moment update

Let $X \sim N(0, \Sigma)$ with $\Sigma \succ 0$, and let $Y = \theta^\top X$, where $\theta^\top \Sigma \theta = 1$. Fix an ordered cover $J_1, \dots, J_n$ of $[m]$ with $|J_j| = d < m$. Every node uses $N > d + 1$ new independent examples to fit its current residual by zero-intercept OLS. The same cover is repeated, with fresh samples at every node. Write $A = \Sigma^{1/2}$ for the symmetric positive square root. If the current residual is $a^\top X$, its coefficient in the globally whitened coordinates $Z = A^{-1}X$ is $b = Aa$. Its population risk is $\|b\|_2^2$.

Let E_J be the coordinate-selection matrix for J. The columns of AE_J span the expert's predictable linear subspace in whitened coordinates, whose orthogonal projector and complement are

$$P_J = AE_J\Sigma_{JJ}^{-1}E_J^\top A, \qquad C_J = I - P_J.$$

Choose $U \in \mathbb{R}^{m\times d}$ with orthonormal columns spanning this subspace, so $UU^\top = P_J$. Using $U^\top Z$ as the local predictors is an invertible change of coordinates from X_J, so the local regression returns exactly the same fitted prediction under this coordinate change.

Condition on all previous training data, making b fixed. The fresh response decomposes as

$$b^\top Z = (U^\top b)^\top U^\top Z + (C_J b)^\top Z.$$

The two terms use orthogonal Gaussian components. Consequently, the local design $G \in \mathbb{R}^{N\times d}$ has iid standard Gaussian entries, while the residual response vector ε is independent of G, conditionally centered, and has conditional covariance $\|C_J b\|_2^2 I_N$. The fitted coefficient and the new global residual coefficient satisfy

$$\widehat{\gamma} = U^\top b + \xi, \qquad \xi = (G^\top G)^{-1}G^\top \varepsilon, \qquad b' = C_J b - U\xi.$$

Thus, conditional on G and the past,

$$\mathbb{E}[\xi \mid G, b] = 0, \qquad \mathbb{E}[\xi\xi^\top \mid G, b] = \|C_J b\|_2^2 \, (G^\top G)^{-1}.$$

For completeness, the inverse-Gram expectation follows from a scalar calculation. Partition $G = [g_1, G_2]$ and let Π_2 project onto the columns of G_2. The Schur complement gives

$$((G^\top G)^{-1})_{11} = \left(g_1^\top (I - \Pi_2) g_1\right)^{-1}.$$

Conditional on G_2, the denominator has distribution χ^2_{N-d+1}. For $V \sim \chi^2_r$ with $r > 2$, integration against its density gives

$$\mathbb{E}[V^{-1}] = \frac{\Gamma(r/2 - 1)}{2\Gamma(r/2)} = \frac{1}{r-2}.$$

Permutation symmetry yields the same diagonal expectation for every column. Off-diagonal entries are integrable by the positive-semidefinite inequality $|B_{ij}| \le \sqrt{B_{ii} B_{jj}}$, and column sign symmetry makes their expectations zero. Hence, with $\beta = (N - d - 1)^{-1}$,

$$\mathbb{E}[(G^\top G)^{-1}] = \beta I_d.$$

The argument also applies to $d = 1$, taking $\Pi_2 = 0$.

This local calculation is also a specialization of the Gaussian sketch-and-solve covariance in Bartan & Pilanci (2023, Lemma 2.1): take the deterministic design to be U, the response to be b, and the Gaussian sketch to be the fresh sample matrix. The covering-product argument below determines what happens when these local fits repeatedly change one another's targets.

The conditional mean-zero identity removes the cross terms in $b'b'^\top$. Averaging first over the current sample and then over the past proves (24):

$$H' = \mathcal{M}_J(H) = C_J H C_J + \beta \operatorname{tr}(C_J H) P_J, \qquad H = \mathbb{E}[bb^\top]. \tag{91}$$

Here H is an uncentered second-moment matrix; the residual coefficient need not have mean zero. The map is positive and finite on positive semidefinite matrices. Starting from a fixed target, finiteness therefore propagates through every finite training sequence.

S.2 The covering-sweep threshold

Write $P_j = P_{J_j}$, $C_j = I - P_j$, and define

$$\mathcal{T} = \mathcal{M}_{J_n} \circ \cdots \circ \mathcal{M}_{J_1}, \qquad F = C_n \cdots C_1, \qquad c = \beta(m - d).$$

Since $\operatorname{tr}(C_j) = m - d$, each node map obeys

$$\mathcal{M}_{J_j}(I) = I + (c - 1) P_j.$$

Telescoping the composition yields

$$\mathcal{T}(I) - I = (c-1) D, \qquad D = \sum_{j=1}^{n} \mathcal{M}_{J_n} \circ \cdots \circ \mathcal{M}_{J_{j+1}}(P_j), \tag{92}$$

where an empty composition is the identity. For $H \succeq 0$, equation (91) gives $\mathcal{M}_{J_j}(H) \succeq C_j H C_j$. Applying positivity successively through each suffix gives

$$D \succeq \sum_{j=1}^{n} C_n \cdots C_{j+1} P_j C_{j+1} \cdots C_n = I - F F^\top. \tag{93}$$

The equality is the telescoping identity for the maps $H \mapsto C_j H C_j$, using $I - C_j^2 = P_j$. The order of every factor is preserved.

The cover implies $\bigcap_j \operatorname{ran}(C_j) = \{0\}$ because A is invertible and the columns of the matrices AE_{J_j} together span $\mathbb{R}^m$. Consequently, $\|F\|_{\mathrm{op}} < 1$. Indeed, if a unit vector v satisfied $\|Fv\|_2 = 1$, equality would hold at every orthogonal projection in the product. Each projection would then fix v, contradicting the trivial intersection. Compactness makes the strict inequality uniform over the unit sphere, yielding the required operator-norm bound.

Suppose first that $c < 1$. Multiplication of (93) by $c - 1$ reverses the order, so

$$0 \preceq \mathcal{T}(I) \preceq cI + (1 - c) F F^\top \preceq \kappa I, \qquad \kappa = c + (1 - c)\|F\|^2_{\mathrm{op}} < 1.$$

Positivity then gives $\mathcal{T}^K(I) \preceq \kappa^K I$. For any unit initial whitened coefficient $b_0 = A\theta$, we have $b_0 b_0^\top \preceq I$. The expected risk after K sweeps therefore satisfies

$$R_K(b_0) := \mathbb{E}\,\|b_K\|_2^2 = \operatorname{tr} \mathcal{T}^K(b_0 b_0^\top) \le m\kappa^K. \tag{94}$$

At $c = 1$, equation (92) gives $\mathcal{T}(I) = I$ and hence $\mathcal{T}^K(I) = I$. If $c > 1$, the same identity and (93) instead imply

$$\mathcal{T}(I) \succeq gI, \qquad g = 1 + (c-1)(1 - \|F\|_{\mathrm{op}}^2) > 1.$$

Thus $\mathcal{T}^K(I) \succeq g^K I$ in the latter case. Set $g = 1$ at the endpoint. To exhibit a fixed target that fails to converge, choose an orthonormal basis $e_1, \dots, e_m$ once and take $\theta_i = A^{-1}e_i$. Every θ_i has unit Σ-norm, and

$$\sum_{i=1}^{m} R_K(e_i) = \operatorname{tr} \mathcal{T}^K(I) \ge mg^K.$$

For each K, some index satisfies $R_K(e_i) \ge g^K$. One of the finitely many indices occurs for infinitely many K. For that fixed target, $\limsup_K R_K(e_i) \ge 1$ when $c = 1$, and $\limsup_K R_K(e_i) = \infty$ when $c > 1$. This proves necessity for consistency for every fixed unit target. Finally,

$$c < 1 \quad \Longleftrightarrow \quad N > m + 1,$$

which completes the threshold claim of Theorem 10 within its stated domain $N > d + 1$.

The spectral formulation follows on the finite-dimensional space of symmetric matrices. A positive linear map $\mathcal{L}$ has induced spectral-norm operator norm $\|\mathcal{L}\| = \|\mathcal{L}(I)\|_{\mathrm{op}}$: apply positivity to $-I \preceq H \preceq I$ and then take $H = I$. The preceding bounds imply $\rho(\mathcal{T}) \le \kappa < 1$ for $c < 1$, $\rho(\mathcal{T}) = 1$ for $c = 1$, and $\rho(\mathcal{T}) \ge g > 1$ for $c > 1$. Conversely, if every fixed unit target is consistent, summing over a fixed basis gives $\operatorname{tr} \mathcal{T}^K(I) \to 0$. Positivity implies that the operator powers tend to zero, which in finite dimension is equivalent to $\rho(\mathcal{T}) < 1$.

S.3 A SUFFICIENT TRAINING COST

For independent standard Gaussian features, the coordinate complement projections commute and a covering sweep has $F = 0$. The geometric bound specifies a sufficient batch size and a number of local residual fits that attain the desired prediction accuracy.

Corollary 23 (Fresh-sample OLS cost)**.** *Let $X \sim N(0, I_m)$, $1 \le d < m$, and let a fixed ordered cover contain n sets of size d. For any $\delta \in (0,1)$, choose*

$$N = 2m - d + 1, \qquad K = \left\lceil \log_2 \frac{m}{\delta} \right\rceil, \qquad L = nK, \qquad S = nNK. \tag{95}$$

Actual residual OLS with N fresh examples per node has expected population risk at most δ after L fitted reports, using S labeled examples, for every fixed $\|\theta\|_2 = 1$.

Proof. The integer choice satisfies $N - d - 1 = 2(m - d) > 0$, so the local second moments exist and $c = 1/2$. Since $F = 0$, equation (94) gives $R_K(\theta) \le m2^{-K} \le \delta$. Each of the K sweeps contains n fits, and every fit uses a separate batch of N examples. □

The same batch choice applies to correlated Gaussian features. Its geometric factor is $\kappa = (1 + \|F\|_{\mathrm{op}}^2)/2 < 1$, so replacing K in (95) by

$$K = \left\lceil \frac{\log(m/\delta)}{\log(1/\kappa)} \right\rceil$$

suffices, with the same formulas $L = nK$ and $S = nNK$. Correlation affects this cost through the geometry of the expert subspaces, while the stability threshold remains $N > m + 1$.

For comparison, Corollary 7 gives oracle minimax risk 2^{-K} at $L = nK$ under homogeneous projected-relative tolerance $\eta^2 = 1/2$. Corollary 23 gives the learned iid Gaussian procedure the upper bound $m2^{-K}$ at that report count. These bounds have the same geometric factor across sweeps, but quantify different error classes: the oracle condition constrains every admissible response, whereas the OLS bound averages over training samples. The corollary is a sufficient cost for this fitting procedure, without a sample-optimality claim or a pathwise projected-relative guarantee.

What the resource counts measure. The count L records local residual-fitting nodes; S counts unique fresh labeled examples. Both exclude the additional evaluation of earlier predictors needed to construct labels. Since the fitted functions are linear, the coordinator can retain one accumulated predictor per expert when forming later targets. The final reports from repeated visits can likewise be combined locally before deployment. Accordingly, L is not a lower bound on deployment communication, and S is not a lower bound for more general distributed training. If all raw coordinates are centrally available, a full-rank noiseless pooled design already recovers the target coefficients.

Gaussian correlation curves. Figure 3(a) uses $m = 9$, star ownership $J_i = \{1, i+1\}$, and covariance $\Sigma_{ij} = a^{|i-j|}$ at $a = 0$ and $a = 0.8$. Its curves are spectral radii of the exact full-sweep map (24), not empirical measurements. The linear map is represented on matrix entries using $C_J \otimes C_J + \beta \operatorname{vec}(P_J) \operatorname{vec}(C_J)^\top$, then composed in query order. This calculation keeps the feature correlation while retaining the same ownership and local sample count.

T A PAIRED RIDGE SWEEP UNDER A FIXED FEATURE LAW

We vary the ridge penalty while holding the primitive law, sample size, target, and ownership pattern fixed. The primitives are $m = 9$ independent Rademacher variables and the target is $Y = \frac{1}{3}\sum_{j=1}^{9} Z_j$. Every learner owns two coordinates and uses $N = 4$ fresh complete examples. We compare $\lambda \in \{0.25, 1, 2, 4, 16\}$ under both ownership patterns. The star visits $\{1, j\}$ in increasing order for $j = 2, \ldots, 9$; the ring visits $\{j, j+1\}$ for $j = 1, \ldots, 8$, followed by $\{9, 1\}$.

Each configuration contains $1{,}000$ independent training trajectories of 12 complete sweeps. Within a pattern, corresponding runs use the same full training batch at each stage for all five penalties. Batches are independent across stages and runs, and the two patterns use separate random streams. Thus the penalty comparisons are paired, while each marginal trajectory follows the fresh-data protocol. With current residual coefficient vector r, a learner forms the actual response $y = Zr$ on its newly drawn batch and fits

$$\widehat{b} = (G^\top G + \lambda I_2)^{-1} G^\top y.$$

Here G consists of the two owned columns of Z. The learner subtracts $\widehat{b}$ from those residual coefficients, implementing zero-intercept ridge with summed squared loss. The initial residual is $r_0 = \mathbf{1}/3$, and its population risk is one. At each stage we evaluate population risk as $\|r\|_2^2$, so the measured variation comes entirely from training.

Stability crossing. For two-feature Rademacher designs, $G^\top G$ has diagonal entries 4 and off-diagonal entry $S = 2K - 4$, where $K \sim \mathrm{Binomial}(4, 1/2)$. We compute α, γ, β by the exact finite binomial sum, using rational arithmetic, then evaluate energy-matrix products and eigenvalues in floating-point arithmetic. In increasing penalty order, the common full owned-row sums are

$$c = \alpha + \gamma + 7\beta \in \{1.660238,\ 1.177536,\ 0.894271,\ 0.704896,\ 0.724863\}.$$

Both ownership patterns therefore cross from expected-risk instability at $\lambda = 1$ to stability at $\lambda = 2$. For the star, the sweep spectral radius decreases from 1.375915 to 0.824979; for the ring it decreases from 1.470465 to 0.768549. Increasing the penalty further need not accelerate contraction: both patterns have a larger spectral radius at $\lambda = 16$ than at $\lambda = 4$.

Finite trajectory distributions. Table 7 retains all ten configurations. At $\lambda = 0.25$, the star mean is 2.41×10^5 against expectation 4.03×10^5, and the ring mean is 1.18×10^6 against expectation 2.29×10^6. The largest sampled trajectory supplies 23.4% and 34.4% of their respective observed totals. Concentration also occurs on the stable side: at $\lambda = 2$, the ring mean is 0.0305, its expectation is 0.0439, and its median is 0.000553. The exact recurrence establishes the expected-risk crossing; the observed means and sample standard errors can underrepresent the expectation at this trajectory budget. The medians and largest-run shares describe the concentration of the measured trajectories.

Cost and reproduction. The star performs eight fits and consumes 32 fresh examples per sweep; the ring performs nine fits and consumes 36. Over 12 sweeps, each trajectory uses 384 or 432 fresh examples, respectively. Across all penalties and trajectories, the experiment performs $1{,}020{,}000$ ridge

Table 7: Paired Rademacher ridge experiment with $N = 4$, after 12 sweeps. Each configuration has 1,000 training trajectories. Expectation is computed from the exact energy recurrence; mean and median summarize measured population risks. Largest share is the maximum individual risk divided by the sum of all observed risks. Within each ownership pattern, each penalty uses the same feature and outcome pairs while forming its own current residual labels.

Pattern	λ	c	Expectation	Mean	Median	Largest share
Star	0.25	1.6602	4.03×10^{5}	2.41×10^{5}	1.10×10^{3}	23.4%
Star	1	1.1775	44.6	30.8	0.49	17.6%
Star	2	0.8943	0.101	0.0654	0.00327	9.3%
Star	4	0.7049	0.00272	0.00211	4.53×10^{-4}	3.9%
Star	16	0.7249	0.015	0.0147	0.0118	0.7%
Ring	0.25	1.6602	2.29×10^{6}	1.18×10^{6}	2.55×10^{3}	34.4%
Ring	1	1.1775	98	59	0.419	19.8%
Ring	2	0.8943	0.0439	0.0305	5.53×10^{-4}	16.8%
Ring	4	0.7049	6.39×10^{-5}	5.77×10^{-5}	3.89×10^{-6}	16.7%
Ring	16	0.7249	3.47×10^{-4}	3.67×10^{-4}	2.26×10^{-4}	2.2%

fits. Pairing requires 816,000 distinct example draws, which account for 4,080,000 example uses across the five penalties. Rates per sweep therefore compare different training costs across patterns.

The script `experiments/matched_ridge.py` uses NumPy PCG64 with base seed 20260915. Seeds derived by `SeedSequence([base, pattern])`, with pattern indices 0 and 1, are 1157426631 for the star and 1072642311 for the ring. The recorded run used Python 3.13.9 and NumPy 2.5.2 on a CPU, with numerical-library thread counts set to one. Its measured simulation and output time was 4.01 seconds, excluding the subsequent archive validation. The saved summaries include every sweep's mean, sample standard error, median, quantiles, maximum, and exact expectation; the compressed trajectory file retains every node's population risk and run identifier. Validation exhaustively enumerates all 4,096 complete Rademacher datasets at $N = 4, m = 3$ for each penalty, giving coordinate-energy discrepancies below 3.6×10^{-14}. Reaggregation of the 1,020,000 saved node risks agrees with all reported means and quantiles in the sweep summaries.

U Allocating a fixed label budget

We fix $B = 2048$ unique labeled examples and compare allocations across local sample size and repeated sweeps. The population has $m = 9$ independent unit-variance Gaussian coordinates, target $Y = \theta^\top Z$ with $\theta = (1, \dots, 1)/3$, and eight experts with coordinate sets $J_i = \{1, i + 1\}$, $i = 1, \dots, 8$. The coefficients are unknown to the fitted regressions. Each node fits zero-intercept OLS to its two coordinates and the actual residual labels, obtained by subtracting previous fitted predictions. Experts run in the order $1, \dots, 8$ during every sweep. The six fresh-data schedules use $N \in \{8, 16, 32, 64, 128, 256\}$ examples per node and $K = 256/N$ complete sweeps, so $L = 8K$ fitted nodes consume exactly $LN = 2048$ fresh labels.

We run 1000 independent repetitions. Within a repetition, all configurations use the same ordered stream of 2048 full Gaussian examples; fresh schedules partition that stream into consecutive blocks. Population risk is evaluated directly as $\|\theta - \widehat{\theta}\|_2^2$, with no holdout sampling. For fresh data, the exact expected risk is computed by successive coordinate-energy transport: a node replaces each owned coordinate energy by the sum of unowned energies divided by $N - 3$ and leaves the other energies at their current values throughout that local fitting step.

Table 8 exhibits an allocation tradeoff: $N = 32$, $K = 8$ has the smallest exact expected risk among the tested schedules, whereas increasing N further leaves fewer correction sweeps. At $N = 8$, the unstable transport produces expected risk 5.60×10^9, although the 1000-run mean is only 4.75×10^5. The largest run contributes 72.1% of the observed total, and the median is 2.22. Thus, in this regime, the finite-run mean and its sample SE do not resolve the expected risk. The $N = 16$ schedule also has appreciable concentration, with one trajectory contributing 34.2% of the observed total.

For comparison, one additional configuration reuses the same 2048-example set at all eight nodes of one sweep. It preserves the two-coordinate regression access at each expert, but changes data

Table 8: Final population risk at a budget of 2048 unique labels. Mean and SE use 1000 repetitions. Share is the largest trajectory's fraction of the sum of final risks. L counts fitted nodes. The reused-data row has eight times the example-node visits of every fresh-data row.

Data	N	K	L	Exact expectation	Mean	SE	Median	Share
Fresh	8	32	256	5.60×10^{9}	4.75×10^{5}	3.45×10^{5}	2.22	72.1%
Fresh	16	16	128	1.17×10^{-8}	7.78×10^{-9}	3.26×10^{-9}	1.63×10^{-11}	34.2%
Fresh	32	8	64	4.83×10^{-9}	3.94×10^{-9}	5.04×10^{-10}	4.54×10^{-10}	6.70%
Fresh	64	4	32	1.85×10^{-6}	2.01×10^{-6}	1.74×10^{-7}	6.14×10^{-7}	5.64%
Fresh	128	2	16	4.60×10^{-4}	5.06×10^{-4}	1.83×10^{-5}	3.27×10^{-4}	1.17%
Fresh	256	1	8	1.26×10^{-2}	1.27×10^{-2}	2.39×10^{-4}	1.12×10^{-2}	0.421%
Reused	2048	1	8	—	1.55×10^{-3}	2.83×10^{-5}	1.38×10^{-3}	0.444%

availability by allowing every node to revisit the labels. It processes 16384 example-node pairs and 32768 local feature values, compared with 2048 pairs and 4096 local feature values for each fresh schedule. These counts exclude previous predictor evaluations used to form residual labels. Its mean risk is 1.55×10^{-3}; the fresh-data expectation formula does not apply because fitted residuals depend on the reused examples. This experiment compares allocations and reuse within local sequential fitting; it does not establish an optimal learning rate over all estimators.

The implementation is `experiments/fixed_budget.py`. Run r uses NumPy's `SeedSequence([20260916,r])` to generate its ordered 2048×9 Gaussian array, which is shared across the seven configurations. The recorded run used Python 3.13.9 and NumPy 2.5.2 on CPU. The directory `experiments/results/fixed_budget/` contains the final summaries, every fitted node's population risk, and the sample and processing counts for each schedule.

V Heterogeneous relative precision and query order

This section proves Theorem 6. Expert i has a known tolerance $\eta_i \geq 0$, and every visit to that expert satisfies

$$\|e_t\| \leq \eta_i \|P_i H_t\|, \qquad P_i e_t = e_t.$$

The minimax class consists of the priors in (1), arbitrary repeated-expert linear target graphs with finite coefficients, and measurable decoders. The independent public seed fixes the graph before any signal is observed. The adversary chooses one prior before the seed. Its local error functions may depend on the seed. Write $\rho_i = \min\{\eta_i^2, 1\}$ and let

$$\mathcal{C}_L = \left\{ N \in \mathbb{Z}_+^n : \sum_i N_i \leq L \right\}.$$

We first identify the exact allocation problem,

$$\mathcal{R}^*_{L,\mathrm{rel},\mathrm{rep}}(\boldsymbol{\eta}) = \min_{\mu \in \Delta(\mathcal{C}_L)} \max_i \mathbb{E}_\mu \rho_i^{N_i}, \qquad 0^0 := 1, \tag{96}$$

and then evaluate it through the piecewise-linear formula in (15). The upper bound requires an order that controls how local errors move between overlapping information spaces.

V.1 A private-information lower bound for every graph

Fix a protocol and denote its seed-dependent visit counts by N_i. Choose any $v_i \geq 0$ with $\sum_i v_i = 1$. Give expert i independent private Gaussian primitives $Z_{i,1}, \ldots, Z_{i,L}, V_i$ with variances

$$\mathrm{Var}(Z_{i,h}) = v_i(1-\rho_i)\rho_i^{h-1}, \qquad \mathrm{Var}(V_i) = v_i \rho_i^L.$$

All primitives across experts are independent. Set $X_i = \sum_{h=1}^{L} Z_{i,h} + V_i$ and $Y = \sum_i X_i$. Then $\mathrm{Var}(Y) = 1$. For $0 \leq h \leq L$, define the tail

$$W_{i,h} = \sum_{j=h+1}^{L} Z_{i,j} + V_i, \qquad \mathrm{Var}(W_{i,h}) = v_i \rho_i^h.$$

This prior is fixed for every seed and every realized graph of the protocol.

Condition on a seed. At the hth visit to expert i, choose $e_t = -a_t W_{i,h}$. Inductively, all preceding reports by i depend linearly on its first $h-1$ increments. Reports from other experts are centered and independent of $\mathcal{F}_i$, so their conditional means vanish. Consequently,

$$P_i H_t = a_t W_{i,h-1} + M_t,$$

where M_t depends linearly on the first $h-1$ increments and is independent of $W_{i,h-1}$. Hence

$$\|P_i H_t\|^2 \geq a_t^2 v_i \rho_i^{h-1}, \qquad \|e_t\|^2 = a_t^2 v_i \rho_i^h \leq \eta_i^2 \|P_i H_t\|^2 .$$

The new report is $M_t + a_t Z_{i,h}$, completing the induction. The transcript therefore reveals at most the first N_i increments of each expert. Its remaining independent target component has variance $\sum_i v_i \rho_i^{N_i}$. Conditional variance bounds the risk of every measurable decoder, giving

$$\mathbb{E}[(g_Z(R) - Y)^2] \geq \sum_i v_i \mathbb{E}_Z \rho_i^{N_i}. \tag{97}$$

Choose v concentrated on an expert maximizing $\mathbb{E}_Z \rho_i^{N_i}$. This choice uses the protocol's distribution of counts, not its realized seed. Taking the infimum over protocols proves the lower bound in (96). The argument also covers $\rho_i = 0$ and $\rho_i = 1$ through degenerate Gaussian increments.

V.2 An attaining order on overlapping information

Assume first that every $\rho_i < 1$. For a fixed count vector N, form one token (i,k) for each $1 \leq k \leq N_i$, with level ρ_i^k. Sort the tokens by increasing level and query their expert indices in that order. Break ties within an expert by decreasing k, and all remaining ties deterministically. At each node use the actual residual target and add the report to the aggregate. The following bound is uniform over admissible local errors:

$$\|Y - A\|^2 \leq \sum_\ell \mathrm{Var}(X_\ell) \min_{i \in T_\ell} \rho_i^{N_i}. \tag{98}$$

To prove it, work backward from the terminal quadratic potential $Q = I$, with scalar bounds $w_i = 1$. We maintain

$$Q \succeq 0, \qquad QP_i = P_i Q, \qquad P_i Q P_i \preceq w_i P_i \quad \text{for every } i. \tag{99}$$

Here Q is a bounded operator on the full population L^2 space. The projections commute on that space by (25). No linearity of the error functions is required.

Consider one reversed query to expert i, and put $C_i = I - P_i$. Its forward residual update is $r' = C_i r - e$, where $P_i e = e$ and $\|e\|^2 \leq \rho_i \|P_i r\|^2$. Since Q commutes with P_i, the two terms are orthogonal in the quadratic form defined by Q. Thus

$$\begin{aligned} \langle r', Q r' \rangle &= \langle C_i r, Q C_i r \rangle + \langle e, Qe \rangle \\ &\leq \big\langle r, \big(C_i Q C_i + \rho_i w_i P_i\big) r \big\rangle . \end{aligned} \tag{100}$$

Replace Q by $Q' = C_i Q C_i + \rho_i w_i P_i$ and w_i by $w_i' = \rho_i w_i$. Positivity and commutation are preserved, and $P_i Q' P_i = w_i' P_i$. For $j \neq i$, restriction to the P_i and C_i blocks of $\mathrm{ran}(P_j)$ gives

$$P_j Q' P_j \preceq w_j P_j \quad \text{whenever} \quad w_i' \leq w_j.$$

This condition follows from the token order. Backward processing visits the levels in nonincreasing order. When token (i,k) is processed, its expert's current bound is ρ_i^{k-1} and its new bound is ρ_i^k. Every other current bound is either one or a previously processed level, so it is at least ρ_i^k. This proves (99) throughout the backward recursion.

At the beginning of the forward protocol, the resulting potential Q_0 therefore satisfies $P_i Q_0 P_i \preceq \rho_i^{N_i} P_i$. Iterating (100) bounds the final risk by $\langle Y, Q_0 Y \rangle$. The operator Q_0 is a polynomial in the commuting projections. Each centered primitive X_ℓ is a simultaneous eigenvector by (26), and its eigenvalue under Q_0 is bounded by $\rho_i^{N_i}$ for every owner $i \in T_\ell$. The centered primitive signals are mutually orthogonal, so their weighted energies sum to the bound in (98).

Now draw N from any distribution μ on $\mathcal{C}_L$ and use its sorted schedule. For each primitive choose an owner $i(\ell) \in T_\ell$, independently of the public seed. Averaging (98) yields

$$\mathbb{E}_Z \|Y - A\|^2 \le \sum_\ell \mathrm{Var}(X_\ell) \mathbb{E}_\mu \rho_{i(\ell)}^{N_{i(\ell)}} \le \max_i \mathbb{E}_\mu \rho_i^{N_i}.$$

The protocol depends only on the known precisions and the budget. Together with (97), this proves (96) when all $\rho_i < 1$. If some $\rho_i = 1$, its term in the allocation objective is always one. Concentrating the private prior there proves a lower bound of one, and zero prediction attains it.

V.3 Exact allocation by piecewise-linear interpolation

Let $g_i : [0, \infty) \to [0, 1]$ interpolate the points (k, ρ_i^k) linearly. Its successive slopes $\rho_i^{k+1} - \rho_i^k$ are nondecreasing, so g_i is convex and nonincreasing. For any random count vector,

$$\mathbb{E}\rho_i^{N_i} = \mathbb{E} g_i(N_i) \ge g_i(\mathbb{E} N_i).$$

Therefore the allocation value is at least

$$\min_{x_i \ge 0,\ \sum_i x_i \le L} \max_i g_i(x_i). \tag{101}$$

Conversely, an optimizer of (101) can be chosen with $\sum_i x_i = L$: any unused budget can be added to a coordinate without increasing the objective. Write $x_i = k_i + p_i$, where $k_i = \lfloor x_i \rfloor$ and $0 \le p_i < 1$. Then $R = \sum_i p_i = L - \sum_i k_i$ is an integer. Fixed-cardinality rounding with these marginals is a standard convex-decomposition construction. The elementary argument is included here: a distribution on R-subsets $S \subseteq [n]$ with $\Pr(i \in S) = p_i$ follows directly from the extreme points of

$$\left\{ p \in [0,1]^n : \sum_i p_i = R \right\}.$$

Every extreme point is an indicator of an R-subset. Two fractional coordinates permit opposite perturbations, and a sole fractional coordinate contradicts the integer sum. Expressing p as a convex combination of these extreme points supplies the distribution. Set $N_i = k_i + \mathbf{1}\{i \in S\}$. Then $\sum_i N_i = L$ almost surely and

$$\mathbb{E}\rho_i^{N_i} = (1 - p_i)\rho_i^{k_i} + p_i \rho_i^{k_i + 1} = g_i(x_i).$$

Thus (101) equals (96), establishing (15).

The resulting separable min-max problem has the standard inverse-cost resource-allocation form (Wang & Liang, 2022). Define $h_i(t) = \inf\{x \ge 0 : g_i(x) \le t\}$, with the infimum of the empty set equal to infinity. A common risk level t is feasible precisely when $\sum_i h_i(t) \le L$. For $0 < \rho_i < 1$ and $\rho_i^{k+1} \le t \le \rho_i^k$,

$$h_i(t) = k + \frac{\rho_i^k - t}{\rho_i^k - \rho_i^{k+1}}. \tag{102}$$

For a perfect expert, $\rho_i = 0$ gives $g_i(x) = (1 - x)_+$ and $h_i(t) = 1 - t$. For $\rho_i = 1$, $h_i(t) = \infty$ if $t < 1$ and $h_i(1) = 0$. If every $\rho_i < 1$ and some $\rho_i > 0$, the value is the unique $t \in (0, 1]$ solving $\sum_i h_i(t) = L$. Here $h_i(0) = \infty$ for each $\rho_i > 0$. If every expert is perfect, the value is $(1 - L/n)_+$. At $L = 0$, no report is available and the minimax risk equals the original unit target variance.

For homogeneous $0 < \rho < 1$, all inverse functions coincide and the optimal expected counts equal L/n. Adjacent rounding recovers the balanced counts and value in Corollary 7. With heterogeneous precisions, the same calculation allocates more visits to experts whose residual information contracts slowly. For example, at $\boldsymbol{\rho} = (0.04, 0.25, 0.64)$ and $L = 6$, the value is 0.1857703, attained by expected counts $(0.8481560, 1.3425587, 3.8092854)$. One attaining distribution assigns probabilities $(0.1518440, 0.6574413, 0.1907146)$ to count vectors $(0, 2, 4)$, $(1, 1, 4)$, and $(1, 2, 3)$, respectively. Two visits per expert have private-information lower bound $0.64^2 = 0.4096$.

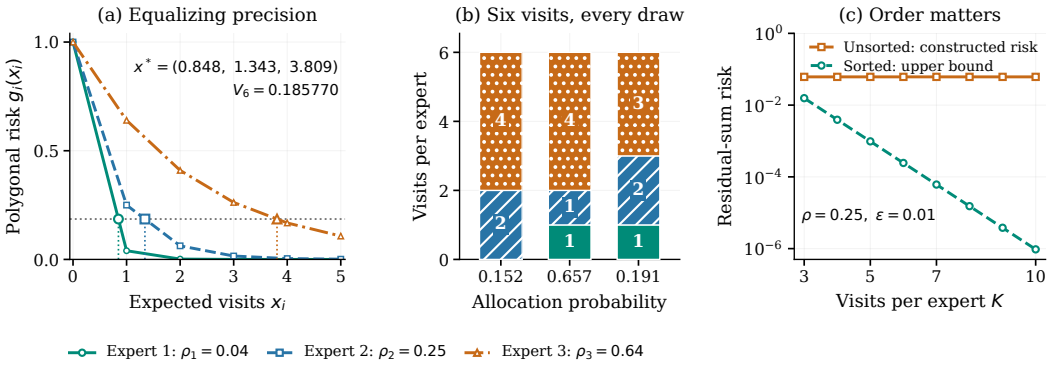

Figure 7: Exact allocation and the effect of query order. (a) At $\boldsymbol{\rho} = (0.04, 0.25, 0.64)$ and $L = 6$, the polygonal curves meet the minimax level $V_6 = 0.185770$ at expected counts $(0.848156, 1.342559, 3.809285)$. (b) An attaining distribution over three count vectors, each using six visits. Colors identify the experts in (a). (c) With common $\rho = 0.25$ and $\varepsilon = 0.01$, the unsorted construction has residual-sum risk $\rho^2(1-2\varepsilon) = 0.06125$, while full sweeps with the same counts guarantee ρ^K. The curves compare these residual-sum protocols; all values are analytical.

V.4 Why equal visit counts do not determine residual risk

The schedule affects which local error directions remain correctable. Even with homogeneous precision, equal counts can give different guarantees for the residual-sum decoder.

Proposition 24 (An order gap with identical counts)**.** *Fix $0 < \rho < 1$, an integer $K \geq 2$, and $0 < \varepsilon < 1/2$, with expert tolerances $\eta_A = \eta_B = \sqrt{\rho}$. There is a two-expert instance in (1) for which residual querying in the order*

$$A^{K-1},\ B,\ A,\ B^{K-1}$$

admits local projected-relative errors with final risk $\rho^2(1-2\varepsilon)$. In contrast, K full residual sweeps have risk at most ρ^K for every admissible instance and error sequence. The schedules use identical visit counts, with K reports from each of the two experts.

Proof. Take independent centered Gaussian W, U_A, U_B with variances $\varepsilon, \varepsilon, 1-2\varepsilon$. Expert A observes (W, U_A), expert B observes (W, U_B), and $Y = W + U_A + U_B$. Use zero errors for the first $K-1$ visits to A. After the first such visit the residual is U_B, and the remaining visits leave it unchanged. At the first visit to B, choose

$$e_B = \sqrt{\frac{\rho(1-2\varepsilon)}{\varepsilon}}\, W.$$

Its squared norm is $\rho \|P_B U_B\|^2$, so it saturates the relative budget. The residual becomes $-e_B$. At the last visit to A, choose

$$e_A = \rho\sqrt{\frac{1-2\varepsilon}{\varepsilon}}\, U_A.$$

This again saturates the budget, since $\|e_A\|^2 = \rho \|P_A(-e_B)\|^2$. The residual is now $-e_A$. Every remaining visit to B has zero predictable residual and zero error, leaving final risk $\rho^2(1-2\varepsilon)$. The full-sweep guarantee follows by applying Theorem 5 K times. □

For $K > 2$ and sufficiently small ε, the ratio between this attained risk and the full-sweep guarantee is $\rho^{2-K}(1-2\varepsilon)$. The comparison uses actual residual targets and the sum decoder in both protocols. It isolates query order as the cause of the different worst-case guarantees.