SHARP MINIMAX RATES FOR BOUNDED LIPSCHITZ GAUSSIAN FUNCTIONALS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study recovery of a bounded scalar Lipschitz functional under an infinite-dimensional Gaussian input law with covariance eigenvalues of order $j^{-\tau}$, $\tau > 1$. For the common unit sup-norm and global-Lipschitz class, the sharp expected integrated norm error after m evaluations is $\Theta([\log(m+2)]^{-(\tau+1)/2})$; squared risk has exponent $\tau + 1$. For a known law, passive noisy regression, adaptive noisy queries and adaptive exact queries share this rate. We then attain the same sample order from finite noisy physical sensors, with unknown mean, covariance and tuning parameters, using only the original passive pairs and conditionally unbiased finite-variance labels. The squared-risk upper bound adds the omitted-covariance and sensor-noise operator norms. Under a coverage condition it yields a sharp sample–resolution–noise minimax law; cell averages of elliptic Gaussian curves provide a concrete example. The procedure combines estimated centering, empirical Hermite features, projected matching-loss gradients and independent validation, with expected polynomial cost in exact real arithmetic. Classical partitions and Gaussian spectral methods supply the core ingredients.

1 INTRODUCTION

Learning a scalar response to a function-valued input arises whenever an observation is a trajectory, field, or other high-dimensional object. A global Lipschitz condition gives a direct form of stability: nearby inputs must have nearby responses. For Gaussian inputs, the covariance spectrum also specifies how much the input varies in each direction. Understanding how these two restrictions combine is a basic question about the information needed to recover the response over the input distribution.

We consider the full common bounded Lipschitz class, with no additional sparsity, compositional structure, or finite-dimensional dependence. Our initial comparison fixes a known input law; a separate passive theorem then removes knowledge of its covariance and eigenbasis. This makes it possible to isolate the role of point information: how accurately can a learner recover the entire functional after m evaluations? We compare ordinary noisy regression with much stronger access that permits adaptive queries anywhere in the Hilbert space and even exact answers. Error is integrated over the same Gaussian law in all three experiments.

For covariance eigenvalues of order $j^{-\tau}$, $\tau > 1$, we prove that all three experiments have expected integrated norm error of order $(\log m)^{-(\tau+1)/2}$. The squared-loss exponent is $\tau+1$. The comparison follows from a lower bound for the strongest experiment and an upper bound for passive noisy regression. For example, eigenvalues $\lambda_j = j^{-2}$ give norm error of order $(\log m)^{-3/2}$, even with adaptive exact queries.

The quantitative gap. For a same-model comparison, fix scalar output, exact eigenvalues $\lambda_j = j^{-\tau}$ with $\tau > 1$, the common bounds $B = L = 1$, unit Gaussian noise, and expected $L^2(\mu)$ norm loss. Let $R_m^{\mathrm{iid},*}$ optimize over iid design laws and all estimators, independently of the unknown function, as in Section 4.1. Theorem 2.9 of Adcock et al. (2026c) implies

$$c_1 e^{-c_2\sqrt{\log m}} \le R_m^{\mathrm{iid},*} \le c_3 \left(\frac{\log m}{\log\log m}\right)^{-(\tau-1)/2}$$

for all sufficiently large m, with positive constants and a threshold that may depend on the fixed τ. The discussion following their Theorem 2.10 suggests the sharper order $(\log m)^{-\tau/2}$. In this scalar

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.

iid setting, Theorem 1 and (6) instead establish

$$R_m^{\mathrm{iid},*} = \Theta\Big((\log m)^{-(\tau+1)/2}\Big).$$

Our passive upper bound already improves on the suggested order; our lower bound shows sharpness even for adaptive exact queries. This qualifies the conjectural order in these instances, without contradicting the proved bounds or settling deterministic fixed designs and general output regimes.

A second contribution attains the same sample rate from finite noisy physical sensors, with unknown Gaussian mean and covariance and without supplied tuning parameters. The labels still depend on the full latent input. The additional squared-risk terms are an omitted-covariance operator norm and a sensor-noise operator norm. Under a physical coverage condition these terms give a sharp minimax law, and a cell-average example shows how the condition can follow from geometry.

The contribution is the statistical characterization and its attainment under these information restrictions. Padded partitions, Gaussian spectral approximation, PCA, matching losses and validation are established tools. The new sensor procedure has a finite expected polynomial operation bound in exact real arithmetic; numerical stability and finite-bit complexity are not established.

Section 2 places these results in the literature. Sections 3–4 state and explain the sharp known-law result; Sections 5–6 treat unknown covariance and finite observations. Table 1 summarizes the information assumptions and rates; complete proofs appear in Appendices A–I.

2 Related Work

Uniform classes and information restrictions. Kovachki, Lanthaler and Mhaskar (Kovachki et al., 2024) develop operator-learning data-complexity lower bounds, including a Gaussian polynomial-spectrum setting. Their Section 2.1 sampling width fixes sample locations and a decoder before taking the worst target. This differs from our randomized adaptive setting, where the target is chosen independently of the realized random seed. We do not infer a randomized lower bound merely from a deterministic width. Lanthaler (Lanthaler, 2024) also studies the information complexity and metric entropy of Lipschitz operator classes.

The Gaussian Sobolev-bounded, individually Lipschitz class of Adcock et al. (2026a) does not impose one common Lipschitz bound, and its linear information can include coefficient integrals. Those distinctions prevent direct transfer of its lower bounds or information-width algorithms to our common class and point-query model. The product-measure entropy theorem of Bizeul & Klartag (2025) concerns $\varepsilon > d^{-1/4}$; our sufficient whitened bump scale is $d^{-1/2}$.

Adcock, Griebel and Maier (Adcock et al., 2026b) already combine empirical PCA, Hermite regression and unknown Gaussian covariance. Their Section 2.2 and equation (3.4) use labels at a constructed Christoffel sampling law, whereas our unknown-covariance result (Theorem 3) uses only the original passive pairs. We claim the stated uniform statistical rate, not invention of PCA–Hermite learning. Our bounded globally Lipschitz scalar result does not settle their general Sobolev-regularity question. Likewise, the boundedness of our bumps is not itself a distinction from Bizeul–Klartag: their entropy constructions can also be bounded. The relevant comparison is the scale and the resulting infinite-dimensional statistical statement.

Quantitative comparison. The comparison in the introduction uses the scalar optimized-iid model (2.3) of Adcock et al. (2026c). It addresses their suggested sharp order without contradicting their proved bounds. The extensions to two-sided spectral envelopes and adaptive point access are conclusions of our theorem.

Recovery versus integration. Adaptive full-space point information also appears in Creutzig, Dereich, Müller-Gronbach and Ritter's infinite-dimensional quadrature model (Creutzig et al., 2008). There the target is the single number $\int f \, d\mu$. Here the learner must recover f in integrated error. Quadrature and quantization results are therefore relevant precedents, but are not full-function recovery upper bounds without a further reduction.

Noisy functional observations. Joint sample/resolution questions have been studied for functional linear regression with noisy point measurements (Cai & Li, 2026). Nonlinear noisy-test functional prediction is also established: Sang & Li (2026, Section 6, Theorem 20 and Corollary 21) reconstruct

test curves in a nested RKHS framework. Our result concerns the common bounded global-Lipschitz scalar class, Gaussian minimax risk and bounded linear average sensors. It does not claim the first nonlinear noisy-test procedure or compare rates across incompatible regularity assumptions. Appendix F records the scope of these source comparisons.

3 Problem setup and main result

Let $H = \ell^2$ over the reals, with its usual norm and Borel structure. Fix $\tau > 1$ and constants $0 < c_\lambda \le C_\lambda < \infty$. The known input law is

$$\mu = \mathcal{L}(X), \qquad X = \sum_{j\ge 1} \sqrt{\lambda_j} Z_j e_j, \qquad Z_j \overset{\text{iid}}{\sim} N(0,1), \qquad c_\lambda j^{-\tau} \le \lambda_j \le C_\lambda j^{-\tau}. \tag{1}$$

The actual eigenvalues and orthonormal eigenbasis are known, not merely their envelope. Since $\sum_j \lambda_j < \infty$, the series defines an H-valued random variable. Our common function class is

$$\mathcal{F} = \left\{ f : H \to \mathbb{R} : \sup_{x\in H} |f(x)| \le 1, \quad |f(x) - f(y)| \le \|x - y\|_H \text{ for all } x, y \in H \right\}. \tag{2}$$

The truth is a specified continuous function, so point evaluations are well defined. We use three experiments:

Experiment	Information available with budget m
$\mathsf{P}, 1$	Independent full inputs $X_t \sim \mu$ and labels $Y_t = f(X_t) + \xi_t$, $t = 1, \dots, m$.
$\mathsf{A}, 1$	At most m adaptively selected points $x_t \in H$, with answers $f(x_t) + \xi_t$.
$\mathsf{A}, 0$	At most m adaptively selected points $x_t \in H$, with exact answers $f(x_t)$.

In the noisy experiments, the ξ_t are fresh independent $N(0,1)$ variables, independent of the seed and the information available before each response. Procedures may use an independent random seed. Active queries and stopping decisions are measurable functions of this seed and previous queries and answers; the budget holds almost surely. The seed distribution and procedure cannot depend on the unknown truth. The output is a measurable function of the available data and seed, taking values in the Borel space $L^2(\mu)$; it need not lie in $\mathcal{F}$. Known-law sampling is allowed, but derivative and integral oracles are not.

For each experiment B, define two separate minimax risks:

$$R_m^{\mathsf{B}} = \inf_{\widehat{f}\in \mathsf{B}} \sup_{f\in\mathcal{F}} \mathbb{E}_f \|\widehat{f} - f\|_{L^2(\mu)}, \qquad Q_m^{\mathsf{B}} = \inf_{\widehat{f}\in \mathsf{B}} \sup_{f\in\mathcal{F}} \mathbb{E}_f \|\widehat{f} - f\|_{L^2(\mu)}^2. \tag{3}$$

The expectation includes design, label noise, and procedure randomization. In particular, Q_m^{B} is not defined as $(R_m^{\mathsf{B}})^2$.

Theorem 1 (sharp rates under polynomial covariance). Under (1), there are positive constants c_R, C_R, c_Q, C_Q, depending only on $\tau, c_\lambda, C_\lambda$, such that for every integer $m \ge 1$ and every $\mathsf{B} \in \{(\mathsf{P}, 1), (\mathsf{A}, 1), (\mathsf{A}, 0)\}$,

$$\begin{aligned} c_R[\log(m+2)]^{-(\tau+1)/2} \le R_m^{\mathsf{B}} \le C_R[\log(m+2)]^{-(\tau+1)/2}, \\ c_Q[\log(m+2)]^{-(\tau+1)} \le Q_m^{\mathsf{B}} \le C_Q[\log(m+2)]^{-(\tau+1)}. \end{aligned} \tag{4}$$

Logarithms are natural. The constants are uniform over known covariance sequences in the fixed envelope. Each estimator may use its own known covariance; this uniformity is not an adaptation result for an unknown input law. The constants need not be uniform as τ or the envelope changes. The exact-power case $\lambda_j = j^{-2}$ gives norm exponent $3/2$ and squared-loss exponent 3. Theorem 3 below treats unknown covariance in a separate passive model; it does not change the information supplied in Theorem 1.

Table 1: Information supplied to the learner and squared-risk conclusions. Write $a_m = [\log(m+2)]^{-(\tau+1)}$. Here β is the omitted-covariance operator norm, ν the sensor-noise operator norm, and v its class bound (Section 6). Norm risks are stated separately in each result.

Observation model	Population information	Squared-risk result
Full inputs / point queries (Theorem 1)	Known centered law; passive noisy, adaptive noisy or adaptive exact access	Minimax $\asymp a_m$
Passive full inputs (Theorem 3)	Unknown covariance; known zero mean and spectral envelope	Minimax $\asymp a_m$
First d eigen-coordinates (Corollary 4)	Known centered law and eigen-coordinates; passive labels from full input	Minimax $\asymp a_m + (d+1)^{-\tau}$
Noisy physical sensors (Theorem 5; Corollary 6)	Unknown mean, covariance and tuning; original passive pairs	Upper $K(a_m + \beta + \nu)$; coverage-class minimax $\asymp a_m + (d+1)^{-\tau} + v$

4 Proof overview

We prove Theorem 1 by comparing the observation models, constructing a lower bound for adaptive exact queries, and matching it with a passive Hermite estimator.

4.1 Comparison of observation models

For both losses, the experiments satisfy

$$R_m^{\mathsf{A},0} \le R_m^{\mathsf{A},1} \le R_m^{\mathsf{P},1}, \qquad Q_m^{\mathsf{A},0} \le Q_m^{\mathsf{A},1} \le Q_m^{\mathsf{P},1}. \tag{5}$$

An exact-query procedure can simulate any noisy-query procedure by adding independent Gaussian noise before feeding each answer to its policy. This also reproduces adaptive queries and stopping decisions. A noisy-query procedure can simulate passive regression by drawing its queries independently from the known law μ. Thus an exact-query lower bound and a passive-regression upper bound imply (4) throughout the chain. Equality of rates does not assert equality of risks or constants.

The same reasoning covers optimization over iid design laws: allow the learner to choose an input law independently of f, draw m iid queries from it, and use any measurable estimator with unit-noise responses. Denoting these risks by $R_m^{\mathrm{iid},*}$ and $Q_m^{\mathrm{iid},*}$, we have

$$R_m^{\mathsf{A},1} \le R_m^{\mathrm{iid},*} \le R_m^{\mathsf{P},1}, \qquad Q_m^{\mathsf{A},1} \le Q_m^{\mathrm{iid},*} \le Q_m^{\mathsf{P},1}. \tag{6}$$

Consequently these optimized-iid risks also obey (4). The upper comparison uses both the admissibility of μ as a design law and the availability of the estimator in Section 4.3. It is not a conclusion for arbitrary restricted estimator families. Nor does the passive upper bound yield a matching deterministic fixed-design upper bound: averaging over a random design cannot interchange the worst-case target and design expectation.

4.2 Lower bound via Gaussian partitions

Padded partitions and signed distance functions are established tools for constructing Lipschitz families (Gupta et al., 2003, Sections 3.1 and 4.2). The following self-contained conic version supplies the geometry.

Lemma 2 (Gaussian partition). Write $\gamma_k = N(0, I_k)$, $\delta = 10^{-3}$, and $h_k = \delta/(4\sqrt{k})$. For every integer $k \ge 10000$, there are 2^k closed conic Voronoi cells $C_i \subset \mathbb{R}^k$, covering the space with disjoint interiors, such that

$$\max_i \gamma_k(C_i) \le e^{-k/64}, \qquad \gamma_k\{z : \max_i \operatorname{dist}(z, C_i^c) \ge h_k\} \ge 1/16. \tag{7}$$

Both properties hold for the same deterministic partition. Appendix B proves the simultaneous selection from random spherical codebooks. Signed sums of truncated distances to cell complements are globally Lipschitz, including across cell boundaries. Transferring the cells to coordinates $k+1,\dots,2k$ and scaling by the covariance gives bounded unit-Lipschitz bumps of amplitude comparable to $k^{-(\tau+1)/2}$. Their total squared mass is at least one-sixteenth of the squared amplitude, while each cell has exponentially small mass.

Assign independent signs to these bumps. An adaptive exact query can reveal at most one sign. Conditional on the transcript, unvisited signs remain fair and independent, so most of their squared energy remains unknown for $k \asymp \log(m+2)$. For the separate norm lower bound, clipping an estimate to the bump amplitude gives $Z^2 \le 2aZ$ for its norm error Z and amplitude a. Appendix C proves both losses and the all-budget constants, including adaptive stopping and independent randomization.

4.3 Upper bound via Hermite approximation

For the upper bound, normalized covariance-scaled Hermite products H_γ form an orthonormal basis. Set $c_\gamma = \langle f, H_\gamma\rangle$ and $\nu_\gamma = \sum_j \gamma_j/\lambda_j$. The global Lipschitz condition implies the derivative-energy bound

$$\sum_\gamma \nu_\gamma c_\gamma^2 \le 1. \tag{8}$$

Finite-coordinate conditioning preserves Lipschitzness; Gaussian integration by parts and Parseval give the bound before passage to the limit. This is a derivative seminorm bound, not a unit bound on the full Sobolev norm. Gaussian Sobolev and Hermite approximation methods are established (Adcock et al., 2026a).

Retain $S_E = \{\gamma : \nu_\gamma \le E\}$ and estimate c_γ by $m^{-1}\sum_t Y_t H_\gamma(X_t)$, a classical projection estimator (Bizeul & Klartag, 2025, Equation (9)). Conditional label second moment at most two gives

$$\sup_f \mathbb{E}\|\widehat{f}_E - f\|_2^2 \le E^{-1} + 2|S_E|/m. \tag{9}$$

A generating-function bound and one-dimensional integral comparison yield

$$\log|S_E| \le B_\tau (C_\lambda E)^{1/(\tau+1)}, \quad B_\tau = (\tau+1)(G_\tau/\tau)^{\tau/(\tau+1)}, \quad G_\tau = \Gamma(1+1/\tau)\zeta(1+1/\tau). \tag{10}$$

For $m \ge 3$, choosing $E = C_\lambda^{-1}(\log m/(2B_\tau))^{\tau+1}$ gives

$$\sup_f \mathbb{E}\|\widehat{f}_E - f\|_2^2 \le C_\lambda (2B_\tau)^{\tau+1}(\log m)^{-(\tau+1)} + 2m^{-1/2}. \tag{11}$$

The zero estimator handles $m = 1, 2$; Jensen gives the norm upper bound. Appendix D supplies the complete regularity, counting and all-budget proof. Together with the exact-query lower bound and (5), this proves Theorem 1.

5 Unknown covariance and finite observations

We next consider two departures from the known-law, full-input model: learning the covariance from passive data and observing only finitely many known eigencoordinates.

5.1 Passive regression with unknown covariance

Now let the centered covariance Σ be unknown. Its positive eigenvalues, in nonincreasing order with multiplicity, satisfy the envelope (1), with $\tau, c_\lambda, C_\lambda$ fixed and known; ambient null directions are allowed. The learner receives only m iid full-input pairs (X_t, Y_t), with $Y_t = f(X_t) + \xi_t$ and unit independent Gaussian noise. It uses a single jointly Borel predictor of the data and test input, independent of Σ. Define R_m^{unk} and Q_m^{unk} by taking the infimum over these predictors and the supremum over both Σ and $f \in \mathcal{F}$, with norm and squared-norm losses evaluated under μ_Σ, respectively.

Theorem 3 (unknown-covariance passive rates). There are positive constants depending only on $\tau, c_\lambda, C_\lambda$ such that, for every integer $m \geq 1$,

$$R_m^{\mathrm{unk}} \asymp [\log(m+2)]^{-(\tau+1)/2}, \qquad Q_m^{\mathrm{unk}} \asymp [\log(m+2)]^{-(\tau+1)}.$$

Neither the eigenbasis, the individual eigenvalues, nor an eigengap is supplied. The lower bounds follow from Theorem 1 by fixing one admissible covariance and even revealing it to the learner. The upper bound uses a different estimator from the known-law projection estimator above.

The estimator splits the original passive sample, estimates covariance from the first group, retains empirical Hermite features below a prescribed energy, and fits a clipped approximate empirical squared-loss minimizer on the second group. A feature-count guard and a covariance operator-norm event control the finite-dimensional comparison. Gaussian conditioning bounds the unexplained variance by the discarded operator tail. Appendix G gives the complete estimator, measurability argument and proof; all original pairs, including discarded labels, are charged to the budget.

5.2 Finite-coordinate observations

Return to the known law (1), and let P_d project onto $e_1, \ldots, e_d$. The learner observes only $(P_d X_t, Y_t)_{t=1}^m$, where $Y_t = f(X_t) + \xi_t$, and only $P_d x$ at prediction. Labels therefore depend on the full latent input. Define $R_{m,d}$ and $Q_{m,d}$ by (3), with infima over jointly measurable predictors $h(\mathrm{data}, U, P_d x)$ and losses integrated against $f(x)$ under the original law μ.

Corollary 4 (known-law observation budget). Uniformly for integers $m, d \geq 1$, with constants depending only on $\tau, c_\lambda, C_\lambda$,

$$\begin{aligned} R_{m,d} &\asymp [\log(m+2)]^{-(\tau+1)/2} + (d+1)^{-\tau/2}, \\ Q_{m,d} &\asymp [\log(m+2)]^{-(\tau+1)} + (d+1)^{-\tau}. \end{aligned} \tag{12}$$

Thus retaining the full-input minimax order requires and suffices, up to fixed constants at the level of rates, that d be at least of order $[\log(m+2)]^{(\tau+1)/\tau}$; for $\tau = 2$, the threshold is $[\log(m+2)]^{3/2}$.

The upper bound estimates the conditional mean $g_d(y) = \mathbb{E}[f(y + X_{>d})]$ using only Hermite indices supported in $[d]$. Gaussian Poincare bounds the unexplained squared error by $\sup_{j>d} \lambda_j$, rather than the trace tail. Theorem 1 supplies the sample-information lower bounds; a clipped omitted coordinate supplies a separate, pathwise test-information floor. Appendix H proves both losses separately. This is a consequence for exact known eigen-coordinates, not an extension of Theorem 3 to truncated inputs or a claim for arbitrary grids.

6 Learning from finite noisy physical sensors

Full trajectories and population eigen-coordinates are idealized observations. We next use a fixed, known physical map $P : H \to \mathbb{R}^d$ with $PP^* = I_d$, without assuming that its sensor space is a covariance eigenspace. Write $X = b + X_0$, where the mean $b \in H$ and covariance Σ of the centered Gaussian X_0 are unknown. For fixed $0 < c \leq C < \infty$ and $\tau > 1$, the ordered positive eigenvalues satisfy $cj^{-\tau} \leq \lambda_j(\Sigma) \leq Cj^{-\tau}$; ambient null directions are allowed. The learner receives exactly m iid original pairs

$$U_i = PX_i + \eta_i, \qquad Y_i = f(X_i) + \xi_i, \qquad \eta_i \sim N(0, \Gamma), \quad \eta_i \perp X_i,$$

where $f \in \mathcal{F}$, $\Gamma \succeq 0$ is arbitrary with $\operatorname{tr} \Gamma \leq B_{\mathrm{noise}} < \infty$, and

$$\mathbb{E}[\xi_i \mid X_i, \eta_i] = 0, \qquad \mathbb{E}[\xi_i^2 \mid X_i, \eta_i] \leq s_2 < \infty.$$

Here $s_2 \geq 0$; label noise may be heavy-tailed and heteroskedastic. A marginal variance bound alone does not suffice. At test time only a fresh $U = PX + \eta$ is observed, and the target is the full latent response $f(X)$. Define

$$\beta = \|(I - P^*P)\Sigma(I - P^*P)\|_{\mathrm{op}}, \qquad \nu = \|\Gamma\|_{\mathrm{op}}, \qquad \ell(h) = \left[\mathbb{E}_{X,\eta}|h(PX + \eta) - f(X)|^2\right]^{1/2}.$$

The test expectation is inside the norm; training expectation is outside it. Thus $\mathbb{E}_D \ell(\widehat{h}_D)$ and $\mathbb{E}_D \ell(\widehat{h}_D)^2$ remain separate risks.

Theorem 5 (passive finite-sensor upper bound). There is one deterministic jointly Borel procedure, using only $(U_i, Y_i)_{i=1}^m$ and m, d, such that for every fixed $\tau > 1$, $C > 0$, $B_{\mathrm{noise}} < \infty$ and $s_2 \geq 0$, simultaneously for all $m, d \geq 1$ and every admissible P, Σ, Γ,

$$\sup_{b \in H,\ f \in \mathcal{F},\ \mathcal{L}_\xi} \mathbb{E}_D \ell(\widehat{h}_D)^2 \leq K(\tau, C, B_{\mathrm{noise}}, s_2) \left\{ [\log(m+2)]^{-(\tau+1)} + \beta + \nu \right\}. \tag{13}$$

The label supremum is over iid kernels with the displayed conditional moments. The same rule has expected norm risk at most $\sqrt{K}\{[\log(m+2)]^{-(\tau+1)/2} + \sqrt{\beta} + \sqrt{\nu}\}$. It requires no mean, covariance, eigenbasis, spectral exponent or envelope, noise bound, eigengap, or lower sensor-variance bound. Its expected operation count is polynomial in m, d for fixed model bounds in exact real arithmetic.

This is a passive result; it does not remove the known-law assumption from the active-query comparisons. The covariance compression and noise may be singular, and the retained space may be empty. The constant and rate onset depend on the fixed model bounds, but not on the unknown mean or sensor dimension. The terms β and ν need not be matching lower floors for every individual model.

Corollary 6 (sharp rate under physical coverage). Fix $\tau, c, C, B_{\mathrm{noise}}, s_2$, $D_0 \geq C$, P, and $0 \leq v \leq B_{\mathrm{noise}}$. Let $\mathcal{C}_d(v)$ consist of the covariance pairs above with $\beta \leq D_0(d+1)^{-\tau}$ and $\|\Gamma\|_{\mathrm{op}} \leq v$. Taking infima over measurable sensor predictors and suprema over $\mathcal{C}_d(v)$, all $b \in H$, $f \in \mathcal{F}$ and admissible label kernels defines $Q_{m,d,v}$ and $R_{m,d,v}$ for squared and norm risk, respectively. Uniformly in m, d, v,

$$\begin{aligned} Q_{m,d,v} &\asymp [\log(m+2)]^{-(\tau+1)} + (d+1)^{-\tau} + v, \\ R_{m,d,v} &\asymp [\log(m+2)]^{-(\tau+1)/2} + (d+1)^{-\tau/2} + \sqrt{v}. \end{aligned} \tag{14}$$

Constants depend only on the fixed class bounds. The lower bounds hold even if those bounds are supplied. They use zero-mean, noiseless-label subclasses, valid for every $s_2 \geq 0$, and distinct sample, omitted-coordinate and rank-one-noise witnesses.

6.1 Estimator using the original passive pairs

For $m < 8$ return zero. Otherwise split into a training pool of size $M = \lfloor m/2 \rfloor$ and validation pool of size $L = m - M$. Split training into $n = \lfloor M/2 \rfloor$ covariance inputs and $N = M - n$ fitting pairs. Estimate $\widehat{u}$ and the centered covariance from those n inputs only, charging their discarded labels to the budget. Store $\widehat{u}$ for centering all later inputs. Maximal-pivot Jacobi rotations give a positive semidefinite surrogate within operator error $n^{-1/4}$ of this sample covariance.

Use dyadic energies 2^j, $-k_m \leq j \leq k_m$, with k_m the smallest integer such that $2^{k_m} \geq M$, plus an explicit zero candidate. At energy E retain surrogate eigenvalues $\widehat{\lambda}_i \geq 1/E$ and Hermite indices satisfying $\sum_i \alpha_i / \widehat{\lambda}_i \leq E$. Enumerate lazily and reject if their number exceeds the global guard $D_{\max} = \lfloor \sqrt{M} \rfloor$. Empty retention uses the constant feature. Write clip for projection onto $[-1, 1]$. For an accepted feature vector ϕ of actual dimension D, set $\theta_1 = 0$ and run

$$\theta_{t+1} = \Pi_{\|\theta\|_2 \leq 1} \left[\theta_t - \frac{\{\mathrm{clip}(\theta_t^* \phi(U_{n+t})) - Y_{n+t}\} \phi(U_{n+t})}{\sqrt{10DN}} \right].$$

Return the clipped prediction from the average of the N pre-update iterates. Original labels are not clipped. Select among all slots by the independent mean score $L^{-1} \sum_{i>M} \{h(U_i)^2 - 2Y_i h(U_i)\}$; fixed earliest-slot ties make the rule deterministic. Appendix I.3 specifies every threshold, rotation, enumeration and tie rule.

The proof selects one good energy with actual $D_* \leq \lfloor M^{1/4} \rfloor$. This smaller count is used only for its estimation remainder; the algorithm's global guard remains $\lfloor \sqrt{M} \rfloor$. Its step uses the actual D, so adding more candidates does not change that good candidate's bound. Bounded predictions let validation compare to that slot even when other candidates have poor statistical properties.

6.2 Proof overview

On the covariance/mean good event for the proof candidate, under the conditions of Appendix I.4, conditioning directly on the retained noisy coordinates bounds the residual covariance norm by

$\beta + 1/E + 3n^{-1/4}$. The conditional response is Lipschitz on that retained space, even when no uniform Lipschitz bound is available on the entire measured space. Estimated centering and covariance perturb both polynomial moments and raw approximation error. Appendix I.6 proves the required shifted Hermite moment and coupling inequalities. The coefficient comparator lies in the unit ball; the matching-loss gradient converts its raw approximation error into clipped prediction error. Conditional label moments control the SGD second moment, and a second-moment validation bound suffices for the eventual logarithmic rate. Matching losses and related GLMtron procedures are established tools (Kakade et al., 2011; Gollakota et al., 2023).

6.3 Example: cell-average sensors

For a concrete sensor geometry, let $H = L^2(0,1)$ and $\Sigma = L_a^{-1}$, where L_a is the Dirichlet operator associated with $\int_0^1 a(x)|u'(x)|^2 dx$ and unknown measurable $0 < a_{\min} \le a(x) \le a_{\max} < \infty$. Its eigenvalues have exponent $\tau = 2$. Equal-cell orthonormal averages $\langle X, \sqrt{d}\,\mathbf{1}_{I_j}\rangle$ satisfy $\beta \le (a_{\min}\pi^2 d^2)^{-1}$ by cellwise Poincare. If raw averages $d\int_{I_j} X$ have independent sensor-noise variance σ^2_{sensor}, dividing the reports by $\sqrt{d}$ gives $\Gamma = (\sigma^2_{\text{sensor}}/d)I_d$. For fixed variance bound, the sharp squared-risk order over this elliptic family is

$$[\log(m+2)]^{-3} + (d+1)^{-2} + \sigma^2_{\text{sensor}}/d.$$

The expected norm order is the sum of the corresponding square roots. Appendix I.12 proves coverage and lower bounds within this physical family. Perfectly correlated raw-average noise instead has operator norm σ^2_{sensor} after normalization. Average sensors are bounded L^2 functionals; this example makes no claim about arbitrary point grids.

7 Conclusion

We established sharp minimax rates for the common bounded Lipschitz class under Gaussian inputs with polynomial covariance decay. For a known input law, passive noisy regression already attains the worst-case rate available to adaptive exact point queries. A separate passive construction extends rate attainment to finite noisy physical sensors with unknown mean, covariance and tuning parameters. Under the stated physical coverage condition, matching bounds separate the contributions of sample size, sensor resolution and sensor noise.

Future work should develop numerically stable implementations with finite-bit guarantees and useful finite-sample bounds, extend the analysis beyond Gaussian input laws, and study sensor selection with provable coverage. Equality of worst-case rates leaves open improvements in constants and gains from adaptive queries on more structured target classes.

AI USE STATEMENT

AI tools assisted with formalizing and check proofs, literature retrieval and comparison, paper drafting and polishing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ben Adcock, Michael Griebel, and Gregor Maier. The Sample Complexity of Learning Lipschitz Operators with respect to Gaussian Measures, 2026a. URL https://arxiv.org/abs/2410.23440v4. arXiv:2410.23440v4, 4 September 2026.

Ben Adcock, Michael Griebel, and Gregor Maier. Near-Optimal Learning of Gaussian Sobolev Operators, 2026b. URL https://arxiv.org/abs/2607.11921v1. arXiv:2607.11921v1, 8 July 2026.

Ben Adcock, Gregor Maier, and Rahul Parhi. Towards Sharp Minimax Risk Bounds for Operator Learning, 2026c. URL https://arxiv.org/abs/2512.17805v2. arXiv:2512.17805v2, 5 March 2026.

Karine Bertin and Nicolas Klutchnikoff. Adaptive regression with Brownian path covariate, 2020. URL https://arxiv.org/abs/1907.11284v2. arXiv:1907.11284v2, 20 November 2020.

Pierre Bizeul and Boaz Klartag. Entropy and Learning of Lipschitz Functions under Log-Concave Measures, 2025. URL https://arxiv.org/abs/2509.10355v1. arXiv:2509.10355v1, 12 September 2025.

T. Tony Cai and Yicheng Li. The cost of discretization in functional linear regression: Minimax rates and adaptation, 2026. URL https://arxiv.org/abs/2607.09350. Version 1, 10 July 2026.

Jakob Creutzig, Steffen Dereich, Thomas Müller-Gronbach, and Klaus Ritter. Infinite-Dimensional Quadrature and Approximation of Distributions, 2008. URL https://www.uni-muenster.de/Stochastik/dereich/Publikationen/Preprints/InfiniteDimensionalQuadrature.pdf. Author manuscript, February 2008.

Aravind Gollakota, Parikshit Gopalan, Adam Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14685–14704, 2023. doi: 10.52202/075280-0646. URL https://papers.nips.cc/paper_files/paper/2023/hash/2f46ef5725a8eca24f7f24a17955ad1a-Abstract-Conference.html.

Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003. URL https://cs.nyu.edu/~anupamg/papers/focs03b.pdf.

Sham M. Kakade, Adam Tauman Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://proceedings.neurips.cc/paper/2011/file/30bb3825e8f631cc6075c0f87bb4978c-Paper.pdf.

Nikola B. Kovachki, Samuel Lanthaler, and Hrushikesh Mhaskar. Data Complexity Estimates for Operator Learning, 2024. URL https://arxiv.org/abs/2405.15992v2. arXiv:2405.15992v2, 17 October 2024.

David Krieg, Erich Novak, Leszek Plaskota, and Mario Ullrich. Noisy Nonlinear Information and Entropy Numbers. *Journal of Fourier Analysis and Applications*, 32:36, 2026. doi: 10.1007/s00041-026-10249-z. URL https://doi.org/10.1007/s00041-026-10249-z. Article 36, published 16 March 2026.

Samuel Lanthaler. Operator Learning of Lipschitz Operators: An Information-Theoretic Perspective, 2024. URL https://arxiv.org/abs/2406.18794v2. arXiv:2406.18794v2, 2 July 2024.

André Mas. Representation of small ball probabilities in Hilbert space and lower bound in regression for functional data, 2011. URL https://arxiv.org/abs/0901.0264v2. arXiv:0901.0264v2, 2 March 2011.

Alexander Meister. Optimal classification and nonparametric regression for functional data. *Bernoulli*, 22(3):1729–1744, 2016. doi: 10.3150/15-BEJ709. URL https://arxiv.org/abs/1603.09130v1. Version consulted: arXiv:1603.09130v1.

Hrushikesh N. Mhaskar and Nahmwoo Hahm. Neural Networks for Functional Approximation and System Identification. *Neural Computation*, 9(1):143–159, 1997. doi: 10.1162/neco.1997.9.1.143. URL https://doi.org/10.1162/neco.1997.9.1.143.

Leszek Plaskota, Klaus Ritter, and Grzegorz W. Wasilkowski. Average Case Complexity of Weighted Approximation and Integration over $\mathbb{R}_+$, 2000. URL https://www.mathematik.tu-darmstadt.de/media/mathematik/forschung/preprint/ps2pdf/2131.pdf. Author manuscript, 28 November 2000.

Peijun Sang and Bing Li. Nonlinear function-on-function regression by RKHS. *Journal of Machine Learning Research*, 27(12):1–54, 2026. URL https://jmlr.org/papers/v27/25-1017.html.

Yuya Suzuki and Toni Karvonen. Construction of optimal algorithms for function approximation in Gaussian Sobolev spaces. *BIT Numerical Mathematics*, 66:23, 2026. doi: 10.1007/s10543-026-01122-y. URL https://doi.org/10.1007/s10543-026-01122-y. Article 23, published 30 March 2026.

Grzegorz W. Wasilkowski. Liberating the dimension for L_2-approximation. *Journal of Complexity*, 28(3):304–319, 2012. doi: 10.1016/j.jco.2011.12.002. URL https://www.sciencedirect.com/science/article/pii/S0885064X11000987.

A Model and measurable procedures

A.1 Exact model and theorem

Fix $\tau > 1$ and constants $0 < c_\lambda \le C_\lambda < \infty$. The known covariance has a known orthonormal eigenbasis (e_j) and eigenvalues satisfying, for every integer $j \ge 1$,

$$c_\lambda j^{-\tau} \le \lambda_j \le C_\lambda j^{-\tau}. \tag{A.1}$$

Let μ be the law on ℓ^2 of

$$X = \sum_{j \ge 1} \sqrt{\lambda_j} Z_j e_j, \qquad Z_j \overset{\text{iid}}{\sim} N(0,1). \tag{A.2}$$

Since $\sum_j \lambda_j \le C_\lambda \zeta(\tau) < \infty$, $\mathbb{E}\|X\|^2$ is finite, and the series defines an ℓ^2-valued random variable almost surely. Coordinates below are taken in the known eigenbasis; its orientation has no effect on the global Hilbert-space Lipschitz bound. Equivalently, all formulas can be taken on the closed support of μ if the ambient covariance has null directions; orthogonal projection onto that support is contractive.

The class remains

$$\mathcal{F} = \{f : \ell^2 \to \mathbb{R} : \|f\|_\infty \le 1,\ \mathrm{Lip}_{\ell^2}(f) \le 1\}. \tag{A.3}$$

For each integer $m \ge 1$, consider the same three experiments:

- **P,1:** m independent full-input observations $X_t \sim \mu$, with labels $Y_t = f(X_t) + \xi_t$ and independent $\xi_t \sim N(0,1)$.
- **A,1:** at most m adaptive queries at arbitrary points in ℓ^2, with an independent fresh unit Gaussian noise on each returned value.
- **A,0:** at most m adaptive queries at arbitrary points in ℓ^2, returning exact noiseless values.

All learners may randomize using a seed independent of f. Adaptive locations, stopping and output are measurable from that seed and previous answers. Outputs are arbitrary measurable $L^2(\mu)$-valued estimators. The budget is almost sure, and no function-dependent side information or derivative/integral oracle is supplied. The law μ, including its actual eigenvalues and eigenvectors, is known to the learner; knowing only the two constants in (A.1) is not the upper-bound assumption.

For $B \in \{(P,1),(A,1),(A,0)\}$, define separately

$$R_m^B = \inf_{\widehat{f}\ \text{admissible in}\ B} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\widehat{f} - f\|_{L^2(\mu)}, \qquad Q_m^B = \inf_{\widehat{f}\ \text{admissible in}\ B} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\widehat{f} - f\|_{L^2(\mu)}^2. \tag{A.4}$$

The second is not defined as $(R_m^B)^2$.

Theorem. Put $q = \tau + 1$ and $\beta = q/2$. There are positive constants depending only on $\tau, c_\lambda, C_\lambda$ such that, simultaneously for all three experiments and every integer $m \ge 1$,

$$R_m^B \asymp_{\tau,c_\lambda,C_\lambda} [\log(m+2)]^{-\beta}, \qquad Q_m^B \asymp_{\tau,c_\lambda,C_\lambda} [\log(m+2)]^{-q}. \tag{A.5}$$

All logarithms are natural. Exact valid choices are given in Appendices C and D. They are uniform across known covariance sequences obeying (A.1); this does not allow an unknown covariance or a law chosen after observing the data.

A.2 PRECISE CONDITIONAL MODEL

Let μ be the fixed Gaussian probability law on ℓ^2 specified in Appendix A.1. In fact the following argument works for any probability measure on a standard Borel input space. Fix a finite or countable index set J and fixed nonnegative Borel functions b_j satisfying:

1. The nonzero sets are pairwise disjoint **pointwise**: $b_j(x)b_l(x) = 0$ for all x and distinct j, l.
2. $0 \le b_j(x) \le a$ for a finite constant $a > 0$.
3. Every signed sum $f_\sigma(x) = \sum_{j\in J} \sigma_j b_j(x),\ \sigma \in \{-1, +1\}^J$, belongs to the desired globally bounded 1-Lipschitz class. This is a hypothesis, not derived here. The sum is unambiguous because at most one summand is nonzero at any x.
4. With $w_j = \int b_j^2\, d\mu,\ W = \sum_j w_j$ and $w_* = \sup_j w_j$,

$$W \ge pa^2, \qquad w_* \le a^2 e^{-ck},$$

where p, c, k are positive. Disjointness implies $W \le a^2$, so a feasible premise automatically has $p \le 1$. "Max" may be replaced by "sup" for a countable family.

Take independent uniform Rademacher signs σ_j. The learner may know the entire bump family, weights and prior. This additional knowledge only makes the lower bound stronger. Its only function-dependent observations are individual exact values $f_\sigma(x_t)$ at adaptively chosen points x_t. A query is not a linear functional, derivative, simultaneous collection of values or another richer oracle response.

The learner's random seed U is independent of the signs. Each query and stopping decision is a measurable function of U and the previous queries/answers. The output $g(T, x)$ is jointly measurable in the final transcript T and x. These conditions give measurable L^2 losses and regular conditional distributions. Algorithms whose output has infinite loss cannot improve the minimax bound.

The main statement assumes at most m queries almost surely, for a nonnegative integer m. Expected-budget and random-stopping variants are stated in Appendix C.4. The loss and budget include every run; no postselection discards unfavorable transcripts.

Point evaluation of the truth is unambiguous because the function class consists of specified continuous functions, rather than arbitrary representatives of L^2 equivalence classes. Clipping an L^2-valued output is a well-defined 1-Lipschitz map on L^2, so the norm-risk conversion also applies when estimators are formalized as Borel L^2-valued random variables. Pointwise jointly measurable representatives may equivalently be used as in the conditional lemma.

The comparison allows exact drawing from the known infinite-dimensional Gaussian law as an abstract randomized procedure. Such draws belong to ℓ^2 almost surely. This is admissible in a minimax statement without a computation or finite-precision constraint; it is not an implementation claim.

B GAUSSIAN PARTITION AND SIGNED DISTANCE FUNCTIONS

B.1 A SIMULTANEOUS SPHERICAL-PARTITION LEMMA

Write $\gamma_k = N(0, I_k)$, set

$$k_0 = 10^4, \quad \delta = 10^{-3}, \quad N_k = 2^k, \quad h_k = \frac{\delta}{4\sqrt{k}}. \tag{B.1}$$

Lemma B.1. For every integer $k \ge k_0$, there are distinct unit vectors $v_1, \dots, v_{N_k} \in \mathbb{R}^k$ whose closed conic Voronoi cells

$$C_i = \{z : \langle z, v_i\rangle \ge \langle z, v_j\rangle \text{ for every } j\} \tag{B.2}$$

satisfy

$$\max_i \gamma_k(C_i) \le e^{-k/64}, \tag{B.3}$$

and

$$\gamma_k\left\{z : \max_i \operatorname{dist}(z, C_i^c) \ge h_k\right\} \ge \frac{1}{16}. \tag{B.4}$$

Here distance to the complement is zero outside a cell and on its boundary. Distinct centers give disjoint cell interiors, with all ties contained in finitely many hyperplanes of Gaussian measure zero.

B.1.1 Proof: cap estimates

For a uniform point V on S^{k-1} and a fixed unit vector u, let $T_k(t) = \mathbb{P}(\langle u, V\rangle \ge t)$. The density on $(-1, 1)$ is

$$p_k(s) = c_k(1-s^2)^{(k-3)/2}, \qquad \frac{1}{2} \le c_k \le \sqrt{k}. \tag{B.5}$$

For completeness, the normalizing integral is at most 2. On $[-1/\sqrt{k}, 1/\sqrt{k}]$ its integrand is at least $(1-1/k)^{(k-3)/2} \ge 1/2$, so the integral is at least $1/\sqrt{k}$. These prove both bounds in (B.5).

For $k \ge k_0$, the following deliberately loose consequences suffice:

$$\begin{aligned} T_k(1/2) &\ge \frac{1}{200}(0.7399)^{(k-3)/2} \ge e^{-k/4}, \\ T_k(0.95) &\le \sqrt{k}(0.0975)^{(k-3)/2} < 2^{-k}, \\ T_k(1/4) &\le \sqrt{k}(15/16)^{(k-3)/2} \le \sqrt{k}e^{-(k-3)/32} \le e^{-k/64}. \end{aligned} \tag{B.6}$$

The first lower bound integrates (B.5) over $[0.50, 0.51]$; $0.7399 > e^{-1/3}$ verifies its last inequality for these k. The second uses $0.0975 < 1/8$. The last follows from $\frac{1}{2}\log k + 3/32 \le k/64$, true at k_0 and thereafter by differentiation. No asymptotic cap formula is needed.

Let q_k be the unique number with $T_k(q_k) = 2^{-k}$. Equation (B.6) implies

$$1/2 < q_k < 0.95. \tag{B.7}$$

For $x \in [0.5, 0.951]$, integrate the density on $[x, x + 1/(100k)]$. This interval lies below 0.96, and

$$|(\log p_k)'(s)| = \frac{(k-3)s}{1-s^2} \le 13k \quad (0.5 \le s \le 0.96).$$

Consequently

$$T_k(x) \ge \frac{e^{-0.13}}{100k} p_k(x), \qquad \frac{p_k(x)}{T_k(x)} < 120k. \tag{B.8}$$

Integrating $(\log T_k)' = -p_k/T_k$ between q_k and $q_k + \delta/k$ yields

$$T_k(q_k + \delta/k) \ge e^{-0.12} 2^{-k}. \tag{B.9}$$

B.1.2 PROOF: CONSTANT EXPECTED INTERIOR MASS

Choose $V_1, \dots, V_{N_k}$ independently and uniformly on the sphere. For each *fixed* direction u, their projections are independent. The disjoint events in which one projection is greater than $q_k + \delta/k$ and all others are at most q_k have total probability

$$N_k T_k(q_k + \delta/k)(1 - 1/N_k)^{N_k - 1} \geq e^{-0.12} e^{-1} > \frac{1}{4}. \tag{B.10}$$

The winner exceeds every other projection by at least δ/k. No independence of projections for a fixed codebook under a random input is asserted.

Independently write $Z = RU \sim \gamma_k$, with U uniform on the sphere and independent radius R. Since $\mathbb{E}R^2 = k$ and $\mathrm{Var}(R^2) = 2k$, Chebyshev gives

$$\mathbb{P}(R \geq \sqrt{k}/2) \geq 1 - \frac{32}{9k} > \frac{1}{2}. \tag{B.11}$$

On the event in (B.10), a point RU is in the winning cell with distance to the complement

$$\min_{j \neq i} \frac{\langle RU, V_i - V_j \rangle}{\|V_i - V_j\|} \geq \frac{R\delta}{2k}. \tag{B.12}$$

The equality with a minimum of face distances follows because the complement of the intersection of halfspaces is their union of complementary open halfspaces. Distinct centers occur almost surely. On (B.11), (B.12) is at least h_k.

Let $B(V_1, \dots, V_{N_k})$ be the Gaussian probability in (B.4) for the random codebook. Integrating (B.10)–(B.12), using independence of the radius, gives

$$\mathbb{E}B > \frac{1}{8}. \tag{B.13}$$

B.1.3 PROOF: ALL CELLS SMALL, ON THE SAME CODEBOOK

Fix a Euclidean $1/4$-net of the unit sphere of cardinality at most 9^k; the usual disjoint-ball volume argument supplies it. At each net point, the chance that all projections are below $1/2$ is at most

$$(1 - T_k(1/2))^{N_k} \leq \exp\{-\exp[(\log 2 - 1/4)k]\} \leq \exp\{-e^{0.4k}\}.$$

Thus with event A of probability

$$\mathbb{P}(A^c) \leq 9^k e^{-e^{0.4k}} < 1/16, \tag{B.14}$$

every net point has a center with projection at least $1/2$. On A, every unit direction has a center with projection at least $1/4$, because a unit-center linear functional changes by at most the distance to the net point. Therefore a winning cell for v_i is contained in

$$\{z \neq 0 : \langle z/\|z\|, v_i \rangle \geq 1/4\}$$

apart from the zero vector. Its Gaussian probability is at most $T_k(1/4)$, proving (B.3) for every cell on A.

Finally, $0 \leq B \leq 1$, so (B.13)–(B.14) imply

$$\mathbb{E}[B\mathbf{1}_A] \geq \mathbb{E}B - \mathbb{P}(A^c) > 1/16. \tag{B.15}$$

Hence there is a deterministic distinct-center codebook in A with $B \geq 1/16$. This proves (B.3) and (B.4) simultaneously. No high-probability interior-mass assertion, independence between A and B, or equal-cell-volume assertion is used. □

B.2 Bounded signed distance functions

Fix a codebook from Lemma B.1 and set

$$b_i(z) = \min\{h_k, \operatorname{dist}(z, C_i^c)\}, \qquad g_\theta(z) = \sum_{i=1}^{N_k} \theta_i b_i(z), \quad \theta \in \{-1, 1\}^{N_k}. \tag{B.16}$$

Each b_i is continuous and 1-Lipschitz, positive only in its cell's interior; at most one is positive at any point. In particular $\|g_\theta\|_\infty \le h_k$.

Global Lipschitz check, including opposite signs. Within the same cell, the difference is bounded by distance since b_i is 1-Lipschitz. If z and z' are in different cell interiors, open balls centered there with radii $d_i = \operatorname{dist}(z, C_i^c)$ and $d_j = \operatorname{dist}(z', C_j^c)$ lie in disjoint interiors. They cannot intersect, so $d_i + d_j \le \|z - z'\|$. Therefore

$$|\theta_i b_i(z) - \theta_j b_j(z')| \le b_i(z) + b_j(z') \le \|z - z'\|. \tag{B.17}$$

For boundary points the bump is zero and the same conclusion follows by continuity (or distance to the closed boundary). This proves $\operatorname{Lip}(g_\theta) \le 1$ on all of $\mathbb{R}^k$, not just within cells.

All integrals/events here are measurable. Away from the null event of coincident centers, the interior-distance function is the positive part of the minimum of the finitely many normalized linear inequalities above. It is jointly measurable in the centers and z. This permits the use of Fubini and the deterministic selection. The existence proof is nonconstructive, which is permitted by the minimax lower-bound claim.

C Polynomial scaling and adaptive lower bound

C.1 Reuse of the geometry with the correct physical scale

Lemma B.1 fixes the universal numbers

$$k_0 = 10000, \qquad \delta = 10^{-3}, \qquad h_k = \delta/(4\sqrt{k}).$$

For every integer $k \ge k_0$ it supplies finitely many nonnegative continuous functions b_i on $\mathbb{R}^k$, with pointwise disjoint positive sets, such that every signed sum $g_\sigma = \sum_i \sigma_i b_i$ is globally 1-Lipschitz and bounded by h_k. For $Z \sim N(0, I_k)$,

$$\mathbb{P}\{\max_i b_i(Z) = h_k\} \ge 1/16, \qquad \max_i \mathbb{P}\{b_i(Z) > 0\} \le e^{-k/64}. \tag{C.1}$$

These are properties of one deterministic codebook for each k. Its construction and constants are reused unchanged, including the simultaneous mass guarantees and the Lipschitz check for opposite signs across cell boundaries.

Use the covariance block $j = k+1, \dots, 2k$ and define

$$L_k x = \left(\frac{\langle x, e_j \rangle}{\sqrt{\lambda_j}} \right)_{j=k+1}^{2k}. \tag{C.2}$$

Then $L_k X \sim N(0, I_k)$, and (A.1) gives

$$\|L_k\| = \max_{k < j \le 2k} \lambda_j^{-1/2} \le c_\lambda^{-1/2} (2k)^{\tau/2}. \tag{C.3}$$

Define the safe physical Lipschitz scale and its nominal amplitude by

$$s_k^0 = \sqrt{c_\lambda}(2k)^{-\tau/2}, \qquad s_k^0 h_k = A_{\tau,c}\, k^{-\beta}, \qquad A_{\tau,c} = \frac{\delta\sqrt{c_\lambda}}{4\,2^{\tau/2}}. \tag{C.4}$$

For arbitrary large c_λ, the nominal amplitude need not be below one at k_0. To enforce the global class exactly, take a common scalar reduction

$$s_k = \min\{s_k^0, h_k^{-1}\}, \qquad a_k = s_k h_k = \min\{A_{\tau,c}k^{-\beta}, 1\}. \tag{C.5}$$

No geometric bump is altered: all are multiplied by the same positive scalar. Since $s_k \le s_k^0$, equations (C.3)–(C.5) give $s_k\|L_k\| \le 1$. Define physical bumps and signs by

$$\phi_i(x) = s_k b_i(L_k x), \qquad f_\sigma(x) = \sum_i \sigma_i \phi_i(x) = s_k g_\sigma(L_k x). \tag{C.6}$$

Every f_σ belongs to $\mathcal{F}$ on all of ℓ^2: its Lipschitz constant is at most one and its absolute value at most $a_k \le 1$. The positive sets of the physical bumps remain pointwise disjoint, including at arbitrary active-query locations.

Let $w_i = \int \phi_i^2\, d\mu$, $W = \sum_i w_i$ and $w_* = \max_i w_i$. Equation (C.1) and the amplitude bound imply

$$W \ge a_k^2/16, \qquad W \le a_k^2, \qquad w_* \le a_k^2 e^{-k/64}. \tag{C.7}$$

The first inequality uses the unchanged plateau probability; the last uses the unchanged cell probability. There is no equal-weight or positive-minimum-weight assumption. The entire independent-sign family is admissible; no Fano code subset is substituted for it.

C.2 Posterior under adaptive queries

In Appendices C.2–C.4, we use the generic family $(b_j)_{j\in J}$ on ℓ^2, with amplitude bound a and weights w_j, from Appendix A.2. Thus $x \in \ell^2$ in the argument below. The application in Appendix C.5 takes $J = \{1, \dots, N_k\}$, the generic function $b_i = \phi_i$ from (C.6), and $a = a_k$ from (C.5).

Define $j(x)$ to be the unique index with $b_j(x) > 0$, if one exists, and a null symbol otherwise. It is measurable because J is countable and the nonzero sets are measurable and disjoint. Let V_t be the indices queried at a point with positive bump value by time t. Queries at zeros reveal no index, and repeated visits add no new index.

Let $\mathcal{F}_t$ include U and the complete query/answer history through query t. At time zero the signs are independent and fair, conditional on U. Inductively, x_t and $j(x_t)$ are $\mathcal{F}_{t-1}$-measurable. There are three cases:

- If $j(x_t)$ is null, the answer is zero and supplies no sign information.
- If $j(x_t)$ is already in V_{t-1}, its sign is already known, and the answer is determined by the prior transcript.
- Otherwise the answer is $\sigma_{j(x_t)} b_{j(x_t)}(x_t)$, which reveals exactly one currently independent fair sign. Conditional on that answer, all other unvisited signs remain independent and fair.

Consequently, conditional on the transcript, the posterior on the signs indexed by $J \setminus V_t$ is their product Rademacher law. This statement is understood through all finite collections of unvisited coordinates when J is countable. The query location and the selected index are part of the transcript; their adaptivity does not create an additional observation about a sign the learner has not revealed. The induction conditions on the learner's information, not on a trajectory generated with knowledge of hidden signs.

In particular, for any terminal transcript T with visited set V, the squared-loss posterior mean is

$$\bar{f}_T(x) = \sum_{j\in V} \sigma_j b_j(x),$$

and the conditional pointwise variance is

$$\mathrm{Var}(f_\sigma(x) \mid T) = \sum_{j \notin V} b_j(x)^2.$$

For every bounded measurable output g_T, integration of the conditional variance identity gives

$$\mathbb{E}[\|g_T - f_\sigma\|^2_{L^2(\mu)} \mid T] = \|g_T - \bar{f}_T\|^2_{L^2(\mu)} + \sum_{j \notin V} w_j. \tag{C.8}$$

Tonelli justifies the sum/integral interchange. Thus the conditional optimal squared Bayes risk equals the unvisited energy. For any policy using at most m queries,

$$\sum_{j \notin V} w_j = W - \sum_{j \in V} w_j \geq W - |V| w_* \geq W - m w_*. \tag{C.9}$$

This final inequality is pathwise in the transcript, despite adaptive choice of V. It does not require the visited indices to be uniformly distributed or independent of the signs already seen. In particular, choosing the largest-weight bumps first cannot remove more than mw_* energy.

The same integrated lower bound holds for arbitrary measurable outputs by the nonnegative-loss conditional decomposition or by considering finite-loss cases. For the norm-risk proof below it is enough to apply (C.8) to bounded clipped outputs.

C.3 Conversion to expected integrated norm

Disjointness and the amplitude bound imply $f_\sigma(x) \in [-a, a]$ for every x. For any learner output g, let g^a be its pointwise projection onto $[-a, a]$. Then

$$\|g - f_\sigma\|_{L^2} \geq \|g^a - f_\sigma\|_{L^2} =: Z, \qquad 0 \leq Z \leq 2a.$$

Apply the universal squared-risk bound (C.8)–(C.9) to g^a itself. It yields

$$\mathbb{E}Z^2 \geq W - m w_*.$$

Since $Z^2 \leq 2aZ$, it follows that every learner satisfies

$$\mathbb{E}_{\sigma,U}\|g - f_\sigma\|_{L^2} \geq \frac{(W - m w_*)_+}{2a}. \tag{C.10}$$

Here the expectation includes the random transcript determined by the signs and learner's seed. This is a bounded-loss argument, not a reversal of Jensen's inequality. Nor is it a claim that the posterior mean minimizes expected norm loss.

If

$$m e^{-ck} \leq p/2, \tag{C.11}$$

then $W - m w_* \geq p a^2/2$. Equations (C.8)–(C.10) give the conclusions:

$$\mathbb{E}\|g^a - f_\sigma\|^2_{L^2} \geq p a^2/2, \qquad \mathbb{E}\|g - f_\sigma\|_{L^2} \geq p a/4. \tag{C.12}$$

Because every signed function is assumed to lie in the target class, its supremum risk dominates this prior-average risk. Taking an infimum over arbitrary admissible adaptive randomized learners preserves both lower bounds. No estimator-class restriction such as linearity, locality or a finite computational budget was used.

C.4 Stopping, randomization and observation caveats

Bounded stopping time. If the learner stops at a transcript-measurable $T_{\mathrm{stop}} \le m$, the same posterior argument applies on each event $\{T_{\mathrm{stop}} = t\}$. The stopping decision reveals no new sign beyond the observed history. Equations (C.9)–(C.12) remain unchanged. One need not invoke an optional-stopping theorem for a likelihood ratio or pad the transcript with fictitious function-dependent observations.

Expected query budget. If T_{stop} is only almost surely finite, the conditional residual is at least $W - T_{\mathrm{stop}} w_*$. Averaging gives squared risk at least $W - w_* \mathbb{E} T_{\mathrm{stop}}$ and norm risk at least $(W - w_* \mathbb{E} T_{\mathrm{stop}})_+ / (2a)$. Thus an expected budget $\mathbb{E}_{\sigma,U} T_{\mathrm{stop}} \le m$ under this prior also suffices for (C.12). A uniform expected-budget constraint over the entire function class implies that prior-average constraint. Under this weaker budget the inequality $W - m w_*$ is no longer a pathwise statement; only its averaged version is asserted. An unbounded or uncharged query budget is not covered.

Randomized learner. Conditioning initially on an independent seed U is legitimate even if it contains all future random coins: unlike hidden function signs, these coins are independent of the prior. Locations chosen from future unobserved function information, or an oracle-dependent external stopping signal, would not be admissible.

Pointwise disjointness. The one-sign-per-query fact must hold at every location a learner can choose, not just at typical passive inputs. For a Gaussian law with full support in the domain under consideration, continuity upgrades almost-everywhere disjointness to pointwise disjointness on that domain: a positive overlap would contain an open set of positive Gaussian measure. With ambient covariance null directions, this implication holds on the closed support only; our constructed bumps are independently pointwise disjoint on the whole ambient space. Continuity follows here from the assumption on all signed sums, since changing one sign expresses $2b_j$ as the difference of two Lipschitz functions. Stating pointwise disjointness explicitly is the clearest version of the lemma. Merely orthogonal bumps or unspecified small overlap would not support this proof for an exact noiseless oracle.

Passive noisy data. A noiseless active-query learner can simulate the passive unit-noise experiment by drawing X from the known μ, querying $f(X)$, and adding independent $N(0,1)$ noise before passing the sample to a passive estimator. Therefore a lower bound valid for all noiseless active learners also applies to that passive noisy experiment. This comparison is in the correct direction: the active noiseless experiment is more informative.

C.5 Strongest-access lower bound, including every budget

Put independent uniform signs on the physical bumps. The conditional query lemma in Appendices C.2–C.4 applies with $p = 1/16,\ c = 1/64$ and amplitude a_k. After any transcript with at most m exact point queries, its unvisited squared Bayes risk is at least $W - m w_*$. This remains valid for arbitrary adaptive locations, independent randomization and bounded stopping.

For every $m \ge 1$ choose the integer:

$$k(m) = \max\{10000, \lceil 64 \log(32(m+1)) \rceil\}. \tag{C.13}$$

Then $m e^{-k/64} \le m/[32(m+1)] < 1/32$. Consequently the prior-average squared loss of every learner is at least $a_k^2/32$. Applying this bound to its output clipped into $[-a_k, a_k]$, and using clipped L^2 error $Z \le 2a_k$, also gives prior-average norm loss at least $a_k/64$. Supremum risk dominates prior risk, so

$$Q_m^{A,0} \ge a_{k(m)}^2/32, \qquad R_m^{A,0} \ge a_{k(m)}/64. \tag{C.14}$$

The norm statement follows from $\mathbb{E} Z^2 \le 2a_k \mathbb{E} Z$; it is not reverse Jensen. The squared statement follows directly from unvisited sign variance.

For a constant explicit in the envelope, set

$$d_{\tau,c} = \min\{1, A_{\tau,c} k_0^{-\beta}\} > 0. \tag{C.15}$$

Let $L = \log(m+2) > 1$. The inequality $\log 32 < 4$ implies

$$k(m) \le k_0 L.$$

Since $L^{-\beta} \le 1$, equations (C.5) and (C.15) give

$$a_{k(m)} \ge \min\{1, A_{\tau,c} k_0^{-\beta} L^{-\beta}\} \ge d_{\tau,c} L^{-\beta}. \tag{C.16}$$

Thus valid all-m lower constants are

$$\boxed{R_m^{A,0} \ge \frac{d_{\tau,c}}{64} L^{-\beta}, \qquad Q_m^{A,0} \ge \frac{d_{\tau,c}^2}{32} L^{-q}.} \tag{C.17}$$

This handles both small and large covariance scales without choosing an additional unreported dimension threshold. It also shows that the constants depend on c_λ and τ only on this side of the proof.

D Hermite estimator and completion of the theorem

D.1 General-covariance Hermite energy and coefficient estimation

Let h_r denote the orthonormal probabilists' Hermite polynomials in one standard Gaussian coordinate. For every finite-support multi-index γ, let

$$H_\gamma(x) = \prod_j h_{\gamma_j}\left(\frac{\langle x, e_j\rangle}{\sqrt{\lambda_j}}\right), \quad c_\gamma = \langle f, H_\gamma\rangle_{L^2(\mu)}, \quad \nu_\gamma = \sum_j \frac{\gamma_j}{\lambda_j}. \tag{D.1}$$

These product Hermites form a complete orthonormal basis in $L^2(\mu)$. One can see completeness by finite-dimensional Hermite completeness and convergence of conditional expectations onto the first d coordinates.

To justify regularity without an infinite-dimensional differentiability assumption, average the tail:

$$f_d(x_1, \dots, x_d) = \mathbb{E}_{X_{>d}} f\left(\sum_{j\le d} x_j e_j + X_{>d}\right). \tag{D.2}$$

This is bounded by one and Euclidean 1-Lipschitz. Its weak gradient has squared norm at most one almost everywhere. The finite-dimensional Gaussian has a strictly positive density. Gaussian integration by parts for weak derivatives gives, for a d-dimensional Hermite index α,

$$\langle \partial_j f_d, H_\alpha\rangle_{L^2(\mu_d)} = \sqrt{\frac{\alpha_j + 1}{\lambda_j}}\, c_{\alpha+e_j}. \tag{D.3}$$

The coefficients on the right are the original coefficients of f, by conditional expectation. Bounded Lipschitz functions and their bounded weak gradients permit the usual cutoff/smooth-approximation justification of (D.3). Parseval applied to the derivatives yields

$$\sum_{\operatorname{supp}\gamma\subseteq[d]} \nu_\gamma c_\gamma^2 = \sum_{j\le d} \|\partial_j f_d\|^2_{L^2(\mu_d)} \le 1.$$

Increasing d gives the exact energy bound

$$\boxed{\sum_{\gamma} \nu_\gamma c_\gamma^2 \le 1.} \tag{D.4}$$

The weight uses the inverse covariance eigenvalue, not its square root or inverse square root. This is a derivative seminorm statement. It is not an assertion that the full Sobolev norm is at most one; the separate L^2 part is bounded by one.

For $E > 0$, let $S_E = \{\gamma : \nu_\gamma \le E\}$. It is finite: a nonzero γ_j requires $j \le (C_\lambda E)^{1/\tau}$, and each $\gamma_j \le E\lambda_j$. The constant index is always retained, even when E is small. Define from passive noisy data

$$\widehat{c}_\gamma = \frac{1}{m}\sum_{t=1}^{m} Y_t H_\gamma(X_t), \qquad \widehat{f}_E = \sum_{\gamma \in S_E} \widehat{c}_\gamma H_\gamma. \tag{D.5}$$

These estimates are unbiased. Since $\mathbb{E}[Y_t^2 \mid X_t] \le 2$ and $\int H_\gamma^2 \, d\mu = 1$,

$$\mathrm{Var}(\widehat{c}_\gamma) \le 2/m.$$

Orthogonality removes all cross terms in the L^2 error even though different coefficient estimates use the same data. The omitted energy in (D.4) is at most $1/E$. Therefore uniformly over $f \in \mathcal{F}$,

$$\mathbb{E}_f \|\widehat{f}_E - f\|_{L^2(\mu)}^2 \le E^{-1} + 2|S_E|/m. \tag{D.6}$$

This uses ordinary noisy function values and finitely many input coordinates, not derivative labels or exact coefficient access. Unbounded Hermite values cause no missing variance assumption because f is uniformly bounded.

D.2 Counting with arbitrary polynomial exponent

For $t > 0$, nonnegative summation over finite-support indices gives

$$|S_E| \le e^{tE} \prod_{j \ge 1} (1 - e^{-t/\lambda_j})^{-1},$$

$$\log |S_E| \le tE + \sum_{\ell \ge 1} \frac{1}{\ell} \sum_{j \ge 1} e^{-\ell t/\lambda_j}. \tag{D.7}$$

The product converges: (A.1) implies $1/\lambda_j \ge j^\tau / C_\lambda$, and the resulting exponential sequence is summable for every $t > 0$. Because the function $x \mapsto e^{-bx^\tau}$ is decreasing for $b > 0$,

$$\sum_{j \ge 1} e^{-\ell t/\lambda_j} \le \sum_{j \ge 1} e^{-(\ell t/C_\lambda) j^\tau} \le \int_0^\infty e^{-(\ell t/C_\lambda) x^\tau} \, dx = \Gamma(1 + 1/\tau) \left(\frac{C_\lambda}{\ell t}\right)^{1/\tau}. \tag{D.8}$$

Define the finite positive number

$$G_\tau = \Gamma(1 + 1/\tau)\zeta(1 + 1/\tau). \tag{D.9}$$

Equations (D.7)–(D.9) give

$$\log |S_E| \le tE + G_\tau (C_\lambda / t)^{1/\tau}. \tag{D.10}$$

No integer or rational assumption on τ or the eigenvalues is used. This is a Laplace-transform count of finite multi-indices, so it remains valid for nonintegral weights.

Put $u = C_\lambda E$ and $t = C_\lambda \theta$. Minimize $u\theta + G_\tau \theta^{-1/\tau}$ at

$$\theta = \left(\frac{G_\tau}{\tau u}\right)^{\tau/(\tau+1)}.$$

The result is

$$\boxed{\log |S_E| \le B_\tau (C_\lambda E)^{1/q}, \qquad B_\tau = (\tau+1)\left(\frac{G_\tau}{\tau}\right)^{\tau/(\tau+1)}.} \tag{D.11}$$

For $\tau = 2$, $G_\tau = (\sqrt{\pi}/2)\zeta(3/2)$ and $B_\tau = 3(G_\tau/2)^{2/3}$, giving the optimized exact-spectrum constant. The extra power one in $q = \tau + 1$ comes from this weighted multi-index count; omitting it would give the wrong rate.

D.3 Passive upper bound and explicit all-budget constants

For $m \ge 3$ take the real threshold

$$E_m = C_\lambda^{-1}\left(\frac{\log m}{2B_\tau}\right)^q. \tag{D.12}$$

Then (D.11) gives $|S_{E_m}| \le \sqrt{m}$. Substituting into (D.6) gives

$$Q_m^{P,1} \le C_\lambda (2B_\tau)^q (\log m)^{-q} + 2m^{-1/2}. \tag{D.13}$$

There is no requirement $E_m \ge 1$ and no integer rounding of this threshold. For $m = 1, 2$ use the zero estimator, whose squared risk is at most one.

For an explicit single constant, define

$$U_{\tau,C} = \max\left\{(\log 4)^q, C_\lambda (4B_\tau)^q + 2(4q/e)^q\right\}. \tag{D.14}$$

To verify the bound for $m \ge 3$, use $L = \log(m+2) \le 2\log m$. The first term of (D.13), multiplied by L^q, is at most $C_\lambda(4B_\tau)^q$. For the second term write $x = \log m > 0$ and use

$$2m^{-1/2}L^q \le 2^{q+1}x^q e^{-x/2} \le 2^{q+1}(2q/e)^q = 2(4q/e)^q, \tag{D.15}$$

where the maximum over $x > 0$ occurs at $x = 2q$. For $m = 1, 2$, $L^q \le (\log 4)^q$. Therefore, simultaneously for every integer $m \ge 1$,

$$\boxed{Q_m^{P,1} \le U_{\tau,C} L^{-q}, \qquad R_m^{P,1} \le \sqrt{U_{\tau,C}}\, L^{-\beta}.} \tag{D.16}$$

The norm upper bound follows from Jensen applied to this estimator, followed by supremum and infimum in the valid direction. The upper constants depend on τ and C_λ, not on the lower envelope constant. Their noise-related term remains even for small covariance scale because the class contains unknown constant functions.

D.4 Sandwich and completed parameter dependence

An adaptive noiseless learner can reproduce an adaptive noisy policy by adding fresh independent Gaussian noise to each answer. An adaptive noisy learner can reproduce passive sampling by drawing independent inputs from the known μ. Their budgets, transcript measurability and output classes match. Therefore

$$R_m^{A,0} \le R_m^{A,1} \le R_m^{P,1}, \qquad Q_m^{A,0} \le Q_m^{A,1} \le Q_m^{P,1}. \tag{D.17}$$

Combining (C.17), (D.16) and (D.17), for every B among the three stated experiments and every $m \ge 1$,

$$\boxed{\frac{d_{\tau,c}}{64}[\log(m+2)]^{-q/2} \le R_m^B \le \sqrt{U_{\tau,C}}[\log(m+2)]^{-q/2},} \tag{D.18}$$

$$\boxed{\frac{d_{\tau,c}^2}{32}[\log(m+2)]^{-q} \le Q_m^B \le U_{\tau,C}[\log(m+2)]^{-q}.} \tag{D.19}$$

All constants in these displays were defined explicitly in (C.4), (C.15), (D.9), (D.11) and (D.14), with $k_0 = 10000$ and $\delta = 10^{-3}$. This proves (A.5). The lower bound uses only the lower spectral envelope in its physical scaling, while the upper bound uses the upper spectral envelope in counting. Both are needed for the stated matching rate; the proof does not silently replace the known actual eigenvalues in the estimator by an exact power law.

At $\tau = 2$ and $c_\lambda = C_\lambda = 1$, the scaling in (C.4) is exactly $1/(2k)$, the amplitude cap is inactive, and the lower argument reduces to the exact-spectrum active bound. Equations (D.4) and (D.11) reduce to the preceding Hermite energy and optimized partition count. Thus the original case is recovered with the same interpretation of the loss and point-query access.

E Alternative passive proof for the exact inverse-square spectrum

E.1 Exact statement and observation model

Let μ be the centered Gaussian law on ℓ^2 of

$$X = \sum_{j \ge 1} j^{-1} Z_j e_j, \qquad Z_j \overset{\text{iid}}{\sim} N(0,1). \tag{E.1}$$

The series is in ℓ^2 almost surely because $\sum_j j^{-2} < \infty$. Set

$$\mathcal{F} = \{f : \ell^2 \to \mathbb{R} : \|f\|_\infty \le 1,\ \mathrm{Lip}_{\ell^2}(f) \le 1\}. \tag{E.2}$$

The data are m independent pairs (X_t, Y_t), with full X_t observed and

$$Y_t = f(X_t) + \xi_t, \qquad \xi_t \overset{\text{iid}}{\sim} N(0,1), \tag{E.3}$$

where all noises and inputs are mutually independent. The law μ is known and fixed; the learner cannot choose another sampling law or query inputs adaptively. Define

$$R_m = \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\widehat{f} - f\|_{L^2(\mu)}. \tag{E.4}$$

The infimum allows all measurable estimators taking values in $L^2(\mu)$, without an output Lipschitz restriction or computational budget. Estimators with infinite risk do not improve the infimum.

Theorem. There are numerical constants $0 < c < C < \infty$ such that, for every integer $m \geq 1$,

$$c[\log(m+2)]^{-3/2} \leq R_m \leq C[\log(m+2)]^{-3/2}. \tag{E.5}$$

All logarithms are natural. Equation (E.5) is for expected norm; expected squared error has order $[\log(m+2)]^{-3}$, as follows from the same proof, not by reversing Jensen. Constants below are deliberately poor. The theorem provides neither useful numerical thresholds nor a practical estimation claim.

E.2 Bounded signed bumps and physical scaling

Fix a codebook from Lemma B.1 and set

$$b_i(z) = \min\{h_k, \operatorname{dist}(z, C_i^c)\}, \qquad g_\theta(z) = \sum_{i=1}^{N_k} \theta_i b_i(z), \quad \theta \in \{-1,1\}^{N_k}. \tag{E.6}$$

Each b_i is continuous and 1-Lipschitz, positive only in its cell's interior; at most one is positive at any point. In particular $\|g_\theta\|_\infty \leq h_k$.

Global Lipschitz check, including opposite signs. Within the same cell, the difference is bounded by distance since b_i is 1-Lipschitz. If z and z' are in different cell interiors, open balls centered there with radii $d_i = \operatorname{dist}(z, C_i^c)$ and $d_j = \operatorname{dist}(z', C_j^c)$ lie in disjoint interiors. They cannot intersect, so $d_i + d_j \leq \|z - z'\|$. Therefore

$$|\theta_i b_i(z) - \theta_j b_j(z')| \leq b_i(z) + b_j(z') \leq \|z - z'\|. \tag{E.7}$$

For boundary points the bump is zero and the same conclusion follows by continuity (or distance to the closed boundary). This proves $\operatorname{Lip}(g_\theta) \leq 1$ on all of $\mathbb{R}^k$, not just within cells.

Use the block $j = k+1, \ldots, 2k$ in (E.1), and define

$$L_k x = ((k+1)x_{k+1}, \ldots, (2k)x_{2k}), \qquad f_\theta(x) = \frac{1}{2k} g_\theta(L_k x). \tag{E.8}$$

Then $L_k X \sim \gamma_k$ and $\|L_k\|_{\ell^2 \to \mathbb{R}^k} = 2k$. Thus f_θ is globally 1-Lipschitz and

$$\|f_\theta\|_\infty \leq a_k, \qquad a_k = \frac{h_k}{2k} = \frac{\delta}{8k^{3/2}} < 1. \tag{E.9}$$

These are actual members of (E.2), not only functions on a high-probability shell. The uniform cap in (E.6) is essential to the next maximum-weight bound; unbounded conic distance would not give it in this form.

E.3 Weighted packing without minimum cell masses

Define the deterministic weights

$$w_i = \frac{\mathbb{E}[b_i(Z)^2]}{4k^2}, \quad W = \sum_i w_i, \quad w_* = \max_i w_i. \tag{E.10}$$

Equations (B.3), (B.4), and the cap yield

$$\frac{a_k^2}{16} \leq W \leq a_k^2, \qquad w_* \leq a_k^2 e^{-k/64}. \tag{E.11}$$

Every f_θ has squared $L^2(\mu)$ norm exactly W, since only one bump can be positive. Moreover

$$\|f_\theta - f_{\theta'}\|^2_{L^2(\mu)} = 4\sum_i w_i \mathbf{1}_{\{\theta_i \neq \theta'_i\}}. \tag{E.12}$$

Choose two independent uniform sign vectors. Their disagreement indicators are independent Bernoulli(1/2) variables. Hoeffding's elementary bounded-summand inequality gives

$$\begin{aligned}\mathbb{P}\left\{\sum_i w_i \mathbf{1}_{\{\theta_i \neq \theta'_i\}} < W/4\right\} &\leq \exp\left(-\frac{W^2}{8\sum_i w_i^2}\right) \\ &\leq \exp\left(-\frac{W}{8w_*}\right) \leq \exp\left(-\frac{e^{k/64}}{128}\right).\end{aligned} \tag{E.13}$$

The middle inequality is precisely $\sum_i w_i^2 \leq w_* W$. Tiny or zero individual weights do not cause a problem.

Put $D_k = e^{k/64}/128$ and $M_k = \lfloor e^{D_k/4} \rfloor$. Among M_k independent sign vectors, the union bound for all pairs is at most $M_k^2 e^{-D_k} < 1$. Therefore a deterministic choice of M_k vectors has every weighted disagreement at least $W/4$. For $k \geq k_0$,

$$\log M_k \geq \frac{e^{k/64}}{1024}, \qquad \min_{\theta \neq \theta'} \|f_\theta - f_{\theta'}\|_{L^2(\mu)} \geq \sqrt{W} \geq \frac{a_k}{4}. \tag{E.14}$$

This completes the construction: k physical coordinates, 2^k cells, amplitude of order $k^{-3/2}$, constant total plateau probability, maximum cell probability at most $e^{-k/64}$, and **logarithm of packing cardinality** at least $e^{k/64}/1024$. The distinction between cell count, packing count, and log-packing count is crucial.

E.4 Regression lower bound

Let $P_f^{(m)}$ be the law of the full data in (E.3). By conditioning on the inputs,

$$\mathrm{KL}(P_f^{(m)} \| P_0^{(m)}) = \frac{m}{2}\|f\|^2_{L^2(\mu)}. \tag{E.15}$$

In particular every member of (E.14) has KL at most $ma_k^2/2$. The reference function zero belongs to $\mathcal{F}$, although it need not be one of the packed functions. Under a uniform prior on the packing, mutual information is at most this average reference KL.

For any estimator, decode by its nearest packed function in $L^2(\mu)$, resolving ties deterministically. The finite nearest-neighbor rule is measurable. If the estimator's error is less than $a_k/8$, (E.14) implies correct decoding. Fano therefore gives

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \|\widehat{f} - f\|_{L^2(\mu)} \geq \frac{a_k}{8}\left[1 - \frac{ma_k^2/2 + \log 2}{\log M_k}\right]. \tag{E.16}$$

Take the explicit dimension

$$k = \max\{10^4, \lceil 128 \log(1024(m+2))\rceil\}. \tag{E.17}$$

Then (E.14) gives $\log M_k \geq 1024(m+2)^2$; since $a_k \leq 1$, the bracket in (E.16) is greater than $1/2$ for every $m \geq 1$. Hence

$$R_m \geq \frac{a_k}{16} = \frac{\delta}{128k^{3/2}} \geq c[\log(m+2)]^{-3/2}. \tag{E.18}$$

Here $k \leq C_1 \log(m+2)$ with a numerical constant, absorbing the large fixed k_0. The same decoding argument lower-bounds expected squared loss by $(a_k/8)^2/2$; it thus proves its claimed order from below without any reverse-Jensen step.

E.5 Matching uniform upper bound, with infinite-dimensional regularity justified

The following is the original exact-spectrum upper calculation. Let H_γ be the normalized product Hermite basis under (E.1), indexed by finite-support multi-indices, and set

$$c_\gamma = \langle f, H_\gamma \rangle_{L^2(\mu)}, \qquad \nu_\gamma = \sum_j j^2 \gamma_j.$$

The needed energy inequality can be proved without assuming infinite-dimensional Fréchet differentiability. For each d, average over the tail to obtain

$$f_d(x_1, \dots, x_d) = \mathbb{E}\left[f\left(\sum_{j \le d} x_j e_j + \sum_{j > d} j^{-1} Z_j e_j \right) \right].$$

This is bounded by 1 and 1-Lipschitz in ordinary Euclidean $\mathbb{R}^d$, so its weak gradient has norm at most 1 almost everywhere. Its Hermite coefficients are precisely the c_γ supported in the first d coordinates. Finite-dimensional Gaussian integration by parts and Parseval for each weak derivative give

$$\sum_{\operatorname{supp}\gamma \subseteq [d]} \nu_\gamma c_\gamma^2 = \sum_{j \le d} \|\partial_j f_d\|^2_{L^2(\mu_d)} \le 1.$$

Bounded Lipschitz functions belong to this Gaussian Sobolev space; the identity also follows by ordinary smooth approximation. Passing monotonically in d yields

$$\sum_\gamma \nu_\gamma c_\gamma^2 \le 1. \tag{E.19}$$

Cylinder functions are dense in $L^2(\mu)$, so the product Hermites are a complete basis. For $E > 0$, let $S_E = \{\gamma : \nu_\gamma \le E\}$, a finite set, and estimate

$$\widehat{c}_\gamma = \frac{1}{m} \sum_{t=1}^m Y_t H_\gamma(X_t), \qquad \widehat{f}_E = \sum_{\gamma \in S_E} \widehat{c}_\gamma H_\gamma. \tag{E.20}$$

The coefficients are unbiased and, because $\mathbb{E}[Y_t^2 \mid X_t] \le 2$, have variance at most $2/m$. Independence between coefficient estimates is unnecessary. Orthogonality and (E.19) give

$$\mathbb{E}_f \|\widehat{f}_E - f\|^2_{L^2(\mu)} \le E^{-1} + 2|S_E|/m. \tag{E.21}$$

For $t > 0$, the generating product and the decreasing-function integral bound imply

$$\begin{aligned} \log |S_E| &\le tE + \sum_{\ell \ge 1} \frac{1}{\ell} \sum_{j \ge 1} e^{-\ell t j^2} \\ &\le tE + C_0 t^{-1/2}, \qquad C_0 = \frac{\sqrt{\pi}}{2} \zeta(3/2). \end{aligned} \tag{E.22}$$

Taking $t = E^{-2/3}$ gives $\log |S_E| \le C_* E^{1/3}$, with $C_* = 1 + C_0$. For $m \ge 3$, choose

$$E = \left(\frac{\log m}{2C_*} \right)^3.$$

Then $|S_E| \le \sqrt{m}$, and (E.21) is at most $C[\log(m+2)]^{-3}$. Jensen proves the upper half of (E.5). The cases $m = 1, 2$ are absorbed by a constant using the zero estimator. This also supplies the squared-loss upper bound. Each retained polynomial uses finitely many observed coordinates; the calculation does not assume exact coefficient or derivative labels.

E.6 Passive regression lower bound and explicit all-budget constants

For the explicit constants below, write $a = a_k$ and $M = M_k$.

The common input law cancels in the likelihood ratio. Conditioning on all inputs and integrating the Gaussian label KL gives

$$D(P_f^{(m)} \| P_0^{(m)}) = (m/2)\|f\|_{L^2(\mu)}^2 = (m/2)W \le ma^2/2.$$

The average reference KL bounds the mutual information under the uniform finite packing prior. The zero reference need not itself be in the packing. An independent estimator seed can be adjoined to the observations without changing this bound.

Nearest-packing decoding is measurable because the packing is finite and L^2 distance is continuous. A wrong decoding forces norm error at least $a/8$. Thus the displayed Fano inequality (E.16) is valid for arbitrary measurable L^2-valued estimators. In particular it lower-bounds expected norm directly; taking a square root of an expected squared-error lower bound is not used.

Choose exactly the proof's integer

$$k = \max\{10000, \lceil 128 \log(1024(m+2))\rceil\}.$$

Then $\log M \ge 1024(m+2)^2$. Using $a \le 1$ and $\log 2 \le 1$, the fraction subtracted in Fano is at most

$$\frac{m/2 + \log 2}{1024(m+2)^2} \le \frac{1}{2048(m+2)} < 1/2$$

for every integer $m \ge 1$. Therefore

$$R_m \ge a/16 = \delta/(128k^{3/2}).$$

To make the final dimension comparison explicit, put $L = \log(m+2) \ge \log 3 > 1$. Since $\log 1024 < 7$,

$$\lceil 128 \log(1024(m+2))\rceil \le 128(L+7) + 1 \le 1025L \le 10000L.$$

Also $10000 \le 10000L$. Hence $k \le 10000L$, and one valid all-m lower constant is

$$c = \frac{10^{-3}}{128(10000)^{3/2}}.$$

For the minimax expected squared loss Q_m, the same decoding error bound yields

$$Q_m \ge a^2/128 \ge \frac{\delta^2}{8192(10000)^3}[\log(m+2)]^{-3}.$$

This proves the squared-loss lower statement independently of Jensen as well. The tiny constants reflect conservative choices, not a finite-m exception or useful empirical threshold.

For completeness, numerical universal constants cover the upper bound for all m. Since $\zeta(3/2) < 3$ and $\sqrt{\pi}/2 < 1$, $C_* < 4$. For $m \ge 3$, $\log(m+2) \le 2\log m$, so the squared bound, multiplied by $[\log(m+2)]^3$, is at most

$$8(2C_*)^3 + 16 \sup_{x>0} x^3 e^{-x/2} = 8(2C_*)^3 + 3456e^{-3} < 5000.$$

The supremum occurs at $x = 6$. For $m = 1, 2$ the zero estimator has squared risk at most one, and the same constant covers those cases. Consequently

$$Q_m \le 5000[\log(m+2)]^{-3}, \qquad R_m \le 100[\log(m+2)]^{-3/2}$$

for every integer $m \ge 1$, where the second follows from Jensen and $\sqrt{5000} < 100$. The retained coefficients use finitely many ordinary input coordinates and noisy function labels. This calculation is an existence/risk guarantee, not a numerical-conditioning or practical-runtime guarantee.

F EXACT-MODEL COMPARISON DETAILS

The contribution is the uniform-class, known-covariance statistical characterization in Theorem 1. Classical partitions, signed distance bumps, Gaussian derivative bounds, Hermite projection and estimation, and Bayes-risk arguments are established ingredients. This appendix records the exact comparison interfaces and the limits of the source readings. It does not assert exhaustive priority clearance.

F.1 TARGETS, CLASSES, INFORMATION AND QUANTIFIERS

Framework	Exact comparison dimensions
Theorem 1	Target: an entire scalar functional in $L^2(\mu)$. Class: one global sup/Lipschitz unit ball on ℓ^2. Law: known actual covariance and eigenbasis, with $c_\lambda j^{-\tau} \le \lambda_j \le C_\lambda j^{-\tau}$, fixed $\tau > 1$. Information: passive noisy or randomized adaptive noisy/exact point values. Loss: $\sup_f \mathbb{E}_f \|\widehat{f} - f\|_2$ or its separately defined squared counterpart.
Kovachki et al. (2024)	Operator/functional recovery and data-complexity lower bounds. Section 2.1 sampling widths fix sample points and a decoder before the worst target. Theorems 2.6–2.7 and 2.12 include a polynomial-Gaussian setting; they do not provide the sharp randomized adaptive characterization here. A deterministic $\inf_{\text{design}} \sup_f$ statement does not automatically imply an $\inf_{\text{randomized policy}} \sup_f \mathbb{E}$ statement.
Adcock et al. (2026a)	Recovery in a Gaussian Sobolev-bounded class intersected with individually Lipschitz functions; there is no common global Lipschitz bound in class (3.2). Definitions 4.1–4.3 permit adaptive bounded linear information, including coefficient integrals. These are distinct class and information quantifiers from a uniform Lipschitz ball and point queries.
Bizeul & Klartag (2025)	Finite-dimensional integrated regression and entropy under log-concave input measures. The displayed product-measure entropy range in Theorem 1.5 requires $\varepsilon > d^{-1/4}$. The sufficient whitened bump scale in this paper is proportional to $d^{-1/2}$. Their construction can also be bounded; boundedness is not the distinguishing condition.
Adcock et al. (2026c)	Whole-operator minimax recovery with common bounded Lipschitz classes, noisy values, and fixed or optimized iid design. For their random-design risk (2.3), take input ℓ^2, scalar output, $p = 2$, $B = L = \sigma = 1$, and exact Gaussian eigenvalues $\lambda_j = j^{-\tau}$, fixed $\tau > 1$. The passive estimator here is admissible.
Creutzig et al. (2008)	Target: the one real number $\int f \, d\mu$ for a common Lipschitz class on a Banach space. Sequential full-space point queries are allowed. Recovery of a scalar integral is distinct from recovery of the entire functional in integrated norm.

Framework	Exact comparison dimensions
Plaskota et al. (2000)	The unknown function itself is a Gaussian-process trajectory on $\mathbb{R}_+$, with mean-square derivative regularity and point samples of that trajectory. The expectation is over the function prior. Here the Gaussian distribution describes the test input, and the risk is worst case over deterministic unknown functionals.
Suzuki & Karvonen (2026)	Point-value approximation of $f : \mathbb{R} \to \mathbb{R}$ in Gaussian-weighted L^p under $W^{\alpha,q}$ regularity, $p < q$. The input domain is one-dimensional; Corollary 3.3's $n^{-\alpha}$ rate does not supply a Hilbert-input polynomial-covariance result with tracked dimension dependence.
Krieg et al. (2026)	Arbitrary range-bounded nonlinear measurements of the entire unknown object, with worst-case deterministic noise of size δ. Theorems 2–3 relate these errors to entropy numbers. Their permitted measurements are richer than individual point values, and the result does not compute this Gaussian Lipschitz entropy.

F.2 Classical geometry and spectral estimation

The padded-partition theorem and signed-distance construction in Gupta et al. (2003, Theorem 3.2 and Section 4.2) already provide the relevant geometric mechanism. Appendix B retains a self-contained conic proof for checkability, without claiming priority for partitions or signed distances. The uniform Lipschitz entropy and coordinate-CDF construction in Lanthaler (2024, Section 3.3) are operator-learning antecedents.

The Gaussian regularity and Hermite framework in Adcock et al. (2026a, Remark 2.12 and Sections 3–4), and the subsequent Hermite-estimation work of Adcock et al. (2026b), are direct spectral precedents. Our bound $\sum_\gamma \nu_\gamma c_\gamma^2 \le 1$ controls a derivative seminorm. In the scalar case with derivative weights $\boldsymbol{b} = \mathbf{1}$, their full squared Sobolev norm is $\|f\|^2_{L^2(\mu)} + \sum_\gamma \nu_\gamma c_\gamma^2$, which is at most 2 for $f \in \mathcal{F}$. Hence $\{f/\sqrt{2} : f \in \mathcal{F}\}$ is contained in their Sobolev unit Lipschitz ball (3.2). A lower bound over that ball does not transfer by this inclusion to the smaller rescaled common-Lipschitz class; it therefore supplies no lower bound for $\mathcal{F}$ by inclusion alone. Likewise, exact coefficient integrals are not a point-evaluation implementation.

The scale comparison with Bizeul & Klartag (2025) describes the interface of the inspected theorem and construction, not impossibility of an extension. The inconsistent isolated display in their Corollary 1.6 is not used. The boundedness of either construction supplies no distinction.

F.3 The exact-power conjectural comparison

In the exact scalar specialization listed above, our passive upper bound gives

$$R_m^{\mathrm{iid},*} \le R_m^{\mathrm{P},1} = O\big((\log m)^{-(\tau+1)/2}\big) = o\big((\log m)^{-\tau/2}\big).$$

The discussion following Theorem 2.10 in Adcock et al. (2026c) suggests norm exponent $\tau/2$. Thus its universal interpretation as the sharp order needs qualification for these scalar random-design instances. This inference uses only the passive upper bound; the adaptive lower bound supplies the matching order and robustness to stronger point access. The two-sided envelope is the scope of Theorem 1 here, and is not attributed to their exact-power algebraic display. This does not contradict their proved inequalities, establish a matching deterministic fixed-design upper bound, or resolve all output/noise regimes.

For an arbitrary restricted procedure family, an iid comparison requires both admission of the passive Hermite procedure and that every procedure be admissible in the adaptive noisy model without truth-dependent side information. With these conditions its minimax risk is bounded below by the adaptive noisy risk and above by the passive Hermite risk; the same holds for squared loss. A distribution chosen using unknown-function information falls outside that comparison. A passive random-design upper bound cannot be derandomized by interchanging the supremum over targets and the expectation over designs.

F.4 Recovery, integration and functional regression

An integrated-error approximation yields an estimate of $\int f \, d\mu$ by integration, but the converse does not follow. Applying a quadrature procedure separately to Hermite coefficients would require additional integrand, query-allocation and truncation analysis. The access model of Creutzig et al. (2008) is therefore a relevant precedent, not itself a full-function recovery upper bound.

Older functional-regression results also require exact loss and design comparisons. The inspected model of Mas (2011) concerns risk at a fixed input, not integrated risk. The inspected lower theorem of Meister (2016) permits a sequence of design measures rather than fixing the Gaussian law used here. The classes of Bertin & Klutchnikoff (2020) impose Wiener-chaos kernel regularity and truncation/decay conditions rather than a common global input-metric Lipschitz bound. These are bounded source comparisons, not full-proof audits of those works.

F.5 Reading depth and unresolved source access

The comparisons use the specified versions in the bibliography. For Kovachki et al. (2024), the inspected material is Section 2.1, Theorems 2.6–2.7 and 2.12, and nearby discussion; the full Appendix D proof was not audited. The covariance-eigenvalue convention in its displayed theorem and the square-root convention at the opening of Lemma D.1 require reconciliation before a numerical exponent comparison. No such numerical exponent is quoted here.

For Adcock et al. (2026a), the readings cover Remark 2.12, class (3.2), Hermite definitions, Definitions 4.1–4.3, Theorem 4.4, and relevant counting passages. For Adcock et al. (2026b), Algorithm 1, Theorems 4.1/6.1 and the estimation/noise interfaces were read, without a complete proof audit. For Bizeul & Klartag (2025), the regression/entropy model, Theorem 1.5, projection interface and Section 4 construction were read. For Adcock et al. (2026c), the readings cover Sections 2.1–2.2, risks (2.2)–(2.3), Theorem 2.9 and the conjectural discussion after Theorem 2.10.

For Creutzig et al. (2008), the introduction and Sections 2.1–2.2 were available; full proofs of Theorems 8–10 were not read. For Plaskota et al. (2000), access supported only the model/introduction, not theorem-level comparison. The reads of Suzuki & Karvonen (2026) cover equations (1.1)–(1.2), algorithm definitions, Proposition 3.1, Theorem 3.2 and Corollary 3.3; those of Krieg et al. (2026) cover Section 1 and Theorems 2–3, without auditing the later diagonal-operator results.

The actual theorem text of Mhaskar & Hahm (1997) remains unread because full retrieval of its December 1995 manuscript failed. The full weighted-space theorems of Wasilkowski (2012) also remain unread after a publisher retrieval failure; indexed passages about weighted tensor-product spaces and variable costs do not establish a theorem-level distinction or noncollision. The older information-based-complexity books, the complete weighted infinite-variate literature, and all citation descendants have not been exhaustively audited. Retrieval gaps are not evidence that a matching prior theorem does not exist. Accordingly, no unrestricted first-result claim or human expert priority certification is made.

F.6 Matching losses and finite noisy functional observations

Matching-loss and GLMtron-like updates precede this work (Kakade et al., 2011; Gollakota et al., 2023). The bounded-feature and bounded-response assumptions in Section 2 of Kakade et al. (2011) do not directly cover our unbounded Gaussian polynomial features and original finite-variance labels. Appendix I therefore supplies its own shifted-moment estimates, raw-polynomial comparator bound, projected-gradient calculation and mean-validation proof. The optimization principle is not claimed as new.

The discretization question also has established functional-regression precedents. The model of Cai & Li (2026) uses noisy point observations and linear Sobolev slopes, evaluating loss against a full test trajectory. Sang & Li (2026, Section 6, Theorem 20 and Corollary 21) explicitly allow noisy partial test observations and reconstruct the test curve under nested-RKHS and source assumptions. Our common bounded globally Lipschitz scalar class, Gaussian minimax criterion and bounded linear sensor model differ from those assumptions; their rates are not claimed to be improved by a direct comparison. In particular, nonlinear prediction with noisy test curves is not new by itself.

These are bounded model/loss and methodological comparisons, not exhaustive priority or full-proof audits of the cited works.

G PASSIVE LEARNING WITH UNKNOWN COVARIANCE

G.1 MODEL AND ESTIMATOR

Fix $H = \ell^2$, $\tau > 1$, and known constants $0 < c_\lambda \le C_\lambda < \infty$. Let $\mathcal{C}$ consist of the positive trace-class covariance operators Σ whose positive eigenvalues, in nonincreasing order and with multiplicity, satisfy

$$c_\lambda j^{-\tau} \le \lambda_j(\Sigma) \le C_\lambda j^{-\tau}, \qquad j \ge 1. \tag{G.1}$$

Their positive rank is infinite; ambient null directions are allowed. The input law is $\mu_\Sigma = N(0, \Sigma)$, with known zero mean. The data $\mathcal{D}_m = ((X_t, Y_t))_{t=1}^m$ are iid pairs with $X_t \sim \mu_\Sigma$ and $Y_t = f(X_t) + \xi_t$, where $f \in \mathcal{F}$ from (2) and the ξ_t are independent standard Gaussians independent of the inputs. An admissible predictor is one jointly Borel function

$$\widehat{h} : (H \times \mathbb{R})^m \times H \longrightarrow \mathbb{R}$$

whose specification does not depend on Σ. It need not belong to $\mathcal{F}$. With $\widehat{h}_{\mathcal{D}_m}(x) = \widehat{h}(\mathcal{D}_m, x)$, the risks are

$$\begin{aligned} R_m^{\mathrm{unk}} &= \inf_{\widehat{h}} \sup_{\Sigma \in \mathcal{C},\, f \in \mathcal{F}} \mathbb{E}_{\Sigma, f} \|\widehat{h}_{\mathcal{D}_m} - f\|_{L^2(\mu_\Sigma)}, \\ Q_m^{\mathrm{unk}} &= \inf_{\widehat{h}} \sup_{\Sigma \in \mathcal{C},\, f \in \mathcal{F}} \mathbb{E}_{\Sigma, f} \|\widehat{h}_{\mathcal{D}_m} - f\|^2_{L^2(\mu_\Sigma)}. \end{aligned} \tag{G.2}$$

The losses use each covariance's own input law. All m full inputs and labels are charged; there is no additional law-sampling or label oracle.

Write $\mathrm{clip}_{[u,v]}(z) = \min\{v, \max\{u, z\}\}$. For $m < 4$ set $\widehat{h} = 0$. For $m \ge 4$, set $n = \lfloor m/2 \rfloor$ and $N = m - n$, and discard the first n labels while retaining their inputs. Using the constant B_τ in (10), define

$$T = \frac{1}{n} \sum_{a=1}^{n} X_a \otimes X_a, \qquad E = \frac{1}{2C_\lambda} \left(\frac{\log m}{4B_\tau} \right)^{\tau+1}, \qquad \gamma = E^{-1}. \tag{G.3}$$

Let $P = \mathbf{1}_{[\gamma,\infty)}(T)$ and $K = \operatorname{rank} P$. In decreasing order, with multiplicity, denote the retained eigenvalues by $\widehat{\lambda}_1, \dots, \widehat{\lambda}_K$ and a measurably chosen orthonormal eigenbasis by $\widehat{e}_1, \dots, \widehat{e}_K$. Ties are resolved against the fixed ambient basis, as detailed in Section G.6. In particular, eigenvalues equal to γ are retained. Define

$$\begin{aligned} \mathcal{S}_E &= \left\{ \alpha \in \mathbb{N}_0^K : \sum_{j=1}^{K} \frac{\alpha_j}{\widehat{\lambda}_j} \le E \right\}, \qquad D = |\mathcal{S}_E|, \\ \widehat{H}_\alpha(x) &= \prod_{j=1}^{K} h_{\alpha_j} \left(\frac{\langle x, \widehat{e}_j \rangle}{\sqrt{\widehat{\lambda}_j}} \right), \qquad V = \operatorname{span}\{\widehat{H}_\alpha : \alpha \in \mathcal{S}_E\}. \end{aligned} \tag{G.4}$$

Here h_r are the normalized probabilists' Hermite polynomials. Enumerate the indices by a fixed degree-then-lexicographic order. Each retained eigenvalue is finite and positive, and $\alpha_j \le E\widehat{\lambda}_j$, so D is finite. When $K = 0$, the sole index is the empty index, its product is one, $D = 1$, and V is the constant-function space. The features depend on x only through Px.

If $D > m^{1/4}$, output zero. Otherwise, enumerate $\mathbb{Q}^D$ as $(\theta_s)_{s \ge 1}$, let $\phi(x)$ be the ordered vector of features in (G.4), and put

$$\begin{aligned} h_s(x) &= \mathrm{clip}_{[-1,1]}(\theta_s \cdot \phi(x)), \qquad B = 1 + \sqrt{8 \log m}, \qquad Y_{t,B} = \mathrm{clip}_{[-B,B]}(Y_t), \\ L_s &= \frac{1}{N} \sum_{t=n+1}^{m} (h_s(X_t) - Y_{t,B})^2, \qquad s_* = \min \left\{ s : L_s \le \inf_{r \ge 1} L_r + \frac{1}{N} \right\}. \end{aligned} \tag{G.5}$$

Output $\widehat{h} = h_{s_*}$. The set defining s_* is nonempty by the property of the infimum. Continuity in the coefficients makes this a $1/N$-approximate empirical minimizer also over

$$\mathcal{F}_V = \{\mathrm{clip}_{[-1,1]}(v) : v \in V\}.$$

Section G.6 proves joint Borel measurability, including all branches. This is a statistical existence construction, with no finite running-time claim for its countable search. The estimator uses no unknown population quantity and no constructed input with an uncharged label.

G.2 A COVARIANCE EVENT AND THE EMPIRICAL FEATURE COUNT

Put $S = C_\lambda \zeta(\tau)$, so $\operatorname{tr}\Sigma \le S$. Gaussian fourth moments give

$$\mathbb{E}\|X\|^4 = (\operatorname{tr}\Sigma)^2 + 2\operatorname{tr}(\Sigma^2).$$

Indeed, this identity holds for finite covariance truncations by the independent normal fourth moments; monotone convergence yields the Hilbert-space identity. Since $\|X \otimes X\|_{\mathrm{HS}}^2 = \|X\|^4$, the centered, independent Hilbert–Schmidt summands in $T - \Sigma$ satisfy

$$\begin{aligned}\mathbb{E}\|T-\Sigma\|_{\mathrm{HS}}^2 &= \frac{1}{n}\mathbb{E}\|X\otimes X - \Sigma\|_{\mathrm{HS}}^2 \\ &= \frac{(\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)}{n} \le \frac{2S^2}{n}.\end{aligned} \tag{G.6}$$

For $a = n^{-1/4}$, the operator norm is at most the Hilbert–Schmidt norm, so Markov's inequality implies

$$G = \{\|T-\Sigma\|_{\mathrm{op}} \le a\}, \qquad \mathbb{P}(G^c) \le 2S^2 n^{-1/2}. \tag{G.7}$$

This is uniform in the eigenbasis and ambient null directions. Since E is a fixed constant times $(\log m)^{\tau+1}$, there is m_0, depending only on τ, C_λ, such that

$$a \le \gamma/2 = 1/(2E) \qquad (m \ge m_0). \tag{G.8}$$

All uses of relative empirical spectral control below assume both G and (G.8). The finite initial range will be handled by bounded loss.

Pad the eigenvalue sequence of the finite-rank T with zeros. The compact positive-operator min-max principle (Weyl's inequality) gives $|\widehat{\lambda}_j - \lambda_j| \le a$ on G. For each retained index,

$$\lambda_j \ge \widehat{\lambda}_j - a \ge \gamma - a \ge a, \qquad \widehat{\lambda}_j \le \lambda_j + a \le 2\lambda_j \le 2C_\lambda j^{-\tau}. \tag{G.9}$$

Consequently $K \le (2C_\lambda/\gamma)^{1/\tau}$, also trivially when $K = 0$. There is no eigengap assumption or claim of individual eigenvector convergence.

For clarity, the count in Appendix D applies to these finite empirical indices using only their upper envelope. For $t > 0$, extend its nonnegative finite coordinate sum by the infinite envelope sum to obtain

$$\begin{aligned}\log D &\le tE + \sum_{\ell\ge1}\frac{1}{\ell}\sum_{j=1}^{K} e^{-\ell t/\widehat{\lambda}_j} \\ &\le tE + G_\tau(2C_\lambda/t)^{1/\tau}, \qquad G_\tau = \Gamma(1+1/\tau)\zeta(1+1/\tau).\end{aligned}$$

The minimization already proved there, with $B_\tau = (\tau+1)(G_\tau/\tau)^{\tau/(\tau+1)}$, yields

$$\log D \le B_\tau(2C_\lambda E)^{1/(\tau+1)} = \tfrac{1}{4}\log m, \qquad D \le m^{1/4}. \tag{G.10}$$

This argument allows real τ, nonintegral weights and any $E > 0$. For $K = 0$, $D = 1$ obeys the same inequality. Thus the feature guard does not reject on G for $m \ge m_0$.

G.3 Conditional Gaussian regression and the operator tail

Condition on the covariance sample, making P fixed and independent of the regression group. Work on G with $a \le \gamma/2$. For $K > 0$, let $A = P\Sigma P|_{PH}$. Its smallest eigenvalue is at least $\gamma - a > 0$. Define

$$L = \Sigma P A^{-1} : PH \to H, \qquad Z = X - LPX, \qquad Q = \Sigma - \Sigma P A^{-1} P \Sigma. \tag{G.11}$$

The pair (PX, Z) is jointly centered Gaussian and its cross-covariance vanishes, since $\Sigma P - LA = 0$. The two components are therefore independent. In particular, conditional on $PX = y$, X has law $Ly + Z$, with $Z \sim N(0, Q)$.

Set $C_P = (I - P)\Sigma P$. Since P is a spectral projection of T,

$$C_P = (I - P)(\Sigma - T)P, \quad \|C_P\|_{\mathrm{op}} \le a, \quad Ly = y + C_P A^{-1} y.$$

The two terms in the last expression are orthogonal. Hence

$$\|Ly\|^2 = \|y\|^2 + \|C_P A^{-1} y\|^2 \le \left(1 + \frac{a^2}{(\gamma - a)^2}\right) \|y\|^2 \le 2\|y\|^2. \tag{G.12}$$

Thus $g(y) = \mathbb{E}_Z f(Ly + Z)$ is defined on all PH, is bounded by one, and is globally $\sqrt{2}$-Lipschitz in the physical Hilbert norm: use the same Z at two arguments and average the Lipschitz inequality. The looser bound two will suffice.

In the decomposition $H = PH \oplus (I - P)H$, write $B_P = (I - P)\Sigma(I - P)|_{(I-P)H}$. Then

$$Q = \begin{pmatrix} 0 & 0 \\ 0 & B_P - C_P A^{-1} C_P^* \end{pmatrix}.$$

Positivity of this covariance and positivity of $C_P A^{-1} C_P^*$ give

$$0 \le Q \le (I - P)\Sigma(I - P), \qquad \|Q\|_{\mathrm{op}} \le \gamma + a, \tag{G.13}$$

since every eigenvalue of T on the complementary subspace is below γ and $\|T - \Sigma\|_{\mathrm{op}} \le a$.

To justify the infinite-dimensional Gaussian Poincare step without differentiability of f, diagonalize the positive trace-class Q and truncate its Gaussian series to Z_d in its first d coordinates. For fixed y, the restriction $z \mapsto f(Ly + z)$ is 1-Lipschitz. Finite-dimensional Gaussian Poincare bounds $\mathrm{Var}(f(Ly + Z_d))$ by $\|Q\|_{\mathrm{op}}$. The Lipschitz property and the vanishing omitted covariance trace imply L^2 convergence to $f(Ly + Z)$. Taking the limit and then integrating over y gives

$$\mathbb{E}(f(X) - g(PX))^2 \le \gamma + a. \tag{G.14}$$

This uses the conditional covariance's operator norm; the ordinary projection distance $\mathbb{E}\|X - PX\|^2$ would instead incur a trace tail.

When $K = 0$, set $P = 0$, $g(0) = \mathbb{E}f(X)$ and $Q = \Sigma$, with no inverse or L. No empirical eigenvalue reaches γ, so $\|T\|_{\mathrm{op}} < \gamma$ and $\|\Sigma\|_{\mathrm{op}} \le \gamma + a$ on G. The same truncation argument gives (G.14). For either branch, every bounded predictor $h(PX)$ satisfies the exact conditional-expectation identity

$$\mathbb{E}(h(PX) - f(X))^2 = \mathbb{E}(f(X) - g(PX))^2 + \mathbb{E}(h(PX) - g(PX))^2. \tag{G.15}$$

There is no requirement that f itself depend only on PX.

G.4 Finite-marginal transfer and Hermite approximation

For $K > 0$ let $\widehat{A} = PTP|_{PH}$ and compare only the laws $\mu_P = N(0, A)$ and $\nu_P = N(0, \widehat{A})$ on the same finite-dimensional space PH. They both have strictly positive covariance there on G. The matrix $M = \widehat{A}^{-1/2} A \widehat{A}^{-1/2}$ satisfies

$$\|M - I\|_{\mathrm{op}} \le a/\gamma \le 1/2,$$

so its eigenvalues are in $[1/2, 3/2]$. Integrating the finite Gaussian log-density ratio and using $t - 1 - \log t \le 2(t - 1)^2$ on this interval gives

$$\begin{aligned} \mathrm{KL}(\mu_P \| \nu_P) &= \tfrac{1}{2}(\mathrm{tr}\, M - K - \log\det M) \\ &\le \|M - I\|_F^2 \le K(a/\gamma)^2, \\ \mathrm{TV}(\mu_P, \nu_P) &\le \sqrt{K/2}\, a/\gamma. \end{aligned} \tag{G.16}$$

The last inequality is Pinsker's inequality, with $\mathrm{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$. If $K = 0$, both laws are the point mass on the zero space and their TV distance is zero, without using inverses or the KL formula. No full infinite-dimensional Gaussian TV comparison is asserted or used. Indeed, the finite-rank empirical Gaussian and infinite-rank population Gaussian on H can be mutually singular, making such a replacement invalid.

Under ν_P, the physical-coordinate weak gradient of g has squared norm at most two almost everywhere. The finite-dimensional Hermite energy argument in Appendix D therefore gives, for its Hermite coefficients b_α,

$$\sum_\alpha \left(\sum_{j=1}^K \frac{\alpha_j}{\widehat{\lambda}_j} \right) b_\alpha^2 \le 2.$$

Its energy-E truncation $v_E \in V$ consequently satisfies the conservative bound

$$\|g - v_E\|^2_{L^2(\nu_P)} \le 4/E.$$

Clipping to $h_E = \mathrm{clip}_{[-1,1]}(v_E)$ cannot increase distance to $g(y) \in [-1,1]$. The function $(g-h_E)^2$ takes values in $[0,4]$. The layer-cake formula for this nonnegative function transfers its integral with error at most $4\,\mathrm{TV}$, so

$$\|g - h_E\|^2_{L^2(\mu_P)} \le 4/E + 4\,\mathrm{TV}(\mu_P,\nu_P).$$

Combining (G.14)–(G.16) yields

$$\inf_{h\in\mathcal{F}_V} \mathbb{E}(h(PX) - f(X))^2 \le 5/E + a + 4\sqrt{K/2}\,aE. \tag{G.17}$$

For $K = 0$, the constant g is already in the approximation class; the same bound holds with zero final term and slack in $5/E$. The quantities A, L, Q, g, v_E are only witnesses in the analysis, not quantities evaluated by the estimator.

G.5 Response clipping and a complete elementary ERM bound

For bounded predictors h, h' with values in $[-1,1]$, let $R(h) = \mathbb{E}(h(U) - f(X))^2$, where U is any measurable representation of X, and let $R_B(h) = \mathbb{E}(h(U) - Y_B)^2$ with $Y_B = \mathrm{clip}_{[-B,B]}(Y)$. Independent centered unit-variance noise gives

$$R(h) - R(h') = \mathbb{E}[(h-Y)^2 - (h'-Y)^2].$$

Subtracting the corresponding clipped-response difference cancels all response-only terms, giving

$$\begin{aligned} R(h) - R(h') &= R_B(h) - R_B(h') - 2\mathbb{E}[(h-h')(Y - Y_B)] \\ &\le R_B(h) - R_B(h') + 4\mathbb{E}|Y - Y_B|. \end{aligned} \tag{G.18}$$

This is a comparison of risk differences, not an identification of the clipped and original regression targets.

Lemma G.1 (clipped linear-feature ERM). Let $(X_i, Y_i)_{i=1}^N$ be iid with $N \ge 2$, $Y = f(X) + \xi$, $|f| \le 1$, and independent centered noise of variance one. Let $\phi : X \to \mathbb{R}^D$ be measurable, with finite real coordinates at every input and $D \ge 1$. Define

$$\mathcal{H} = \{h_\theta : x \mapsto \mathrm{clip}_{[-1,1]}(\theta \cdot \phi(x)) : \theta \in \mathbb{R}^D\}.$$

For $B \ge 1$, a measurable $1/N$-approximate empirical squared-loss minimizer against Y_B, selected over rational coefficients, satisfies

$$\begin{aligned} \mathbb{E}R(\widehat{h}) \le \inf_{h\in\mathcal{H}} R(h) + 32(B+1)^2\sqrt{\frac{(D+1)\log(2N)}{N}} \\ + 4\mathbb{E}|Y - Y_B| + 1/N. \end{aligned} \tag{G.19}$$

Neither uniformly bounded feature values nor bounded coefficients are assumed. Features may depend on an independent first sample, in which case the result holds conditionally on that sample.

Proof. Fix any N inputs. For each clipped prediction h_i, quantize downward using

$$t_r = -1 + 2r/N \quad (1 \le r \le N), \qquad q_i = -1 + \frac{2}{N}\sum_{r=1}^N \mathbf{1}_{\{h_i \ge t_r\}}.$$

Then $q_i \in [-1, 1]$ and $|h_i - q_i| \le 2/N$. For all these thresholds, including $t_N = 1$,

$$\mathbf{1}_{\{h_i \ge t_r\}} = \mathbf{1}_{\{\theta \cdot \phi(X_i) \ge t_r\}}.$$

Thus the patterns are determined by at most $L = N^2$ affine hyperplanes in coefficient space $\mathbb{R}^D$. Constant comparisons, when an affine expression has no genuine hyperplane, do not increase this count. The total number of relatively open faces satisfies

$$F(L, D) \le \sum_{j=0}^{D} 2^j \binom{L}{j} \le (2L + 1)^D. \tag{G.20}$$

For the first bound, adding a hyperplane cuts some old faces into two pieces and their intersection, increasing the count by two per cut face. The intersections of cut faces are distinct faces of the induced arrangement in dimension $D - 1$. Hence $F(L, D) \le F(L - 1, D) + 2F(L - 1, D - 1)$, with $F(0, D) = F(L, 0) = 1$; induction gives the binomial sum. Degeneracies can only reduce the count. For the second bound, encode each subset of at most D hyperplanes and its binary choices as an ordered list of at most D elements from a $2L$-element set and pad with zero. This injects the counted choices into $(2L + 1)^D$ lists.

All weak-comparison signs, with their equality conventions, are constant on each relatively open face. The number of possible quantized prediction vectors is consequently

$$J \le (2N^2 + 1)^D. \tag{G.21}$$

Counting faces rather than just full-dimensional cells includes equality cases, including the clipping endpoint one.

Let $M = (B + 1)^2$, $\ell_h(X, Y) = (h(X) - Y_B)^2$, and denote population and empirical averaging by $\mathbb{P}$ and $\mathbb{P}_N$. The countable supremum

$$\Delta_N = \sup_{\theta \in \mathbb{Q}^D} |(\mathbb{P}_N - \mathbb{P})\ell_{h_\theta}|$$

is measurable. Introduce an independent ghost sample. Jensen first replaces $\mathbb{P}$ by its ghost empirical mean inside the supremum. Randomly interchanging each original/ghost pair inserts independent fair signs ε_i in their differences. The triangle inequality then bounds this signed difference by two separate suprema, giving

$$\mathbb{E}\Delta_N \le 2\mathbb{E}_{Z,\varepsilon} \sup_{\theta \in \mathbb{Q}^D} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \ell_{h_\theta}(Z_i) \right|, \qquad Z_i = (X_i, Y_i). \tag{G.22}$$

For a fixed sample, squared loss on the clipped domain is $2(B + 1)$-Lipschitz in its prediction argument. Replacing h_i by q_i therefore changes each loss by at most

$$\delta = 4(B + 1)/N.$$

The quantized loss vectors number at most J, have coordinates in $[0, M]$, and depend only on the sample, not on the signs. For any such vector b, the bound $\cosh(s) \le e^{s^2/2}$ gives, for $t > 0$,

$$\mathbb{E}_\varepsilon \exp\left(\frac{t}{N} \sum_i \varepsilon_i b_i \right) \le \exp\left(\frac{t^2 M^2}{2N} \right).$$

Apply this to the $2J$ choices consisting of each vector and its negative. The exponential bound and Jensen imply

$$\mathbb{E}_\varepsilon \max_b \left| N^{-1} \sum_i \varepsilon_i b_i \right| \le \frac{\log(2J)}{t} + \frac{tM^2}{2N}.$$

Optimizing t yields $M\sqrt{2\log(2J)/N}$. Since $\log(2J) \le 3(D + 1)\log(2N)$ for $N \ge 2$, (G.22) gives

$$\mathbb{E}\Delta_N \le 2M\sqrt{\frac{6(D + 1)\log(2N)}{N}} + \frac{8(B + 1)}{N}. \tag{G.23}$$

The proof was conditional on arbitrary finite feature values and is therefore distribution-free.

For each rational-coefficient comparator h, approximate empirical minimization and the definition of Δ_N give, simultaneously,

$$R_B(\widehat{h}) \le R_B(h) + 2\Delta_N + 1/N.$$

Using (G.18) for this same comparator yields

$$R(\widehat{h}) \le R(h) + 2\Delta_N + 1/N + 4\mathbb{E}|Y - Y_B|.$$

Take expectation over the training sample and the infimum over rational coefficients. Every real vector is approximated by rational vectors. Their predictions converge pointwise because feature coordinates are finite and clipping is continuous. Clipping is bounded, so dominated convergence identifies the rational population-risk infimum with the infimum over $\mathcal{H}$. The same coefficient continuity on a finite sample identifies the empirical infima. Finally, (G.23) and $B \ge 1$ imply

$$2\mathbb{E}\Delta_N \le 32M\sqrt{(D+1)\log(2N)/N}.$$

Indeed, the square-root contribution has constant $4\sqrt{6}$, and $16(B+1)/N$ is bounded by $16M$ times the displayed rate; $4\sqrt{6} + 16 < 32$. This proves (G.19).

For application to the empirical features, one may use $U = PX$ but must retain $R(h) = \mathbb{E}(h(PX) - f(X))^2$. The identity $\mathbb{E}(h(PX) - Y)^2 = R(h) + 1$ still holds. No assertion that the conditional mean of f given PX equals f has been used.

For the actual Gaussian noise, put $t = B - 1 = \sqrt{8\log m}$. Since $|f(X)| \le 1$, $|Y| > B$ implies $|\xi| > t$, and $|Y| + 1 \le |\xi| + 2$. With φ and $\overline{\Phi}$ the standard normal density and upper tail,

$$\begin{aligned} \mathbb{E}|Y - Y_B| &\le \mathbb{E}[(|Y|+1)\mathbf{1}_{\{|Y|>B\}}] \\ &\le 2\varphi(t) + 4\overline{\Phi}(t) \le (2 + 4/t)\varphi(t) = O(m^{-4}), \qquad m \ge 4. \end{aligned} \tag{G.24}$$

This bound is uniform over the truth and covariance. On the feature guard, $D \le m^{1/4}$ and $N \ge m/2$, so the learning term in (G.19) is $O(m^{-3/8}(\log m)^{3/2})$. No additional $\log D$ factor is needed.

G.6 A COMMON JOINTLY BOREL PREDICTOR

The map $(X_1, \dots, X_n) \mapsto T$ is continuous into the Hilbert–Schmidt operators. A finite Gram construction makes the eigenvector choices explicit. Let

$$\mathsf{G}_{ab} = \langle X_a, X_b\rangle/n, \qquad U : \mathbb{R}^n \to H, \quad Uv = \sum_{a=1}^{n} v_a X_a. \tag{G.25}$$

The ordered eigenvalues of G are continuous in its entries. On each Borel stratum with a fixed pattern of distinct-eigenvalue multiplicities, the projector Π_λ for a distinct eigenvalue λ is a finite polynomial in G: the product of $(\mathsf{G} - \eta I)/(\lambda - \eta)$ over its other distinct eigenvalues η. There are finitely many multiplicity patterns for each fixed n, so these projectors are Borel across multiplicities. For positive λ, the corresponding empirical eigenspace projector on H is

$$Q_\lambda = \frac{1}{n\lambda} U \Pi_\lambda U^*. \tag{G.26}$$

It is a Borel finite-rank operator-valued function of the covariance sample.

Project the fixed ambient basis vectors onto the eigenspace. Choose the smallest ambient index with nonzero residual norm, normalize the residual, and repeat Gram–Schmidt until the known finite eigenspace dimension is reached. Such an index exists at each step because the projected ambient basis spans the eigenspace. Each smallest-index choice is a countable Borel partition, and normalization is performed only where the norm is positive. This constructs a Borel orthonormal basis without requiring continuity across multiplicities. Concatenating in decreasing eigenvalue order and retaining exactly eigenvalues at least γ makes $P, K, \widehat{\lambda}_j, \widehat{e}_j$ measurable. Threshold ties are included; neither a gap nor a population basis is supplied.

On each stratum with fixed K, enumerate all multi-indices in the fixed degree-then-lexicographic order. Each energy test in (G.4) is Borel, and only finitely many indices pass. Thus D, the feature

guard, and the ordered retained index set are measurable. Each normalized product Hermite is jointly Borel in the covariance sample and the evaluation input x, with finite real coordinates at every x. When $K = 0$ its empty product is one.

For each possible D, fix the rational enumeration in (G.5). Each L_s and the countable infimum $\inf_s L_s$ are measurable. The first index within $1/N$ of the infimum is measurable and exists. The empirical loss is continuous in the real coefficient vector, so its rational infimum equals its full real infimum. The choice is consequently a $1/N$-approximate minimizer for the full clipped span as claimed.

Combining all strata with the feature guard and the $m < 4$ branch yields a single bounded jointly Borel function of all data and test input. For each fixed Σ, its squared risk is a measurable integral. Equivalently, for a fixed law, bounded joint measurability and separability of L^2 give the usual Borel L^2-valued output, but the common predictor formulation in (G.2) directly accommodates unknown laws. These are existence arguments under full-input observation and unconstrained measurable estimation. They do not furnish a finite-precision eigensolver, a coefficient bound, a numerical-conditioning guarantee, or a finite running-time guarantee for the rational search.

G.7 Uniform rates and the separate lower bounds

For $m \geq m_0$, condition on the covariance group and restrict to G. The guard accepts by (G.10). Apply Lemma G.1 to the resulting fixed feature class and insert (G.17). Since $K \leq (2C_\lambda E)^{1/\tau}$, the covariance and finite-marginal transfer terms satisfy

$$a + 4\sqrt{K/2}\, aE \leq C_{\tau,C_\lambda}\, n^{-1/4}\big(1 + E^{1+1/(2\tau)}\big).$$

Together with (G.24) and the learning bound, the conditional risk on G is bounded by the sum of

$$\frac{5}{E}, \qquad C_{\tau,C_\lambda} n^{-1/4}\big(1+E^{1+1/(2\tau)}\big), \qquad O\big(m^{-3/8}(\log m)^{3/2}\big), \qquad O(m^{-4}), \qquad \frac{1}{N}. \tag{G.27}$$

Off G, both prediction and truth remain bounded by one, whether or not the guard rejects. Squared loss is at most four, and (G.7) bounds this contribution by $8S^2 n^{-1/2}$.

Each remainder other than $1/E$ is a negative power of m times a fixed power of $\log m$. More explicitly, $n \geq m/4$ and $N \geq m/2$ for $m \geq 4$, while E is a positive constant times $(\log m)^{\tau+1}$. For any fixed $b > 0$ and real r, $m^{-b}(\log m)^r[\log(m+2)]^{\tau+1}$ is bounded for $m \geq 4$ and tends to zero. This applies to every term in (G.27) and to the failure contribution. The leading $1/E$ is itself at most a constant times $[\log(m+2)]^{-(\tau+1)}$ on this range. The remaining finite range $4 \leq m < m_0$ is absorbed by increasing a constant, using squared loss at most four; for $m < 4$ the zero estimator has squared risk at most one. We have therefore proved

$$\sup_{\Sigma\in\mathcal{C},\, f\in\mathcal{F}} \mathbb{E}_{\Sigma,f}\|\widehat{h} - f\|^2_{L^2(\mu_\Sigma)} \leq C_{\tau,C_\lambda}[\log(m+2)]^{-(\tau+1)}, \qquad m \geq 1. \tag{G.28}$$

The upper constant needs no lower-envelope parameter. Jensen's inequality for this same estimator gives the expected norm upper bound with exponent $(\tau+1)/2$.

For the lower bounds, fix one admissible covariance, for example the diagonal covariance with eigenvalues $c_\lambda j^{-\tau}$. Restricting the covariance supremum can only decrease the unknown-law minimax risk. Revealing that fixed covariance to the learner can only decrease it further. The known-covariance lower theorem therefore supplies both lower bounds in Theorem 3, separately for their respective norm and squared-norm losses. This neither reverses Jensen nor relies on an unknown-law active learner sampling its input law. Theorem 3 concerns only passive pairs; all original active-query comparisons retain the known-law model.

H Known-law finite-coordinate observations

We prove Corollary 4 under the known law (1). Put $q = \tau + 1$, $L = \log(m+2)$, $\mu_d = \mathcal{L}(P_d X)$, and $X_{>d} = (I - P_d)X$. The training data are $\mathcal{D}_d = (P_d X_t, Y_t)_{t=1}^m$, with independent full latent inputs $X_t \sim \mu$ and labels $Y_t = f(X_t) + \xi_t$. The independent noise variables have law $N(0,1)$. An admissible output is a jointly measurable function $h(\mathcal{D}_d, U, y)$, using an independent seed U;

its lifted prediction is $h(\mathcal{D}_d, U, P_d x)$. All test losses are integrated under the original full-input law. Explicitly,

$$R_{m,d} = \inf_h \sup_{f\in\mathcal{F}} \mathbb{E}_f \|h(\mathcal{D}_d, U, P_d\,\cdot) - f\|_{L^2(\mu)},$$
$$Q_{m,d} = \inf_h \sup_{f\in\mathcal{F}} \mathbb{E}_f \|h(\mathcal{D}_d, U, P_d\,\cdot) - f\|^2_{L^2(\mu)}. \tag{H.1}$$

As in (3), these are separate minimax criteria. Predictors with infinite loss cannot improve either infimum. Joint measurability gives a measurable $L^2(\mu)$-valued lift in the finite-loss case, since $L^2(\mu)$ is separable; outputs outside this case may be replaced by zero where necessary.

H.1 Conditional approximation and estimation

Independence of the Gaussian eigen-coordinates gives the conditional mean

$$g_d(y) = \mathbb{E}_{X_{>d}} f(y + X_{>d}), \qquad y \in P_d H. \tag{H.2}$$

It is bounded by one and 1-Lipschitz in the physical norm. Let $\Lambda_d = \sup_{j>d} \lambda_j \le C_\lambda (d+1)^{-\tau}$. For fixed y, apply finite-dimensional Gaussian Poincare to $f(y + \sum_{j=d+1}^n \sqrt{\lambda_j} Z_j e_j)$. The weak gradient has squared Euclidean norm at most one, so the variance is at most Λ_d. These finite tail sums converge in $L^2(H)$; the Lipschitz property gives convergence of the function values in L^2, and therefore convergence of their variances. Averaging the resulting conditional bound over $y \sim \mu_d$ proves

$$\|f - g_d \circ P_d\|^2_{L^2(\mu)} = \mathbb{E}\,\mathrm{Var}(f(X) \mid P_d X) \le \Lambda_d \le C_\lambda (d+1)^{-\tau}. \tag{H.3}$$

The finite-dimensional Poincare step also follows from the Hermite derivative identity in Appendix D.1. It uses the largest tail eigenvalue, not the sum of the tail eigenvalues.

For every $h \in L^2(\mu_d)$, the conditional-expectation projection identity gives

$$\|h \circ P_d - f\|^2_{L^2(\mu)} = \|h - g_d\|^2_{L^2(\mu_d)} + \|g_d \circ P_d - f\|^2_{L^2(\mu)}. \tag{H.4}$$

Use the notation H_γ, ν_γ of (D.1), restricting the Hermite product to the first d coordinates. For $E > 0$, set

$$S_{E,d} = \{\gamma : \mathrm{supp}\,\gamma \subseteq [d],\ \nu_\gamma \le E\}, \qquad \widehat{g}_{E,d} = \sum_{\gamma\in S_{E,d}} \widehat{c}_\gamma H_\gamma, \qquad \widehat{c}_\gamma = \frac{1}{m}\sum_{t=1}^m Y_t H_\gamma(P_d X_t). \tag{H.5}$$

The actual eigenvalues and eigenbasis are known. Thus (H.5) is a finite, jointly measurable predictor using precisely the observed pairs. Conditional expectation shows that each coefficient is unbiased for the corresponding coefficient of g_d. Although the regression residual contains the unobserved tail,

$$\mathbb{E}[Y_t^2 \mid P_d X_t] = \mathbb{E}[f(X_t)^2 \mid P_d X_t] + 1 \le 2, \qquad \mathrm{Var}(\widehat{c}_\gamma) \le 2/m. \tag{H.6}$$

The same derivative-energy argument as Appendix D.1 applies to the bounded 1-Lipschitz function g_d. Its squared Hermite truncation bias is at most E^{-1}. Orthonormality, (H.3) and (H.4) consequently give

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f \|\widehat{g}_{E,d} \circ P_d - f\|^2_{L^2(\mu)} \le C_\lambda (d+1)^{-\tau} + E^{-1} + 2|S_{E,d}|/m. \tag{H.7}$$

Since $S_{E,d} \subseteq S_E$, the count (D.11) holds uniformly in d. For $m \ge 3$, choose $E = C_\lambda^{-1}[\log(m)/(2B_\tau)]^q$. Then $|S_{E,d}| \le \sqrt{m}$, and the all-budget absorption from Appendix D.3 gives

$$Q_{m,d} \le C_\lambda (d+1)^{-\tau} + U_{\tau,C} L^{-q}. \tag{H.8}$$

For $m = 1, 2$, the zero predictor has squared risk at most one, so the same bound holds with $U_{\tau,C}$ from (D.14). Jensen applied to this estimator, followed by $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, gives

$$R_{m,d} \le \sqrt{C_\lambda}(d+1)^{-\tau/2} + \sqrt{U_{\tau,C}} L^{-q/2}. \tag{H.9}$$

No full training coordinates, derivative labels, or additional observations were used.

H.2 Separate lower bounds for sample and test information

Every procedure in (H.1) is available in the full-input passive experiment by discarding the other training and test coordinates. Hence, separately for the two loss criteria,

$$R_{m,d} \geq R_m^{\mathsf{P},1} \geq c_R L^{-q/2}, \qquad Q_{m,d} \geq Q_m^{\mathsf{P},1} \geq c_Q L^{-q}, \tag{H.10}$$

where the positive constants are those of Theorem 1. This is a comparison of admissible procedure classes and preserves the order of infimum, supremum and expectation.

For the independent test-information obstruction, fix the admissible target

$$f_*(x) = \mathrm{clip}_{[-1,1]}(\langle x, e_{d+1}\rangle). \tag{H.11}$$

It is globally bounded by one and 1-Lipschitz. Symmetry and independence give $\mathbb{E}[f_*(X) \mid P_d X] = 0$. For every realized dataset and seed and every finite-loss output h, therefore,

$$\|h\circ P_d - f_*\|^2_{L^2(\mu)} = \|h\|^2_{L^2(\mu_d)} + v_d \geq v_d, \qquad v_d = \mathbb{E}\min\{\lambda_{d+1} Z^2, 1\}, \quad Z \sim N(0,1). \tag{H.12}$$

This is an identity for integrated loss at each fixed output. Training labels may reveal the target, but they do not reveal a new test input's omitted coordinate. In fact, (H.12) would still hold if the entire target were supplied to the learner.

Put $p_0 = \mathbb{P}(|Z| \geq 1) > 0$. Using the lower envelope and $\min\{ca, 1\} \geq \min\{c, 1\}a$ for $0 < a \leq 1$,

$$v_d \geq p_0 \min\{\lambda_{d+1}, 1\} \geq p_0 \min\{c_\lambda, 1\}(d+1)^{-\tau} =: b_0(d+1)^{-\tau}. \tag{H.13}$$

Taking expectations in (H.12) gives the squared-loss floor. Taking square roots pathwise first and then expectations gives the norm-loss floor:

$$Q_{m,d} \geq b_0(d+1)^{-\tau}, \qquad R_{m,d} \geq \sqrt{b_0}(d+1)^{-\tau/2}. \tag{H.14}$$

No reverse Jensen step or identity between the two minimax criteria is assumed. Combining each floor with its corresponding bound in (H.10), and using $\max\{a, b\} \geq (a+b)/2$, completes (12) uniformly in m, d.

H.3 Interpretation and scope

The observation term is no larger than a constant multiple of the sample term exactly when $d+1$ is at least a constant multiple of $L^{(\tau+1)/\tau}$; for integers $d \geq 1$, replacing $d+1$ by d changes only constants. This proves the stated coordinate threshold, including exponent $3/2$ when $\tau = 2$. It provides an observation-budget consequence of the sharp known-law theorem and classical Gaussian conditioning. It does not provide numerical thresholds or a computationally efficient implementation.

Joint sample and resolution questions are also studied in functional linear regression under noisy point measurements by Cai & Li (2026). Their linear-slope model and full-trajectory test risk differ from the nonlinear bounded Lipschitz targets and truncated test information here. Corollary 4 concerns exact known eigen-coordinates, and makes no pointwise path-sampling or arbitrary-grid claim. It does not extend the unknown-covariance result of Theorem 3 to truncated observations.

I Passive learning from finite noisy Gaussian sensors

This appendix proves Theorem 5 and Corollary 6, specifies the measured-data algorithm, and establishes the physical cell-average example. The mean is denoted by b, while μ continues to denote a probability law in the core results. In the polynomial perturbation lemmas b_{sh} denotes a local standardized shift.

I.1 Model, information and exact risks

Let H be a separable infinite-dimensional real Hilbert space, identified with ℓ^2 if convenient. For fixed parameters $\tau > 1$ and $0 < c \leq C < \infty$, write $X = b + X_0$ for arbitrary unknown $b \in H$, with X_0 a centered Gaussian input having a positive trace-class covariance Σ whose ordered positive eigenvalues satisfy

$$cj^{-\tau} \le \lambda_j(\Sigma) \le Cj^{-\tau}, \qquad j \ge 1. \tag{I.1.1}$$

Ambient covariance null directions are allowed. Let $P : H \to \mathbb{R}^d$, $d \ge 1$, be a fixed physical coordinate map with $PP^* = I_d$. Thus P^*P is the orthogonal projection onto the sensor space. Its orthonormal coordinate convention is known; no population eigenbasis is supplied or assumed aligned with this space. Define

$$A = P\Sigma P^*, \quad B_\perp = (I - P^*P)\Sigma(I - P^*P), \quad \beta = \|B_\perp\|_{\mathrm{op}}. \tag{I.1.2}$$

The common target class is

$$\mathcal{F} = \{f : H \to \mathbb{R} : \ |f(x)| \le 1, \quad |f(x) - f(y)| \le \|x - y\|_H \text{ for all } x, y \in H\}. \tag{I.1.3}$$

Training consists of exactly m independent original pairs

$$U_i = PX_i + \eta_i, \qquad Y_i = f(X_i) + \xi_i, \qquad i = 1, \dots, m, \tag{I.1.4}$$

where $X_i = b + X_{0,i}$, $X_{0,i} \sim N(0, \Sigma)$, and $\eta_i \sim N(0, \Gamma)$ is independent of $X_{0,i}$. The triples (X_i, η_i, ξ_i) are iid. Labels satisfy the conditional moment assumptions

$$\mathbb{E}[\xi_i \mid X_i, \eta_i] = 0, \qquad \mathbb{E}[\xi_i^2 \mid X_i, \eta_i] \le s_2 < \infty, \qquad s_2 \ge 0. \tag{I.1.7}$$

The fixed bound s_2 may be unknown to the learner. Label noise may be nonsymmetric, heavy-tailed and heteroskedastic; independence of ξ_i from (X_i, η_i) and Gaussianity of ξ_i are not assumed. A marginal variance bound alone is not (I.1.7). Γ is an arbitrary positive semidefinite d by d matrix, possibly singular or noncommuting with A. Assume

$$\operatorname{tr}\Gamma \le B_{\mathrm{noise}} < \infty, \qquad \nu = \|\Gamma\|_{\mathrm{op}}. \tag{I.1.5}$$

At test time only $U = PX + \eta$ is observed, with a fresh pair $X = b + X_0$ and η having the same independent laws. The target remains $f(X)$. No test-label noise is included in the loss. In particular, labels in (I.1.4) are not evaluations at a reconstructed or projected input.

For a realized training dataset D and an independent procedure seed ω, define

$$\ell_{b,\Sigma,\Gamma,f}(h_{D,\omega}) = \left[\mathbb{E}_{X,\eta}\{h_{D,\omega}(PX + \eta) - f(X)\}^2\right]^{1/2}. \tag{I.1.6}$$

The inner expectation concerns the fresh test pair. The outer training risk is either $\mathbb{E}_{D,\omega}\ell^2$ or $\mathbb{E}_{D,\omega}\ell$, and these are distinct criteria. This is a norm on the joint test law of (X,η); for nonzero sensor noise it is not the ordinary L^2 norm under the latent-input law of a deterministic predictor of X. We never move the square root across the training expectation in a lower bound.

All admissible predictors are jointly Borel in data, seed and test coordinates. Clipping to [-1,1] cannot increase either loss, so minimax infima can be restricted to such predictors. No unbounded or uncharged additional sample, covariance oracle, latent input, conditional mean, derivative label, or constructed-design label is available.

Known and unknown information. The algorithm below uses only the arrays U_i, Y_i and their dimensions m,d, in the supplied physical coordinates. It does not receive b, the measured mean $Pb, \tau, c, C, B_{\mathrm{noise}}, s_2, \beta, \Gamma, \nu, \Sigma$, an eigenbasis, a compact exponent interval, or any lower sensor-variance bound. The theoretical constants depend on the true fixed model bounds. The same rule serves every such fixed class; this does not mean its constants are uniform as τ approaches one or the model bounds diverge. No part of the algorithm estimates or certifies β. It estimates a measured center only from the covariance split. Uniformity over all b is an exact-arithmetic statement, not a finite-bit guarantee for arbitrary offsets.

I.2 SENSOR THEOREM AND SHARP COVERAGE COROLLARY

Write $q = \tau + 1$ and

$$a_m = [\log(m+2)]^{-q}. \tag{I.2.1}$$

Theorem 5 (passive learning from noisy Gaussian sensors). The single deterministic finite procedure in Appendix I.3 satisfies, for every fixed $\tau > 1$, $C > 0$, $B_{\mathrm{noise}} < \infty$ and $s_2 \geq 0$ finite, a bound

$$\sup_{b\in H,\ f\in\mathcal{F},\ \mathcal{L}_\xi} \mathbb{E}_D \ell_{b,\Sigma,\Gamma,f}(\widehat{h}_D)^2 \leq K(\tau, C, B_{\mathrm{noise}}, s_2)\{a_m + \beta + \nu\}, \tag{I.2.2}$$

simultaneously for all integers $m, d \geq 1$, all fixed P with $PP^* = I_d$, and all Σ,Γ satisfying (I.1.1)–(I.1.5). Here the supremum over label kernels $\mathcal{L}_\xi$ is over all iid observation laws satisfying (I.1.7) for the displayed b,Σ,Γ,f. The training law includes this kernel even though test loss has no test-label noise. The lower spectral constant c is not needed in this upper-bound proof, and neither the constant nor its eventual threshold depends on b. The same procedure has

$$\sup_{b\in H,\ f\in\mathcal{F},\ \mathcal{L}_\xi} \mathbb{E}_D \ell_{b,\Sigma,\Gamma,f}(\widehat{h}_D) \leq \sqrt{K}\{a_m^{1/2} + \beta^{1/2} + \nu^{1/2}\}. \tag{I.2.3}$$

Its expected number of operations is polynomial in m,d in the ideal exact-real-arithmetic model described in Appendix I.11, with fixed model-dependent constants. No finite-bit complexity or reliable floating-point high-degree evaluation is established.

The theorem permits singular A and Γ, an empty retained empirical subspace, repeated eigenvalues, and no covariance eigengap. Its β term is an upper bound: β need not equal the conditional residual covariance norm or a pointwise irreducible loss. The proof conditions directly on retained noisy coordinates; it does not presume a uniformly Lipschitz regression function of all measured coordinates.

Corollary 6 (sharp minimax class with physical coverage). Fix $\tau, c, C, B_{\mathrm{noise}}, s_2$ as above and $D_0 \geq C$. For fixed P,d and a number $0 \leq v \leq B_{\mathrm{noise}}$, let $\mathcal{C}_d(v)$ contain all pairs (Σ,Γ) satisfying (I.1.1),

$$\beta \leq D_0(d+1)^{-\tau}, \qquad \operatorname{tr}\Gamma \leq B_{\mathrm{noise}}, \qquad \|\Gamma\|_{\mathrm{op}} \leq v. \tag{I.2.4}$$

Let

$$\begin{aligned} Q_{m,d,v} &= \inf_h \sup_{(\Sigma,\Gamma)\in\mathcal{C}_d(v),\ b\in H,\ f\in\mathcal{F},\ \mathcal{L}_\xi} \mathbb{E}\ell_{b,\Sigma,\Gamma,f}(h)^2, \\ R_{m,d,v} &= \inf_h \sup_{(\Sigma,\Gamma)\in\mathcal{C}_d(v),\ b\in H,\ f\in\mathcal{F},\ \mathcal{L}_\xi} \mathbb{E}\ell_{b,\Sigma,\Gamma,f}(h). \end{aligned} \tag{I.2.5}$$

The infima permit measurable procedures with the physical coordinates fixed, but not procedures depending on the unknown b,Σ,Γ,f or label kernel. The supremum includes all label laws satisfying (I.1.7). Even if the fixed class bounds are supplied, the lower bounds hold. The parameter-free rule from Theorem 5 supplies the upper bounds. Uniformly in m,d,v in the stated range,

$$\begin{aligned} Q_{m,d,v} &\asymp a_m + (d+1)^{-\tau} + v, \\ R_{m,d,v} &\asymp a_m^{1/2} + (d+1)^{-\tau/2} + \sqrt{v}. \end{aligned} \tag{I.2.6}$$

Comparison constants depend only on $\tau, c, C, D_0, B_{\mathrm{noise}}, s_2$. Lower bounds use $b = 0$ and $\xi = 0$, an admissible subclass for every $s_2 \geq 0$; no unit-Gaussian-label submodel is silently assumed when $s_2 < 1$. These are worst-case rates over (I.2.4). In particular, the noise term is not a lower bound for every individual correlated Γ with the same operator norm. Appendix I.10 supplies distinct sample, resolution and noise witnesses and handles expected norm separately.

I.3 Fully specified measured-data algorithm

All index orders and ties below are deterministic. Arithmetic, comparisons, square roots, finite loops and integer counters are exact in this algorithmic description. Integer roots needed in stopping rules can equivalently be found by repeated integer comparisons, so a floor or logarithm oracle is unnecessary.

I.3.1 Splits and shared covariance preprocessing

For $m < 8$ return the zero predictor. Otherwise set

$$M = \lfloor m/2 \rfloor, \quad L = m - M, \quad n = \lfloor M/2 \rfloor, \quad N = M - n, \quad a = n^{-1/4}. \tag{I.3.1}$$

Use pairs $1, \dots, M$ as the training pool and $M+1, \dots, m$ as validation. Within training, use inputs $U_1, \dots, U_n$ only for covariance. Their labels are discarded but charged to m. Use the remaining N training pairs only for fitting. None of the validation inputs or labels enters preprocessing or any coefficient update.

Estimate the measured center only from those n covariance inputs:

$$\widehat{u} = n^{-1} \sum_{i=1}^{n} U_i, \qquad \bar{U}_i = U_i - \widehat{u}, \qquad T = n^{-1} \sum_{i=1}^{n} \bar{U}_i \bar{U}_i^*. \tag{I.3.11}$$

Use the same stored $\widehat{u}$ to center all fitting, validation and future test inputs. No other input or label estimates this center. Store $\widehat{u}$ as part of every selected model. Neither $u_0 = Pb$ nor b is supplied. Starting with T and an identity accumulated orthogonal matrix V, apply maximal-pivot Jacobi rotations until the off-diagonal Frobenius norm is at most a. Choose a largest absolute off-diagonal entry; ties use the lexicographically first $(i, j), i < j$. A zero pivot implies the matrix is already diagonal. For a nonzero block with diagonal entries a_0,c_0 and off-diagonal b_0, set

$$t_0 = (c_0 - a_0)/(2b_0), \quad t = \frac{\operatorname{sgn}_+(t_0)}{|t_0| + \sqrt{1 + t_0^2}}, \quad c_J = (1 + t^2)^{-1/2}, \quad s_J = t c_J, \tag{I.3.2}$$

where $\operatorname{sgn}_+(0) = 1$. Let Q_J be the plane matrix with block $\left[\begin{smallmatrix} c_J & s_J \\ -s_J & c_J \end{smallmatrix}\right]$. Apply the explicit updates $T_{\text{work}} \leftarrow Q_J^* T_{\text{work}} Q_J$ and $V \leftarrow V Q_J$, starting from $T_{\text{work}} = T$ and $V = I$. This diagonalizes the pivot block. For $d = 1$ there is no rotation; for $d = 2$ one nonzero rotation suffices.

At termination let D_{diag} be the diagonal of the rotated matrix, and put $\widetilde{T} = V D_{\text{diag}} V^*$. All its entries in D_{diag} are nonnegative. Sort them in decreasing order; equal values retain their pre-sort column order. Apply the same permutation to V. These are exact eigenpairs of $\widetilde{T}$, not asserted eigenpairs of T. The stopping guarantee is

$$\|\widetilde{T} - T\|_{\text{op}} \le a. \tag{I.3.3}$$

The formula and loop make a Borel finite map, including all ties. Its finiteness and cost are proved in Appendix I.11.

I.3.2 Universal finite energy grid and count guard

Let k_m be the smallest nonnegative integer with $2^{k_m} \ge M$, found by repeated doubling, and let $D_{\max}$ be the largest integer D with $D^2 \le M$, so $D_{\max} = \operatorname{floor}(\sqrt{M})$. Candidate order is: an explicit zero predictor first, followed by energies

$$E_j = 2^j, \qquad j = -k_m, \dots, k_m, \tag{I.3.4}$$

in increasing j. The global half-power guard permits more features than the proof candidate needs; its actual feature count D_* remains at most $\lfloor M^{1/4} \rfloor$. All fitting step sizes use the actual D, not the

guard. There are $J = 2k_m + 2$ candidate slots, including zero. A rejected energy remains a zero predictor in its prescribed slot. All validation ties retain the earliest slot. No candidate is chosen using a latent test target or an oracle risk.

For an energy E retain all surrogate eigenvalues $\widehat{\lambda}_i \geq \gamma = 1/E$, including equality. Write k for their number, $\widehat{v}_i$ for their orthonormal eigenvectors, and $\widehat{A} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k)$ in those coordinates. Define

$$\mathcal{I}_E = \left\{ \alpha \in \mathbb{N}_0^k : \sum_{i=1}^{k} \alpha_i / \widehat{\lambda}_i \leq E \right\}. \tag{I.3.5}$$

The zero multi-index is included. If $k = 0$, the feature set consists solely of the constant one. If $k + 1 > D_{\max}$, reject immediately: all unit indices would be retained. Otherwise enumerate (I.3.5) lazily by an iterative depth-first traversal, taking coordinate order 1,...,k and trying successive nonnegative degrees in increasing order while the cumulative energy is at most E. Descend into a prefix only if it is feasible; it always has a feasible all-zero completion. Stop as soon as $D_{\max}$+1 leaves have been found, and reject. Do not allocate a whole potentially huge range of degrees in advance. This is a finite count guard with no retained-eigenvalue upper guard.

For an accepted set write $D = |\mathcal{I}_E| \leq D_{\max}$, and use the enumeration order as its feature order. With $h_0(z) = 1, h_1(z) = z$ and the normalized probabilists' Hermite recurrence

$$h_{r+1}(z) = \{z h_r(z) - \sqrt{r}\, h_{r-1}(z)\}/\sqrt{r+1}, \tag{I.3.6}$$

the feature for α is

$$\phi_\alpha(u) = \prod_{i=1}^{k} h_{\alpha_i}(\langle \widehat{v}_i, u - \widehat{u} \rangle / \sqrt{\widehat{\lambda}_i}). \tag{I.3.7}$$

Thus $\phi(u)$ throughout the fitting and validation formulas includes the stored training-only centering. All retained eigenvalues are strictly positive because $\gamma > 0$. The singular measured-covariance directions are never inverted. On a statistically bad sample a retained direction may fail the later population comparison; its finite features and clipped predictor still exist.

I.3.3 Projected matching-loss SGD

For each accepted candidate start $\theta_1 = 0$ in $\mathbb{R}^D$, put $\eta_{\mathrm{step}} = (10DN)^{-1/2}$, and traverse its N original fitting pairs once in their original order. With $\mathrm{clip}(z) = \max(-1, \min(1, z))$, update

$$\begin{aligned} G_t &= \{\mathrm{clip}(\theta_t^* \phi(U_{n+t})) - Y_{n+t}\}\phi(U_{n+t}), \\ \theta_{t+1} &= \Pi_{\{\|\theta\|_2 \leq 1\}}(\theta_t - \eta_{\mathrm{step}} G_t), \qquad t = 1, \ldots, N. \end{aligned} \tag{I.3.8}$$

Projection divides by $\max(1, \|\theta\|_2)$. Return the average of **pre-update** iterates,

$$\bar{\theta} = N^{-1} \sum_{t=1}^{N} \theta_t, \qquad h_E(u) = \mathrm{clip}(\bar{\theta}^* \phi(u)). \tag{I.3.9}$$

Labels are not clipped or replaced. The empty-retention case has $D = 1$ and uses exactly the same routine. Fits are completed without using validation data.

I.3.4 Independent validation

For every slot h_j, including zeros, compute

$$\widehat{S}_j = L^{-1} \sum_{i=M+1}^{m} \{h_j(U_i)^2 - 2Y_i h_j(U_i)\}. \tag{I.3.10}$$

Select the earliest minimizer, updating the current selected slot only on a strict score improvement. Output that fitted predictor. All candidates may be dependent through common training data; independence among candidates is neither assumed nor required.

I.4 Covariance event, retained spectrum and Hermite count

Write $u_0 = Pb$, $K_U = A + \Gamma$, $S = C\zeta(\tau) + B_{\text{noise}}$ and $\Delta_u = \widehat{u} - u_0$. For analysis only, put

$$T_0 = n^{-1} \sum_{i=1}^{n} (U_i - u_0)(U_i - u_0)^*. \tag{I.4.9}$$

For a centered Gaussian Z with covariance K_U, $\mathbb{E}\|ZZ^* - K_U\|_{\text{HS}}^2 = (\operatorname{tr} K_U)^2 + \operatorname{tr}(K_U^2)$. Indeed $\mathbb{E}\|Z\|^4 = (\operatorname{tr} K_U)^2 + 2\operatorname{tr}(K_U^2)$, and expanding the centered squared norm subtracts $\operatorname{tr}(K_U^2)$. Independent centered summands give $\mathbb{E}\|T_0 - K_U\|_{\text{HS}}^2 \le 2S^2/n$. Also $\mathbb{E}\|\Delta_u\|^2 = \operatorname{tr}(K_U)/n \le S/n$. A union bound and Markov's inequality imply, with $a = n^{-1/4}$,

$$\begin{aligned} \mathcal{G} &= \{\|T_0 - K_U\|_{\text{op}} \le a,\ \|\Delta_u\| \le a\}, \\ \mathbb{P}(\mathcal{G}^c) &\le \min\{1, (2S^2 + S)/\sqrt{n}\}, \qquad e = 3a. \end{aligned} \tag{I.4.1}$$

No independence of the two events is needed. The actual centered covariance has the exact identity $T = T_0 - \Delta_u \Delta_u^*$. On $\mathcal{G}$,

$$\|T - K_U\|_{\text{op}} \le a + a^2 \le 2a, \qquad \|\widetilde{T} - K_U\|_{\text{op}} \le 3a = e. \tag{I.4.10}$$

The last step uses the unchanged Jacobi tolerance a. These bounds do not depend on the unknown mean. $\mathcal{G}$, T_0 and u_0 are analysis objects, not algorithm inputs. In particular an uncentered covariance bound depending on $\|u_0\|$ is not substituted for (I.4.10).

Fix one grid energy E for the following argument, and assume

$$e \le \gamma/4, \qquad \nu \le \gamma/4, \qquad \gamma = 1/E. \tag{I.4.2}$$

Compression monotonicity gives $\lambda_j(A) \le \lambda_j(\Sigma) \le Cj^{-\tau}$, padding the finite spectrum with zeros. Weyl's inequality gives, for each retained index j,

$$\widehat{\lambda}_j \le Cj^{-\tau} + \nu + e \le Cj^{-\tau} + \widehat{\lambda}_j/2, \qquad \widehat{\lambda}_j \le 2Cj^{-\tau}. \tag{I.4.3}$$

Consequently

$$k \le \lfloor (2CE)^{1/\tau} \rfloor =: \bar{k}, \quad r := \max_{\alpha \in \mathcal{I}_E} |\alpha| \le \lfloor 2CE \rfloor =: \bar{r}, \quad \operatorname{tr} \widehat{A} \le 2C\zeta(\tau). \tag{I.4.4}$$

The maximum degree is zero for the constant-only case. For the degree bound, $\sum_i \alpha_i \le \widehat{\lambda}_1 \sum_i \alpha_i/\widehat{\lambda}_i \le 2CE$.

Here is the exact inherited counting calculation, restated with its proof. For positive weights $\rho_j \le C' j^{-\tau}$, define $\nu_\alpha = \sum_j \alpha_j/\rho_j$ for finite-support indices, and $D(E) = \#\{\alpha : \nu_\alpha \le E\}$. For each $t > 0$, nonnegative summation and the geometric series give

$$
\begin{aligned}
D(E) &\le e^{tE} \prod_{j\ge 1} (1 - e^{-t/\rho_j})^{-1}, \\
\log D(E) &\le tE + \sum_{\ell \ge 1} \ell^{-1} \sum_{j \ge 1} e^{-\ell t/\rho_j} \\
&\le tE + \Gamma(1 + 1/\tau)\zeta(1 + 1/\tau)(C'/t)^{1/\tau}.
\end{aligned}
\tag{I.4.5}
$$

The last inequality compares the decreasing summand with its integral on $[0,\infty)$. The product converges for $t > 0$. Finite spectra are bounded by the same infinite product. Put

$$
G_\tau = \Gamma(1 + 1/\tau)\zeta(1 + 1/\tau), \qquad B_\tau = (\tau + 1)(G_\tau/\tau)^{\tau/(\tau+1)}. \tag{I.4.6}
$$

Minimizing $tE + G_\tau(C'/t)^{1/\tau}$, or setting $t = C'[G_\tau/(\tau C'E)]^{\tau/(\tau+1)}$, proves

$$
\log D(E) \le B_\tau (C'E)^{1/q}. \tag{I.4.7}
$$

No integer exponent or lattice-valued eigenvalue assumption is used. This is the counting argument of Appendix D, equations (D.7)–(D.11). For (I.4.3), take $C' = 2C$. In particular, the analysis cutoff

$$
E_0 = \frac{1}{2C} \left[\frac{\log M}{4B_\tau} \right]^q \tag{I.4.8}
$$

satisfies $|\mathcal{I}_E| \le M^{1/4}$ whenever $E \le E_0$ and (I.4.2) holds. Since the count is an integer, the good candidate has actual $D_* \le \lfloor M^{1/4} \rfloor \le D_{\max} = \operatorname{floor}(\sqrt{M})$, and the global guard accepts. This is a bound on the good candidate, not on every accepted candidate. The learner does not evaluate (I.4.6)–(I.4.8).

I.5 Direct Gaussian conditioning, including singular and empty cases

Condition throughout this section on the covariance sample, so its retained projection R on $\mathbb{R}^d$ is fixed. Let $Q = I_d - R$, $K_R = RK_UR$ restricted to $\operatorname{ran} R$. Under (I.4.2) and $\mathcal{G}$,

$$
K_R \succeq (\gamma - e)I, \qquad \|QK_UR\|_{\mathrm{op}} \le e. \tag{I.5.1}
$$

The second fact uses that R is the spectral projection of $\widetilde{T}$, so $Q\widetilde{T}R = 0$. If $R = 0$, use no inverse and set the conditional mean to the constant $\mathbb{E}f(X)$. We treat nonempty R first.

I.5.1 Conditional Gaussian formula

Let $V = R(U - \widehat{u})$, $\delta = R(u_0 - \widehat{u})$ and $L_R = \Sigma P^* R K_R^{-1} : \operatorname{ran} R \to H$. Then V has mean δ and covariance K_R, with $\|\delta\| \le a$ on $\mathcal{G}$. The pair $(X_0, V - \delta)$ is centered jointly Gaussian, and $X_0 - L_R(V - \delta)$ has cross-covariance with $V - \delta$

$$
\Sigma P^* R - L_R K_R = 0.
$$

It is therefore independent of V. Its covariance is

$$
V_R = \Sigma - \Sigma P^* R K_R^{-1} R P \Sigma \succeq 0. \tag{I.5.2}
$$

All finite collections of coordinates have the joint Gaussian independence property, which gives the H-valued assertion by separability. We use only the finite-dimensional inverse K_R. Even if A or K_U is singular, no inverse on its full space is needed. Write $Z_R \sim N(0, V_R)$ independent of V. Then the everywhere-defined conditional mean version

$$g_R(v) = \mathbb{E} f(b + L_R(v - \delta) + Z_R), \qquad v \in \operatorname{ran} R \tag{I.5.3}$$

is bounded by one. Its Lipschitz constant is at most $\|L_R\|_{\mathrm{op}}$.

In the orthogonal decomposition $H = \operatorname{ran} P^* \oplus (\operatorname{ran} P^*)^\perp$, write Σ in blocks $\left[\begin{smallmatrix} A & \mathrm{Cross}^* \\ \mathrm{Cross} & B_\perp \end{smallmatrix}\right]$. The observed block of L_R is ARK_R^{-1}. Since $A = K_U - \Gamma$ and the R and Q components of $K_U R K_R^{-1}$ are orthogonal,

$$\begin{aligned} \|ARK_R^{-1}\|_{\mathrm{op}} &\le \sqrt{1 + \{e/(\gamma - e)\}^2} + \nu/(\gamma - e) \\ &\le (\sqrt{10} + 1)/3 < \sqrt{2}. \end{aligned} \tag{I.5.4}$$

Positivity of the joint covariance of the unobserved centered signal and V, or its Schur complement, gives

$$\mathrm{Cross}\, RK_R^{-1} R\, \mathrm{Cross}^* \preceq B_\perp.$$

Thus

$$\|\mathrm{Cross}\, RK_R^{-1}\|_{\mathrm{op}}^2 \le \|\mathrm{Cross}\, RK_R^{-1/2}\|_{\mathrm{op}}^2 \|K_R^{-1/2}\|_{\mathrm{op}}^2 \le \beta/(\gamma - e).$$

The two output blocks are orthogonal, so a valid squared Lipschitz bound is

$$L_g^2 := 2 + \tfrac{4}{3}\beta E \ge \|L_R\|_{\mathrm{op}}^2. \tag{I.5.5}$$

It may grow with E. No uniform full-sensor conditional-map bound is required.

I.5.2 Residual operator norm and latent-target loss

For covariance analysis consider the affine predictor $b + P^* R(U - u_0)$ of X, and let $\Pi = P^* R P$. This predictor is not computed by the learner. Its centered error is

$$X - \{b + P^* R(U - u_0)\} = (I - \Pi)X_0 - P^* R\eta. \tag{I.5.6}$$

The summands are independent and supported on orthogonal subspaces. Their summed covariance norm is the maximum of the two covariance norms. Decompose the first subspace into the physical complement and discarded Q sensor directions. The diagonal covariance-block norms are at most β and $\gamma + e$: indeed $QAQ \preceq QK_UQ$ and $\|QK_UQ\|_{\mathrm{op}} \le \gamma + e$. For a positive semidefinite two-block covariance with diagonal-block bounds b_0, d_0, Cauchy–Schwarz for its positive bilinear form bounds its quadratic form by

$$(\sqrt{b_0}\|x\| + \sqrt{d_0}\|y\|)^2 \le (b_0 + d_0)(\|x\|^2 + \|y\|^2).$$

Hence $\|(I - \Pi)\Sigma(I - \Pi)\|_{\mathrm{op}} \le \beta + \gamma + e$. The retained-noise norm is at most $\nu \le \gamma/4$. Gaussian linear conditional expectation minimizes error covariance in Loewner order: subtracting its error covariance from that of any other linear predictor leaves $(L_R - B)K_R(L_R - B)^*$. Applying this to (I.5.6) proves

$$\|V_R\|_{\mathrm{op}} \le \beta + \gamma + e. \tag{I.5.7}$$

For completeness, the Gaussian Poincare bound used here can be obtained without an infinite-dimensional differentiability assertion. Diagonalize V_R, truncate its Gaussian series at finite rank, and fix the shift $b + L_R(v - \delta)$. The resulting function of the finite physical coordinates is 1-Lipschitz, so its weak gradient has Euclidean norm at most one. Finite-dimensional Hermite integration by parts and Parseval (as in Appendix I.6.1) give variance at most $\|V_R\|_{\mathrm{op}}$ times this gradient energy.

The finite Gaussian series converges in $L^2(H)$, and the Lipschitz property gives L^2 convergence of the function values and their variances. Integrating over v proves

$$r_f(R) := \mathbb{E}\,\mathrm{Var}(f(X) \mid V) \le \|V_R\|_{\mathrm{op}} \le \beta + 1/E + e. \tag{I.5.8}$$

For every fitted predictor h depending only on V, conditional expectation gives the exact, training-pathwise identity

$$\ell_{b,\Sigma,\Gamma,f}(h)^2 = \|h - g_R\|^2_{L^2(\mathcal{L}(V))} + r_f(R). \tag{I.5.9}$$

All appearances of u_0,b,δ and L_R in this conditional representation define proof witnesses; none is evaluated by the algorithm. If $R = 0$, all coordinates are discarded. The same two-block argument bounds $\|\Sigma\|_{\mathrm{op}} \le \beta + \gamma + e$ because $\|A\| \le \|K_U\| \le \gamma + e$. Conditional mean is $\mathbb{E}f(b + X_0)$, $V_R = \Sigma$, and (I.5.8)–(I.5.9) hold. Formulas (I.5.1)–(I.5.5) involving an inverse are simply bypassed.

I.6 Hermite approximation and polynomial covariance transfer

We give the finite-dimensional polynomial lemmas needed for fitting with raw comparators, including perturbations of both covariance and mean.

I.6.1 Physical derivative energy and truncation

Let $\rho_1, \dots, \rho_k > 0$ define the centered Gaussian law $\mu_\rho = N(0, \mathrm{diag}(\rho_1, \dots, \rho_k))$, and let g be bounded by one and L_g-Lipschitz in its physical Euclidean coordinates. Let $H_\alpha(y) = \prod_i h_{\alpha_i}(y_i/\sqrt{\rho_i})$, $c_\alpha = \langle g, H_\alpha\rangle$ and $\nu_\alpha = \sum_i \alpha_i/\rho_i$. The Hermite products form an orthonormal basis. The weak gradient of g has squared norm at most L_g^2 almost everywhere. Gaussian integration by parts gives

$$\langle \partial_i g, H_\alpha\rangle = \sqrt{(\alpha_i + 1)/\rho_i}\, c_{\alpha + e_i}.$$

A smooth approximation with cutoffs justifies the identity for bounded Lipschitz g and its bounded weak derivatives; the Gaussian density is positive and polynomial moments are finite. Parseval, first on finite coefficient sets and then by monotone convergence, yields

$$\sum_\alpha \nu_\alpha c_\alpha^2 = \sum_i \|\partial_i g\|_2^2 \le L_g^2. \tag{I.6.1}$$

The energy-E raw projection $v = \sum_{\nu_\alpha \le E} c_\alpha H_\alpha$ consequently has

$$\|g - v\|_2^2 \le L_g^2/E, \qquad \|\theta_*\|_2^2 := \sum_{\nu_\alpha \le E} c_\alpha^2 \le 1. \tag{I.6.2}$$

The second inequality is Bessel plus $|g| \le 1$. For $k = 0$ the law is a point mass, g is constant, and the approximation error is zero. Also (I.6.1) gives finite-dimensional Poincare: if the covariance operator norm is $\rho_{\max}$, the nonconstant coefficients satisfy $\nu_\alpha \ge 1/\rho_{\max}$, giving $\mathrm{Var}(g) \le \rho_{\max}\mathbb{E}\|\nabla g\|^2$. This supplies the finite-dimensional step in (I.5.8).

I.6.2 Shifted polynomial moment comparison

Lemma I.1 (shifted moments). Let $Z \sim N(0, I_k)$, M_0 be positive definite with $\|M_0 - I\|_{\mathrm{op}} \le \epsilon < 1$, $b_{\mathrm{sh}} \in \mathbb{R}^k$, and p a polynomial of total degree at most r. Then

$$\mathbb{E}p(M_0^{1/2}Z + b_{\mathrm{sh}})^2 \le \Lambda_{\mathrm{shift}}\mathbb{E}p(Z)^2, \quad \Lambda_{\mathrm{shift}} = \exp\left\{\frac{\epsilon r(1+\sqrt{k})}{1-\epsilon} + 2\|b_{\mathrm{sh}}\|\sqrt{\frac{r}{1-\epsilon}}\right\}. \tag{I.6.3}$$

For $r = 0$ use factor one; for $k = 0$ the assertion is the constant identity.

Proof. Put $\Delta = M_0 - I$, $A_t = I + t\Delta$ and $F(t) = \mathbb{E}_{N(tb_{\mathrm{sh}}, A_t)} p^2$. Differentiating finite Gaussian polynomial moments gives

$$F'(t) = \mathbb{E}\{\nabla p^* \Delta \nabla p\} + \mathbb{E}\{p \operatorname{tr}(\Delta \operatorname{Hess} p)\} + 2\mathbb{E}\{p b_{\mathrm{sh}}^* \nabla p\}. \tag{I.6.4}$$

The covariance derivative identity follows by differentiating the Gaussian characteristic function and identifying its polynomial moments, equivalently by $\frac{1}{2}\sum_{i,j} \Delta_{ij} \mathbb{E}\partial_{ij}(p^2)$. The mean derivative contributes $\mathbb{E} b_{\mathrm{sh}}^* \nabla(p^2)$. Both differentiations are finite polynomial identities.

A degree-r polynomial w under standard Gaussian measure satisfies the Hermite lowering identities

$$\mathbb{E}\|\nabla w\|^2 = \sum_\alpha |\alpha| w_\alpha^2 \le r\mathbb{E}w^2, \quad \mathbb{E}\|\operatorname{Hess} w\|_F^2 = \sum_\alpha |\alpha|(|\alpha|-1) w_\alpha^2 \le r(r-1)\mathbb{E}w^2. \tag{I.6.5}$$

For the Hessian, squared lowering factors over ordered pairs sum to $\sum_{i,j} \alpha_i(\alpha_j - \mathbf{1}_{i=j}) = |\alpha|(|\alpha|-1)$. For each derivative or ordered pair the lowering map is injective on surviving indices, so orthogonality removes cross terms.

Whiten around the translated mean using $w(z) = p(tb_{\mathrm{sh}} + A_t^{1/2} z)$. Translation does not increase degree. Chain rule and $\lambda_{\min}(A_t) \ge 1 - \epsilon$ give

$$\mathbb{E}\|\nabla p\|^2 \le rF(t)/(1-\epsilon), \quad \mathbb{E}\|\operatorname{Hess} p\|_F^2 \le r(r-1)F(t)/(1-\epsilon)^2.$$

With $\|\Delta\|_{\mathrm{op}} \le \epsilon$ and $\|\Delta\|_F \le \sqrt{k}\,\epsilon$, the first two terms in (I.6.4) have total absolute value at most $\epsilon r(1+\sqrt{k})F(t)/(1-\epsilon)$. The last is at most $2\|b_{\mathrm{sh}}\|\sqrt{r/(1-\epsilon)}F(t)$ by Cauchy–Schwarz. Integrate this logarithmic differential inequality. A nonzero polynomial has positive $F(t)$ under a nonsingular Gaussian law; the zero polynomial is immediate. This proves (I.6.3).

For an empirical-orthonormal Hermite feature vector of degree at most r, apply (I.6.3) to every linear combination. If $\Lambda_{\mathrm{shift}} \le 2$, its Gram matrix under the shifted true law is at most $2I_D$ and

$$\mathbb{E}\|\phi(U)\|^2 \le 2D. \tag{I.6.6}$$

The bound concerns polynomial moments, not pointwise feature bounds or a full-space Gaussian total-variation replacement.

I.6.3 Multiplication and shifted polynomial coupling

Lemma I.2. For a degree-r standard-Gaussian polynomial w,

$$\mathbb{E}\{\|Z\|^2 \|\nabla w(Z)\|^2\} \le 2r(k+2r)\mathbb{E}w(Z)^2. \tag{I.6.7}$$

If $\epsilon \le 1/2$ and $\Lambda_{\mathrm{shift}} \le 2$ in Lemma I.1, then

$$\|p(M_0^{1/2} Z + b_{\mathrm{sh}}) - p(Z)\|_2^2 \le \{16\epsilon^2 r(k+2r) + 8\|b_{\mathrm{sh}}\|^2 r\}\|p(Z)\|_2^2. \tag{I.6.8}$$

Proof. Multiplication by Z_j is the sum of Hermite raising and lowering. For a polynomial v of degree at most $r-1$,

$$\begin{aligned}\sum_j \|Z_j v\|_2^2 &\le 2\sum_j \{\|\operatorname{raise}_j v\|_2^2 + \|\operatorname{lower}_j v\|_2^2\} \\ &= 2\sum_\alpha (k + 2|\alpha|) v_\alpha^2 \le 2(k+2r-2)\|v\|_2^2.\end{aligned}$$

Apply this to each partial derivative of w and sum with the first identity in (I.6.5); weakening $k + 2r - 2$ to $k + 2r$ proves (I.6.7). For $r = 0$ both sides involving derivatives vanish.

Locally set $D_0 = M_0^{1/2} - I$, $B_t = I + tD_0$ and $w_t(Z) = p(B_t Z + tb_{\mathrm{sh}})$. Eigenvalues of B_t^2 lie in $[1 - \epsilon, 1 + \epsilon]$, while

$$\|B_t^{-1} D_0\|_{\mathrm{op}} \le \sqrt{2}\epsilon, \qquad \|B_t^{-1} b_{\mathrm{sh}}\| \le \sqrt{2}\|b_{\mathrm{sh}}\|.$$

Indeed $\|D_0\|_{\mathrm{op}} \le \epsilon$ and $\lambda_{\min}(B_t) \ge \sqrt{1 - \epsilon}$. Lemma I.1 applied to covariance B_t^2 and shift tb_{sh} yields $\|w_t\|_2^2 \le 2\|p\|_2^2$. Chain rule gives

$$\frac{dw_t}{dt} = \nabla w_t(Z)^* B_t^{-1}(D_0 Z + b_{\mathrm{sh}}).$$

Using the squared-sum inequality and the two operator bounds,

$$\|dw_t/dt\|_2^2 \le 4\epsilon^2 \mathbb{E}\{\|Z\|^2 \|\nabla w_t\|^2\} + 4\|b_{\mathrm{sh}}\|^2 \mathbb{E}\|\nabla w_t\|^2.$$

Apply (I.6.7), (I.6.5) and $\|w_t\|_2^2 \le 2\|p\|_2^2$ to bound this by the right side of (I.6.8), uniformly in t. Cauchy–Schwarz in $t \in [0, 1]$ proves (I.6.8). The case $b_{\mathrm{sh}} = 0$ is included.

I.6.4 Raw approximation transfer under estimated centering

Let $\widehat{A}$ be the positive surrogate covariance on the retained space, $A_{\mathrm{true}} = K_R$ the true retained covariance, and δ its true retained mean after centering by $\widehat{u}$. Set

$$M_0 = \widehat{A}^{-1/2} A_{\mathrm{true}} \widehat{A}^{-1/2}, \quad b_{\mathrm{sh}} = \widehat{A}^{-1/2}\delta, \quad \epsilon = eE, \quad \|M_0 - I\|_{\mathrm{op}} \le \epsilon, \quad \|b_{\mathrm{sh}}\|^2 \le Ea^2. \tag{I.6.9}$$

The last two bounds hold on $\mathcal{G}$ because $\widehat{A} \succeq E^{-1} I$ and $\|\delta\| \le a$. The quantities b_{sh} and δ are used only in this proof, not as algorithm inputs.

Couple $V_0 = \widehat{A}^{1/2} Z$ with $V_1 = \delta + \widehat{A}^{1/2} M_0^{1/2} Z$. Their laws are $N(0, \widehat{A})$ and $N(\delta, A_{\mathrm{true}})$. Since Z is centered, the cross term in the mean squared displacement vanishes. For L_g-Lipschitz g,

$$\mathbb{E}\{g(V_1) - g(V_0)\}^2 \le L_g^2\{\epsilon^2 \operatorname{tr} \widehat{A} + \|\delta\|^2\}. \tag{I.6.10}$$

Take the raw energy-E Hermite projection v of g under $N(0, \widehat{A})$. Its coefficient norm is at most one by (I.6.2), so $p(z) = v(\widehat{A}^{1/2} z)$ has squared Gaussian norm at most one. Decompose $g(V_1) - v(V_1)$ into $g(V_1) - g(V_0)$, $g(V_0) - v(V_0)$ and $v(V_0) - v(V_1)$. Apply $(x + y + z)^2 \le 3(x^2 + y^2 + z^2)$, (I.6.8) and (I.6.10) to conclude

$$\begin{aligned} \|g - v\|^2_{L^2(N(\delta, A_{\mathrm{true}}))} \le\ & 3\|g - v\|^2_{L^2(N(0, \widehat{A}))} \\ & + 3L_g^2\{\epsilon^2 \operatorname{tr} \widehat{A} + \|\delta\|^2\} \\ & + 48\epsilon^2 r(k + 2r) + 24\|b_{\mathrm{sh}}\|^2 r. \end{aligned} \tag{I.6.11}$$

This requires $\epsilon \le 1/2$ and $\Lambda_{\mathrm{shift}} \le 2$. It is a raw polynomial L^2 comparison under the shifted true law; a bounded-loss-only comparison would not justify the matching-loss comparator. For $k = 0$ take $v = g$, $r = 0$, $\delta = b_{\mathrm{sh}} = 0$ and trace zero, so the actual approximation error is zero.

Apply the result to the everywhere-defined g_R from (I.5.3). Its squared Lipschitz bound gives

$$L_g^2/E = 2/E + \tfrac{4}{3}\beta. \tag{I.6.12}$$

Thus estimated centering adds the controlled shifted-law terms in (I.6.11) without changing the leading approximation order.

I.7 PROJECTED SGD WITH ORIGINAL LATENT LABELS

Define the differentiable convex potential

$$H(z) = \begin{cases} z^2/2, & |z| \le 1, \\ |z| - 1/2, & |z| > 1. \end{cases}$$

Its derivative is $\mathrm{clip}(z)$ and is 1-Lipschitz. For every $z_0 \in [-1, 1]$,

$$(\mathrm{clip}(z) - z_0)^2 \le 2\{H(z) - z_0 z + z_0^2/2\} \le (z - z_0)^2. \tag{I.7.1}$$

For $|z| \le 1$ both inequalities are equalities. For $z > 1$ the left-hand difference is $2(z-1)(1-z_0) \ge 0$; for $z < -1$ it is $2(-z-1)(1+z_0) \ge 0$. The upper bound also follows by 1-smoothness of H at z_0, since $H(z_0) = z_0^2/2$ and $H'(z_0) = z_0$.

Condition on the first n covariance inputs, including $\widehat{u}$, and a good event for the fixed analysis candidate where (I.6.6) holds. Let $\phi(U)$ include the stored centering and put

$$\Phi(\theta) = \mathbb{E}[H(\theta^* \phi(U)) - Y \theta^* \phi(U)].$$

Fresh fitting examples are independent of the current iterate. Conditional on (X,η) and the previous fitting history, (I.1.7) gives

$$\mathbb{E}[(\mathrm{clip}(\theta_t^* \phi(U)) - Y)^2 \mid X, \eta, \mathrm{past}] \le 4 + s_2.$$

Hence on this event

$$\mathbb{E}\|G_t\|^2 \le (4 + s_2)\mathbb{E}\|\phi(U)\|^2 \le 2(4 + s_2)D. \tag{I.7.2}$$

Also $\mathbb{E}[Y|V] = g_R(V)$ and $\mathbb{E}[Y^2|X, \eta] \le 1 + s_2$. Conditional second-moment control guarantees integrability and unbiased gradients even when noise correlates with X or η. Marginal $\mathbb{E}\xi^2 \le s_2$ alone would not control feature-weighted gradient moments. Neither Gaussian nor independent label noise is invoked here.

Let θ_* be the raw Hermite coefficient vector from (I.6.2). It has norm at most one and is fixed after conditioning on the first group and the model; it does not depend on fitting labels. Nonexpansiveness of Euclidean projection gives

$$\|\theta_{t+1} - \theta_*\|^2 \le \|\theta_t - \theta_*\|^2 - 2\eta_{\mathrm{step}} G_t^* (\theta_t - \theta_*) + \eta_{\mathrm{step}}^2 \|G_t\|^2.$$

Conditional unbiasedness, convexity, telescoping over t, $\theta_1 = 0$ and $\|\theta_*\| \le 1$ show

$$\frac{1}{N} \sum_{t=1}^{N} \mathbb{E}\{\Phi(\theta_t) - \Phi(\theta_*)\} \le \frac{1}{2\eta_{\mathrm{step}} N} + (4 + s_2) D \eta_{\mathrm{step}} = \frac{9 + s_2}{\sqrt{10}} \sqrt{D/N}. \tag{I.7.3}$$

Convexity at the pre-update average $\bar{\theta}$ gives the same bound for $\mathbb{E}\Phi(\bar{\theta}) - \Phi(\theta_*)$. For retained-coordinate $g = g_R$, $\mathbb{E}[Y \mid V] = g$ implies

$$\Phi(\theta) + \tfrac{1}{2}\mathbb{E}g(V)^2 = \mathbb{E}\{H(\theta^* \phi) - g\theta^* \phi + g^2/2\}.$$

Apply the first half of (I.7.1) to $\bar{\theta}$, then (I.7.3), then the second half of (I.7.1) to θ_* to conclude

$$\mathbb{E}_{\mathrm{fit}} \| \mathrm{clip}(\bar{\theta}^* \phi) - g_R \|_{L^2(\mathcal{L}(V))}^2 \le \|\theta_*^* \phi - g_R\|_2^2 + \frac{2(9 + s_2)}{\sqrt{10}} \sqrt{D/N}. \tag{I.7.4}$$

The comparator is the raw polynomial approximation. There is no claim of the same approximation factor against an arbitrary clipped feature-span optimum. The $D = 1$ constant branch obeys exactly the same proof.

I.8 Mean validation with only conditional second moments

Condition on the complete training pool, including $\widehat{u}$ and all fitted parameters. Each of the J predictors is then fixed and bounded by one. For a fresh validation observation define $Z_j = h_j(U)^2 - 2Yh_j(U)$, with stored centering included in h_j. Decompose

$$Z_j = B_j - 2\xi h_j(U), \qquad B_j = h_j(U)^2 - 2f(X)h_j(U) \in [-1, 3].$$

Its bounded component has variance at most four. The other component has conditional mean zero by (I.1.7), so its mean and its covariance with B_j vanish. Its second moment is at most $4s_2$. Therefore

$$\mathrm{Var}(Z_j) \le 4(1 + s_2). \tag{I.8.1}$$

For L independent validation examples, let $\Delta_j = \widehat{S}_j - S_j$ with $S_j = \mathbb{E}Z_j$. Then $\mathbb{E}\Delta_j^2 \le 4(1 + s_2)/L$. Without independence between scores of different candidates,

$$\begin{aligned}\mathbb{E}\max_j |\Delta_j| &\le \mathbb{E}\left(\sum_j \Delta_j^2\right)^{1/2} \le \left(\sum_j \mathbb{E}\Delta_j^2\right)^{1/2} \\ &\le 2\sqrt{J(1 + s_2)/L}.\end{aligned} \tag{I.8.2}$$

If $\widehat{j}$ minimizes empirical score and j_0 its population version, then $S_{\widehat{j}} \le S_{j_0} + 2\max_j |\Delta_j|$ pathwise. The latent-target squared risk equals $S_j + \mathbb{E}f(X)^2$, with a common last term, so

$$\mathbb{E}_{\mathrm{val}}\ell(\widehat{h})^2 \le \min_j \ell(h_j)^2 + 4\sqrt{J(1 + s_2)/L}. \tag{I.8.3}$$

All candidates, including statistically bad or guard-rejected slots, are bounded. Mean validation never squares Y itself in the score. Integrating over the training pool and comparing its random minimum with a fixed analysis-candidate slot yields

$$\mathbb{E}\ell(\widehat{h})^2 \le \mathbb{E}\ell(h_{j_*})^2 + 4\sqrt{J(1 + s_2)/L}. \tag{I.8.4}$$

The learner does not need s_2 and does not use a robustification parameter or a population center. Conditional mean-zero and bounded conditional second moment suffice for this eventual logarithmic-rate argument.

I.9 Good grid candidate, explicit finite-sample interface and theorem proof

For fixed true model bounds, define the proof-only E_0 in (I.4.8), and

$$E_{\mathrm{good}} = \min\{E_0, 1/(4\nu)\}, \tag{I.9.1}$$

omitting the second entry when $\nu = 0$. For all sufficiently large M depending only on $\tau, C, B_{\mathrm{noise}}$, E_{good} lies in $[2^{-k_m}, 2^{k_m}]$. Indeed E_0 tends to ∞ but is a fixed multiple of a power of $\log M$, hence eventually less than $M \le 2^{k_m}$. If $B_{\mathrm{noise}} > 0$, eventual $E_0 \ge 1/(4B_{\mathrm{noise}})$ makes $E_{\mathrm{good}} \ge 1/(4B_{\mathrm{noise}})$; if $B_{\mathrm{noise}} = 0$ then $\nu = 0$ and the lower endpoint also follows from $E_0 \to \infty$. Since $2^{-k_m} \le 1/M$, both conclusions are uniform in $0 \le \nu \le B_{\mathrm{noise}}$.

Choose j_* to be the largest grid index with $E_* = 2^{j_*} \le E_{\mathrm{good}}$. This index is deterministic for the model, not for the observed sample, and is used only for analysis. Apart from equality at a grid endpoint, $E_{\mathrm{good}}/2 < E_* \le E_{\mathrm{good}}$; in all cases

$$E_* \le E_0, \qquad \nu \le 1/(4E_*), \qquad 1/E_* \le 2/E_0 + 8\nu. \tag{I.9.2}$$

Since E_0 is a fixed log power, eventually $eE_0 \le 1/4$, uniformly in d, ν, b. Conditions (I.4.2) then hold for E_*. Let $\epsilon = eE_*$. With $\bar{k}, \bar{r}$ from (I.4.4),

$$\epsilon \le 1/2, \qquad \frac{\epsilon \bar{r}(1+\sqrt{\bar{k}})}{1-\epsilon} + 2a\sqrt{E_* \bar{r}/(1-\epsilon)} \le \log 2 \tag{I.9.3}$$

also hold eventually: their left sides are bounded by fixed multiples of $aE_0^2(1 + E_0^{1/(2\tau)}) + aE_0$, hence tend to zero. The bound $\|b_{\mathrm{sh}}\| \le a\sqrt{E_*}$ makes (I.9.3) sufficient for $\Lambda_{\mathrm{shift}} \le 2$. Equations (I.4.7)–(I.4.8) give the actual good-candidate count $D_* \le \lfloor M^{1/4} \rfloor$, which the larger half-power guard accepts. Its fitting step uses this actual D_*, so the SGD remainder retains the quarter-power count bound. No assertion about every accepted candidate is needed.

To make the eventual argument inspectable, consider any deterministic grid E satisfying (I.4.2), (I.9.3), and the count condition

$$Q_\tau(2CE) := \#\{\alpha \in \mathbb{N}_0^{(\mathbb{N})} : \sum_j \alpha_j j^\tau \le 2CE\} \le D_{\mathrm{an}} \le D_{\max}. \tag{I.9.4}$$

Here D_{an} is an analysis count bound on this candidate; take $D_{\mathrm{an}} = \lfloor M^{1/4} \rfloor$ for E_*. It is distinct from the global guard $D_{\max} = \lfloor \sqrt{M} \rfloor$. In (I.9.3) replace E_* by E for a generic analysis energy. The count includes the constant; if $2CE < 1$ then $\bar{k} = \bar{r} = 0$. Set $\bar{L}^2 = 2 + (4/3)\beta E$. Equations (I.5.8), (I.6.2), (I.6.11), (I.7.4) and the retained trace bound give, on $\mathcal{G}$ and after averaging the independent fitting sample,

$$\begin{aligned} \mathcal{R}_{\mathrm{good}}(E) = \beta + 1/E + e + 3\bar{L}^2/E \\ + 6C\zeta(\tau)\bar{L}^2\epsilon^2 + 3\bar{L}^2 a^2 \\ + 48\epsilon^2\bar{r}(\bar{k} + 2\bar{r}) + 24Ea^2\bar{r} + \frac{2(9+s_2)}{\sqrt{10}}\sqrt{D_{\mathrm{an}}/N} \end{aligned} \tag{I.9.5}$$

as a valid upper bound. The empty-retention branch satisfies this same loose bound without inverses or nonconstant features. Both fitted predictions and the target are bounded by one on $\mathcal{G}^c$, giving loss at most four. By (I.4.1) and (I.8.4),

$$\mathbb{E}\ell(\widehat{h})^2 \le \min\left\{4,\ \mathcal{R}_{\mathrm{good}}(E) + 4\min\{1, (2S^2+S)/\sqrt{n}\} + 4\sqrt{J(1+s_2)/L}\right\}. \tag{I.9.6}$$

The outer minimum follows separately from bounded loss. This is a sufficient-bound interface, not an observable confidence interval: its model constants, β and conditions are not learned by the procedure.

For $E = E_*$, the leading terms in (I.9.5) are

$$\beta + 1/E_* + 3\bar{L}^2/E_* = 5\beta + 7/E_*.$$

Using $\beta \le \|\Sigma\|_{\mathrm{op}} \le C$ and $E_* \le E_0$, the other covariance terms are bounded by fixed constants times $n^{-1/4}$ or $n^{-1/2}$ times fixed log powers. The good candidate's SGD term is $O((1+s_2)M^{-3/8})$, since $D_* \le D_{\mathrm{an}} = \lfloor M^{1/4} \rfloor$ and $N \ge M/2$. Replacing this count by the global half-power guard would lose that remainder; we do not make this replacement. The failure contribution is $O(n^{-1/2})$ for fixed S. Since $J = O(\log(m+2))$, validation contributes $O(\sqrt{(1+s_2)\log(m+2)/m})$. Each

is eventually at most a fixed multiple of a_m: for every fixed $a_0 > 0$ and finite b_0, q, $m^{-a_0}[\log(m + 2)]^{b_0+q}$ is bounded and tends to zero. The fixed-factor splits compare $\log M$ and $\log(m + 2)$ by constants.

Finally (I.9.2) makes $1/E_* \le K_0 a_m + 8\nu$. This proves (I.2.2) beyond a threshold depending only on $\tau, C, B_{\text{noise}}, s_2$, uniformly in d,b,Σ,Γ,f and the label kernel satisfying (I.1.7). Before that threshold, loss is at most four and a_m is bounded below over the finite set of budgets; increasing the same model-dependent constant proves the all-m statement. For $m < 8$ the explicit zero predictor is included in this reasoning. Jensen for the final selected predictor, then $\sqrt{x + y + z} \le \sqrt{x}+\sqrt{y}+\sqrt{z}$, proves (I.2.3).

Uniformity over arbitrary b also has a direct exact-arithmetic invariance check. Translating all measured covariance, fitting, validation and future test inputs by one common vector translates $\widehat{u}$ by that vector and leaves every centered array, candidate, score and selected prediction unchanged. Under latent shift b, $f_b(x) = f(b + x)$ belongs to the same target class, and the conditional label assumptions persist under that change of variables. This invariance agrees with the mean-independent event above; it does not replace the shifted-moment proof needed for estimated centering. Arbitrarily large offsets can lose information in finite precision, which is outside this result.

This completes the upper proof. The constants need not give a useful numerical rate onset.

I.10 Inherited sample lower bound and separate sensor lower bounds

I.10.1 Explicit inherited lemma: adaptive exact-query sample information

Lemma I.3 (adaptive exact-query lower bound, Appendix C). Under (I.1.1) at zero mean, for the common class (I.1.3), the known-law model permitting an independent random seed and at most m adaptive exact point evaluations has separate expected squared and expected norm minimax risks

$$Q_m^{\mathsf{A},0} \ge \frac{d_{\tau,c}^2}{32} a_m, \qquad R_m^{\mathsf{A},0} \ge \frac{d_{\tau,c}}{64} a_m^{1/2}, \tag{I.10.1}$$

where

$$k_0 = 10000, \quad \delta_0 = 10^{-3}, \quad A_{\tau,c} = \frac{\delta_0\sqrt{c}}{4\,2^{\tau/2}}, \quad d_{\tau,c} = \min\{1, A_{\tau,c} k_0^{-q/2}\} > 0. \tag{I.10.2}$$

These bounds are proved in Appendix C, equations (C.4), (C.13)–(C.17), for the adaptive-exact access model. Appendices A–C give the padded-cell/signed-bump construction and the separate expected-norm clipping argument. We now reduce the sensor experiment to that same access model.

Fix and reveal an admissible zero-mean law Σ and a known sensor covariance Γ. Set $\xi = 0$, which satisfies (I.1.7) for every $s_2 \ge 0$. An exact-query learner can draw m iid full inputs X_i from the revealed Gaussian law using its independent seed, query their exact $f(X_i)$ values, simulate $U_i = PX_i + \eta_i$ using independent sensor noise, and run any sensor procedure on these pairs $(U_i, f(X_i))$. Exactly m function evaluations are used. Any seed of that procedure can be included in the independent exact-query seed.

Given its fitted clipped sensor predictor h_D, the exact-query learner outputs

$$\bar{h}_D(x) = \mathbb{E}_\eta h_D(Px + \eta). \tag{I.10.3}$$

This is bounded and jointly measurable, as an integral of a bounded Borel function against a fixed Gaussian law. For each simulated training realization and seed, conditional Jensen over fresh test sensor noise gives

$$\|\bar{h}_D - f\|_{L^2(\mu_\Sigma)}^2 \le \ell_{0,\Sigma,\Gamma,f}(h_D)^2. \tag{I.10.4}$$

Taking square roots before training expectation gives the corresponding norm comparison. Thus for each sensor learner its supremum risk on this noiseless-label submodel is at least the respective

adaptive-exact minimax lower bound (I.10.1). Revealing the law and noise cannot make the original unknown-law problem harder, so these are valid lower bounds for the original class. The reduction averages fresh test noise, rather than using independent seeds at uncountably many test inputs.

This lower-bound route is valid even at $s_2 = 0$. A unit-Gaussian-label subclass would not be available for $s_2 < 1$ and is not used anywhere in this minimax proof.

I.10.2 Gaussian lost-information sine bound

For the lower bounds in this section restrict to the admissible zero-mean subclass $b = 0$. For any fixed Gaussian linear observation $U = PX + \eta$, define $K_U = A + \Gamma$ and its finite-dimensional Moore–Penrose inverse $K_U^\dagger$, extended by zero on its kernel. If $z \in \ker K_U$, nonnegativity of its signal and noise variances implies $\Sigma P^* z = 0$. Consequently

$$W = \Sigma P^* K_U^\dagger U, \quad M_{\exp} = \Sigma P^* K_U^\dagger P \Sigma, \quad V = \Sigma - M_{\exp} \tag{I.10.5}$$

satisfy $X = W + Z$, with $Z \sim N(0, V)$ independent of U, $\mathrm{Cov}(W) = M_{\exp}$ and $0 \preceq V \preceq \Sigma$. This follows by the same zero cross-covariance argument as (I.5.2), restricted to the support of U. $M_{\exp}$ has rank at most d, even with singular A or Γ.

For a unit v let $r = \langle v, Vv\rangle$, $s = \langle v, \Sigma v\rangle$, and use the admissible target $f_v(x) = \sin\langle v, x\rangle$. For $r > 0$, $0 < r \le s \le C$. If $a = \langle v, W\rangle$, then $a \sim N(0, s - r)$, independently of the residual coordinate. The Gaussian characteristic function gives

$$\mathbb{E}[f_v(X) \mid U] = e^{-r/2}\sin(a), \qquad \mathbb{E} f_v(X)^2 = (1 - e^{-2s})/2.$$

The squared conditional-mean expectation is $e^{-r}(1 - e^{-2(s-r)})/2$. Subtracting proves

$$\begin{aligned} \mathbb{E}\,\mathrm{Var}(f_v(X) \mid U) &= \frac{(1 - e^{-r})\{1 + e^{-(2s-r)}\}}{2} \\ &\ge \frac{1 - e^{-r}}{2} \ge b_C r, \qquad b_C = \frac{1 - e^{-C}}{2C} > 0. \end{aligned} \tag{I.10.6}$$

The last step uses monotonicity of $(1 - e^{-r})/r$ for $r > 0$. For $r = 0$ the floor is zero. Every fixed fitted predictor has squared integrated error at least (I.10.6), by conditional projection. Its norm error is therefore at least $\sqrt{b_C r}$ before training expectation. This holds even if the target and channel are supplied. It also holds for any fixed more informative Gaussian test observation, with its corresponding residual covariance.

I.10.3 Proof of Corollary 6

For the upper bounds use Theorem 5 with $\beta \le D_0(d + 1)^{-\tau}$ and $\nu \le v$. The same universal predictor works throughout the class.

For lower bounds fix $b = 0$ and $\xi = 0$, and an orthonormal basis whose first d elements span the prescribed physical sensor space. Set the covariance eigenvalues in this basis to $cj^{-\tau}$. It belongs to the class because $c \le C$ and $D_0 \ge C$. Revealing this covariance is allowed in a lower-bound reduction.

For the sample term choose $\Gamma = 0$, apply the adaptive-exact simulation (I.10.1)–(I.10.4), and obtain separate positive multiples of a_m and $a_m^{1/2}$. For the resolution term at the same Γ choose the omitted eigenvector e_{d+1}; its residual variance is $c(d + 1)^{-\tau}$. Equation (I.10.6) gives the squared floor $b_C c(d + 1)^{-\tau}$ and its pathwise square root.

For the noise term choose $\Gamma = v(Pe_1)(Pe_1)^*$ in the orthonormal physical coordinates. It has trace and operator norm $v \le B_{\mathrm{noise}}$, so it is admissible even when $dv > B_{\mathrm{noise}}$. The first latent coordinate then has conditional variance

$$r = \frac{cv}{c + v} \ge \frac{c}{c + B_{\mathrm{noise}}} v. \tag{I.10.7}$$

For $v = 0$ the noise floor is simply zero; for $v > 0$ use $f(x) = \sin\langle e_1, x\rangle$ and (I.10.6). Each floor is valid pathwise at a fitted output, so square roots precede the training expectation for R. The supremum in (I.2.5) dominates all three witnesses, although they can use different targets and noise covariances. The maximum of three nonnegative lower bounds is at least one-third their sum. This proves (I.2.6) without identifying R^2 with Q or using reverse Jensen.

I.11 Finiteness, operation count and measurable edge cases

I.11.1 Jacobi preprocessing

Each exact maximal-pivot Jacobi step decreases the squared off-diagonal Frobenius norm by twice the squared pivot. Orthogonal congruence preserves the full Frobenius norm, all off-block paired entries preserve their squared sums, and the pivot pair is set to zero. If $d \geq 3$ and $F_{\text{off}}^2 > 0$, the maximum squared off-diagonal entry is at least $F_{\text{off}}^2/[d(d-1)]$. One step therefore multiplies F_{off}^2 by at most $1 - 2/[d(d-1)]$. The process reaches $F_{\text{off}} \leq a$ after at most a constant times

$$d^2\{1 + \log_+(\|T\|_F/a)\}$$

rotations. For $d = 1$ it has already terminated; for $d = 2$ one nontrivial rotation terminates. The deterministic choice of (I.3.2) covers equal diagonals as well as unequal ones. Positive semidefiniteness is preserved, so the terminal diagonal is nonnegative and $\widetilde{T}$ is positive semidefinite.

Scanning the whole off-diagonal array costs $O(d^2)$ per pivot; matrix and accumulated-column updates cost at most this amount. Hence the conservative preprocessing count is

$$O\left(nd^2 + d^4\{1 + \log_+(\|T\|_F/n^{-1/4})\}\right). \tag{I.11.1}$$

As the actual sample-centered T is positive semidefinite, $\|T\|_F \leq \operatorname{tr} T$ and $\mathbb{E}\operatorname{tr} T = (1 - 1/n)\operatorname{tr} K_U \leq S$, independently of b. Since $\log_+ x \leq \log(1 + x)$, Jensen bounds the expected log factor by $\log(1 + n^{1/4}S)$. Thus the expected count is polynomial under fixed S. Estimating the center and subtracting it from the measured arrays costs $O(md)$, and storing it costs $O(d)$. The expected arithmetic bound remains independent of b. This does not assert a deterministic worst-case polynomial bound in m,d independent of arbitrary observed magnitudes or a finite-bit subtraction guarantee.

Each step is Borel, and the finite first stopping time is a countable disjoint union of Borel step-count events. Sorting with fixed ties is Borel. The output eigenbasis is that of the surrogate $\widetilde{T}$; no continuity at a repeated population eigenvalue is required.

I.11.2 Count guard bounds dimension and degree on every sample

For a retained coordinate i, $E\widehat{\lambda}_i \geq 1$, so its unit multi-index is in $\mathcal{I}_E$. Therefore every accepted candidate has

$$k + 1 \leq D \leq D_{\max}. \tag{I.11.2}$$

For $k > 0$ the maximum total degree is exactly $\lfloor E\widehat{\lambda}_{\max} \rfloor$: energy implies degree at most this number, while every power along a largest-eigenvalue coordinate up to that degree is feasible. Those powers and the constant are distinct indices, so

$$r + 1 \leq D \leq D_{\max}. \tag{I.11.3}$$

For $k = 0$, $r = 0$ and $D = 1$. These assertions hold on every sample, without an upper eigenvalue guard. The immediate rejection when $k + 1 > D_{\max}$ is valid. In lazy traversal every feasible prefix has a zero completion, and the ancestor paths of the first $D_{\max} + 1$ visited leaves have at most $1 + k(D_{\max} + 1)$ nodes. Bounded bookkeeping and feasible-child comparisons use $O(d(D_{\max} + 1))$

operations before acceptance or rejection, without allocating a huge degree range. Energy comparisons are exact.

Center each input by stored $\widehat{u}$, then compute its retained coordinates in $O(d + dk)$ operations per example. The recurrence (I.3.6) through degree r in each coordinate and all product features cost $O(kr + Dk) = O(D_{\max}^2)$ for an accepted candidate, in addition to projection. An SGD update and coefficient projection use $O(D)$; averaging pre-update vectors costs $O(ND)$. Validation and prediction use the same features. Rejected and explicit zero slots need only a zero prediction.

There are $J = 2k_m+2 = O(\log(m+2))$ slots. Combining shared preprocessing with a conservative bound for all training and validation evaluations gives expected real-arithmetic count at most the expectation of

$$O\left(md + nd^2 + d^4\{1 + \log_+(\|T\|_F n^{1/4})\} + J\{d(D_{\max} + 1) + m(dD_{\max} + D_{\max}^2)\}\right). \tag{I.11.4}$$

Here $D_{\max} = \mathrm{floor}(\sqrt{M})$ is the global algorithmic guard; it is permitted in the operation/memory bounds, unlike the sharper D_* bound used only for the good candidate's risk. This is polynomial in m,d for fixed model bounds, uniformly over b in exact arithmetic. Retained degree and dimension bounds also keep the feature descriptions and coefficient vectors finite. Streaming feature evaluation is possible. Store the sorted d by d eigenvector matrix once; every retained basis is a prefix/view into that shared matrix, rather than a separate d by k copy for each candidate. With this shared representation, a conservative implementation may retain the measured dataset plus $O(d^2+JD_{\max}^2)$ preprocessing/model memory.

The maximum grid energy 2^{k_m} is less than $2M$ for $M > 1$, so every retained eigenvalue is greater than or equal to $1/(2M)$. This deterministic whitening cutoff does not control all polynomial feature magnitudes or floating-point errors. A finite-bit statement would additionally need precision, condition-number and numerical-stability analysis, which is absent here.

I.11.3 No omitted singular or budget branches

Small m uses zero. The stored center is always estimated only on the covariance split and subtracted from fitting, validation and future test inputs. A zero measured covariance has zero sample-centered and surrogate eigenvalues, so every energy has only the constant feature. A singular covariance in a nonempty retained space is harmless in the algorithm because it inverts only positive surrogate eigenvalues; on the analysis good event the true retained covariance is also strictly positive. The main proof never inverts A itself. At threshold equality directions are retained; exact energy equality retains an index. Equal scores keep the earliest slot. Every allocated label belongs to an original pair in (I.1.4), including discarded covariance-split labels. These conventions give one common measurable predictor over the model class.

I.12 Physical example: cell averages of unknown elliptic Gaussian curves

This is a concrete covariance family within Theorem 5; it is not a theorem for arbitrary grids or pointwise path samples. Let $H = L^2(0,1)$, $X = b + X_0$ with arbitrary unknown $b \in H$, and centered $X_0 \sim N(0,\Sigma)$. For an unknown measurable coefficient $a(x)$ with fixed bounds $0 < a_{\min} \le a(x) \le a_{\max} < \infty$, define the positive Dirichlet operator L_a by the closed form

$$\mathfrak{q}_a(v) = \int_0^1 a(x)|v'(x)|^2 dx, \qquad v \in H_0^1(0,1), \qquad \Sigma = L_a^{-1}. \tag{I.12.1}$$

The form is coercive. Comparison with the Dirichlet Laplacian, compact embedding of H_0^1 into L^2, and the min–max principle give a positive compact inverse with

$$\frac{1}{a_{\max}\pi^2 j^2} \le \lambda_j(\Sigma) \le \frac{1}{a_{\min}\pi^2 j^2}. \tag{I.12.2}$$

The inverse is trace class because $\sum_j j^{-2} < \infty$. Thus $\tau = 2$, $c = (a_{\max}\pi^2)^{-1}$, $C = (a_{\min}\pi^2)^{-1}$. The covariance and eigenfunctions vary with unknown a. At $a = 1$ the inverse kernel is $\min(s,t) - st$, giving the Brownian-bridge example; the theorem does not rely on that specific member.

I.12.1 Fixed average sensors and coverage

Partition (0,1) into d cells I_j of equal length $h = 1/d$ and use $\psi_j = h^{-1/2}\mathbf{1}_{I_j}$. These are orthonormal bounded linear L^2 sensors. The physical coordinate map is $P_d x = (\langle x, \psi_j\rangle)_j$; let $Q_d = I - P_d^* P_d$. For v in H_0^1, cellwise mean-zero Poincare gives

$$\|Q_d v\|_2^2 \le h^2\pi^{-2}\|v'\|_2^2. \tag{I.12.3}$$

For $v = \Sigma^{1/2} z$, spectral calculus gives $\mathfrak{q}_a(v) = \|z\|_2^2$, so $\|v'\|_2^2 \le a_{\min}^{-1}\|z\|_2^2$. It follows that

$$\beta_d = \|Q_d\Sigma Q_d\|_{\mathrm{op}} = \|Q_d\Sigma^{1/2}\|_{\mathrm{op}}^2 \le \frac{1}{a_{\min}\pi^2 d^2} \le \frac{4}{a_{\min}\pi^2}(d+1)^{-2}. \tag{I.12.4}$$

This is the needed operator-norm coverage. A trace-tail estimate would yield a weaker order d^{-1}. No lower bound on $P_d\Sigma P_d^*$ is needed for Theorem 5. Its sensor space is fixed independently of a and need not be invariant under Σ. The universal learner does not receive a, its eigenfunctions, or even the envelope constants to run.

For completeness, the physical family also satisfies an optional compression lower bound. If u has constant value u_j on cell (l_j, r_j), put $w(x) = u_j(x - l_j)(r_j - x)/2$ there. Then $w \in H_0^1$, and direct integration gives $\langle u, w\rangle = \|w'\|_2^2 = h^2\|u\|_2^2/12$. The inverse-form identity

$$\langle \Sigma u, u\rangle = \sup_{v \in H_0^1}\{2\langle u, v\rangle - \mathfrak{q}_a(v)\}$$

at $v = w/a_{\max}$ gives $\langle \Sigma u, u\rangle \ge h^2\|u\|_2^2/(12a_{\max})$. This stronger compression property is not used by the tail-only theorem.

I.12.2 Upper bound and noise units

For arbitrary independent Gaussian coordinate noise with $\operatorname{tr}\Gamma \le B_{\mathrm{noise}}$ and any iid label law satisfying (I.1.7), Theorem 5 and (I.12.4) imply

$$\sup_f \mathbb{E}\ell(\widehat{h})^2 \le K\{[\log(m+2)]^{-3} + (d+1)^{-2} + \|\Gamma\|_{\mathrm{op}}\}, \tag{I.12.5}$$

uniformly over the fixed ellipticity family, all $b \in H$ and the label-law class. The cell-average covariance and coverage calculations are independent of b; stored centering handles the observed mean. Constants depend on the fixed bounds and s_2.

If a device reports the raw average $d\int_{I_j} X$ plus independent noise of variance $\sigma^2_{\mathrm{sensor}}$, divide that report by $\sqrt{d}$. It becomes the orthonormal coordinate $\langle X, \psi_j\rangle$ with noise covariance $\Gamma = (\sigma^2_{\mathrm{sensor}}/d)I_d$. Its trace is $\sigma^2_{\mathrm{sensor}}$, and its operator norm is $\sigma^2_{\mathrm{sensor}}/d$. For fixed $\sigma^2_{\mathrm{sensor}} \le B_{\mathrm{noise}}$, (I.12.5) becomes

$$K\{[\log(m+2)]^{-3} + (d+1)^{-2} + \sigma^2_{\mathrm{sensor}}/d\}. \tag{I.12.6}$$

In contrast, perfectly correlated raw-average noise of covariance $\sigma^2_{\mathrm{sensor}}\mathbf{1}\mathbf{1}^*$ becomes $\Gamma = (\sigma^2_{\mathrm{sensor}}/d)\mathbf{1}\mathbf{1}^*$. Its trace is still $\sigma^2_{\mathrm{sensor}}$ but its operator norm is $\sigma^2_{\mathrm{sensor}}$, so no $1/d$ operator-norm reduction follows. The bound is not a claim that every fixed correlated pattern attains its worst-case noise floor.

I.12.3 LOWER BOUNDS INSIDE THIS SMALLER PHYSICAL FAMILY

A lower bound over a larger abstract coverage class does not establish a lower bound for this elliptic family. Fix instead $b = 0, \xi = 0$ and a constant coefficient a_0 in $[a_{\min}, a_{\max}]$, reveal its law, and write $\lambda_j = (a_0 \pi^2 j^2)^{-1}$.

Sample term. For this fixed zero-mean, noiseless-label law, the adaptive-exact simulation (I.10.3)–(I.10.4) applies for any fixed known sensor covariance, including the isotropic law in (I.12.6). Lemma I.3 gives squared order $[\log(m+2)]^{-3}$ and norm order $[\log(m+2)]^{-3/2}$.

Resolution term. For the actual d-dimensional observation, its explained covariance $M_{\exp}$ in (I.10.5) has rank at most d and is positive. On the span of the first $d+1$ signal eigenvectors, choose a unit vector v in the kernel of its restriction. Then $M_{\exp} v = 0$, and

$$\langle v, Vv \rangle = \langle v, \Sigma v \rangle \geq \lambda_{d+1}. \tag{I.12.7}$$

One way to choose v is to restrict the rank-at-most-d map $M_{\exp}$ to that $(d+1)$-dimensional span and take a nonzero kernel vector. The sine bound (I.10.6) gives a constant times $(d+1)^{-2}$, with a pathwise norm square root. This works for noiseless or noisy sensing without aligning cell averages with eigenvectors.

Noise-class term. Let e_1 be the leading signal eigenvector, $\lambda = \lambda_1$, $p = P_d e_1$. Give the learner all other latent eigen-coordinates at test time and reveal both law and target. This extra information only makes prediction easier. After subtracting the other coordinates, the noisy observation is $pZ_1 + \eta$ with $Z_1 \sim N(0, \lambda)$. If p is nonzero, choose $\Gamma = vww^*$ where $w = p/\|p\|$ and $0 < v \leq B_{\text{noise}}$. Its trace and norm are v. The conditional variance of Z_1 is

$$\frac{\lambda v}{v + \lambda \|p\|^2} \geq \frac{\lambda v}{v + \lambda} \geq \frac{\lambda}{B_{\text{noise}} + \lambda} v. \tag{I.12.8}$$

If $p = 0$ the coordinate is entirely hidden, with variance λ, which is also at least the last expression. Apply the same sine calculation even in this easier oracle experiment. This proves the squared v floor in the original experiment. For $v = 0$ the claimed noise term is zero and needs no division.

Specified independent raw-average noise. If $\Gamma = (\sigma^2_{\text{sensor}}/d) I_d$, projection of η onto $p/\|p\|$ has variance $\nu = \sigma^2_{\text{sensor}}/d$; perpendicular coordinates contain no information about Z_1 and are independent. The extra-information calculation gives the same first expression in (I.12.8) with $v = \nu$. This lower bound therefore holds for that particular isotropic noise model, not merely after substituting a rank-one law from a larger class. The case $p = 0$ is again immediate.

The maximum of the three separate floors controls their sum up to constants. Combining them with (I.12.5) gives the squared/norm sharp orders over the elliptic family and bounded-trace/operator-norm noise class. Combining with (I.12.6) gives the squared order $[\log(m+2)]^{-3} + (d+1)^{-2} + \sigma^2_{\text{sensor}}/d$ for independent raw-average noise, and the sum of square roots for expected norm. Every norm observation floor is taken pathwise first; Lemma I.3 supplies the separate norm sample lower bound.

The classical form comparison, cellwise Poincare and variational inverse calculation establish this example's sensor geometry. They do not validate arbitrary real-world sensor coverage. Point evaluation is not a bounded functional on general $L^2(0,1)$, and no arbitrary point grid follows from this average-sensor result.