THE STATISTICAL COST OF FINDING SMOOTH COORDINATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Being smooth in the right coordinates is not the same as knowing them. We study which frame differences matter for approximation and how unknown coordinates affect function-estimation risk. Our geometric analysis shows that a positive-dimensional frame space can be covered by a single representation neighborhood at coarse resolution. For a fixed two-block matrix-dilation anisotropic Besov profile with rough dimension k and inverse-smoothness gap Δ, let $m = k(d-k)$. At resolution r and a fixed distortion threshold $c > 0$, the representation entropy (log covering number) is zero when $\Delta \log(1/r) \le c/2$ and equals $m\Delta \log(1/r) + O_c(m)$ when $\Delta \log(1/r) \ge c$, with constants independent of k and d. We derive this transition from an exact finite-resolution metric formula; rotation transfer and explicit witnesses connect the metric to approximation and separation in function space. In a separate regression model with standard Gaussian design and independent Gaussian noise of fixed variance, consider $f(x) = g(W^\top x)$ with unknown $W \in \mathrm{St}(d,k)$ and g in the unit $B^s_{2,q}(\mathbb{R}^k)$-ball, $s > k/2$, $1 \le q \le \infty$. We prove matching local entropy bounds and the minimax rate $R^\star_n \asymp n^{-2s/(2s+k)} + \frac{m}{n}\log(1+\frac{n}{m})$ for $d > k$ and $m = o(n)$, with constants uniform in the ambient dimension. The first term is the known-subspace estimation rate; the second is the representation cost, whose logarithm is unavoidable when that term dominates. Unknown coordinates can therefore be the leading contribution to function-estimation risk as d grows.

1 INTRODUCTION

A function can be much smoother in some directions than in others. When these directions are known, estimation can exploit their different regularities. When they are unknown, the same regularity profile is compatible with many coordinate frames. Anisotropic Besov classes formalize this profile through a smoothness vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$. Classical fixed-frame theory describes how the profile governs estimation rates (Birgé, 1983; Lepski, 2015; Suzuki, 2019). It does not, by itself, quantify the additional cost of an unknown orientation.

We ask two questions. At a given approximation accuracy, which differences between frames need to be distinguished? And how much does an unknown frame add to the risk of estimating the function? We answer the first through finite-resolution geometry and the second through minimax analysis in a separate Gaussian-design model. Both distinguish two sources of complexity,

$$\boxed{\text{function estimation complexity} \;+\; \text{representation identification complexity.}} \tag{1}$$

The first concerns estimating a function within a frame; the second concerns distinguishing the representations themselves. The effective number of representations depends on resolution, not just on the dimension of the frame space.

Which frame differences matter? At a given approximation accuracy, an unknown frame need not be specified exactly. We count the frame neighborhoods needed at that accuracy, rather than coordinate bases individually. Consider a two-block profile in the matrix-dilation Besov model, with a rough block of dimension k and smoothness s, and a smooth block of smoothness $t > s$. The rough subspaces form a Grassmannian of dimension $m = k(d-k)$, yet a single representation neighborhood can cover the entire space at coarse resolution. With contrast $\Delta = 1/s - 1/t$ and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

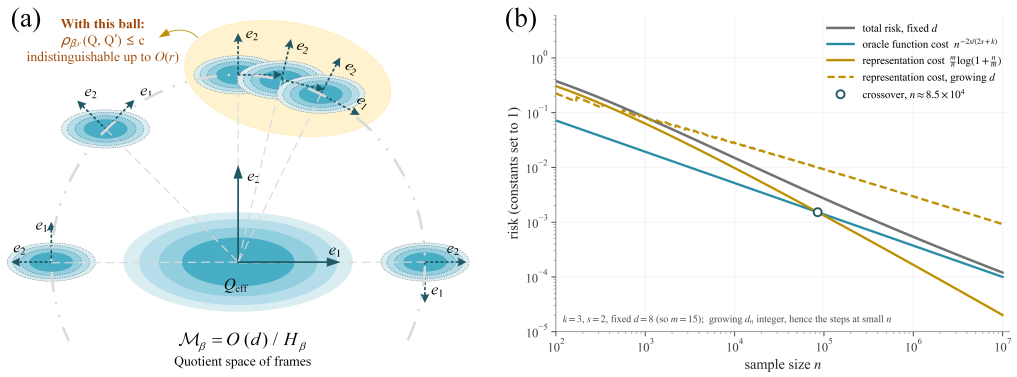

Figure 1: **Regularity in unknown coordinates.** (a) One anisotropic profile $\boldsymbol{\beta}$, many candidate frames, forming the quotient $\mathcal{M}_{\boldsymbol{\beta}} = O(d)/H_{\boldsymbol{\beta}}$ (dashed). Frames inside one $\rho_{\boldsymbol{\beta},r}$-ball (shaded) are equivalent up to $O(r)$ at L_2-resolution (Theorem 1) and act as the single effective representation Q_{eff}, drawn larger for emphasis only; distinguishable representations are counted by a covering number of the quotient. (b) The resulting split of the risk, evaluating the rate of Theorem 4 directly (no data, no fitting): function term $n^{-2s/(2s+k)}$ (teal), representation term $(m/n)\log(1+n/m)$, $m = k(d-k)$ (amber), and their sum (grey). For $k = 3$, $s = 2$, $d = 8$, so that $s > k/2$ as Theorem 3 requires, the sum settles onto the function term past the crossover and the representation is lower-order; with d growing so that $k(d-k) \asymp n^{k/(2s+k)}$ (dashed) it stays on top and is first-order; the small steps are the integer values of d.

$\zeta = \Delta \log(1/r)$, the representation entropy—the log covering number at a fixed distortion threshold $c > 0$—is exactly zero for $\zeta \le c/2$ and equals $m\zeta + O_c(m)$ for $\zeta \ge c$, with constants independent of k and d (Theorem 2).

To make these neighborhoods precise, we remove changes of basis that preserve the regularity representation, obtaining the quotient $\mathcal{M}_{\boldsymbol{\beta}} = O(d)/H_{\boldsymbol{\beta}}$. We equip it with a scale-dependent metric $\rho_{\boldsymbol{\beta},r}$ measuring distortion between the multiscale systems of two frames up to resolution r. Our rotation-transfer theorem turns metric proximity into $O(r)$ approximation in L_2, with a fixed inflation of the Besov norm (Theorem 1). For two-block profiles, an exact metric formula gives the entropy transition (Proposition 1), while witness constructions realize sufficiently separated frames by functions separated at the same L_2 scale (Propositions 2 and 3).

The fine-resolution slope $k(d-k)\Delta$ measures the dispersion of the inverse smoothnesses, rather than their sum. This sum defines the aggregate $\widetilde{\beta} = \left(\sum_i \beta_i^{-1}\right)^{-1}$ governing the fixed-frame rate. An isotropic profile, for example, has zero dispersion regardless of its aggregate smoothness. For general profiles, we prove the tangent-space mechanism (Lemma 1) and conjecture the slope $\mathfrak{A}(\boldsymbol{\beta}) = \sum_{i<j} |\beta_i^{-1} - \beta_j^{-1}|$ (Conjecture 1).

What does an unknown representation cost? For the statistical question, consider an unknown k-dimensional active subspace,

$$f_{W,g}(x) = g(W^\top x), \qquad W \in \mathrm{St}(d,k), \quad \|g\|_{B^s_{2,q}(\mathbb{R}^k)} \le 1, \tag{2}$$

with $d > k$, $s > k/2$, and $1 \le q \le \infty$. The design is $N(0, I_d)$, with independent mean-zero Gaussian noise of fixed variance; statistical L_2 distances use this design distribution. This model shares the Grassmannian representation geometry of the two-block setting, but the two function classes are not nested.

The representation has only $m = k(d-k)$ degrees of freedom, suggesting a parametric cost of m/n. What this count misses is that small functions can retain fine directional structure. Our witness functions narrow as their L_2 norm shrinks: witnesses of size r still distinguish subspaces at angular scale $r^{1/s}$. Even inside an L_2-ball of radius r about the origin, these directionally distinct functions form a packing with separation proportional to r and log-cardinality of order $m\log(1/r)$. The logarithm therefore persists in the *local* entropy; it is not merely an artifact of covering the entire class.

We prove the dimension-uniform local entropy characterization (Theorem 3), for sufficiently small r,

$$H_{\mathrm{loc}}(r) = \Theta_{k,s,q}\Big(r^{-k/s} + k(d-k)\log(1/r)\Big). \tag{3}$$

Applying the local-entropy minimax characterization of Prasadan & Neykov (2025) gives the matching rate (Theorem 4)

$$R_n^\star \asymp n^{-2s/(2s+k)} + \frac{k(d-k)}{n}\log\Big(1+\frac{n}{k(d-k)}\Big) \qquad (k(d-k)=o(n)), \tag{4}$$

with constants independent of the ambient dimension. The first term is the oracle rate for a known subspace; the second is the cost of the unknown representation. The two terms cross when $m\log(1+n/m) \asymp n^{k/(2s+k)}$: the representation term is lower-order for fixed d, but can dominate as d grows (Figure 1b).

The separation of function and representation costs, and the importance of the latter in high dimension, already appear in composite and single-index bounds (Baraud & Birgé, 2014; Alquier & Biau, 2013). Our contribution is a matching characterization on the stated Gaussian–Besov class, uniform in d. Because the local entropy bounds match, the logarithm cannot be removed where the representation term dominates.

Section 6 provides numerical illustrations of the finite-resolution transition and exploratory tests of the risk crossover.

2 Setup: oriented classes and the representation quotient

We first define the function class in a fixed frame, then identify the rotations that leave its regularity representation unchanged. Throughout, $p = 2$ and $1 \le q \le \infty$; all further parameter restrictions are stated in each result. Write $\lambda_i = \beta_i^{-1}$, $\Lambda_{\boldsymbol{\beta}} = \mathrm{diag}(\lambda_1,\dots,\lambda_d)$, $\lambda_{\min} = \min_i \lambda_i > 0$, and $A_{\boldsymbol{\beta}} = 2^{\Lambda_{\boldsymbol{\beta}}}$, so that $|\det A_{\boldsymbol{\beta}}|^{\widetilde{\beta}} = 2$.

Matrix-dilation anisotropic Besov classes. For the geometric analysis, we use the expansive-matrix formulation of anisotropic Besov spaces (Bownik, 2005; 2013). A frame $Q \in O(d)$ orients the dilation through $\Lambda_Q = Q\Lambda_{\boldsymbol{\beta}}Q^\top$, with frequency ellipsoids $E_t^Q = \{\xi \in \mathbb{R}^d : \|e^{-t\Lambda_Q}\xi\|_2 \le 1\}$, increasing in t. To define regularity relative to these ellipsoids, fix a smooth radial φ with $\mathbf{1}_{B(0,1)} \le \varphi \le \mathbf{1}_{B(0,2)}$ and set $\Phi_j^Q(\xi) = \varphi(e^{-j\tau\Lambda_Q}\xi)$, where $\tau = \log 2$. The Littlewood–Paley blocks are the Fourier multipliers $P_0^Q = \Phi_0^Q$ and $P_j^Q = \Phi_j^Q - \Phi_{j-1}^Q$ for $j \ge 1$. The oriented class is the unit ball

$$\mathcal{K}_Q = \Big\{ f \in L_2(\mathbb{R}^d) :\ \big\|\big(2^j\|P_j^Q f\|_{L_2}\big)_{j\ge 0}\big\|_{\ell^q} \le 1\Big\}, \tag{5}$$

the scalar-smoothness-$\widetilde{\beta}$ ball for $A_{\boldsymbol{\beta},Q} = QA_{\boldsymbol{\beta}}Q^\top$, since $2^j = |\det A_{\boldsymbol{\beta},Q}|^{j\widetilde{\beta}}$. Allowing the frame to vary gives the unknown-coordinate class $\mathcal{F}_{\boldsymbol{\beta}} = \bigcup_{Q\in O(d)} \mathcal{K}_Q$.

Remark 1 (relation to coordinatewise anisotropic Besov spaces)**.** The geometric results concern the class (5). Coordinatewise formulations based on directional moduli of smoothness (Suzuki & Nitanda, 2021) provide a related description of anisotropic regularity; no norm equivalence between the two formulations is assumed or established here.

The representation quotient. Changing the basis within a block of equal smoothness leaves the oriented class unchanged. These redundant rotations form the centralizer $H_{\boldsymbol{\beta}} = \{R \in O(d) : RA_{\boldsymbol{\beta}} = A_{\boldsymbol{\beta}}R\} \simeq O(m_1)\times\cdots\times O(m_L)$, where $m_1,\dots,m_L$ are the multiplicities of the distinct λ_i. Identifying Q and QR for $R \in H_{\boldsymbol{\beta}}$ gives the compact homogeneous space

$$\mathcal{M}_{\boldsymbol{\beta}} = O(d)/H_{\boldsymbol{\beta}}, \qquad \dim \mathcal{M}_{\boldsymbol{\beta}} = \sum_{a<b} m_a m_b. \tag{6}$$

Isotropic profiles give a point ($L = 1$); two-block profiles give the Grassmannian $\mathrm{Gr}(k,d)$ of rough subspaces ($L = 2$, $m_1 = k$); fully distinct profiles give the real flag manifold. The quotient indexes regularity representations; the next section asks which of them remain distinguishable at a given resolution.

3 FINITE-RESOLUTION GEOMETRY OF FRAMES

The quotient $\mathcal{M}_{\boldsymbol{\beta}}$ distinguishes regularity representations; the next step is to compare them at a given approximation accuracy. The effect of a rotation depends on the shape of the frequency ellipsoids: a fixed angle can cause little distortion for a nearly spherical ellipsoid and much more for an elongated one. Since the relative axis lengths depend on the smoothness contrasts and the scale, angular separation alone does not specify the distortion at a chosen resolution. We therefore compare the two frames' frequency ellipsoids over all scales resolved at L_2-accuracy r.

Definition 1 (representation metric)**.** For $r \in (0,1)$ let $T_r = \log(1/r)$. For $[Q], [Q'] \in \mathcal{M}_{\boldsymbol{\beta}}$ set

$$\rho_{\boldsymbol{\beta},r}([Q],[Q']) = \sup_{0\le t\le T_r} \log \kappa\big(e^{-t\Lambda_{\boldsymbol{\beta}}}\, Q^\top Q'\, e^{t\Lambda_{\boldsymbol{\beta}}}\big), \tag{7}$$

where κ is the condition number. The quantity is symmetric, invariant under $H_{\boldsymbol{\beta}}$ on either side, and nondecreasing in T_r.

At scale t, the conjugated matrix is the change of basis between the two anisotropic multiscale systems. The supremum records the largest ellipsoid distortion over the resolved scales. The next theorem turns this control into L_2 approximation with a fixed inflation of the Besov ball.

Theorem 1 (rotation transfer)**.** *Fix d, $\boldsymbol{\beta}$ and $1 \le q \le \infty$, and let $c_0 > 0$. There exist constants C_0, C_1 depending only on $(c_0, \lambda_{\min}, q)$ such that if $\rho_{\boldsymbol{\beta},r}([Q],[Q']) \le c_0$ then*

$$\sup_{f\in\mathcal{K}_Q} \inf_{g\in C_0\mathcal{K}_{Q'}} \|f-g\|_{L_2} \;\le\; C_1 r, \tag{8}$$

and symmetrically in Q, Q'.

The proof in Appendix B follows the frequency decomposition. The metric bound and $|\det(Q^\top Q')| = 1$ place every singular value of the change-of-basis matrix in $[e^{-c_0}, e^{c_0}]$. This yields two-sided ellipsoid inclusions with a fixed level shift $L = \lceil c_0/(\tau\lambda_{\min})\rceil$. Truncation at level $\sim \log_2(1/r)$ leaves an $O(r)$ tail, while finite band overlap bounds the Besov norm of the truncation in the other frame. Thus a fixed-radius $\rho_{\boldsymbol{\beta},r}$-net gives a representation cover at L_2-resolution r, using the inflated balls $C_0\mathcal{K}_Q$.

The two-block case is exactly solvable. Let $\boldsymbol{\beta}$ be two-block with rough dimension k, smoothnesses $t > s$, and contrast $\Delta = 1/s - 1/t > 0$; identify $[Q] \in \mathcal{M}_{\boldsymbol{\beta}}$ with the rough subspace $U \in \mathrm{Gr}(k,d)$, and let $\theta_{\max}(U,V)$ be the largest principal angle, so $\sin\theta_{\max} = \|P_U - P_V\|_{\mathrm{op}}$.

Proposition 1 (exact two-block metric)**.** *For all $U, V \in \mathrm{Gr}(k,d)$ and all $r \in (0,1)$, with $\zeta = \zeta(r) = \Delta\log(1/r)$,*

$$\rho_{\boldsymbol{\beta},r}(U,V) = 2\,\mathrm{arsinh}\big[\sinh(\zeta)\,\sin\theta_{\max}(U,V)\big]. \tag{9}$$

The proof in Appendix A uses the CS decomposition to reduce the metric to a 2×2 calculation in each principal plane; the largest angle determines the condition number. Monotonicity in ζ and $\theta_{\max}$ then converts a fixed distortion threshold into an angular threshold. This gives both the visibility transition and the covering radius needed to count representations in Section 4.

Corollary 1 (visibility threshold)**.** *Fix a distortion threshold $c > 0$ and set $a_r(c) = \sinh(c/2)/\sinh(\zeta(r))$. (i) If $\zeta(r) < c/2$, then $\rho_{\boldsymbol{\beta},r}(U,V) < c$ for* every *pair: the maximal possible distortion at resolution r is $\rho_{\max} = 2\zeta(r) < c$, attained whenever $\theta_{\max} = \pi/2$. (ii) If $\zeta(r) \ge c/2$, then $\rho_{\boldsymbol{\beta},r}(U,V) \ge c$ if and only if $\sin\theta_{\max}(U,V) \ge a_r(c)$. In either case the covering relation*

$$\rho_{\boldsymbol{\beta},r}(U,V) \le c \quad\Longleftrightarrow\quad \sin\theta_{\max}(U,V) \le \epsilon_r(c) := \min\{1,\, a_r(c)\} \tag{10}$$

holds for all pairs. In particular, for $c = 1$, no pair of frames reaches distortion 1—at any angle—once $\zeta(r) < 1/2$.

For fixed c, the covering radius satisfies $\epsilon_r(c) \asymp e^{-\zeta} = r^{\Delta}$ once ζ is sufficiently large. Near isotropy, reaching distortion 1 requires $\Delta\log(1/r) \ge 1/2$: the required resolution is exponentially fine as $\Delta \to 0$. The exact covering relation (10) describes both this transition and the power-law regime. Section 6 compares the metric identity and visibility threshold with numerical calculations.

From frame separation to function separation. Theorem 1 gives approximation between nearby frames. In the two-block case, sufficiently separated frames admit a converse: a function in one oriented ball stays separated from the entire competing ball, even after a prescribed fixed norm inflation.

Proposition 2 (orientation witness)**.** *Let $\boldsymbol{\beta}$ be two-block, $1 \le q \le \infty$, and $c_2 \ge 1$. There exist ρ^* and $a > 0$ depending only on (s, t, q, c_2) such that for all $r \in (0, \frac{1}{2})$ and all $U, V \in \mathrm{Gr}(k, d)$ with $\rho_{\boldsymbol{\beta},r}(U, V) \ge \rho^*$ there is $f \in \mathcal{K}_{Q_U}$ with $\|f\|_{L_2} = a\, r$ and $\inf_{g \in c_2 \mathcal{K}_{Q_V}} \|f - g\|_{L_2} \ge \frac{a}{2} r$. Neither ρ^* nor a depends on d or r.*

This witness is a spectral bump placed where the two frames' level ellipsoids disagree, using the top singular direction of the change-of-basis matrix. It depends on the pair (U, V). To build a function packing from a subspace packing, we instead need one choice of f_U per subspace that works simultaneously for all pairs. The next construction provides such a family with an exact correlation law.

Proposition 3 (canonical spectral witness)**.** *Let $\boldsymbol{\beta}$ be two-block, $1 \le q \le \infty$. There is $a = a(s, t, q) > 0$ such that for every $r \in (0, 1)$ and every $U \in \mathrm{Gr}(k, d)$ the function f_U defined by $\widehat{f_U}(\xi) = c_{r,d} \exp\big(-d\, \|e^{-T_r \Lambda_U} \xi\|^2 / 2\big)$, with $c_{r,d}$ normalizing $\|f_U\|_{L_2} = a\, r$, lies in $\mathcal{K}_{Q_U}$; and for all U, V with principal angles $\theta_1, \dots, \theta_k$,*

$$\mathrm{Corr}(f_U, f_V) = \prod_{i:\, \theta_i > 0} \big[1 + \sinh^2(\zeta) \sin^2 \theta_i\big]^{-1/2} \;\le\; \frac{1}{\cosh\big(\rho_{\boldsymbol{\beta},r}(U, V)/2\big)}. \tag{11}$$

In particular $\rho_{\boldsymbol{\beta},r}(U, V) \ge \rho^*$ gives $\|f_U - f_V\|_{L_2}^2 \ge 2(1 - \mathrm{sech}(\rho^*/2)) a^2 r^2$.

Under the normalized spectral energy, the factor d fixes the expected squared radius of $e^{-T_r \Lambda_U} \xi$ at $1/2$. This keeps a independent of d and leaves the correlation unchanged. The sech bound follows by retaining the largest-angle factor in (11) and applying Proposition 1; proofs are in Appendix C.4. The correlation law is the exact finite-t analogue of the Gaussian-design witness correlation in Section 5.1. A subspace packing at fixed positive $\rho_{\boldsymbol{\beta},r}$-separation therefore yields an L_2-packing at scale r; Section 4 uses this family to realize the representation entropy.

4 Representation entropy

Section 3 identifies frame differences that matter at resolution r. We now count the representations needed at that resolution.

Definition 2 (representation entropy)**.** For a threshold $c > 0$, let $\mathsf{E}_{\boldsymbol{\beta}}(r; c) = \log N\big(c,\ \mathcal{M}_{\boldsymbol{\beta}},\ \rho_{\boldsymbol{\beta},r}\big)$ be the log covering number of the frame space by $\rho_{\boldsymbol{\beta},r}$-balls of radius c.

By Theorem 1, a covering with $\exp(\mathsf{E}_{\boldsymbol{\beta}}(r; c))$ centers approximates the unknown-frame class at L_2-resolution r by the corresponding oriented balls, each with a fixed norm inflation. For two-block profiles, Proposition 1 and (10) identify this covering number exactly with a Grassmann covering number at angular radius $\arcsin \epsilon_r(c)$. Applying the two-sided bounds with universal constants of Szarek (1998, Theorem 8 and Remark (iii)) gives the following.

Theorem 2 (two-block representation entropy)**.** *Let $\boldsymbol{\beta}$ be two-block with contrast Δ, let $m = k(d - k)$, and fix $c > 0$. There exist constants $c_1(c), C_1(c)$, independent of k and d, such that $\mathsf{E}_{\boldsymbol{\beta}}(r; c) = 0$ whenever $\zeta(r) \le c/2$, while for all $r \in (0, 1)$, $\mathsf{E}_{\boldsymbol{\beta}}(r; c) \le m\Big[\log\big(1/\epsilon_r(c)\big) + C_1\Big]$, and, whenever $\epsilon_r(c) \le 1/\sqrt{2}$, $\mathsf{E}_{\boldsymbol{\beta}}(r; c) \ge m\Big[\log\big(1/\epsilon_r(c)\big) - c_1\Big]$. Consequently, for fixed c and $\zeta(r) \ge c$,*

$$\mathsf{E}_{\boldsymbol{\beta}}(r; c) = m\, \zeta(r) + O_c(m) = \mathfrak{A}(\boldsymbol{\beta}) \log(1/r) + O_c(m), \qquad \mathfrak{A}(\boldsymbol{\beta}) = k(d - k)\Delta. \tag{12}$$

The proof is in Appendix C. The quotient dimension m sets the exponent of the Grassmann covering number, while the contrast Δ controls how its required radius shrinks with r. Their product gives the fine-resolution slope in (12). When $\zeta(r) \le c/2$, a single ball covers the representation space, even though the Grassmannian itself has dimension m.

Realization in function space. To realize this count in function space, apply Proposition 3 to a subspace packing at fixed positive $\rho_{\boldsymbol{\beta},r}$-separation. Its canonical family gives one function per subspace, with simultaneous pairwise L_2 separation at scale r:

Corollary 2 (the canonical family realizes the representation entropy)**.** *Let $\mathcal{F}_{\boldsymbol{\beta}}^{\mathrm{2b}} = \bigcup_Q \mathcal{K}_Q$ be the two-block unknown-frame class, fix $\rho^* > 0$, let a be as in Proposition 3 and $c_3 = a\sqrt{2(1-\operatorname{sech}(\rho^*/2))}$. There is $C = C(s,t,q,\rho^*)$ such that for all $r \in (0, \frac{1}{2})$ with $\zeta(r) \geq C$,*

$$\log \mathcal{M}\big(c_3 r,\, \mathcal{C}_r\big) \;\geq\; \mathfrak{A}(\boldsymbol{\beta}) \log(1/r) - O(m), \qquad \mathcal{C}_r \subset \{f : \|f\|_{L_2} \leq ar\} \cap \mathcal{F}_{\boldsymbol{\beta}}^{\mathrm{2b}},$$

for the canonical family $\mathcal{C}_r = \{f_U : U \in \mathrm{Gr}(k,d)\}$ of Proposition 3; $\mathcal{M}$ is the $L_2(\mathbb{R}^d)$ packing number.

The proof is in Appendix C.4. This packing realizes the representation slope within an L_2-ball of radius ar, but does not establish a gap from fixed-frame entropy. In $L_2(\mathbb{R}^d, dx)$, translation invariance of (5) makes each $\mathcal{K}_Q$ fail to be totally bounded. The union-covering bound (Proposition 5, Appendix C) therefore has an infinite frame-wise entropy term here. We compare the two costs where both are finite in the separate Gaussian-design model of Section 5 (Theorem 3).

Beyond two blocks. For a general profile, each pair of unequal-smoothness blocks has its own contrast. The following tangent identity isolates the corresponding local weights. Decompose $\mathfrak{so}(d) = \mathfrak{h} \oplus \mathfrak{m}$ with $\mathfrak{h}$ the commutant of $\Lambda_{\boldsymbol{\beta}}$.

Lemma 1 (weighted tangent identity)**.** *For antisymmetric A let $S_t(A) = \frac{1}{2}\big(e^{-t\Lambda_{\boldsymbol{\beta}}} A e^{t\Lambda_{\boldsymbol{\beta}}} + (e^{-t\Lambda_{\boldsymbol{\beta}}} A e^{t\Lambda_{\boldsymbol{\beta}}})^\top\big)$. Then $[S_t(A)]_{ij} = \sinh\big(t(\lambda_j - \lambda_i)\big) A_{ij}$ entrywise, and the one-sided derivative of the metric at the identity along A equals $F_t(A) = \lambda_{\max}(S_t(A)) - \lambda_{\min}(S_t(A))$:*

$$\lim_{\epsilon \downarrow 0} \tfrac{1}{\epsilon} \log \kappa\big(e^{-t\Lambda_{\boldsymbol{\beta}}} e^{\epsilon A} e^{t\Lambda_{\boldsymbol{\beta}}}\big) = F_t(A).$$

Directions within equal-smoothness blocks are annihilated, while each of the $m_a m_b$ directions mixing blocks $a \neq b$ is amplified by $\sinh(t|\lambda_a - \lambda_b|)$. At $t = T_r$, these weights suggest a contribution of order $\log\sinh(T_r|\lambda_a - \lambda_b|) \approx T_r|\lambda_a - \lambda_b|$ per tangent coordinate to the representation entropy. Summing over block pairs motivates the following conjectured slope.

Conjecture 1 (general representation entropy)**.** *For fixed $d, \boldsymbol{\beta}$ and fixed threshold c, as $r \downarrow 0$,*

$$\mathsf{E}_{\boldsymbol{\beta}}(r; c) \;=\; \mathfrak{A}(\boldsymbol{\beta}) \log(1/r) + O_{d,\boldsymbol{\beta}}(1), \qquad \mathfrak{A}(\boldsymbol{\beta}) = \sum_{a<b} m_a m_b \, |\lambda_a - \lambda_b|.$$

Theorem 2 proves the conjecture for $L = 2$. For general profiles, passing from tangent weights to covering numbers requires uniform control of the exponential chart and of the local equivalence between $\rho_{\boldsymbol{\beta},r}$ and F_{T_r} beyond first order, followed by the volume-to-covering step. These steps remain unproved; Conjecture 1 is not used in any subsequent result.

5 High-dimensional active subspaces

Sections 3–4 identify which frame differences matter at a given approximation accuracy. We now ask what an unknown representation adds to the risk of estimating a function from noisy data. For this comparison, we use a separate Gaussian-design model in which both function and representation entropies are finite. The target depends on an unknown k-dimensional subspace. Knowing that subspace leaves a function of k variables to estimate; allowing the subspace to vary introduces representation uncertainty.

For $d > k$, let $\mathrm{St}(d,k) = \{W \in \mathbb{R}^{d\times k} : W^\top W = I_k\}$ be the Stiefel manifold and set $f_{W,g}(x) = g(W^\top x)$. The model class is

$$\mathcal{F}_{k,s,d} = \Big\{ f_{W,g} :\; W \in \mathrm{St}(d,k),\; \|g\|_{B^s_{2,q}(\mathbb{R}^k)} \leq 1 \Big\}, \qquad s > k/2,\; 1 \leq q \leq \infty. \tag{13}$$

The pair (W, g) is redundant: $(WO, g \circ O)$ defines the same function for $O \in O(k)$. Since the isotropic link ball is $O(k)$-invariant, the representation parameter descends to $\mathrm{St}(d,k)/O(k) =$

$\mathrm{Gr}(k,d)$. We index it by $U = \mathrm{span}(W)$, with $P_U = WW^\top$, and use aligned frames only for comparisons in the proofs (Appendix F). We observe $Y_i = f^\star(X_i) + \xi_i$, with $X_i \sim N(0, I_d)$ and independent noise $\xi_i \sim N(0,\sigma^2)$. All L_2 distances below use the Gaussian design measure γ_d. The assumption $s > k/2$ gives the uniform bound $\sup_{f\in\mathcal{F}_{k,s,d}} \|f\|_\infty \le C(k,s,q)$ and the Hölder control used below (Lemma 5). Constants in the entropy and risk bounds depend only on (k,s,q,σ), not on d; each theorem specifies its dependence.

The shared geometric object with Sections 3–4 is $\mathrm{Gr}(k,d)$, but the function classes are not nested. The present model is not obtained by setting $t=\infty$ in the earlier classes: those require $\lambda_{\min} > 0$, whereas nonzero ridge functions are constant along $U^\perp$ and do not belong to $L_2(\mathbb{R}^d, dx)$.

5.1 The Gaussian subspace witness

The local lower bound needs functions whose L_2 norm is small but whose shape still distinguishes nearby subspaces. We use Gaussian ridge bumps of width R^{-1}: for $R \ge 1$,

$$g_R(z) = c_w\, R^{k/2-s} e^{-R^2\|z\|^2/2}, \qquad v_{R,U}(x) = g_R(W^\top x) \quad (U = \mathrm{span}\, W), \tag{14}$$

which is frame-independent since g_R is radial.

Proposition 4 (witness). *There is $c_w = c_w(k,s,q) > 0$ such that for all $R \ge 1$ and all $U, V \in \mathrm{Gr}(k,d)$ with principal angles $\theta_1,\dots,\theta_k$:*

(a) $\sup_{R\ge1} \|g_R\|_{B^s_{2,q}(\mathbb{R}^k)} \le 1$, *and* $\|v_{R,U}\|^2_{L_2(\gamma_d)} = c_w^2 R^{k-2s}(1+2R^2)^{-k/2} \asymp_k R^{-2s}$;

(b) $\mathrm{Corr}(v_{R,U}, v_{R,V}) = \prod_{i=1}^k \Big[1 + \dfrac{R^4 \sin^2\theta_i}{1+2R^2}\Big]^{-1/2}$;

(c) if $\sin\theta_{\max}(U,V) \ge c/R$ *then* $\|v_{R,U} - v_{R,V}\|_{L_2(\gamma_d)} \ge \tau(c)\, \|v_{R,U}\|_{L_2(\gamma_d)}$ *with* $\tau(c) = \big(2 - 2(1+c^2/3)^{-1/2}\big)^{1/2} > 0$.

Appendix D proves the Besov membership by a dyadic calculation and derives the norm and correlation identities from Gaussian integrals. Setting $r = R^{-s}$ gives witnesses of L_2-size comparable to r, separated by a fixed fraction of that size when the subspace angle is of order $R^{-1} = r^{1/s}$. Thus shrinking the functions preserves the relative separation needed for a local packing.

5.2 Local entropy and minimax rate

The full entropy bound combines this directional complexity with the cost of estimating the link. The upper bound covers the entire class; the lower bound places both sources of complexity inside a ball about zero.

Theorem 3 (active-subspace local entropy; dimension-uniform). *Fix k, s, q with $s > k/2$, $1 \le q \le \infty$, and let $m = k(d-k)$. There exist $\epsilon_0, c_i, C_i > 0$ depending only on (k,s,q) such that for all $d > k$ and all $\epsilon \in (0,\epsilon_0]$, in $L_2(\gamma_d)$:*

$$\log N\big(\epsilon,\ \mathcal{F}_{k,s,d}\big) \ \le\ C_1\, \epsilon^{-k/s} + C_2\, m \log(C_3/\epsilon), \tag{15}$$

$$\log \mathcal{M}\big(\epsilon/c_4,\ B(0,\epsilon) \cap \mathcal{F}_{k,s,d}\big) \ \ge\ c_5\, \epsilon^{-k/s} + c_6\, m \log(1/\epsilon), \tag{16}$$

where N is the covering number, $\mathcal{M}$ the packing number and $B(0,\epsilon)$ the $L_2(\gamma_d)$-ball. Consequently the local metric entropy in the sense of Prasadan & Neykov (2025) satisfies $H_{\mathrm{loc}}(\epsilon) = \Theta_{k,s,q}\big(\epsilon^{-k/s} + m\log(1/\epsilon)\big)$ on $(0,\epsilon_0]$.

Upper bound: covering links and subspaces. Appendix E proves $\log N(\epsilon, \{\|g\|_{B^s_{2,q}} \le 1\}, L_2(\gamma_k)) \asymp \epsilon^{-k/s}$ by weighted localization on $\mathbb{R}^k$, with no logarithmic loss. For subspace approximation, Procrustes alignment gives $\|W_U - W_V\|_{\mathrm{op}} \le 2\delta$ whenever $\|P_U - P_V\|_{\mathrm{op}} \le \delta$. The Hölder embedding then bounds the function error by $O(\delta^\alpha)$ for fixed $0 < \alpha < \min\{1, s - k/2\}$ (Appendix F). The Gaussian moment bound involves only k projected coordinates, keeping the constants independent of d. A Grassmann net at radius $\delta \asymp \epsilon^{1/\alpha}$ and a link net for each of its centers give (15): their log-cardinalities add, with the Grassmann term supplied by Szarek (1998).

Lower bound: a local product packing. At $R \asymp \epsilon^{-1/s}$, choose a subspace packing at angular separation of order R^{-1} and a link bump code from Appendix E. For each pair of indices, form $f_{U,\omega} = a\, v_{R,U} + b\, w_{\omega,U}$, where $w_{\omega,U}(x) = w_\omega(W_U^\top x)$. Fixed coefficients $a, b > 0$ keep the family in $\mathcal{F}_{k,s,d}$ and, after a fixed rescaling, in $B(0,\epsilon)$. Within one subspace, the witness terms cancel and the link code supplies the separation. Across distinct subspaces, choosing b small enough relative to a makes the reverse-triangle inequality preserve the witness separation, regardless of the link indices. The family therefore realizes the product of the two code cardinalities in one local packing, giving (16) (Appendix G). In particular, the representation logarithm persists inside the small ball, not just in a global cover.

From local entropy to risk. The characterization of Prasadan & Neykov (2025, Theorem 3.15 in arXiv v5) converts these matching local bounds into a minimax rate. In the small-risk regime, the balance is $n\epsilon^2 \asymp \epsilon^{-k/s} + m\log(1/\epsilon)$; its two terms yield the function-estimation and representation contributions below.

Theorem 4 (active-subspace minimax rate; dimension-uniform)**.** *In the model above there are constants depending only on (k, s, q, σ) such that, for every n and every $d = d_n > k$,*

$$R_n^\star \;\asymp\; \Big[\, n^{-\frac{2s}{2s+k}} + \frac{m}{n}\, W\Big(\frac{n}{m}\Big)\Big] \wedge 1, \qquad m = k(d-k), \tag{17}$$

with W the Lambert function; in the regime $m = o(n)$ this reads

$$R_n^\star \;\asymp\; n^{-\frac{2s}{2s+k}} \;+\; \frac{k(d-k)}{n}\, \log\Big(1 + \frac{n}{k(d-k)}\Big). \tag{18}$$

Appendix H checks star-shapedness, completeness, bounded diameter, the moment condition and Gaussian noise; closedness is proved in Lemma 8, and the uniform sup-norm bound supplies the moment control. It handles saturation with a fixed-scale packing lower bound and the zero estimator. The resulting constants depend only on (k, s, q, σ), so the same bounds hold for every pair (n, d_n).

In the regime $m = o(n)$, the two terms cross when $m\log(1 + n/m) \asymp n^{k/(2s+k)}$. For fixed d, the representation term is a lower-order correction to the known-subspace rate; when d grows so that the left side exceeds the right side in order, representation identification dominates the risk.

6 Numerical audit

We compare the geometric formulas with independent numerical evaluations and examine the risk crossover in Gaussian-design regression. The study uses double precision and comprises 1.53×10^6 metric evaluations and 423 regression runs, with three seeds per regression configuration. Appendix J gives the protocol, estimators, and numerical safeguards.

Finite-resolution geometry. Independent metric computations closely track the closed form (9) (Figures 3 and 4). Figure 2a shows the contrast-dependent critical-angle scaling and its pre-asymptotic departure from a power law, as described by (10). For the Gaussian witness, the critical angle decays approximately as R^{-1} across the tested (d, k) values (Figure 2b). Table 3 reports additional packing results for $m = 1, 2, 3, 4$.

Risk crossover. The adaptive estimator learns a subspace and fits a Matérn kernel regressor within it; the oracle uses the true subspace (Appendix J). In the rough-link experiments, relative excess risk rises through one within an order of magnitude of $\eta = 1$ (Figure 2c). This reference compares the two rate terms in Theorem 4, leaving their constants unspecified. Figure 6 further shows subspace-error scaling near $(m/n)^{1/2}$ and excess risk approximately proportional to $\sin^2\widehat{\theta}$ over the tested range. These observations illustrate the subspace-estimation mechanism; they do not isolate the logarithmic factor in (18) or establish minimax optimality of the experimental estimator.

7 Related work and discussion

Prior work already studies unknown directions and their estimation cost. We compare how representation complexity depends on resolution and whether the statistical bounds match. Table 1 records the differences in models, losses, and guarantees.

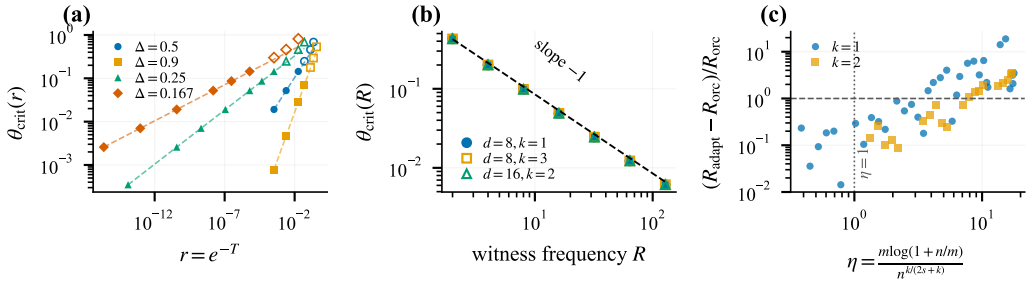

Figure 2: **Geometry and risk crossover.** (a) Critical angle at distortion $c = 1$ versus r for four contrasts. Dashed lines show the asymptotic slopes Δ from (10); open symbols mark $\zeta < 2$. (b) Critical subspace angle versus witness frequency R for the tested (d, k) values. The finite-range fitted slope is -1.0006 ± 0.0002 (estimate ± 1.96 regression standard errors), compared with the asymptotic exponent -1 (Proposition 4(c)). The interval describes fit precision, not finite-R departures from the asymptotic exponent. (c) Relative excess test risk $(R_{\mathrm{adapt}} - R_{\mathrm{orc}})/R_{\mathrm{orc}}$ versus $\eta = m\log(1 + n/m)/n^{k/(2s+k)}$ for rough-link (n, d) cell means. The dashed horizontal line marks relative excess risk 1; the dotted vertical line marks $\eta = 1$.

Anisotropic Besov learning. Suzuki (2019) and Suzuki & Nitanda (2021) established minimax-optimal deep-ReLU rates over (anisotropic) Besov classes, the latter for affine composition models whose learned first layer identifies smooth directions, with lower bounds for response-linear estimators (also Hayakawa & Suzuki, 2020). Architectures thus *adapt* to favorable affine representations; we isolate the representation itself, with intrinsic metric (7) and entropy (Theorem 2), dominant in the rate in identified regimes (Theorem 4).

Ridge and multi-index classes: the closest prior work. Table 1 positions our results against this line. Mayer et al. (2015) study ridge classes $g(a \cdot x)$ whose entropy numbers asymptotically match the univariate profile, ambient dimension affecting only how long one waits for that decay—a preasymptotic range parallel to our visibility threshold, which Corollary 1 quantifies exactly at all scales and for $k \geq 1$ (Table 1 records the differences in setting). The cost split, and its weight in high dimension, are not new. Baraud & Birgé (2014) bound composite-model risk by approximation, link and component terms, their multi-index corollary carrying an explicit ambient-dimension factor with constants free of it, and ask there whether its logarithm is necessary; Alquier & Biau (2013) give a sparse single-index oracle bound and note that the direction cost, negligible for fixed d, is not when d grows. Steffen (2026), Klock et al. (2021) and Serdyukova (2018) likewise treat unknown reduction matrices (also Hristache et al., 2001; Constantine et al., 2014; Lepski, 2015). What we add is a *matching* decomposition on a stated class: Theorem 3 is two-sided, so the logarithm cannot be removed once the representation term dominates, and Theorem 4 has constants uniform in d. The counting $m = k(d - k)$ is the right quotient dimension, not a sharper rate: it is comparable to kd when constants may depend on k. On the geometry side the new phenomenon is *resolution-dependent activation*: at coarse scales the same nontrivial Grassmannian carries zero representation entropy before the $k(d - k)\Delta \log(1/r)$ regime of Theorem 2 begins.

Quotient geometry, entropy, and minimax. Grassmann covering with universal constants: Szarek (1998); manifold geometry: Edelman et al. (1998). Our minimax theorem applies Prasadan & Neykov (2025), sharpening the global-entropy program (Birgé, 1983; Yang & Barron, 1999; Tsybakov & Zaiats, 2009; Wainwright, 2019).

Limitations. The entropy bounds (15)–(16) and risk rate (18) hold up to multiplicative constants. Computationally efficient minimax attainment remains open. The multi-block law remains Conjecture 1, and we do not establish a norm equivalence with coordinatewise Besov classes (Remark 1). The statistical analysis uses $s > k/2$ for boundedness and Hölder control; we prove no response-linear separation (Appendix I).

Perspective. Representation dimension alone does not determine the complexity relevant to these two settings. In the two-block model, coarse resolution can reduce the representation cover to one neighborhood. In the separate Gaussian model, the logarithmic representation cost can dominate the known-subspace estimation rate as d grows.

AI USE STATEMENT

The research question originated with the authors, motivated by their observation of coordinate-dependent smoothness and their reading of prior work on Besov learning (Suzuki, 2019; Suzuki & Nitanda, 2021). The theorem statements, Conjecture 1, and the proof ideas are the authors'; Claude Fable 5 and Claude Opus 5 (Anthropic), together with GPT-6 Astra and GPT-5.6 Sol (OpenAI), suggested refinements to some statements, assisted with derivations, and audited assumptions and intermediate steps, and the status of each claim (theorem versus conjecture) was decided by the authors after discussion with these tools. The synthetic data of Appendix J were sampled by AI-assisted code from distributions specified by the authors; the numerical audit was designed jointly, the code was written mainly with Claude assistance, and the LLMs helped organize the results, whose interpretation is the authors'. AI tools also supported manuscript revision, typesetting, partial translation, and reference retrieval and formatting; figures were produced by the authors, and dataset cleaning and qualitative analysis do not arise in this work. The authors verified the proofs, validated the code by unit tests and independent re-runs, and take responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

The model definitions and assumptions are given in Sections 2–5, and complete proofs of the theorems are provided in Appendices A–H. External mathematical results are cited where they are used. Appendix J documents the synthetic-data generation, estimators, parameter settings, random-seed protocol, and software environment for the numerical experiments. Code will be released upon publication.

REFERENCES

Pierre Alquier and Gérard Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(8):243–280, 2013. URL http://jmlr.org/papers/v14/alquier13a.html.

Yannick Baraud and Lucien Birgé. Estimating composite functions by model selection. *Annales de l'I.H.P. Probabilités et statistiques*, 50(1):285–314, 2014. doi: 10.1214/12-AIHP516. URL https://www.numdam.org/articles/10.1214/12-AIHP516/.

Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237, 1983.

Marcin Bownik. Atomic and molecular decompositions of anisotropic besov spaces. *Mathematische Zeitschrift*, 250(3):539–571, 2005.

Marcin Bownik. Erratum to: Atomic and molecular decompositions of anisotropic besov spaces. *Mathematische Zeitschrift*, 273(3):1297–1299, 2013.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138. URL https://doi.org/10.1137/130916138.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954. URL https://doi.org/10.1137/S0895479895290954.

David Eric Edmunds and Hans Triebel. *Function spaces, entropy numbers, differential operators*, volume 120. Cambridge University Press Cambridge, 1996.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021. doi: 10.1214/20-AOS1990. URL https://doi.org/10.1214/20-AOS1990.

Satoshi Hayakawa and Taiji Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.12.014. URL https://www.sciencedirect.com/science/article/pii/S089360801930406X.

Marian Hristache, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny. Structure Adaptive Approach for Dimension Reduction. *The Annals of Statistics*, 29(6):1537 – 1566, 2001. doi: 10.1214/aos/1015345954. URL https://doi.org/10.1214/aos/1015345954.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

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

Oleg Lepski. Adaptive estimation over anisotropic functional classes via oracle approach. *The Annals of Statistics*, 43(3):1178 – 1242, 2015. doi: 10.1214/14-AOS1306. URL https://doi.org/10.1214/14-AOS1306.

Sebastian Mayer, Tino Ullrich, and Jan Vybiral. Entropy and sampling numbers of classes of ridge functions. *Constructive Approximation*, 42(2):231–264, 2015.

Akshay Prasadan and Matey Neykov. Characterizing the minimax rate of nonparametric regression under bounded star-shaped constraints. *Electronic Journal of Statistics*, 19(2), January 2025. ISSN 1935-7524. doi: 10.1214/25-ejs2419. URL http://dx.doi.org/10.1214/25-EJS2419.

Nora Serdyukova. Lower bounds in estimation at a point under multi-index constraint. *Statistics & Probability Letters*, 138:183–189, 2018. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2018.02.062. URL https://www.sciencedirect.com/science/article/pii/S016771521830107X.

Maximilian F Steffen. Pac-bayes estimation for high-dimensional multi-index models with unknown active dimension. *Japanese Journal of Statistics and Data Science*, 9:285–314, 2026. doi: 10.1007/s42081-025-00309-x.

Taiji Suzuki. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3609–3621. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1dacb10f0623c67cb7dbb37587d8b38a-Paper.pdf.

Stanisław Szarek. Metric entropy of homogeneous spaces. *Banach Center Publications*, 43(1):395–410, 1998. URL http://eudml.org/doc/208860.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564 – 1599, 1999. doi: 10.1214/aos/1017939142. URL https://doi.org/10.1214/aos/1017939142.

APPENDIX ROADMAP

Appendix A: exact two-block metric. Appendix B: rotation transfer. Appendix C: representation entropy (two-block theorem, sandwich bound, tangent lemma); Appendix C.4 contains the proofs of Propositions 2 and 3 and Corollary 2. Appendix D: Gaussian subspace witness, including uniform Besov membership. Appendix E: link entropy under the Gaussian measure. Appendix F: aligned representation covering; embedding lemma. Appendix G: product packing; proof of Theorem 3. Appendix H: proof of Theorem 4. Appendix I: discussion of response-linear estimators. Appendix J: numerical audit. Table 1 positions the results against prior work; Table 2 summarizes selected statements, their status, and downstream use.

Table 1: **Positioning against prior work.** ✓ yes; – no; ∼ partial or implicit. Every symbol refers to the cited work's own setting, loss and constants, never to an inclusion between models. Columns 7–8 refer to the representation (frame or subspace), not to the link function; column 9 to each work's own main complexity bound. Except for the fixed-frame baseline, every prior row treats unknown directions; the difference lies in the form of the price, not its existence. Our two-block row marks "separates cost" as partial because the two-sided function/representation split is proved only in the Gaussian-design setting, and its blind regime is a two-block, finite-t result (Corollary 1) not proved in the Gaussian-design limit.

Work	Model / class	Unk. rep.?	Main theorem type	Main rate / complexity	d-unif.?	Sep. cost?	Blind reg.?	Sharp 2-sided?	Relative to ours
Classical aniso. theory (Bownik, 2005; Lepski, 2015)	fixed-frame aniso. classes (matrix Besov / Nikolskii)	–	function-space theory; entropy numbers; adaptive minimax	dense-zone rate via harmonic smoothness $\widetilde{\beta}$; regime-dependent in general	–	–	–	∼	within-frame function theory we build on; no representation to price
Suzuki & Nitanda (2021); Suzuki (2019)	aniso. Besov on $[0,1]^d$; affine comp. $h(Ax+b)$	✓	deep-ReLU upper bounds; matching minimax and response-linear lower bounds	$n^{-2\widetilde{\beta}/(2\widetilde{\beta}+1)}$ up to $\log^3 n$; hull-elevated linear exponent	–	–	–	✓	already learns unknown A; representation cost not isolated; fixed d
Structural adaptation (Hristache et al., 2001; Serdyukova, 2018)	$g(W^\top x)$, W unknown; fixed d	✓	index-space estimation (Hristache et al.); pointwise minimax lower bounds (Serdyukova)	$n^{-1/2}$ index-space rate for effective dim. ≤ 3; pointwise lower bounds with a log factor	–	∼	–	–	already estimates unknown W; no class-level entropy or two-sided minimax price
Composite and single-index upper bounds (Baraud & Birgé, 2014; Alquier & Biau, 2013)	composite $g(Ax)$ on a cube, ℓ_1-constrained directions (B–B); sparse single index (A–B)	✓	model-selection risk bound with an explicit component term (B–B); PAC-Bayes oracle bound (A–B)	approximation + link + component terms, the component one carrying an explicit ambient-dimension factor; direction cost negligible at fixed d, not when d grows	∼	✓	–	–	the explicit cost split is already known; ours matches local entropy and minimax risk in the Gaussian–Besov class, with a logarithm that cannot be dropped once that term dominates
Mayer et al. (2015)	fixed ridge class on the Euclidean ball; L_∞ metric	✓	entropy and sampling numbers for direction and profile	asymptotically univariate entropy after a d-dependent preasymptotic range; sampling can stay cursed	∼	∼	∼	✓	a preasymptotic plateau, not a zero-entropy phase; neither Gaussian regression nor a resolution-dependent quotient metric
Klock et al. (2021)	multi-index regression; estimated index space	✓	subspace estimation and downstream prediction bounds	$N^{-1/2}$ index estimation plus a subspace-perturbation term in the prediction error	–	∼	–	–	the propagation is explicit, but its prediction constants are not uniform over our Gaussian–Besov class

Table 1, continued.

Work	Model / class	Unk. rep.?	Main theorem type	Main rate / complexity	d-unif.?	Sep. cost?	Blind reg.?	Sharp 2-sided?	Relative to ours
Steffen (2026)	multi-index, unknown active dimension; sparsity-adaptive, dense matrices included	✓	PAC-Bayes oracle inequality	explicit link term plus a nonzero-entry penalty $\|I\| \log(p \vee n)/n$	$\sim$	✓	–	–	an explicit upper-bound split under bounded-design assumptions, adaptive in the active dimension; no matching lower bound for our class
Ours: two-block geometry (§§3–4)	$\mathcal{K}_Q$, two-block $\boldsymbol{\beta}$; frame Q unknown	✓	exact metric (Prop. 1); transfer + converse (Thm. 1, Props. 2–3); entropy (Thm. 2)	$\mathsf{E}_{\boldsymbol{\beta}}(r;c) = k(d-k)\zeta(r) + O(m)$; 0 for $\zeta \le c/2$	✓	$\sim$	✓	✓	—
Ours: active-subspace minimax (§5)	$g(W^\top x)$, $W \in \mathrm{St}(d,k)$, Besov g; Gaussian design	✓	two-sided local entropy (Thm. 3); minimax (Thm. 4)	$n^{-2s/(2s+k)} + \frac{m}{n}\log(1+\frac{n}{m})$	✓	✓	–	✓	—
Ours: multi-block (Conj. 1)	$L \ge 3$ blocks; frame unknown; fixed d	✓	conjectured slope; tangent identity proved (Lemma 1)	$\mathfrak{A}(\boldsymbol{\beta})\log(1/r) + O(1)$, $\mathfrak{A} = \sum_{a<b} m_a m_b \lvert\lambda_a - \lambda_b\rvert$	–	n/a	open	open	not used downstream

Statement	Setting	Type	Status	Used later?
Def. 1: representation metric $\rho_{\boldsymbol{\beta},r}$	any profile	definition	—	Thms. 1, 2; Cor. 1
Thm. 1: rotation transfer	class (5), fixed $\boldsymbol{\beta}$	theorem (one-way cover)	proved, App. B; d-free	Thm. 2 (covering direction); Prop. 5
Prop. 1: exact two-block metric	two-block	identity	proved, App. A	Cor. 1; Thm. 2; Prop. 3
Cor. 1: visibility threshold, blind regime	two-block	corollary	proved (from Prop. 1)	Thm. 2; Cor. 2; §6
Prop. 2: orientation witness (pairwise)	two-block, finite t	converse to Thm. 1	proved, App. C.4	no (independent converse)
Prop. 3: canonical spectral witness (per frame)	two-block, finite t	converse, exact correlation law	proved, App. C.4; membership needs the factor d in the exponent	Cor. 2
Thm. 2: two-block representation entropy	two-block, fixed threshold c	theorem, two-sided at every r	proved, App. C (with Szarek, 1998); d-free	Cor. 2; §6
Cor. 2: the canonical family realizes the representation entropy	two-block, $\zeta \geq C$	corollary (lower bound on a packing)	proved, App. C.4; not a separation from fixed-frame entropy	no (interpretive)
Lemma 1: weighted tangent identity	general flag	lemma	proved, App. C	Conj. 1 only (heuristic support)
Conj. 1: general representation entropy	general flag, fixed d	**conjecture**	**open** (proved for $L = 2$ by Thm. 2)	**not used anywhere**
Prop. 5: union entropy sandwich (upper half)	generic	auxiliary	proved, App. C; vacuous over $L_2(\mathbb{R}^d, dx)$, where both sides are infinite	interpretation of §4 only
Prop. 4: Gaussian witness	active subspace	auxiliary (identities, Besov membership)	proved, App. D	Thm. 3 (lower bound)
Prop. 6: link entropy in $L_2(\gamma_k)$	$B^s_{2,q}(\mathbb{R}^k)$, $s > k/2$	auxiliary, two-sided	proved, App. E (upper bound via Edmunds & Triebel, 1996)	Thm. 3
Lemmas 4, 5–7: bump code, embedding, Procrustes, cover	active subspace	auxiliary	proved, App. E–F; d-free	Thm. 3
Thm. 3: active-subspace local entropy	Gaussian design, $s > k/2$	theorem, two-sided	proved, App. F–G; d-uniform	Thm. 4
Lemma 8: class is closed	active subspace	auxiliary	proved, App. H	Thm. 4 (hypothesis of Prasadan & Neykov, 2025)
Thm. 4: active-subspace minimax rate	Gaussian design and noise	theorem, two-sided	proved, App. H (via Prasadan & Neykov, 2025); d-uniform	§6 (crossover)
Response-linear barrier (App. I)	affine composition, uniform design	discussion of a theorem of Suzuki & Nitanda (2021)	**no theorem of ours**; Gaussian-design analogue open	no
Coordinatewise bridge (Remark 1)	class (5) vs. coordinatewise spaces	—	**not claimed**	no

Table 2: **Claim-status roadmap.** Selected statements of the paper, with their settings, logical types, proof status, and downstream use. The two-block finite-resolution geometry and the active-subspace entropy and minimax theorems are proved; the general multi-block law is conjectural and nothing downstream depends on it.

A EXACT TWO-BLOCK METRIC (PROPOSITION 1)

Write the two-block generator as $\Lambda_{\boldsymbol{\beta}} = \mathrm{diag}(\lambda_r I_k, \lambda_s I_{d-k})$ with $\lambda_r - \lambda_s = \Delta > 0$. Let $Q_U, Q_V \in O(d)$ be orthogonal matrices whose first k columns span U, V. For $t \geq 0$, with $b = (\lambda_r + \lambda_s)/2$ we have $e^{-t\Lambda_{\boldsymbol{\beta}}} = e^{-tb} D_t$ and $e^{t\Lambda_{\boldsymbol{\beta}}} = e^{tb} D_t^{-1}$, so the two scalars cancel and

$$e^{-t\Lambda_{\boldsymbol{\beta}}}\, Q_U^\top Q_V\, e^{t\Lambda_{\boldsymbol{\beta}}} = D_t\, Q_U^\top Q_V\, D_t^{-1} =: M_t, \qquad D_t = \mathrm{diag}\left(e^{-t\Delta/2} I_k,\ e^{t\Delta/2} I_{d-k}\right).$$

Step 1 (gauge invariance). Replacing $Q_U \mapsto Q_U B$ with block-diagonal $B = \mathrm{diag}(B_1, B_2)$, $B_1 \in O(k)$, $B_2 \in O(d-k)$, replaces M_t by $B^\top M_t$ (as B commutes with D_t), leaving singular values unchanged; similarly for Q_V. Hence $\kappa(M_t)$ depends only on the subspaces, and the metric descends to $\mathrm{Gr}(k, d)$.

Step 2 (CS decomposition). There exist block-diagonal orthogonal B, B' such that $B^\top (Q_U^\top Q_V) B'$ consists of 2×2 rotation blocks $R(\theta_i) = \left(\begin{smallmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{smallmatrix}\right)$, one for each *nonzero* principal angle θ_i, each pairing one rough with one smooth coordinate, plus identity entries on the remaining $d - 2m_0$ coordinates, where $m_0 = \#\{i : \theta_i > 0\} \leq \min\{k, d-k\}$; only nonzero angles consume a smooth direction, so no constraint such as $d \geq 2k$ is needed. By Step 1 the singular values of M_t are the union, over i with $\theta_i > 0$, of the singular values of $m_t(\theta_i) = \mathrm{diag}(e^{-t\Delta/2}, e^{t\Delta/2})\, R(\theta_i)\, \mathrm{diag}(e^{t\Delta/2}, e^{-t\Delta/2})$, together with 1's.

Step 3 (2×2 block). $\det m_t(\theta) = 1$ and $\mathrm{tr}\ m_t(\theta) m_t(\theta)^\top = 2 + 4\sin^2\theta\, \sinh^2(t\Delta)$; hence the two singular values are $e^{\pm\rho(\theta,t)/2}$ with $\cosh\rho(\theta, t) = 1 + 2\sin^2\theta\, \sinh^2(t\Delta)$, i.e. $\rho(\theta, t) = 2\,\mathrm{arsinh}\left(\sinh(t\Delta)\sin\theta\right)$.

Step 4 (assembly). Since $\rho(\cdot, t)$ is increasing on $[0, \pi/2]$, $\kappa(M_t) = e^{\rho(\theta_{\max}, t)}$, and since $t \mapsto \rho(\theta_{\max}, t)$ is increasing, the supremum over $t \in [0, T_r]$ in (7) is attained at $t = T_r$, giving (9) with $\zeta = \Delta T_r$. □

B ROTATION TRANSFER (THEOREM 1)

Recall the objects of Section 2: $\tau = \log 2$, ellipsoids E_t^Q, cutoffs Φ_j^Q, blocks P_j^Q, and the class (5). Throughout fix Q, Q' with $\rho_{\boldsymbol{\beta},r}([Q], [Q']) \leq c_0$ and write $W = Q^\top Q'$.

Lemma 2 (all singular values are controlled)**.** *For every $t \in [0, T_r]$, all singular values of $M_t = e^{-t\Lambda_{Q'}} e^{t\Lambda_Q}$ lie in $[e^{-c_0}, e^{c_0}]$.*

Proof. $M_t = Q' e^{-t\Lambda_{\boldsymbol{\beta}}}(W^\top) e^{t\Lambda_{\boldsymbol{\beta}}} Q^\top \cdot$(orthogonal factors), so its singular values are those of $N_t = e^{-t\Lambda_{\boldsymbol{\beta}}} W^\top e^{t\Lambda_{\boldsymbol{\beta}}}$. By hypothesis $\kappa(N_t) \leq e^{c_0}$; moreover $\det N_t = \det W^\top = \pm 1$, so $\prod_i \sigma_i(N_t) = 1$. Then $\sigma_{\max}^d = \prod_i \left(\sigma_{\max}/\sigma_i\right) \leq \kappa^{d-1} \leq e^{c_0(d-1)}$, whence $\sigma_{\max} \leq e^{c_0(d-1)/d} \leq e^{c_0}$; and since the geometric mean of the σ_i is 1 we have $\sigma_{\max} \geq 1$, so $\sigma_{\min} = \sigma_{\max}/\kappa \geq e^{-c_0}$. □

Lemma 3 (ellipsoid transfer)**.** *Let $L = \lceil c_0/(\tau\lambda_{\min}) \rceil$ and $a_0 = \lceil 1/\lambda_{\min} \rceil$. For all $t \in [0, T_r]$ and $\mu \in \{1, 2\}$: $\mu E_t^Q \subseteq E_{t+(L+(\mu-1)a_0)\tau}^{Q'}$ whenever $t \leq T_r$, and symmetrically with $Q \leftrightarrow Q'$.*

Proof. Let $\|e^{-t\Lambda_Q}\xi\| \leq \mu$. Then

$$\|e^{-(t+L'\tau)\Lambda_{Q'}}\xi\| \leq \|e^{-L'\tau\Lambda_{Q'}}\|_{\mathrm{op}}\, \|M_t\|_{\mathrm{op}}\, \|e^{-t\Lambda_Q}\xi\| \leq e^{-L'\tau\lambda_{\min}}\, e^{c_0}\, \mu,$$

using Lemma 2. For $\mu = 1$ take $L' = L$; for $\mu = 2$ take $L' = L + a_0$, noting $e^{-a_0\tau\lambda_{\min}} \leq \frac{1}{2}$. □

Proof of Theorem 1. Set $L^* = L + a_0 + 1$ and $j_r = \lfloor T_r/\tau \rfloor - L^*$ (if $j_r \leq 0$ the claim is trivial with $C_1 = 2^{L^*+1}$, since $\|f\|_2 \leq \sum_j 2^{-j} \leq 2$ and $r \geq 2^{-L^*-1}$; assume $j_r \geq 1$).

Truncation. For $f \in \mathcal{K}_Q$, the definition (5) with $\ell^q \hookrightarrow \ell^\infty$ gives $\|P_j^Q f\|_2 \leq 2^{-j}$, hence $g := \Phi_{j_r}^Q f = \sum_{j \leq j_r} P_j^Q f$ satisfies $\|f - g\|_2 \leq \sum_{j > j_r} 2^{-j} = 2^{-j_r} \leq 2^{L^*+1} r$.

Band overlap. The symbol of P_j^Q ($j \geq 1$) is supported in $A_j^Q = 2E_{j\tau}^Q \setminus E_{(j-1)\tau}^Q$. Let $j \leq j_r$ and suppose $A_j^Q \cap A_i^{Q'} \neq \emptyset$, $i \geq 1$. From $\xi \in 2E_{j\tau}^Q \subseteq E_{(j+L+a_0)\tau}^{Q'}$ (Lemma 3; note $j\tau \leq T_r$) and $\xi \notin E_{(i-1)\tau}^{Q'}$ we get $i - 1 < j + L + a_0$, i.e. $i \leq j + L^*$. From $\xi \in 2E_{i\tau}^{Q'} \subseteq E_{(i+a_0)\tau}^{Q'}$ and $\xi \notin E_{(j-1)\tau}^Q \supseteq E_{(j-1-L)\tau}^{Q'}$ (Lemma 3 applied with the roles of Q, Q' exchanged, valid since $(j-1-L)\tau \leq T_r$) we get $i + a_0 > j - 1 - L$, i.e. $i \geq j - L^*$. (The case $i = 0$ is covered by the same outer bound.) Hence

$$P_i^{Q'} g = \sum_{j \leq j_r,\ |j-i| \leq L^*} P_i^{Q'} P_j^Q f,$$

and $P_i^{Q'} g = 0$ for $i > j_r + L^*$.

Norm transfer. Each $P_i^{Q'}$ has multiplier bounded by 1, so $\|P_i^{Q'} g\|_2 \leq \sum_{|j-i| \leq L^*} \|P_j^Q f\|_2$ and

$$2^i \|P_i^{Q'} g\|_2 \leq 2^{L^*} \sum_{|j-i| \leq L^*} 2^j \|P_j^Q f\|_2.$$

Taking ℓ^q norms in i and using the triangle inequality over the $2L^* + 1$ shifted copies, $\|g\|_{B_{Q'}} \leq (2L^* + 1)2^{L^*} =: C_0$. With $C_1 = 2^{L^*+1}$ the theorem follows. □

All constants depend only on $(c_0, \lambda_{\min})$; no step used the dimension d beyond Lemma 2, whose conclusion is dimension-free.

C Representation entropy

C.1 Proof of Theorem 2

By the covering relation (10) of Corollary 1 (valid at every ζ, unlike the threshold relation), $\rho_{\boldsymbol{\beta},r}(U, V) \leq c$ iff $\sin\theta_{\max}(U, V) \leq \epsilon_r(c)$. Hence covering $\mathcal{M}_{\boldsymbol{\beta}} = \mathrm{Gr}(k, d)$ by $\rho_{\boldsymbol{\beta},r}$-balls of radius c is covering $\mathrm{Gr}(k, d)$ by $\theta_{\max}$-balls of radius $\theta_r := \arcsin \epsilon_r(c)$:

$$\mathsf{E}_{\boldsymbol{\beta}}(r; c) = \log N\big(\mathrm{Gr}(k, d),\ \theta_r,\ \theta_{\max}\big).$$

If $\zeta \leq c/2$ then $\epsilon_r(c) = 1$, $\theta_r = \pi/2 = \mathrm{diam}$, and one ball covers: $\mathsf{E}_{\boldsymbol{\beta}} = 0$.

For the covering bounds, Szarek (1998, Theorem 8) applies to $M = \mathrm{Gr}(k, d) = SO(d)/S(O(k) \times O(d-k))$ with the intrinsic metric induced by the operator norm: by Remark (iii) there, $\kappa(M) = 1$, the metric equals the largest principal angle, $\mathrm{diam} = \pi/2$, and $\theta(M) = \pi$. The theorem gives, with universal constants c', C',

$$\big(c'\pi/\theta\big)^m \leq N(\mathrm{Gr}, \theta) \ \text{ for } \theta \leq \pi/4, \qquad N(\mathrm{Gr}, \theta) \leq \big(C'\pi/2\theta\big)^m \ \text{ for } \theta \leq \pi/2.$$

Since $\frac{2}{\pi}\theta \leq \sin\theta \leq \theta$ on $[0, \pi/2]$, $\log(1/\theta_r) = \log(1/\epsilon_r(c)) + O(1)$ with an absolute $O(1)$. The upper bound holds for all $r \in (0, 1)$, while the lower bound holds whenever $\theta_r \leq \pi/4$, equivalently $\epsilon_r(c) \leq \sin(\pi/4) = 1/\sqrt{2}$. For (12): when $\zeta \geq c$,

$$\log \frac{1}{\epsilon_r(c)} = \log\sinh\zeta - \log\sinh(c/2) = \zeta - \log\big(2\sinh(c/2)\big) + \log\big(1 - e^{-2\zeta}\big),$$

and the last term is bounded in absolute value by a constant depending only on c for $\zeta \geq c$; finally $m\Delta$ per unit $\log(1/r)$ equals $\sum_{i<j} |\lambda_i - \lambda_j| = k(d-k)\Delta$ for the two-block profile. □

C.2 A generic union bound

Proposition 5 (union entropy sandwich, upper half)**.** *Let $\{K_\omega\}_{\omega \in \Omega}$ be subsets of a normed space, let ρ be a metric on Ω, and suppose there are c_0, C_0, C_1 such that $\rho(\omega, \omega') \leq c_0$ implies $K_\omega \subseteq C_0 K_{\omega'} + C_1 r\, B_{L_2}$. Then for every $\epsilon > 0$,*

$$\log N\Big(C_1 r + \epsilon,\ \bigcup_\omega K_\omega\Big) \ \leq\ \log N(c_0, \Omega, \rho) \ +\ \sup_\omega \log N\big(\epsilon,\ C_0 K_\omega\big).$$

Proof. Pick a c_0-net $\{\omega_l\}$ of Ω and, for each l, an ϵ-net of $C_0 K_{\omega_l}$; the union of the latter is a $(C_1 r + \epsilon)$-net of the union class. □

With Theorem 1 providing the hypothesis and Theorem 2 counting the net of $\Omega = \mathcal{M}_{\boldsymbol{\beta}}$, the entropy of the unknown-frame class exceeds that of a single frame by at most the representation entropy. Over $L_2(\mathbb{R}^d, dx)$ both sides are infinite, since (5) is translation invariant, so the bound is a covering statement rather than a finite decomposition; the additive reading is established in the active-subspace setting, where both terms are finite and the lower bound is proved by explicit construction (Appendix G).

C.3 Proof of Lemma 1

Entrywise, $(e^{-t\Lambda_{\boldsymbol{\beta}}} A e^{t\Lambda_{\boldsymbol{\beta}}})_{ij} = e^{t(\lambda_j - \lambda_i)} A_{ij}$; for antisymmetric A the symmetrization has entries $\frac{1}{2}\big(e^{t(\lambda_j-\lambda_i)}A_{ij} + e^{t(\lambda_i-\lambda_j)}A_{ji}\big) = \sinh\big(t(\lambda_j-\lambda_i)\big)A_{ij}$. For the derivative: write $e^{-t\Lambda_{\boldsymbol{\beta}}} e^{\epsilon A} e^{t\Lambda_{\boldsymbol{\beta}}} = I + \epsilon C_t(A) + O(\epsilon^2)$ with $C_t(A) = e^{-t\Lambda_{\boldsymbol{\beta}}} A e^{t\Lambda_{\boldsymbol{\beta}}}$. First-order perturbation of singular values of $I + \epsilon M$ gives $\sigma_i = 1 + \epsilon\, \lambda_i\big(\frac{M+M^\top}{2}\big) + O(\epsilon^2)$; hence $\log \kappa = \epsilon\big(\lambda_{\max}(S_t(A)) - \lambda_{\min}(S_t(A))\big) + O(\epsilon^2)$. □

C.4 Orientation witness and the operational lower bound

Proof of Proposition 2. Write $\tau = \log 2$, $j_r = \lfloor T_r/\tau \rfloor$, $t_0 = j_r \tau$, and $a_0 = \lceil t \rceil = \lceil 1/\lambda_{\min} \rceil$, where t is the smooth-block smoothness. Constants L', a, ρ^* are fixed in the course of the proof.

Step 1 (metric at the rounded time). For a scale variable $u \geq 0$, Proposition 1 gives $\rho(u) := 2\,\mathrm{arsinh}(\sinh(u\Delta)\sin\theta_{\max})$ with $\rho'(u) \leq 2\Delta$. Hence $\rho(t_0) \geq \rho_{\boldsymbol{\beta},r}(U,V) - 2\Delta\tau \geq \rho^* - 2\Delta\tau$.

Step 2 (top singular direction). Let $M = e^{-t_0\Lambda_V} e^{t_0\Lambda_U}$. By Appendix A (Steps 2–3) the singular values of M are $\{e^{\pm\rho(\theta_i,t_0)/2}\} \cup \{1\}$, so $\sigma_{\max}(M) = e^{\rho(t_0)/2}$. Pick a unit η attaining $\sigma_{\max}$ and set $\xi_0 = e^{t_0\Lambda_U}\eta$. Then $\|e^{-t_0\Lambda_U}\xi_0\| = 1$, and for any integer $L' \geq 0$,

$$\|e^{-(t_0+L'\tau)\Lambda_V}\xi_0\| \;\geq\; e^{-L'\tau/s}\,\|e^{-t_0\Lambda_V}\xi_0\| \;=\; e^{-L'\tau/s}\,\sigma_{\max}(M) \;\geq\; \exp\Big(\tfrac{\rho^*}{2} - \Delta\tau - \tfrac{L'\tau}{s}\Big),$$

since the smallest eigenvalue of $e^{-L'\tau\Lambda_V}$ is $e^{-L'\tau/s}$. Choose $\rho^* = 2\big(\Delta\tau + L'\tau/s + \log 4\big)$, so the right side is ≥ 4.

Step 3 (spectral bump). By continuity there is a ball $B(\xi_0,\delta)$ on which simultaneously $\|e^{-(t_0-\tau)\Lambda_U}\xi\| > 1$ (at ξ_0 the value is $\geq e^{\tau\lambda_{\min}} > 1$), $\|e^{-t_0\Lambda_U}\xi\| < 2$, and $\|e^{-(t_0+L'\tau)\Lambda_V}\xi\| > 2$. Let $\widehat{f}$ be a nonzero smooth bump supported in $B(\xi_0,\delta) \cup B(-\xi_0,\delta)$, even and real, so f is a real Schwartz function. Its spectrum lies in $2E^U_{t_0} \setminus E^U_{t_0-\tau}$, so (as in Appendix B) at most $2a_0 + 2$ blocks $P^U_i f$ are nonzero, each of L_2 norm at most $\|f\|_2$, with $i \leq j_r + a_0$. Normalizing $\|f\|_2 = a\,r$ with $a := \big[(2a_0+2)\,2^{a_0+2}\big]^{-1}$ gives, using $2^{j_r} \leq 1/r$,

$$\|f\|_{B_U} \leq (2a_0+2)\,2^{j_r+a_0}\,\|f\|_2 \leq \tfrac{1}{2} \leq 1, \qquad \text{so } f \in \mathcal{K}_{Q_U}.$$

Step 4 (separation). Set $J = j_r + L'$ and let $g \in c_2\mathcal{K}_{Q_V}$. The multiplier Φ^V_J is supported in $\{\|e^{-J\tau\Lambda_V}\xi\| \leq 2\}$, which is disjoint from $\operatorname{supp}\widehat{f}$ by Step 3; hence $\langle f, \Phi^V_J g\rangle = 0$ by Plancherel, and

$$\|f-g\|_2 \;\geq\; \Big\langle \tfrac{f}{\|f\|_2},\, f-g \Big\rangle \;=\; \|f\|_2 - \Big\langle \tfrac{f}{\|f\|_2},\, (1-\Phi^V_J)g \Big\rangle \;\geq\; a\,r - \big\|(1-\Phi^V_J)g\big\|_2.$$

Finally $(1-\Phi^V_J)g = \sum_{i>J} P^V_i g$, so $\|(1-\Phi^V_J)g\|_2 \leq c_2 \sum_{i>J} 2^{-i} = c_2 2^{-J} \leq 2c_2 2^{-L'} r$, using $2^{-j_r} \leq 2r$. Choosing $L' = \lceil \log_2(4c_2/a) \rceil$ makes this at most $\frac{a}{2}r$. (The circular-looking order of choices is resolved by noting a depends only on (s,t), so L' and then ρ^* are well defined.) □ □

Proof of Proposition 3. Write $T = T_r$, $\zeta = T\Delta$, $J = T/\tau$, and $A_U = d\,e^{-2T\Lambda_U}$, so $\widehat{f_U}(\xi)^2 = c^2_{r,d} e^{-\xi^\top A_U \xi}$.

Step 1 (norms and correlation). By Plancherel, $\|f_U\|_2^2 = (2\pi)^{-d} c^2_{r,d} \int e^{-\xi^\top A_U \xi} d\xi = (2\pi)^{-d} c^2_{r,d} \pi^{d/2} \det(A_U)^{-1/2}$, and $\det A_U = d^d e^{-2T\,\mathrm{tr}\,\Lambda_{\boldsymbol{\beta}}}$ is frame-independent, so $c_{r,d}$

is well defined by $\|f_U\|_2 = ar$ and does not depend on U. Similarly $\langle f_U, f_V\rangle = (2\pi)^{-d} c_{r,d}^2 \pi^{d/2} \det\big(\frac{A_U+A_V}{2}\big)^{-1/2}$, whence

$$\mathrm{Corr}(f_U, f_V) = \frac{\det(A_U)^{1/4}\det(A_V)^{1/4}}{\det\big(\frac{A_U+A_V}{2}\big)^{1/2}}.$$

Write $\Lambda_U = \frac{1}{t} I + \Delta P_U$, so $A_U = d\, e^{-2T/t}(I - \mu P_U)$ with $\mu = 1 - e^{-2\zeta}$; the scalar factor, d included, cancels in the ratio, leaving $\mathrm{Corr}^2 = 2^d(1-\mu)^k \big/ \det\big(2I - \mu(P_U + P_V)\big)$. The spectrum of $P_U + P_V$ consists of the eigenvalue 2 on $U \cap V$ (dimension $k - m_0$, $m_0 = \#\{\theta_i > 0\}$), the eigenvalue 0 on $U^\perp \cap V^\perp$, and the pair $1 \pm \cos\theta_i$ on a two-dimensional plane for each nonzero principal angle. Hence

$$\det\big(2I - \mu(P_U + P_V)\big) = (2-2\mu)^{\,k-m_0}\, 2^{\,d-k-m_0} \prod_{\theta_i > 0} \big[(2-\mu)^2 - \mu^2 \cos^2\theta_i\big],$$

and per nonzero angle $\frac{4(1-\mu)}{(2-\mu)^2 - \mu^2\cos^2\theta} = \big[1 + \frac{\mu^2}{4(1-\mu)}\sin^2\theta\big]^{-1}$ with $\frac{\mu^2}{4(1-\mu)} = \frac{(1-e^{-2\zeta})^2 e^{2\zeta}}{4} = \sinh^2\zeta$. This gives (11); the sech bound follows by keeping only the $\theta_{\max}$ factor and using $\sinh(\zeta)\sin\theta_{\max} = \sinh(\rho/2)$ (Proposition 1) together with $1 + \sinh^2 = \cosh^2$.

Step 2 (Besov membership, uniform in d). Write the normalized energy $|\widehat{f_U}|^2 / \|\widehat{f_U}\|_2^2$ as the law of $\xi = e^{T\Lambda_U} Z$ with $Z \sim N(0, I_d/(2d))$, so that $\mathbb{E}\|Z\|^2 = \frac{1}{2}$ for every d. Fix an integer $p > t$. Since $2d\|Z\|^2 \sim \chi^2_d$,

$$\mathbb{E}\|Z\|^{2p} = \prod_{\ell=0}^{p-1}\Big(\tfrac{1}{2} + \tfrac{\ell}{d}\Big) \;\le\; C_p := \prod_{\ell=0}^{p-1}\Big(\tfrac{1}{2} + \ell\Big),$$

a constant depending on p alone. For $j \le J+1$ use $\|P_j^U f_U\|_2 \le \|f_U\|_2$, so, since $2^J = 1/r$, $\sum_{j \le J+1} 2^j \|P_j^U f_U\|_2 \le 2^{J+2}\|f_U\|_2 = 4a$. For $j > J+1$ the symbol of P_j^U vanishes where $\|e^{-(j-1)\tau\Lambda_U}\xi\| \le 1$, and $e^{-(j-1)\tau\Lambda_U}\xi = e^{-((j-1)\tau - T)\Lambda_U} Z$ with $(j-1)\tau \ge T$, so on its support $\|Z\| \ge e^{((j-1)\tau - T)\lambda_{\min}} = 2^{\lambda_{\min}(j-1-J)}$. Plancherel and Markov's inequality then give

$$\|P_j^U f_U\|_2 \;\le\; \|f_U\|_2 \sqrt{C_p}\, 2^{-p\lambda_{\min}(j-1-J)}.$$

Because $p\lambda_{\min} = p/t > 1$, $\sum_{j > J+1} 2^j 2^{-p\lambda_{\min}(j-1-J)} \le C(t)\, 2^J$, so the tail contributes at most $C(t)\, a$. Neither bound involves d or r; as $\|\cdot\|_{\ell^q} \le \|\cdot\|_{\ell^1}$, choosing $a = a(t)$ small enough gives $\|f_U\|_{B_U} \le 1$. □

Proof of Corollary 2. Let $\mathcal{P} \subset \mathrm{Gr}(k,d)$ be a maximal family with pairwise $\rho_{\boldsymbol{\beta},r} \ge \rho^*$. By Corollary 1 this is a $\theta_{\max}$-packing at angle $\arcsin a_r(\rho^*)$, so by the two-sided Grassmann bounds (Szarek, 1998), for $\zeta(r) \ge C(\rho^*)$, $\log|\mathcal{P}| \ge m\zeta(r) - O(m) = \mathfrak{A}(\boldsymbol{\beta})\log(1/r) - O(m)$. Assign to each $U \in \mathcal{P}$ its *canonical* witness f_U of Proposition 3—a function of U alone. All f_U lie in $\mathcal{F}^{2\mathrm{b}}_{\boldsymbol{\beta}}$ with $\|f_U\|_2 = ar$, and for every pair $U \ne V$ in $\mathcal{P}$ the correlation bound gives $\|f_U - f_V\|_2 \ge c_3 r$ with $c_3 = a\sqrt{2(1 - \mathrm{sech}(\rho^*/2))}$. □ □

D WITNESS PROPOSITION (PROPOSITION 4)

Norm and correlation. With $Z = W^\top X \sim N(0, I_k)$,

$$\|v_{R,U}\|^2_{L_2(\gamma_d)} = c_w^2 R^{k-2s}\, \mathbb{E}\, e^{-R^2\|Z\|^2} = c_w^2 R^{k-2s}(1+2R^2)^{-k/2},$$

which lies in $[c_w^2 2^{-k} R^{-2s},\, c_w^2 R^{-2s}]$ for $R \ge 1$. For the correlation, with $P = P_U + P_V$, $\mathbb{E}[v_{R,U} v_{R,V}] = c_w^2 R^{k-2s}\det(I_d + R^2 P)^{-1/2}$; the nonzero eigenvalues of P are $1 \pm \cos\theta_i$ on the principal planes, so $\det(I + R^2 P) = \prod_i \big[(1+R^2)^2 - R^4\cos^2\theta_i\big] = \prod_i \big[1 + 2R^2 + R^4\sin^2\theta_i\big]$, giving (b). These are frame-independent computations: only $P_U = WW^\top$ enters.

Part (c). If $\sin\theta_{\max} \ge c/R$ then the $\theta_{\max}$ factor in (b) is at most $\big[1 + \frac{R^2c^2}{1+2R^2}\big]^{-1/2} \le (1 + c^2/3)^{-1/2}$ for $R \ge 1$, and all factors are ≤ 1; hence $\mathrm{Corr} \le (1 + c^2/3)^{-1/2}$ and $\|v_{R,U} - v_{R,V}\|^2 \ge 2\big(1 - (1 + c^2/3)^{-1/2}\big)\|v_{R,U}\|^2$ (the two witnesses have equal norms).

Part (a): uniform Besov membership. With the convention $\widehat{f}(\xi) = \int f(z)e^{-i\langle z,\xi\rangle}dz$,

$$\widehat{g_R}(\xi) = c_w(2\pi)^{k/2}R^{-k/2-s}e^{-\|\xi\|^2/(2R^2)}.$$

Let $\{P_j\}_{j\geq 0}$ be an isotropic Littlewood–Paley decomposition of $\mathbb{R}^k$ (annuli $\{2^{j-1} \leq \|\xi\| \leq 2^{j+1}\}$ for $j \geq 1$). By Plancherel,

$$\|P_j g_R\|_2^2 \leq C_k R^{-k-2s} \int_{2^{j-1}\leq\|\xi\|\leq 2^{j+1}} e^{-\|\xi\|^2/R^2} d\xi \leq C'_k\, R^{-k-2s}\, 2^{jk}\, e^{-4^{\,j-1}/R^2}.$$

Hence $2^{js}\|P_j g_R\|_2 \leq C\, u_j^{s+k/2} e^{-u_j^2/8}$ with $u_j = 2^j/R$, up to the factor $R^{-s-k/2} \cdot R^{s+k/2} = 1$. The sequence $\big(u^{s+k/2}e^{-u^2/8}\big)_{u=2^j/R,\, j\geq 0}$ is, for every $R \geq 1$, dominated by the two-sided geometric envelope $\min\{u_j,\, 1/u_j\}^{\,\delta} \cdot C(s,k)$ for some $\delta > 0$ near the peak $u \asymp 1$, so its ℓ^q norm is bounded by a constant $C(k,s,q)$ *uniformly in* R, for every $q \in [1,\infty]$. The $j = 0$ block obeys $\|P_0 g_R\|_2 \leq \|g_R\|_2 \leq CR^{-s} \leq C$. Choosing $c_w = 1/C(k,s,q)$ gives $\sup_{R\geq 1}\|g_R\|_{B^s_{2,q}} \leq 1$. □

E Link entropy under the Gaussian measure

Proposition 6. *Let $\mathcal{G} = \{g : \|g\|_{B^s_{2,q}(\mathbb{R}^k)} \leq 1\}$ with $s > k/2$, $1 \leq q \leq \infty$. There are constants c, C and $\epsilon_0 > 0$ depending only on (k,s,q) such that for $\epsilon \in (0,\epsilon_0]$,*

$$c\,\epsilon^{-k/s} \;\leq\; \log N\big(\epsilon,\ \mathcal{G},\ L_2(\gamma_k)\big) \;\leq\; C\,\epsilon^{-k/s}.$$

Two classical external inputs are used; (E1) only in the *upper* bound and (E2) on both sides, since Lemma 4 estimates the Besov norm of bump combinations through it. (E1) the entropy of Besov balls on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^k$ in $L_2(\Omega)$ satisfies $\log N(\delta) \asymp \delta^{-k/s}$ (Edmunds & Triebel, 1996, Chapter 3); (E2) the equivalence of the Littlewood–Paley and modulus-of-smoothness characterizations of $B^s_{2,q}(\mathbb{R}^k)$, and boundedness of pointwise multiplication by a fixed C^∞_c function; see Edmunds & Triebel (1996, Section 2.8) and references therein.

Lemma 4 (single-scale bump code)**.** *Fix $s > 0$, $1 \leq q \leq \infty$ and an integer $m > s$. There are constants $R_0, c_1, C_2 > 0$ depending only on (k,s,q,m) such that for every $R \geq R_0$ there is a family $\{w_\omega : \omega \in \Omega\}$ of functions supported in the unit cube of $\mathbb{R}^k$, with $\log|\Omega| \geq N/8$ for $N = \lfloor (R/R_0)^k \rfloor$, such that* (i) $\|w_\omega\|_{B^s_{2,q}} \leq \frac{1}{2}$; (ii) $\|w_\omega\|_{L_2(\gamma_k)} \leq C_2 R^{-s}$; (iii) $\|w_\omega - w_{\omega'}\|_{L_2(\gamma_k)} \geq c_1 R^{-s}$ *for $\omega \neq \omega'$.*

Proof. Put $L = \lceil R/R_0 \rceil$, so $R/R_0 \leq L \leq 2R/R_0$ for $R \geq R_0$, and partition the unit cube into the $L^k \geq N$ boxes of side $1/L$ of the regular grid; keep N of them and place in box i the L_2-normalized bump $\psi_i(z) = L^{k/2}\prod_a \psi\big(L(z_a - z_{i,a})\big)$, $\psi \in C^\infty_c((0,1))$, $\|\psi\|_2 = 1$. For $\omega \in \{0,1\}^N$ set $f_\omega = \sum_i \omega_i \psi_i$.

Besov norm of combinations. For $|h| \leq t$ the m-th difference $\Delta_h^m \psi_i$ is supported in an $O(mt)$-enlargement of box i, so for $t \leq R_0/R$ the supports of $\{\Delta_h^m \psi_i\}_i$ have overlap multiplicity at most $C(m,k)$, whence $\omega_m(f_\omega,t)_2^2 \leq C\sum_i \omega_i\, \omega_m(\psi_i,t)_2^2 \leq CN\,\omega_m(\psi_{\mathrm{sc}},t)_2^2$, where ψ_{sc} is a single scaled atom; for $t > R_0/R$ use $\omega_m(f_\omega,t)_2 \leq 2^m\|f_\omega\|_2 \leq 2^m\sqrt{N}$. The scaling identity $\omega_m(\psi_{\mathrm{sc}},t)_2 \leq C\min\{1,(Rt/R_0)^m\}$ and the modulus-of-smoothness characterization (E2) give

$$\|f_\omega\|_{B^s_{2,q}} \leq C_b(k,s,q,m)\,\sqrt{N}\,(R/R_0)^s.$$

Set $w_\omega = c_b f_\omega$ with $c_b = \frac{1}{2}\,(R/R_0)^{-s}/(C_b\sqrt{N})$: this gives (i).

Separation and size. By Varshamov–Gilbert there is $\Omega \subset \{0,1\}^N$ with $\log|\Omega| \geq N/8$ and pairwise Hamming distance $\geq N/4$. The ψ_i have disjoint supports inside the unit cube, where the γ_k-density lies in $[c_k, C_k]$; hence for $\omega \neq \omega'$, $\|w_\omega - w_{\omega'}\|_{L_2(\gamma_k)} \geq \sqrt{c_k}\,c_b\sqrt{N/4} \geq c_1 R^{-s}$, and $\|w_\omega\|_{L_2(\gamma_k)} \leq \sqrt{C_k}\,c_b\sqrt{N} \leq C_2 R^{-s}$. □

Proof of Proposition 6. Lower bound. Given $\epsilon \leq \epsilon_0$, apply Lemma 4 with R chosen so that $c_1 R^{-s} = 2\epsilon$: the family $\{w_\omega\}$ lies in the ball $\mathcal{G}$, is 2ϵ-separated in $L_2(\gamma_k)$, and has log-cardinality $\geq N/8 \gtrsim \epsilon^{-k/s}$. No bounded-domain entropy theorem is needed on this side.

Upper bound. Fix a smooth partition of unity $\{\chi_\nu\}_{\nu\in\mathbb{Z}^k}$, $\chi_\nu = \chi(\cdot - \nu)$, $\operatorname{supp}\chi \subset [-1,1]^k$, $\sum_\nu \chi_\nu = 1$, with overlap multiplicity 2^k. For $g \in \mathcal{G}$ write $g = \sum_\nu \chi_\nu g$; by (E2), $\|\chi_\nu g\|_{B^s_{2,q}} \le C_\chi$ for all ν. Also, by the embedding $\mathcal{G} \subset \{\|g\|_\infty \le C_\infty\}$ (Lemma 5), $\|\chi_\nu g\|_{L_2(Q_\nu)} \le 2^{k/2} C_\infty =: D$.

On the cube Q_ν (side 2, center ν) the Gaussian density is bounded above by $C_k e^{-\|\nu\|^2/4}$. Allocate to cube ν the Lebesgue-approximation budget $\delta_\nu = \epsilon\, e^{\|\nu\|^2/16}$; approximate $\chi_\nu g$ within δ_ν in $L_2(Q_\nu)$ using an (E1)-net of the ball $C_\chi \mathcal{G}|_{Q_\nu}$ when $\delta_\nu < D$, and by the zero function otherwise (possible since the local diameter is $\le D$; this happens once $\|\nu\|^2 \ge 16\log(D/\epsilon)$, so only finitely many cubes carry nontrivial nets). Summing the weighted errors over cubes, using overlap 2^k,

$$\Big\| \sum_\nu \mathrm{err}_\nu \Big\|^2_{L_2(\gamma_k)} \le 2^k \sum_\nu C_k e^{-\|\nu\|^2/4}\, \delta_\nu^2 = 2^k C_k\, \epsilon^2 \sum_\nu e^{-\|\nu\|^2/8} = C' \epsilon^2.$$

The total log-cardinality is

$$\sum_{\nu:\ \delta_\nu < D} \log N_{(E1)}(\delta_\nu) \le C \sum_\nu \big(\epsilon\, e^{\|\nu\|^2/16}\big)^{-k/s} = C\, \epsilon^{-k/s} \sum_\nu e^{-\|\nu\|^2 k/(16s)} = C''\, \epsilon^{-k/s},$$

with no logarithmic factor. Rescaling ϵ by $\sqrt{C'}$ finishes. □

F Aligned representation covering

Lemma 5 (embedding)**.** *For $s > k/2$, $1 \le q \le \infty$ and $0 < \alpha < \min\{1, s - k/2\}$ there is $C = C(k,s,q,\alpha)$ with $\|g\|_\infty + \sup_{z\neq z'} \frac{|g(z)-g(z')|}{\|z-z'\|^\alpha} \le C\|g\|_{B^s_{2,q}(\mathbb{R}^k)}$.*

Proof. With isotropic LP blocks, Bernstein's inequality gives $\|P_j g\|_\infty \le C 2^{jk/2}\|P_j g\|_2 \le C 2^{j(k/2-s)}\|g\|_B$, summable since $s > k/2$; and $|P_j g(z) - P_j g(z')| \le C \min\{2^j\|z - z'\|, 1\}\, 2^{j(k/2)}\|P_j g\|_2$, so summing over j splits at $2^j \approx \|z - z'\|^{-1}$ and yields the Hölder bound for any $\alpha < s - k/2$, $\alpha \le 1$. □

Lemma 6 (Procrustes alignment)**.** *If $\|P_U - P_V\|_{\mathrm{op}} = \sin\theta_{\max} \le \delta$, there exist frames W_U, W_V of U, V with $\|W_U - W_V\|_{\mathrm{op}} \le 2\delta$.*

Proof. Choose principal-angle frames, so that $W_U^\top W_V = \operatorname{diag}(\cos\theta_i)$. Then $(W_U - W_V)^\top (W_U - W_V) = 2I - 2\operatorname{diag}(\cos\theta_i)$, whose eigenvalues are $4\sin^2(\theta_i/2)$, so $\|W_U - W_V\|_{\mathrm{op}} = 2\sin(\theta_{\max}/2) \le \sqrt{2}\,\sin\theta_{\max} \le 2\delta$, the middle step because $\theta_{\max} \le \pi/2$. This uses no complement of U and so holds for every $d > k$, including $d < 2k$, where $U^\perp$ is too small to carry k orthonormal columns. □

Lemma 7 (representation cover)**.** *Fix $\alpha < \min\{1, s - k/2\}$. For any U, V with $\|P_U - P_V\|_{\mathrm{op}} \le \delta$ and any g with $\|g\|_{B^s_{2,q}} \le 1$, the aligned frames of Lemma 6 satisfy*

$$\|g(W_U^\top \cdot) - g(W_V^\top \cdot)\|_{L_2(\gamma_d)} \ \le\ C(k,s,q,\alpha)\, \delta^\alpha.$$

Proof. By Lemma 5 and Jensen ($\alpha \le 1$),

$$\mathbb{E}\big|g(W_U^\top X) - g(W_V^\top X)\big|^2 \le C^2\, \mathbb{E}\|(W_U - W_V)^\top X\|^{2\alpha} \le C^2 \big(\|W_U - W_V\|_F^2\big)^\alpha \le C^2 \big(4k\delta^2\big)^\alpha,$$

using $\mathbb{E}\|A^\top X\|^2 = \|A\|_F^2$ for $X \sim N(0, I_d)$ and $\|\cdot\|_F \le \sqrt{k}\|\cdot\|_{\mathrm{op}}$ on $d \times k$ matrices. No constant depends on d. □

G Product packing and proof of Theorem 3

Upper bound (15). Let $\delta = (\epsilon/2C)^{1/\alpha}$ with C from Lemma 7, and take a δ-net $\{U_l\}$ of $\mathrm{Gr}(k,d)$ in $\|P_U - P_V\|_{\mathrm{op}}$ of log-cardinality $\le m\log(C'/\delta)$ (Szarek, 1998) (operator-norm and largest-principal-angle metrics are equivalent up to absolute factors). Fix aligned frames per net point. For each l, take an $(\epsilon/2)$-net of the link ball in $L_2(\gamma_k)$ of log-cardinality $\le C\epsilon^{-k/s}$ (Proposition 6);

note $\|g(W^\top\cdot)\|_{L_2(\gamma_d)} = \|g\|_{L_2(\gamma_k)}$ exactly. By Lemma 7 every $f_{W,g}$ is within $\epsilon/2 + \epsilon/2$ of some $g'(W_{U_l}^\top\cdot)$: this proves (15) with $C_2 = 1/\alpha$-inflated constants.

Lower bound (16). Fix $\epsilon \le \epsilon_0$, put $R = (c_\star/\epsilon)^{1/s}$ with $c_\star = c_\star(k,s,q)$ chosen below, and let $r_\star = c_\star R^{-s} = \epsilon$.

(1) Subspace family. Let $\{U_j\}_{j\le M_{\mathrm{Gr}}}$ be a packing of $\mathrm{Gr}(k,d)$ at largest principal angle $\ge \theta_\star = \arcsin(c/R)$ (with the absolute c of Proposition 4(c)), of log-cardinality $\log M_{\mathrm{Gr}} \ge m\log(c''R)$ (Szarek, 1998), valid for $\theta_\star \le \pi/4$.

(2) Link code. Apply Lemma 4 at frequency R: a family $\{w_\omega\}_{\omega\in\Omega}$ of links supported in the unit cube, $\log|\Omega| \ge N/8$ with $N = \lfloor (R/R_0)^k \rfloor$, satisfying $\|w_\omega\|_{B^s_{2,q}} \le \frac{1}{2}$, $\|w_\omega\|_{L_2(\gamma_k)} \le C_2 R^{-s}$, and pairwise $\|w_\omega - w_{\omega'}\|_{L_2(\gamma_k)} \ge c_1 R^{-s}$.

(3) Product family. For a frame W_j of each U_j set

$$f_{j,\omega} \;=\; a\, v_{R,U_j} \;+\; b\, w_\omega(W_j^\top\cdot), \qquad a = \tfrac{1}{2}, \quad b = \min\Big\{\tfrac{1}{2},\; \frac{a\,\tau(c)\,c_w 2^{-k/2}}{4C_2}\Big\},$$

with $\tau(c)$ from Proposition 4(c). Each $f_{j,\omega}$ lies in $\mathcal{F}_{k,s,d}$: its link is $a\,g_R + b\,w_\omega$, of Besov norm $\le a\cdot 1 + b\cdot\frac{1}{2} \le 1$ and has $\|f_{j,\omega}\|_{L_2(\gamma_d)} \le (a+b)C_3R^{-s} \le C_4\epsilon$ for suitable $c_\star$. Separations: for $j = j'$, $\omega \ne \omega'$, the witness terms cancel and $\|f_{j,\omega} - f_{j,\omega'}\| \ge b\,c_1 R^{-s}$; for $j \ne j'$, by the reverse triangle inequality and Proposition 4(c),

$$\|f_{j,\omega} - f_{j',\omega'}\| \;\ge\; a\,\|v_{R,U_j} - v_{R,U_{j'}}\| - 2bC_2R^{-s} \;\ge\; \big(a\,\tau(c)\,c_w 2^{-k/2} - 2bC_2\big)R^{-s} \;\ge\; \tfrac{a\tau c_w}{2^{k/2+1}}R^{-s}.$$

Hence the family is $c_5\epsilon$-separated, of log-cardinality $\log M_{\mathrm{Gr}} + \log|\Omega| \ge c_6\, m\log(1/\epsilon) + c_7\,\epsilon^{-k/s}$, and contained in $B(0, C_4\epsilon)$. Rescaling ϵ by fixed factors gives (16). □

H MINIMAX RATE (THEOREM 4)

We apply the minimax characterization of Prasadan & Neykov (2025) (Theorem 3.15 in arXiv:2401.07968v5, whose statement we quote below): for a class $\mathcal{F} \subseteq L_2(\mathbb{P}_X)$ that is star-shaped and complete, has $L_2(\mathbb{P}_X)$-diameter $D < \infty$, satisfies their moment condition, and for Gaussian noise of fixed level σ, the minimax risk is $\asymp \epsilon^{*2} \wedge D^2$ with $\epsilon^* = \sup\{\epsilon : n\epsilon^2 \le \log \mathcal{M}^{\mathrm{loc}}(\epsilon, c)\}$, where $\mathcal{M}^{\mathrm{loc}}(\epsilon, c) = \sup_f \mathcal{M}(\epsilon/c, B(f,\epsilon)\cap\mathcal{F})$ and c is a sufficiently large absolute constant; the surrounding constants depend only on (c, L, σ).

Hypotheses. (i) *Star-shaped at* 0: $tf_{W,g} = f_{W,tg}$ and $\|tg\| \le 1$. (ii) *Completeness*: $\mathcal{F}_{k,s,d}$ is closed in $L_2(\gamma_d)$ (Lemma 8 below), hence complete. (iii) *Diameter*: $D \le 2C_\infty(k,s,q)$ by Lemma 5. (iv) *Moment condition*: their condition is implied by the uniform sup-norm bound C_∞, with $L = L(C_\infty)$. (v) *Noise*: $N(0,\sigma^2)$, σ fixed. All resulting constants depend on (k,s,q,σ) only; since the cited theorem is non-asymptotic with absolute constants, it applies verbatim to each pair (n, d_n).

Lemma 8 (the class is closed)**.** *$\mathcal{F}_{k,s,d}$ is closed in $L_2(\gamma_d)$.*

Proof. Let $f_n = g_n(W_n^\top\cdot) \to f$ in $L_2(\gamma_d)$ with $W_n \in \mathrm{St}(d,k)$, $\|g_n\|_{B^s_{2,q}} \le 1$. By compactness of $\mathrm{St}(d,k)$ pass to a subsequence with $W_n \to W$ in operator norm. By Lemma 5 and the computation in the proof of Lemma 7 (which uses only $\mathbb{E}\|A^\top X\|^2 = \|A\|_F^2$ and needs no alignment, the frames here being convergent),

$$\|g_n(W_n^\top\cdot) - g_n(W^\top\cdot)\|_{L_2(\gamma_d)} \le C\big(\sqrt{k}\,\|W_n - W\|_{\mathrm{op}}\big)^\alpha \longrightarrow 0.$$

Hence $g_n(W^\top\cdot) \to f$ in $L_2(\gamma_d)$; since $\|g(W^\top\cdot)\|_{L_2(\gamma_d)} = \|g\|_{L_2(\gamma_k)}$ for every g, the sequence (g_n) is Cauchy in $L_2(\gamma_k)$, with limit g satisfying $f = g(W^\top\cdot)$ a.e.

It remains to check $\|g\|_{B^s_{2,q}} \le 1$. First, $\|g_n\|_{L_2(\mathrm{Leb})} \le \sum_j \|P_j g_n\|_2 \le \sum_j 2^{-js} \le C_s$, so (g_n) is bounded in $L_2(\mathrm{Leb})$; and since the γ_k-density is bounded below on compacts, $g_n \to g$ in $L_{2,\mathrm{loc}}$, so every weak $L_2(\mathrm{Leb})$ subsequential limit equals g: $g_n \rightharpoonup g$ in $L_2(\mathrm{Leb})$. Each Littlewood–Paley block P_j is a bounded self-adjoint Fourier multiplier, so $P_j g_n \rightharpoonup P_j g$, and weak lower semicontinuity of the norm gives $\|P_j g\|_2 \le \liminf_n \|P_j g_n\|_2$. For $q < \infty$, Fatou's lemma in j yields $\sum_j (2^{js}\|P_j g\|_2)^q \le \liminf_n \sum_j (2^{js}\|P_j g_n\|_2)^q \le 1$; for $q = \infty$ the bound holds termwise. Hence $f \in \mathcal{F}_{k,s,d}$. □

Local entropy. Fix the local-packing ratio once and for all as $c_{\mathrm{pack}} \geq \max\{c_{\mathrm{PN}}, c_4\}$, so that both displays below are used at the same ϵ and no rescaling of the radius is needed. Writing $H_{\mathrm{loc}}(\epsilon) = \log \mathcal{M}^{\mathrm{loc}}(\epsilon, c_{\mathrm{pack}})$ and $p = k/s$, for $\epsilon \leq \epsilon_0$,

$$a_{\mathrm{ent}}\big(\epsilon^{-p} + m\log(1/\epsilon)\big) \;\leq\; H_{\mathrm{loc}}(\epsilon) \;\leq\; U_{\mathrm{ent}}\big(\epsilon^{-p} + m\log(1/\epsilon)\big), \tag{19}$$

the upper bound from (15) via $H_{\mathrm{loc}}(\epsilon) \leq \log N(\epsilon/(2c_{\mathrm{pack}}), \mathcal{F})$ after absorbing the fixed $\log(C/\epsilon)$ factor, the lower bound from (16) via $H_{\mathrm{loc}}(\epsilon) \geq \log \mathcal{M}(\epsilon/c_{\mathrm{pack}}, B(0,\epsilon) \cap \mathcal{F})$. Both $a_{\mathrm{ent}}, U_{\mathrm{ent}}$ depend only on (k, s, q).

Saturation. For $\epsilon > \epsilon_0$ we do not extend (19); the two regimes are separated by (20) instead. If $m < \rho_0 n$ then $H_{\mathrm{loc}}(\epsilon_0) \leq U_{\mathrm{ent}} m[\epsilon_0^{-p} + \log(1/\epsilon_0)] < \frac{1}{2} n\epsilon_0^2$, and since the local-packing function is non-increasing in ϵ (Prasadan & Neykov, 2025, Lemma 3.12) no fixed point can lie at $\epsilon \geq \epsilon_0$: the rate is decided by the small-scale bounds already proved. If instead $m \geq \rho_0 n$, the rate saturates, and we bound it below directly. Fix $\epsilon_b \leq \epsilon_0$ and take the family of (16) at scale ϵ_b: its members have norm at most ϵ_b, pairwise separation at least ϵ_b/c_4, and $\log \mathcal{M} \geq c_6\, m \log(1/\epsilon_b)$. Against the zero function, each Kullback–Leibler divergence is $n\|f\|^2_{L_2(\gamma_d)}/(2\sigma^2) \leq n\epsilon_b^2/(2\sigma^2)$, so if $m \geq \rho_0 n$ for the fixed threshold

$$\rho_0 = \min\left\{\tfrac{1}{2},\; \frac{\epsilon_0^2}{2U_{\mathrm{ent}}\left[\epsilon_0^{-p} + \log(1/\epsilon_0)\right]}\right\} > 0, \tag{20}$$

choosing $\epsilon_b = \epsilon_b(k, s, q, \sigma)$ small enough that $\epsilon_b^2/(2\sigma^2) \leq \frac{1}{4} c_6 \rho_0 \log(1/\epsilon_b)$ and that $\log \mathcal{M} \geq 4 \log 2$ makes Fano's inequality give $R_n^\star \geq \epsilon_b^2/(8c_4^2)$, a fixed positive constant; the zero estimator supplies the matching upper bound, and the target rate is itself bounded below there, $\frac{m}{n} W(n/m) \geq \rho_0 W(1/\rho_0) > 0$. The two branches cover every n and every $d > k$.

Fixed point. Both bounds are, up to constants, the same function $\phi(\epsilon) = \epsilon^{-k/s} + m\log(1/\epsilon)$, which is decreasing; hence ϵ^{*2} is, up to constants, the solution of $n\epsilon^2 = \phi(\epsilon)$, and the solution of a sum equation is comparable to the larger of the two individual solutions: $n\epsilon^2 = \epsilon^{-k/s}$ gives $\epsilon^2 = n^{-2s/(2s+k)}$, while $n\epsilon^2 = m\log(1/\epsilon)$ is solved exactly by $\epsilon^2 = \frac{m}{2n} W\big(\frac{2n}{m}\big)$: writing $x = 2n/m$, $\log(1/\epsilon) = \frac{1}{2}\log\big(x/W(x)\big) = \frac{1}{2}W(x)$ because $x/W(x) = e^{W(x)}$. Since $W(2x) \asymp W(x)$ for $x \geq 1$ this is $\asymp \frac{m}{n} W\big(\frac{n}{m}\big)$, and $W(x) = \log x - \log\log x + O(1)$ for $x \geq e$ yields the readable form $\frac{m}{n}\log(1 + \frac{n}{m})$ when $m \leq n$; for $m \gtrsim n$ the fixed point exceeds constants and the rate saturates: $R_n^\star \asymp 1 \wedge D^2$. Combining, (17) and (18) follow. We note that the achieving procedure of Prasadan & Neykov (2025) is an explicit packing-tree algorithm: constructive, though not polynomial-time. Theorem 4 therefore asserts attainability by an explicit estimator; the computational estimator of Appendix J.2 is a separate heuristic used only to probe the predicted mechanism. □

I Response-linear estimators

Estimators linear in the responses (Jacot et al., 2018; Chizat et al., 2019; Ghorbani et al., 2021) face an ambient-dimension barrier under unknown directions—a theorem of Suzuki & Nitanda (2021) for affine composition under uniform design. Under Gaussian design, Figure 5 shows higher test risk for the plotted ambient-space baselines than for the adaptive estimator and known-subspace oracle over the tested sample-size range. We do not establish a corresponding response-linear separation theorem in this Gaussian-design model.

J Numerical audit

J.1 Protocol

The numerical study evaluates the finite-resolution geometric predictions and examines the risk crossover in Gaussian-design regression. The six diagnostic checks address the metric identity, critical-angle scaling, witness correlations, subspace-resolution scaling, packing growth with m, and non-cancellation in the product-packing construction. We cross-check closed-form quantities through independent computational routes: SVD, closed-form, and projector-norm evaluations for the metric; determinant formulas, principal-angle products, and Monte Carlo estimates for witness correlations; and numerical root finding against analytic solutions for critical angles.

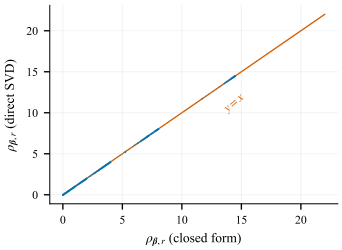

Figure 3: Direct SVD evaluation of (7) against the closed form (9) (1.53×10^6 configurations); all points lie on $y = x$.

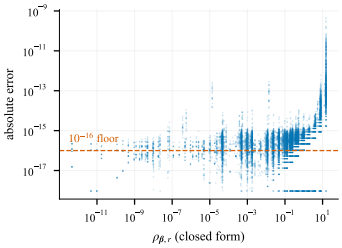

Figure 4: Absolute difference between closed-form and direct SVD metric evaluations, versus the closed-form value. The dashed line marks 10^{-16} for reference. Discrepancies increase at larger metric values; Appendix J discusses numerical limitations of direct SVD.

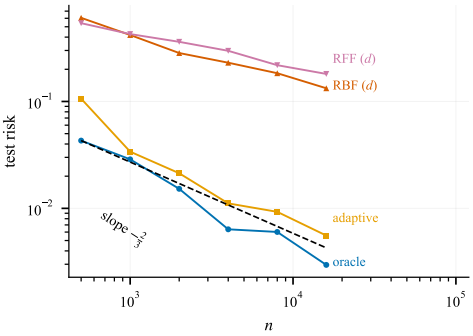

Figure 5: Test risk vs. n ($k = 1$, rough link, $d = 8$): the oracle follows the $n^{-2/3}$ rate, the adaptive estimator converges to it from above, ambient kernel baselines stagnate.

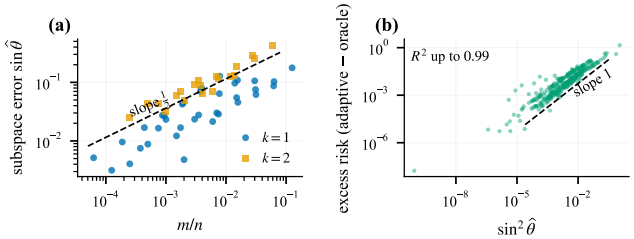

Figure 6: **Statistical audit** (423 runs, three seeds), illustrating the m/n-scale subspace-error mechanism. (a) Subspace error of the estimator of Appendix J.2 vs. m/n, dashed slope $1/2$. (b) Excess risk over the known-U oracle vs. $\sin^2\hat{\theta}$, dashed slope 1; fitted slopes in Table 3.

Computations use Python 3.13, NumPy 2.1, SciPy 1.15, and PyTorch 2.7, with float64 arithmetic throughout, on a system equipped with eight RTX A6000 GPUs. Regression experiments use three random seeds per configuration. The study comprises 1.53×10^6 metric evaluations and 423 regression runs, with a total computational cost of approximately 11 GPU-hours.

J.2 The adaptive estimator and experimental protocol

Scope disclaimer. The computational estimator below is used only to probe the mechanism predicted by Theorem 4 (subspace error $\asymp (m/n)^{1/2}$, excess risk $\asymp \sin^2\hat{\theta}$). Theorem 4 does *not* claim that

this estimator attains the minimax rate; the rate-attaining procedure is the explicit (non-polynomial-time) scheme of Prasadan & Neykov (2025) (Appendix H).

Data protocol. For each configuration $(k, s, \text{link}, n, d)$ and seed: $X_i \sim N(0, I_d)$, $Y_i = g^\star(W^{\star\top} X_i) + \xi_i$, $\sigma = 0.5$; links frozen across runs and normalized to unit variance; train/validation split $80/20$; risk evaluated against the noiseless target on $50{,}000$ fresh points. k and s are known to all estimators; three seeds per configuration.

Oracle. Knows $W^\star$; fits kernel ridge regression on $W^{\star\top} X$ with a Matérn kernel whose smoothness index $\nu \in \{1/2, 3/2, 5/2\}$ is chosen so that $\nu + k/2$ is closest to s; lengthscale grid $0.3\sqrt{k} \cdot \{1, 2, 4, 8\}$ and ridge grid $\{10^{-8}, \ldots, 10^{-1}\}$, both selected on the validation split.

Adaptive estimator. (i) *Candidates*: top-k eigenvectors of the SIR and SAVE matrices (10 response slices) and of their sum, plus two Haar-random Stiefel points. (ii) *Profile objective*: for $A \in \mathbb{R}^{d\times k}$, $J(A)$ is the validation mean-squared error of the oracle-type Matérn KRR fitted on $(\mathrm{QR}(A)^\top X, Y)$ over an inner fit/validation split ($n_{\text{fit}} \le 3000$, $n_{\text{val}} \le 2000$), with (ℓ, λ) selected once at the SIR+SAVE candidate and then frozen. (iii) *Optimization*: Adam on A (learning rate 0.02, 250 steps), retracting $A \leftarrow \mathrm{QR}(A)$ after each step and differentiating through the Cholesky-based KRR solve; the best-validation iterate along each trajectory is retained, and the best candidate wins. (iv) *Refit*: with $\widehat{W}$ fixed, the k-dimensional Matérn KRR is refitted on the full training set with fresh grid selection. Reported quantities: $\sin\widehat{\theta} = \|W^\star W^{\star\top} - \widehat{W}\widehat{W}^\top\|_{\text{op}}$ and the test risk.

Baselines. Ambient RBF-KRR, ambient Matérn-$1/2$ KRR, and random Fourier features (4096 features) with ridge; identical grid/validation protocol, no access to $W^\star$.

J.3 Numerical pitfalls

Four failure modes that would silently corrupt a naive replication: (1) direct SVD evaluation of (7) loses $\sigma_{\min}$ to underflow once $T_r\Delta \gtrsim 40$ and returns a wrong finite answer with no warning (observed 136.7 vs. true 200.0); a log-space power iteration exploiting the reciprocal singular-value pairing of Appendix A is exact to $T_r\Delta = 2000$. (2) Gauss–Hermite weight computation overflows to NaN beyond $\sim$200 nodes in standard implementations. (3) Measuring packing separation on a subsample of a combinatorial code can overestimate the full code's minimum pairwise distance. If this bias varies with scale, it can distort the fitted exponent. We include an explicit design-distance pair as a check. (4) Witness Monte Carlo relative variance grows like R^k; fixed sample sizes are vacuous at large R (we scale to 4×10^8 samples).

J.4 Slope table

Quantity	Prediction	Measured	Regime
Metric evaluation discrepancy	0	magnitude-dependent (Fig. 4)	1.53M evaluations
Critical-angle exponent (Δ=0.5)	0.5	0.5054 ± 0.0024	$\zeta \geq 2$
Critical-angle exponent (Δ=0.9)	0.9	0.9009 ± 0.0005	$\zeta \geq 2$
Critical-angle exponent (Δ=0.25)	0.25	0.2507 ± 0.0003	$\zeta \geq 2$
Critical-angle exponent (Δ=0.167)	0.167	0.1671 ± 0.0002	$\zeta \geq 2$
Critical-angle exponent (Δ=0.024)	0.024	0.02387 ± 0.00003	$r \lesssim e^{-84}$
Witness norm exponent (s=1, 1.5, 2)	$-s$	within 0.02 of $-s$	$R \in [2, 64]$
Subspace resolution exponent	-1	-1.0006 ± 0.0002	$R \geq 16$
Packing slope m=1, 2, 3, 4	m	1.00, 1.98, 2.98, 3.87	pool-stable
Subspace error exponent	$1/2$	0.48–0.59	all links
Excess-risk vs. $\sin^2 \widehat{\theta}$	1	0.82–1.03	R^2 up to 0.99
Oracle rate exponent (k=1, s=1)	$-2/3$	-0.674 ± 0.25 / -0.682 ± 0.24	d=4 / 16

Table 3: Theoretical references and finite-range numerical measurements. Predicted exponents are asymptotic, whereas the reported slopes are fitted over the listed finite ranges. Reported $\pm$ values are 1.96 times the standard errors of the log–log fits; they describe fit precision, not finite-scale departures from the asymptotic exponents. Regression experiments use three seeds per configuration. The table does not identify the logarithmic factor in (18), a response-linear asymptotic exponent, or packing slopes for $m \geq 7$.