ONE ATTENTION HEAD MAKES WHERE TO LOOK ALMOST FREE: THE COST OF UNKNOWN SUPPORT WITH AND WITHOUT ADAPTIVE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A transformer must decide where to look before it decides what to compute. We measure the statistical price of that decision in nonparametric regression. The target depends on an unknown set of k of its D input coordinates, one token each, through a smooth function of k variables, namely an anisotropic Besov ball with a sup-norm envelope; the design is i.i.d. with a density bounded above and below uniformly in D and n, the noise is Gaussian, and k is fixed. Without adaptive selection the price is combinatorial. Every estimator that is affine in the training responses, with offset and coefficients fixed before the responses are read, as in kernel ridge regression at a label-independent regularisation level, random-feature ridge, or lazy/NTK rules, has worst-case risk at least $c\min\{1,(M/n)^{\beta}\}$ with $M=\binom{D}{k}$, which does not vanish once $M \geq n$; for integrability index $1 \leq p < 2$ the exponent β is the smaller, shifted one. One positional attention head makes the price additive and polylogarithmic in D. The exact clipped least-squares minimiser over an explicitly budgeted class of one-head transformers with positional keys attains $n^{-2\bar{s}/(2\bar{s}+1)}$ in the effective smoothness $\bar{s}$, up to logarithmic factors in n and with a constant free of D, plus a single additive selection term polylogarithmic in D. The mechanism is a whitening of the positional codes: for every support, the head can attend equally to the k relevant tokens, with only a controlled remainder elsewhere, inside one parameter budget that does not depend on which tokens they are. At $D = n$ with all structural parameters fixed, the response-affine worst-case risk stays bounded away from zero while the attention bound vanishes. The saving is not specific to attention, since a direct search over supports attains the same leading bound; what attention supplies is a compact, continuous realisation of it. The upper bound concerns the exact minimiser, and no optimisation guarantee is claimed.

1 INTRODUCTION

A transformer decides where to look before it decides what to compute: its attention weights pick out the input tokens that matter, and everything downstream acts on what they selected. When a target depends on only k of D input tokens and nobody tells the learner which k, that first decision is itself a statistical problem with $\binom{D}{k}$ candidate answers. This paper asks three questions about it in nonparametric regression. How much does not knowing where to look cost a learner that cannot adapt its selection to the labels? How much does it cost one positional attention head? And what mechanism inside the head accounts for the difference?

The setting is the following. Inputs $x \in [0,1]^D$ are observed with each coordinate treated as one token. The target has the form $f(x) = g(x_S)$ for an unknown set S of k coordinates, where g is a smooth function of k variables: it lies in an anisotropic Besov ball of fixed radius with a fixed sup-norm envelope, with smoothness s_a along the a-th active direction and effective smoothness $\bar{s} = (\sum_{a \leq k} 1/s_a)^{-1}$. The n design points are i.i.d. with a density bounded above and below uniformly in D and n, the noise is Gaussian, k is fixed, and the risk is $R(\hat{f}, f) = \mathbb{E}\|\hat{f} - f\|^2_{L^2(P_X)}$. Section 3 states the class precisely; here it suffices that $p \in [1, \infty]$ is the integrability index of the Besov norm, that $\bar{s} > 1/p$, and that both results concern the same class $\mathcal{F}$ and the same data.

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.

Two theorems answer the first two questions. The first concerns learners that cannot adapt their selection to the labels. We call an estimator *response-affine* if it has the form $a + \sum_i w_i\, y_i$ with an offset and coefficients that may depend on the design and on label-independent randomisation but not on the responses; kernel ridge regression at a label-independent regularisation level, random-feature ridge, and lazy or NTK rules $f_0 + A^\top y$ with label-independent f_0, A are all of this form. Theorem 5.2 shows that for every such rule the number of candidate supports $M = \binom{D}{k}$ enters the worst-case risk multiplicatively. The second concerns one positional attention head. Theorem 5.1 shows that the exact least-squares fit over an explicitly budgeted class of one-head transformers with positional keys, each followed by a small ReLU network, pays for the unknown support additively and polylogarithmically:

$$\inf_{\hat f \in \mathcal{L}_{\mathrm{aff}}} \sup_{f \in \mathcal{F}} R(\hat f, f) \;\geq\; c\, \min\Big\{1, \big(M/n\big)^{\beta}\Big\},$$

$$\sup_{f \in \mathcal{F}} R(\hat f_n, f) \;\lesssim\; n^{-\alpha} \log^{3\alpha}(2n) \;+\; \frac{\log^2(2D)\,\big[\log^2(2n) + \log\log(e^e D)\big]}{n}.$$

Here $\alpha = 2\bar s/(2\bar s + 1)$ is the exponent of the k-dimensional nonparametric term; $\beta = \alpha$ for $p \geq 2$, while for $1 \leq p < 2$ the exponent is the smaller $\beta = 2s_*/(2s_* + 1) < \alpha$ with $s_* = \bar s - (1/p - 1/2)$; and the constants depend on the structural parameters $(k, s, p, q, F, F_\infty, \sigma, c_\pm)$ but not on D, n or S. Read together, the two displays say what one attention head buys. Without adaptive selection the ambient dimension enters through the combinatorial factor M, and once $M \geq n$ the bound saturates at a constant; with one head it enters only through the second term, additively and polylogarithmically, never in the exponent or the constant of the first. This is the sense in which where to look becomes almost free. It is not free: the selection term still depends on D, and when D grows quickly with n it can be the larger of the two. Its exact form, with the selector count P_{sel} and the envelope $\bar\Xi_{n,D}$, is in Theorem 5.1; Figure 1 draws both sides.

The mechanism behind the second display answers the third question. Each token i carries a fixed positional code $c_i \in \mathbb{R}^m$ with entries $\pm 1/\sqrt{m}$ and length $m = O(k^2 \log D)$; these codes are the keys. The trainable part of the head is a query, two $m \times m$ matrices and a $k \times m$ decoder, $P_{\mathrm{sel}} = O(k^4 \log^2 D)$ parameters in all, confined to one box. For any support S, whitening the k active codes gives a query that makes their logits exactly equal while every other logit is at most a third as large; the softmax then places equal weight on the k relevant tokens, up to a controlled off-support mass, and the decoder turns that equal split back into x_S. Because the codes are nearly orthogonal, the query and decoder that do this for one support have the same size as for any other, so a single box valid for every S contains a good selector for each of the $\binom{D}{k}$ supports (Lemma E.9); the least-squares fit is never told S, and the lemma only certifies that the class it searches contains what it needs. The lower bound has a matching intuition: a rule whose weights are fixed before the responses are read must decide in advance how much to trust each of the M candidate blocks, and the prior we construct activates one block per draw yet spreads its covariance evenly over all M of them (Lemmas F.7 and F.5).

Taking $D = n$ with $n \geq 2k$ and all structural parameters fixed makes the contrast concrete: $M = \binom{n}{k} \geq n$ keeps the response-affine bound at a positive constant, while $P_{\mathrm{sel}} = O(\log^2 n)$ and $\bar\Xi_{n,n} = O(\log^2 n)$ give an upper bound $O(n^{-\alpha} \log^{3\alpha} n + \log^4 n/n)$, which tends to zero.

The paper sits where two lines of work meet. Sparse token selection by attention has been studied at the level of representation, sample complexity and gradient-descent convergence on noiseless population losses (Edelman et al., 2022; Sanford et al., 2023; Wang et al., 2024); the exponents α and β are the anisotropic Besov rates of Suzuki (2019); Suzuki & Nitanda (2021), and an additive selection cost for unknown support is known for the all-estimator minimax risk on sparse Hölder classes (Yang & Tokdar, 2015). What is new is the pairing on one class: (i) an upper bound for the exact least-squares fit of a one-head positional-attention transformer with an explicit D-independent architecture budget and an explicit selector count, resting on the whitening construction; (ii) an explicit factor M in the worst-case risk of every response-affine rule on the coordinate-support union, which neither a single-support lower bound nor a lower bound for a larger affine-composition class yields; (iii) the certified separation between the two, including the $D = n$ reading. Section 2 gives the theorem-level comparison.

Two limits belong up front; Section 6 collects the rest. The saving is not specific to attention: exact least squares over the union of clipped feed-forward networks on all supports attains the

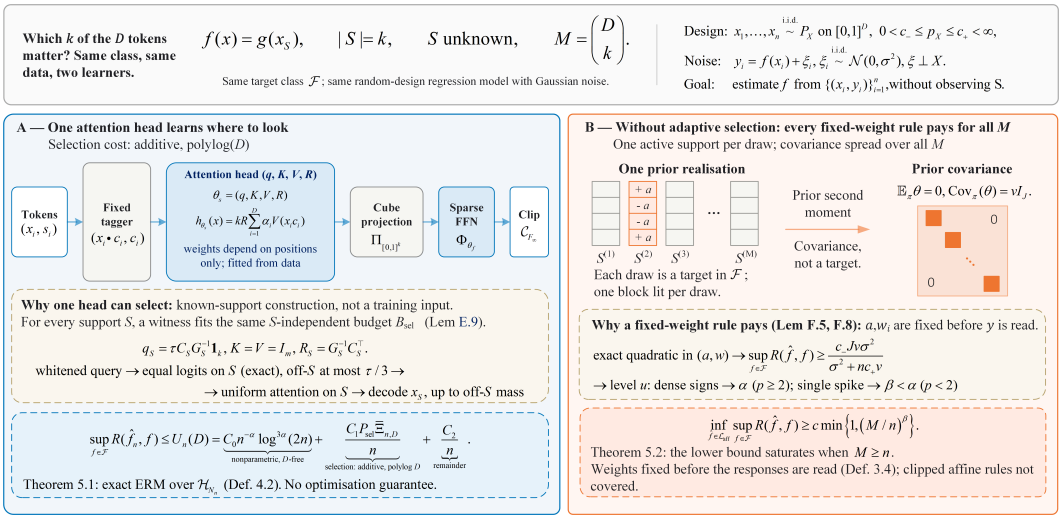


Figure 1: Where to look: the cost of unknown support with and without one attention head. Top: the task; both panels concern the same class and the same data. **A** (Theorem 5.1): the map fitted by the exact ERM of Definition 4.2; selector and FFN trainable, attention weights positional. The lower box is the whitening mechanism of Lemma E.9, a known-support construction and not a training input. **B** (Theorem 5.2, for $\mathcal{L}_{\text{aff}}$ only): one prior draw activates a single support block while the prior covariance is spread over all M blocks; the lower box gives the construction behind the bound (Lemmas F.5 and F.8). Theoretical schematic, not measured data.

same leading bound with $\log\binom{D}{k}/n$ in place of the selection term. What attention supplies is a continuous realisation of that search, one box of P_{sel} parameters in place of $\binom{D}{k}$ discrete candidates at a polylogarithmic price; nothing here says that attention is necessary or optimal, and neither estimator comes with an optimisation guarantee, both bounds concerning exact minimisers. Conversely, the lower bound covers response-affine rules as defined; label-adaptive tuning, early stopping or model selection, and the pointwise clipping of an affine rule are not automatically covered.

2 Related Work

Attention as token selection. That a single attention head is a natural device for functions of a few tokens is known at the level of representation and sample complexity. Edelman et al. (2022) show that a bounded-norm self-attention head represents a sparse function of the input sequence with sample complexity growing only logarithmically in the context length, and Sanford et al. (2023) exhibit a sparse averaging task on which transformers need complexity logarithmic in the input size while feedforward and recurrent networks need polynomial complexity. The closest architectural precedent is the sparse token selection task of Wang et al. (2024), in which the indices to be selected are supplied with each input; under their positional-encoding and initialisation assumptions, their Theorems 3.1 and 3.3 prove gradient-descent convergence for a one-layer transformer on a noiseless population loss, and their Theorem 3.5 gives a width lower bound for fully connected networks on the same task. Both papers select tokens by a query against fixed positional keys. The questions differ: here the support is a hidden global parameter rather than an input, the responses carry Gaussian noise, the upper bound is a worst-case risk bound for the exact least-squares fit over a nonparametric class, and the lower bound restricts response-affinity rather than width or architecture. Positional attention enters as a compact, analysable realisation with explicitly counted parameters (Section 4), not through an optimisation guarantee or a necessity result for content-dependent routing.

Unknown support and the price of linearity over smoothness classes. Two ingredients of our bounds are known separately. The additive selection cost for an unknown support appears in Yang & Tokdar (2015): for a joint d-sparse Hölder-type function of p predictors, not an additive function, their Remark 3.2 gives the all-estimator minimax L^2 risk $\lambda^2(\sqrt{n}\lambda/\sigma)^{-4\alpha/(2\alpha+d)} + (\sigma^2 d/n)\log(p/d)$. Our additive selector term instead bounds a specified estimator on an anisotropic Besov class; it is

not an optimal all-estimator selection rate, and the exact all-estimator selection term for our class remains undetermined. Their result does not characterise the extra cost of response-affinity. The exponent shift for $p < 2$ is the linear-estimator phenomenon of Suzuki (2019), Proposition 3, on a one-dimensional Besov unit ball under uniform design, and of Suzuki & Nitanda (2021), Theorem 5(i), whose anisotropic linear minimax lower bound $n^{-(2\bar{s}-v)/(2\bar{s}+1-v)}$, $v = 2(1/p-1/2)_+$, on a unit ball under uniform design has exactly our exponent β. Their Theorem 5(ii) gives an ambient-dimensional exponent for the composition model $\{h(Ax+b)\}$ with arbitrary A. Neither result yields the support multiplicity M: a lower bound for the larger composition class does not descend to the coordinate-support union it contains, and the single-ball result does not produce the factor by class inclusion; ours obtains $\min\{1,(M/n)^\beta\}$ from a direct random-design prior supported on one block at a time with covariance across all blocks. Finally, Hayakawa & Suzuki (2020), Theorem 3.3, show under uniform design that replacing a target class by its L^2-closed convex hull leaves the affine minimax risk unchanged. That classical reduction applies here as well, since Definition 3.4 permits offsets and label-independent randomisation and the bounded design density identifies the relevant closures (Section 6); the contribution is the quantitative support factor, not a new convex-hull principle.

Feature learning versus fixed features. A separate line shows that learning the features beats fixing them when the target depends on a few directions or coordinates of a high-dimensional input: single-hidden-layer networks with non-Euclidean regularisation adapt to unknown low-dimensional linear structure and select variables among exponentially many (Bach, 2017); kernel methods suffer a curse of dimensionality that two-layer networks escape by learning a low-dimensional representation (Ghorbani et al., 2020); gradient descent learns the relevant directions of $f^\star(x) = g(Ux)$ with far fewer samples than the kernel regime needs (Damian et al., 2022); for sparse functions of binary inputs, linear methods on any fixed feature map, the neural tangent kernel included, cannot learn efficiently while mean-field SGD can (Abbe et al., 2022); and random features cannot learn a single ReLU neuron without exponentially many of them (Yehudai & Shamir, 2019). Those results fix the feature side through a particular kernel, feature map or regime and obtain the gain through particular dynamics; here the fixed-feature side is the whole response-affine class, the target is nonparametric, and the price of fixed features is the explicit factor $\min\{1,(M/n)^\beta\}$ in a pair of worst-case bounds on one class.

3 The task, the target class and the fixed-weight comparator

Section 1 described the task in words; this section fixes it. Each of the D input coordinates is one token; the target depends on k of them through a smooth function; the learner sees noisy responses and is not told which k. The class of estimators whose weights are fixed before the responses are read, the comparator of Theorem 5.2, is defined at the end of the section.

Assumption 3.1 (Standing hypotheses)**.** $k \in \mathbb{N}$ is *fixed* (it does not grow with n or D, and no statement below is uniform in k) and $D \geq 2k$. The smoothness parameters are $p, q \in [1,\infty]$ (with $1/p := 0$ at $p = \infty$) and $s = (s_1,\dots,s_k) \in \mathbb{R}^k_{++}$, whose *effective smoothness* is

$$\bar{s} := \Big(\sum_{a=1}^{k} \tfrac{1}{s_a}\Big)^{-1}, \qquad s_{\max} := \max_a s_a, \qquad s_{\min} := \min_a s_a, \qquad \text{and we assume} \quad \boxed{\bar{s} > 1/p}.$$

Two radii $F > 0$ and $F_\infty > 0$ are fixed class parameters, independent of each other and of n and D.

Smoothness is measured by the *joint* anisotropic Besov norm of appendix (11), $\|f\|_{B^s_{p,q}} = \|f\|_{L^p} + \big\|\big(2^v \sup_{|h_a|\leq 2^{-v/s_a}\,\forall a} \|\Delta^r_h f\|_{L^p}\big)_{v\geq 0}\big\|_{\ell^q}$, where the r-th order difference with $r = \max_a \lfloor s_a \rfloor + 1$ is taken along one and the same vector h and is set to 0 unless x and $x + rh$ both lie in $[0,1]^k$. It is neither a decoupled per-direction modulus nor dominating mixed smoothness.

Definition 3.2 (Target class; appendix Definition B.1)**.**

$$\mathcal{G} := \big\{g : [0,1]^k \to \mathbb{R} \;:\; \|g\|_{B^s_{p,q}([0,1]^k)} \leq F,\ \ \|g\|_{L^\infty} \leq F_\infty\big\},$$

$$\mathcal{F} := \bigcup_{S\subseteq[D],\,|S|=k} \big\{x \mapsto g(x_S) \;:\; g \in \mathcal{G}\big\}.$$

Writing $S = \{i_1 < \cdots < i_k\}$, the a-th smallest index of S carries the smoothness s_a; the class is not enlarged to arbitrary assignments of s to coordinates.

The envelope F_∞ is a separate class parameter, not inferred from the Besov radius; it bounds the target and every clipped candidate in the oracle inequality and is carried in every constant (appendix Remark B.2).

Definition 3.3 (Observation model; appendix Definition B.3)**.** We observe n i.i.d. pairs

$$y_i = f^\star(x_i) + \xi_i, \qquad x_1, \dots, x_n \overset{\text{iid}}{\sim} P_X \text{ on } [0,1]^D, \qquad \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \text{ independent of } X,$$

with $f^\star \in \mathcal{F}$ and $\sigma > 0$ fixed, and P_X having a Lebesgue density p_X with $0 < c_- \le p_X \le c_+ < \infty$, where c_- and c_+ do not depend on D or n. Performance is the population integrated squared error $R(\hat{f}, f^\star) = \mathbb{E}_{X,\xi,\omega}\|\hat{f} - f^\star\|^2_{L^2(P_X)}$. The learner *does not know* S; it may use $k, s, p, q, F, F_\infty, \sigma, n, D$, the budgets of Section 4 and a randomisation ω as in Definition 3.4.

Definition 3.4 (Response-affine estimators; appendix Definition B.4)**.**

$$\mathcal{L}_{\text{aff}} := \Big\{ \hat{f}(\cdot) = a(X, \omega, \cdot) + \sum_{i=1}^{n} w_i(X, \omega, \cdot)\, y_i \Big\},$$

where the randomisation law of ω, conditional on X, is independent of the target and the noise (that is, $\omega \perp (f^\star, \xi) \mid X$), and where the offset a and the coefficients w_i are measurable in (X, ω, x) and *do not depend on* y.

These are the learners of the introduction whose selection is fixed before the labels are seen. $\mathcal{L}_{\text{aff}}$ contains kernel ridge regression at a label-independent regularisation level, ridge regression on fixed or randomly drawn features at a label-independent level, and lazy or NTK-type rules $f_0 + A^\top y$ with label-independent f_0 and A, including a non-zero initialisation. It does not contain rules whose coefficients read the responses (label-adaptive tuning, stopping or hyperparameter selection, or feature learning), and pointwise-clipped affine rules are not automatically covered by Theorem 5.2 either. None of this asserts that two classes are disjoint; the zero rule, for one, lies in $\mathcal{L}_{\text{aff}}$ and in its clipped image.

Constants below may depend on $(k, s, p, q, F, F_\infty, \sigma, c_+)$, and on c_- for the lower bound, but never on D, n, the resolution N, the support S or the target (appendix Remark B.5).

4 One positional attention head

The estimator is the exact least-squares minimiser over an explicitly parameterised class of one-head transformers with a clipped output. In the usual vocabulary: each coordinate x_i is a token; a fixed positional code c_i is its positional encoding and, through a learned matrix, its key; one softmax head with a learned query produces attention weights that depend on positions; a learned decoder reads the attended values out as a vector in $\mathbb{R}^k$; a small ReLU network maps that vector to the prediction; and a fixed clip bounds the output. Which map is optimised is part of the theorem, so each stage is written out below with its parameter count.

4.1 Tokens and positional codes

Each coordinate $i \in [D]$ carries a fixed positional code

$$c_i := s_i/\sqrt{m} \in \mathbb{R}^m, \qquad s_i \in \{-1, +1\}^m, \qquad m := \lceil 64k^2 \log(2D) \rceil,$$
$$\mu := \max_{i \neq j} |\langle c_i, c_j \rangle| \le \frac{1}{4k}. \tag{1}$$

A sign family meeting this coherence threshold exists and is fixed once, before any data are seen; it depends on neither the labels nor the support (appendix Definition C.1).

Tokens are the pairs (x_i, s_i). A *fixed* affine-plus-ReLU block with no trainable parameters emits the tagged value together with the code, $(x_i, s_i) \mapsto (x_i c_i,\, c_i)$, by an exact two-ReLU identity on $[0,1]^D$; the value and key paths below need both outputs (appendix Definition C.10, stage (0), and (26)).

4.2 THE ATTENTION HEAD

The trainable parameters are $\theta_s = (q, K, V, R)$ with $q \in \mathbb{R}^m$, $K, V \in \mathbb{R}^{m\times m}$ and $R \in \mathbb{R}^{k\times m}$, and the forward map is

$$z_i := \langle q,\, Kc_i\rangle, \qquad \alpha := \mathrm{softmax}(z) \in \mathbb{R}^D, \qquad h_{\theta_s}(x) := k\,R\sum_{i=1}^{D}\alpha_i\,V(x_i c_i) \;\in\; \mathbb{R}^k. \tag{2}$$

The softmax is at temperature one; the gain k is a fixed constant of the architecture, not a trainable scalar; and all four groups are constrained by one box $\|\theta_s\|_\infty \le B_{\mathrm{sel}}$. The free-parameter count is therefore exactly

$$P_{\mathrm{sel}} = m + m^2 + m^2 + km = 2m^2 + km + m \qquad (q,\ K,\ V,\ R). \tag{3}$$

The temperature schedule $\tau_{N,D}$ enters only through the magnitude of q, hence through B_{sel}, and contributes no slot to (3) (appendix Remark C.11).

The keys come from the *fixed* positional codes through the *learned* matrix K, so at prediction time the weights α do not depend on the content of the current input; the parameters that produce them are estimated from the training sample.

4.3 THE CANDIDATE CLASS AND THE ESTIMATOR

$$T_\theta(x) = \mathcal{C}_{F_\infty}\Big(\Phi_{\theta_f}\big(\Pi_{[0,1]^k} h_{\theta_s}(x)\big)\Big), \qquad \mathcal{H}_N := \{T_\theta\}, \tag{4}$$

with $\Pi_{[0,1]^k}$ the fixed coordinatewise cube projection, $\mathcal{C}_{F_\infty}(t) = \max(-F_\infty, \min(F_\infty, t))$ the fixed output clip at the class's own sup-norm radius, and Φ_{θ_f} a sparse deep ReLU network whose final scalar affine readout is one of its own layers, counted in the sparsity budget and bounded by the magnitude budget; there is no trainable multiplier after the network, the budgets B_{sel} and B_{ffn} are separate, and the tagger, the cube projection and the clip carry no free parameters.

Remark 4.1 (The architecture budgets)**.** For every N we use *exactly* the integer architecture schedule of appendix Definition C.6, which also fixes B_{sel} and every quantity left unexpanded here; its structural orders are $L_{\mathrm{net}} = O(\log 2N)$, $W_{\mathrm{net}} = O(N)$, $S_{\mathrm{net}} = O(N\log 2N)$ and $\log B_{\mathrm{ffn}} = O(\log 2N)$, with constants depending on the structural parameters only. *The candidate class is the one the schedule defines*; these orders only record its growth.

Definition 4.2 (The estimator; appendix (30) and Definition C.13)**.**

$$\hat f_n \;\in\; \operatorname*{arg\,min}_{T\in\mathcal{H}_{N_n}} \sum_{i=1}^{n}\big(y_i - T(x_i)\big)^2, \qquad N_n = \max\Big\{2, \min\big(n, \lfloor (n/\log^3(2n))^{1/(2\bar s+1)}\rfloor\big)\Big\},$$

the exact global minimiser over the clipped class at the deterministic resolution N_n, which reads $(n, \bar s)$ only; a minimiser exists and a measurable selection is available (appendix Proposition C.12), and nothing here asserts that an optimisation algorithm finds it.

5 MAIN RESULTS: THE COST OF UNKNOWN SUPPORT WITH AND WITHOUT ATTENTION

Write $\alpha := \dfrac{2\bar s}{2\bar s + 1}$ and $M := \binom{D}{k}$. Both theorems concern the same target class $\mathcal{F}$ of Definition 3.2 and the same observation model of Definition 3.3. Theorem 5.1 says what one positional head pays: a nonparametric term in n whose exponent and constant do not depend on D, plus an additive selection term polylogarithmic in D, which can still dominate when D grows with n. Theorem 5.2 says what every response-affine rule pays: the support multiplicity M through the ratio M/n, and for $p < 2$ a smaller exponent. Figure 1 draws both sides.

5.1 WHAT ONE ATTENTION HEAD PAYS

The selector envelope is $\bar\Xi_{n,D} = \sup_{2\le N\le n}\Xi_{N,n,D}$, with $\Xi_{N,n,D}$ as in appendix (32); it has the smooth majorant $\bar\Xi_{n,D} \le \Xi_{\mathrm{safe}}(n, D) := C_\Xi[\log^2(2n) + \log(2k) + \log\log(e^e D)]$ (appendix Proposition E.16), an inequality rather than an identity.

Theorem 5.1 (Upper bound; appendix Theorem D.1)**.** *Under Assumption 3.1 and Definitions 3.2–3.3, let $\hat{f}_n$ be the estimator of Definition 4.2. There are constants $C_0, C_1, C_2 < \infty$ depending on $(k, s, p, q, F, F_\infty, \sigma, c_+)$ and on nothing else (in particular not on D, n, N, S, $f^\star$ or g) such that for* every $n \geq 2$ *and* every $D \geq 2k$*,*

$$\sup_{f \in \mathcal{F}} \mathbb{E}\big\|\hat{f}_n - f\big\|^2_{L^2(P_X)} \;\leq\; U_n(D) := \underbrace{C_0\, n^{-\alpha} \log^{3\alpha}(2n)}_{\textit{nonparametric}} + \underbrace{C_1 \frac{P_{\mathrm{sel}}\, \bar{\Xi}_{n,D}}{n}}_{\textit{selection}} + \underbrace{\frac{C_2}{n}}_{\textit{oracle remainder}}, \quad (5)$$

where $\alpha = 2\bar{s}/(2\bar{s}+1)$. Replacing $\bar{\Xi}_{n,D}$ by $\Xi_{\mathrm{safe}}(n, D)$ gives a weaker bound that also holds.

The dimension enters (5) only through $P_{\mathrm{sel}} = 2m^2 + km + m = \Theta(k^4 \log^2 2D)$ and through $\log\log(e^e D)$ in the envelope: additively, never in the nonparametric exponent or its constant, and with no term $N^a \log D$ for $a > 0$. When D grows with n the selector term may still be the larger of the two. The exponent $3\alpha < 3$ is the optimum of the majorant in the appendix; the display of Section 1 is (5) with P_{sel} and $\bar{\Xi}_{n,D}$ replaced by their orders.

5.2 How one head selects k of D tokens

The approximation step must realise every support S within a single S-independent budget; for each S, whitening constructs such a witness, and the ERM itself is not given S. For a support S with code matrix C_S and Gram $G_S = C_S^\top C_S$, take

$$q_S = \tau\, C_S G_S^{-1} \mathbf{1}_k, \qquad K = V = I_m, \qquad R_S = G_S^{-1} C_S^\top. \quad (6)$$

Because $C_S^\top c_{i_a} = G_S e_a$, the active logits are *exactly equal*, $z_{i_a} = \tau\, \mathbf{1}^\top G_S^{-1} G_S e_a = \tau$ for every a, while off support $|z_i| \leq \tau/3$; equal logits make the softmax split its in-support mass evenly, and $R_S C_S = I_k$ turns that even split back into the selected coordinates. The two identities are exact; the recovery holds *up to the off-support softmax mass*, the decoder returning $(1 - \epsilon_{\mathrm{off}})x_S$ plus a term controlled by that mass. Coherence $\mu \leq 1/(4k)$ gives $\mathrm{spec}(G_S) \subseteq [3/4, 5/4]$, so all of (6) fits inside *one* budget B_{sel} valid for every S: the witness may depend on the support, the class budgets may not (appendix Lemma E.9).

The remaining steps are standard once selection works: Π places the network's argument in $[0,1]^k$, where a sparse ReLU network approximates g at rate $N^{-\bar{s}}$, and the 1-Lipschitz clip supplies the sup-norm envelope the oracle inequality needs (appendix Proposition E.7, Lemma E.8). The estimation term is the metric entropy of *all* of $\mathcal{H}_N$; because the selector's logarithms depend on N, the appendix takes the uniform envelope $\bar{\Xi}_{n,D}$ *before* optimising N, which leaves the second term of (5) free of N and produces α and 3α (appendix Proposition E.16, Section E.6).

5.3 What every response-affine rule pays

Theorem 5.2 (Response-affine lower bound; appendix Theorem D.2)**.** *Under the same Assumption 3.1 and the same Definitions 3.2–3.4, there are constants $c_1, c_2 > 0$ such that for* every $\hat{f} \in \mathcal{L}_{\mathrm{aff}}$*,* every $n \geq 1$ *and* every $D \geq 2k$*,*

$$\sup_{f \in \mathcal{F}} R(\hat{f}, f) \;\geq\; L_n(D) := \begin{cases} c_1 \min\{1, (M/n)^\alpha\}, & p \geq 2, \\ c_2 \min\{1, (M/n)^\beta\}, & 1 \leq p < 2, \end{cases} \quad (7)$$

$$s_* := \bar{s} - \Big(\frac{1}{p} - \frac{1}{2}\Big), \qquad \beta := \frac{2s_*}{2s_* + 1}. \quad (8)$$

No constraint ties n to the number of atoms used in the construction.

The saturation $\min\{1, \cdot\}$ is part of the statement: once $M \geq n$ the bound is a *constant*. Since $\bar{s} > 1/p$ gives $s_* = \bar{s} - (1/p - 1/2) > 1/2$, the strict inequality $\beta < \alpha$ at $1 \leq p < 2$ follows from Assumption 3.1 and is not a separate hypothesis.

5.4 WHY NO FIXED-WEIGHT RULE AVOIDS THE FACTOR M

A response-affine rule must commit its coefficients before seeing the responses, so the question is how much of the target a prior can hide from any such commitment. The construction exploits a contrast: *each realisation of the prior activates a single support block, yet the prior's covariance is spread evenly over all MQ_u candidate coordinates*. Concretely, fix a level u, build Q_u orthonormal atoms ψ_j per support, and draw a uniform support, a uniform relabelling and independent signs. Write $J := MQ_u$, $v := \mathcal{E}_u/J$ and $Z_{ij} := \psi_j(x_i)$, where $\mathcal{E}_u$ is the total squared coefficient energy of one realisation of the prior. Every realisation lies in $\mathcal{F}$ (a deterministic check at each point of the prior's support, not on average), while $\mathrm{Cov} = vI_J$ exactly. Only the first two moments are used; the prior is neither Gaussian nor independent across blocks (appendix Lemma F.7).

Conditioning on the design and on ω, the coefficient risk of a response-affine rule is an exact quadratic in its offset and coefficients, minimised at $\mathrm{tr}[(v^{-1}I + \sigma^{-2}Z^\top Z)^{-1}]$; the harmonic-trace inequality, Jensen, and the density upper bound then give

$$\sup_{f\in\mathcal{F}} R(\hat{f}, f) \;\geq\; \frac{c_- \, J \, v \, \sigma^2}{\sigma^2 + n\, c_+ \, v} \tag{9}$$

(appendix Lemma F.5), which needs neither $n \geq J$ nor an invertible empirical Gram. Finally, two profiles satisfy the Besov *and* the sup-norm constraint, a dense-sign profile and a single spike, each with amplitude $a_0 = \min\{F/C_{\mathrm{B}}, F_\infty/C_\eta^k\}$ consuming both radii, and they carry different energies $\mathcal{E}_u$. Optimising (9) over u turns those energies into α for $p \geq 2$ and β for $1 \leq p < 2$; the spike's admissibility is where $\bar{s} > 1/p$ is spent (appendix Lemma F.8 and Section F.4).

Comparing the two guarantees. Both sides are guarantees, not exact risks: they certify a risk separation whenever $U_n(D) < L_n(D)$, and no converse is claimed. Under two sufficient conditions on the constants, (S10) $C_1 P_{\mathrm{sel}} \bar{\Xi}_{n,D} \leq C_0 n^{1/(2\bar{s}+1)} \log^{3\alpha}(2n)$ and (R) $C_2 \leq C_0 n^{1-\alpha} \log^{3\alpha}(2n)$, one has $U_n(D) \leq 3C_0 n^{-\alpha} \log^{3\alpha}(2n)$ and, for every M and every D,

$$\frac{L_n(D)}{U_n(D)} \;\geq\; \frac{c}{3C_0} \cdot \frac{n^\alpha}{\log^{3\alpha}(2n)} \cdot \min\Big\{1, (M/n)^\beta\Big\}, \tag{10}$$

with $(c, \beta) = (c_1, \alpha)$ when $p \geq 2$ and (c_2, β) as in (8) when $p < 2$ (appendix Corollary D.3; its (40) writes out the branches $M \leq n$ and $M \geq n$). The branch is decided by M against n, not by (S10), and where (S10) fails one compares the full $U_n(D)$ and $L_n(D)$ (sample-complexity form: appendix Corollary D.5).

5.5 A FINITE-SAMPLE ILLUSTRATION

Both theorems hold at finite n and are worst-case over the class; what they do not show is how the support-uncertainty term behaves for one concrete prior at a concrete size. We instantiate the lower-bound construction at level $u = 0$: each support carries the single atom $\psi_S(x) = \eta(x_i)\eta(x_j)$, and $f^* = \varepsilon a \, \psi_{S^*}$ with S^* uniform over the $M = \binom{D}{2}$ supports and ε an independent Rademacher sign. Every rule is given this generative knowledge, so what is measured is the price of not knowing S^*, not the difficulty of approximating an unknown g. The affine comparator minimises the prior-average risk within $\mathcal{L}_{\mathrm{aff}}$; its ridge level $\lambda = M\sigma^2/a^2$ is fixed by the prior, not tuned on labels, so it is the strongest such rule here. Against it we run the exact posterior mean over the $2M$ support and sign hypotheses, and a non-Bayes rule fitting a support and a bounded amplitude by least squares without σ or the prior probabilities.

Risk under support uncertainty. Panel (a) of Figure 2 fixes $D = 32$ ($M = 496$) and varies n, over 100 independent datasets per cell. At $n = 128$ the paired exact gap between the affine and the label-driven rule is 0.880 of the null risk, with 95% dataset bootstrap interval $[0.875, 0.886]$ and all 100 datasets of one sign; the posterior mean gives 0.883 $[0.878, 0.889]$. Across the sweep the affine rule declines only from 0.991 to 0.780 of the null risk and never drops below the $M/(M+n)$ line, while the two label-driven rules fall to 0.0015 and below once n is large enough to locate a support; the same affine family given S^* is at 0.170 already at $n = 8$ and at 0.0028 at $n = 256$. The label-driven rule is *worse* than the null rule at $n = 8$ and 16, where it cannot yet identify a support, and clipping the affine rule to the envelope does not close the gap on these test points (full grid, all seven rules and reproduction files: Appendix H).

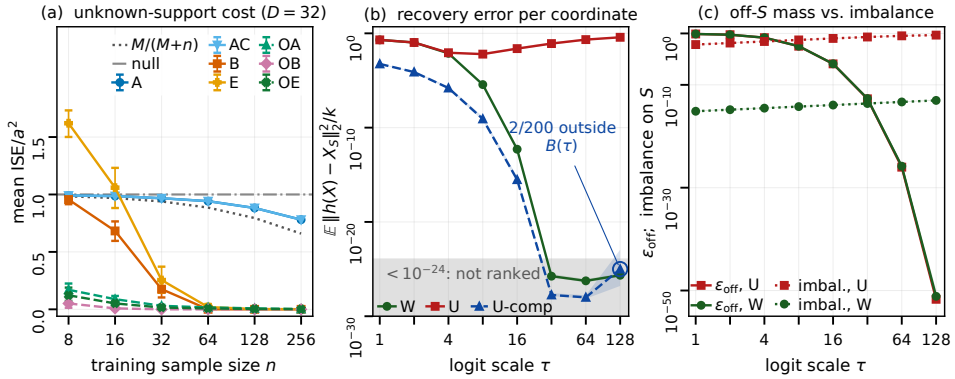


Figure 2: The cost of not knowing where to look, and the whitening mechanism, in two separate studies. **(a)** Finite-prior regression at $D = 32$ ($M = 496$): mean noiseless ISE$/a^2$ over 100 datasets per n on shared test points; the null risk is one. A is the prior-optimal response-affine rule and AC its clipping; B is the posterior mean and E bounded support-search least squares; OA, OB and OE are the same rules given S^*. Bars are 95% dataset bootstrap intervals; the dotted $M/(M+n)$ bounds the prior-average affine risk. **(b, c)** One known-support construction sweep over the logit scale τ at $k = 3$, $D = 32$: W is the whitened query of Lemma E.9, U the raw code sum at the same mean active logit, U-comp the raw query with a rebuilt decoder. Medians with 5–95% bands across ten codebooks (spreads, not confidence intervals); values below 10^{-24} are not ranked; two conditions outside the reference box $B(\tau)$ are retained. Whitening is shown sufficient, not necessary. Protocols and full grids: Appendices H and I.

A separate diagnostic of the whitening mechanism. Lemma E.9 is a known-support construction, not a training procedure; panels (b) and (c) probe it with no training sample and no learned support. On admissible sign codebooks (Definition C.1), with $C, S, K = V = I$ and the decoder R_S fixed, the whitened query (55) is compared with the raw code sum rescaled to the same mean active logit, so the off-support mass is measured rather than held equal. W has equal normalised active weights in exact arithmetic at every τ and U generally does not; at large τ both can have negligible off-support mass while U retains a recovery error, and the compensated decoder also recovers, so whitening is sufficient, not necessary. This is not the sup-norm bound (56) or a calibration of its constants; what whitening supplies is the proof's S-independent budget certificate (all conditions: Appendix I).

6 What this does and does not say about attention

The saving certified by Theorems 5.1 and 5.2 is not specific to attention: exact least squares over the union, across all $\binom{D}{k}$ supports, of clipped feed-forward networks with the budget of Section 4 attains the same leading bound with $\log\binom{D}{k}/n$ in place of the selection term (appendix, after Remark G.1), and the affine convex-hull reduction of Hayakawa & Suzuki (2020) applies to the lower bound (Section 2). What attention supplies is a continuous realisation of that search, one box of $P_{\text{sel}} = O(k^4 \log^2 D)$ parameters in place of $\binom{D}{k}$ discrete candidates, at a price that stays polylogarithmic in D. Lemma E.9 is a sufficiency statement: neither it nor the diagnostic of Section 5.5 shows whitening, or attention, to be necessary or optimal, and the separation is not attributed to a single component (appendix Remark D.8).

The lower bound is a statement about $\mathcal{L}_{\text{aff}}$ as defined, not about all estimators: rules whose coefficients read the responses (label-adaptive tuning, early stopping, model selection) lie outside it, the pointwise clipping of an affine rule is not automatically covered, and extending the bound to clipped affine rules remains open. It uses no fixed-design or sequence-model transfer. Whether the exponent α of the upper bound is minimax over $\mathcal{F}$ against all estimators is a different question, not addressed here.

The upper bound concerns the exact global minimiser of Definition 4.2, whose existence and measurability are proved; nothing asserts that an optimisation algorithm finds it, and the exploratory training of a reduced parameterisation in Appendix J gives mixed results that establish neither practical superiority nor the exact-ERM guarantee. Both theorems hold for fixed k with no uniformity in k; their constants and thresholds are symbolic, and the illustration of Section 5.5 does not calibrate them (appendix Remarks D.8 and G.1).

AI USE STATEMENT

The research question, the idea of this work and its assumptions are the authors' own: generative AI tools were not used to propose or refine Assumption 3.1 or the definitions of Section 3. We used generative AI tools (Claude models, Anthropic: Claude Opus 4.8, Claude Opus 5 and Claude Fable 5.1; and GPT models, OpenAI: GPT-6 and GPT-5.6) for tasks with required disclosure: developing the mathematical derivations from the authors' proof drafts, formulating and refining the theorem statements, writing and auditing the proofs, and designing, implementing and interpreting the numerical illustrations of Section 5.6 and Appendices H–J, including their code. The synthetic data in those illustrations are sampled by that code from the generators specified in the corresponding appendix; no data were produced by a language model directly. Translation, dataset cleaning and qualitative analysis are not applicable to this work. We also used generative AI tools for tasks with recommended disclosure: editing the text for readability and producing the figure sources. We have reviewed all AI-assisted work. The authors took part in the derivations and in the audit of every proof. For the code, we reviewed its logic against Appendices H–I, recomputed key quantities independently, ran the consistency checks shipped with the supplementary material, and re-ran it from the configuration and seeds provided there. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The theoretical claims are reproducible by checking the proofs: every hypothesis is collected in Assumption 3.1 and Definitions 3.2–3.4, the candidate class and the estimator are fixed in Section 4 and in the appendix definitions it references, and the two theorems are proved in the appendix (Lemma E.9 and the entropy and oracle steps behind Theorem 5.1; Lemmas F.7, F.5 and F.8 behind Theorem 5.2). All constants are symbolic, so no numerical value needs reproducing. The numerical illustration of Section 5.5, the diagnostic of Appendix I and the pilot of Appendix J ship their configuration, seeds, outputs and code in the supplementary material, with a single command that regenerates every number and figure from the shipped outputs; the training runs of Appendix J are reproducible from their stored checkpoints, not bit-for-bit on retraining. No proof depends on running any code.

REFERENCES

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4782–4887. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/abbe22a.html.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL http://jmlr.org/papers/v18/14-546.html.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/edelman22a.html.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and

H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14820–14830. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a9df2255ad642b923d95503b9a7958d8-Paper.pdf.

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: 10.1016/j.neunet.2019.12.014. URL https://www.sciencedirect.com/science/article/pii/S089360801930406X.

Christian Leisner. Nonlinear wavelet approximation in anisotropic Besov spaces. *Indiana University Mathematics Journal*, 52(2):437–455, 2003.

Clayton Sanford, Daniel J. Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36677–36707. Curran Associates, Inc., 2023. doi: 10.52202/075280-1596. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/73bf692447f174984f30499ec9b20e04-Paper-Conference.pdf.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL https://doi.org/10.1214/19-AOS1875.

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. Statement numbering in this paper follows arXiv:1910.12799v2.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D. Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51854–51912. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24ca.html.

Yun Yang and Surya T. Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652 – 674, 2015. doi: 10.1214/14-AOS1289. URL https://doi.org/10.1214/14-AOS1289.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. 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/5481b2f34a74e427a2818014b8e103b0-Paper.pdf.

A PROOFS

The appendix gives the complete definitions, architecture budgets and proofs used in the main text.

B THE MODEL

Throughout, $k \in \mathbb{N}$ is *fixed*: it does not grow with the sample size n or with the ambient dimension D. We write $D \geq 2k$, $p, q \in [1, \infty]$ with the convention $1/p := 0$ at $p = \infty$, and $s = (s_1, \ldots, s_k) \in \mathbb{R}^k_{++}$, and set

$$\bar{s} := \Big(\sum_{a=1}^{k} \tfrac{1}{s_a} \Big)^{-1}, \qquad s_{\max} := \max_a s_a, \qquad s_{\min} := \min_a s_a.$$

B.1 ANISOTROPIC BESOV SMOOTHNESS

We use the *joint* anisotropic modulus of smoothness of Suzuki & Nitanda (2021). For $h \in \mathbb{R}^k$ let $\Delta_h f := f(\cdot + h) - f(\cdot)$ and $\Delta_h^r := \Delta_h \circ \Delta_h^{r-1}$ — one and the same vector h at every order. On the cube $\Omega = [0,1]^k$ we adopt the source's boundary convention: $\Delta_h^r f(x)$ is used only when x and $x + rh$ both lie in Ω, and is set to 0 otherwise; *no zero-extension regularity is imposed on the target*. (For the compactly supported atoms of Section F a smooth zero extension is available, and the domain modulus is then bounded by the full-space estimate; that is a proof tool there, not an assumption here.) With $r_{\mathrm{diff}} := \max_a \lfloor s_a \rfloor + 1$ and $w_{r,p}(f,t) := \sup_{|h_a| \leq t_a \, \forall a} \|\Delta_h^r f\|_{L^p}$, the seminorm is

$$|f|_{B^s_{p,q}} := \Big\| \big(2^v \, w_{r_{\mathrm{diff}},p}\big(f, (2^{-v/s_1}, \ldots, 2^{-v/s_k})\big) \big)_{v \geq 0} \Big\|_{\ell^q}, \qquad \|f\|_{B^s_{p,q}} := \|f\|_{L^p} + |f|_{B^s_{p,q}}. \tag{11}$$

This is *not* a decoupled per-direction modulus and *not* mixed smoothness; the distinction is load-bearing in Proposition F.3.

Definition B.1 (Target class)**.** Fix radii $F > 0$ and $F_\infty > 0$ and impose

$$\bar{s} > 1/p. \tag{12}$$

Put $\mathcal{G} := \{ g : [0,1]^k \to \mathbb{R} : \ \|g\|_{B^s_{p,q}([0,1]^k)} \leq F, \ \|g\|_{L^\infty} \leq F_\infty \}$ and

$$\mathcal{F} \ := \bigcup_{S \subseteq [D],\, |S| = k} \big\{ \, x \mapsto g(x_S) \ : \ g \in \mathcal{G} \, \big\}. \tag{13}$$

Writing $S = \{i_1 < \cdots < i_k\}$, the a-th smallest index of S carries the smoothness s_a; this binds one function of k variables to each set S.

Remark B.2 (F_∞ is a parameter, not a consequence)**.** Both oracle inequalities used in Section E require a *common* sup-norm envelope on the target *and* on every candidate. Under (12) the anisotropic Besov space does embed into C^0, but that embedding is stated qualitatively in the source and names no constant, so no numerical F_∞ may be read off it. We therefore carry F_∞ as an independent parameter. Every constant below carries it explicitly. Condition (12) is strictly stronger than $\bar{s} > (1/p - 1/2)_+$, and the difference is exactly what makes the imported approximation hypothesis strict.

Definition B.3 (Observation model)**.** We observe n i.i.d. pairs $(x_i, y_i)_{i \leq n}$ with

$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} P_X \text{ on } [0,1]^D, \qquad y_i = f^\star(x_i) + \xi_i, \qquad \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \ \perp X,$$

where P_X has a Lebesgue density p_X with $0 < c_- \leq p_X \leq c_+ < \infty$, the constants $c_\pm$ being independent of D and n, and $\sigma > 0$ is fixed. Risk is $R(\hat{f}, f^\star) := \mathbb{E}_{X,\xi,\omega} \|\hat{f} - f^\star\|^2_{L^2(P_X)}$. An estimator may use $k, s, p, q, F, F_\infty, \sigma, n, D$ and all budgets declared below, together with any randomisation ω chosen *without looking at the responses*; it does *not* know S. The radii F, F_∞ and the noise level σ are fixed and do not grow with n or D.

Definition B.4 (Response-affine estimators)**.**

$$\mathcal{L}_{\mathrm{aff}} := \Big\{ \ \hat{f}(\cdot) = a(X, \omega, \cdot) + \sum_{i=1}^{n} w_i(X, \omega, \cdot) \, y_i \ \Big\},$$

where a and $w_1, \ldots, w_n$ are measurable in (X, ω) and in the evaluation point, ω is independent of $(f^\star, \xi)$ given X, and *neither a nor any w_i is a function of y*. This class contains kernel ridge regression at a label-independent regularisation level and lazy/NTK rules $f_0 + A^\top y$ with label-independent f_0, A, including non-zero initialisation. It does *not* contain rules whose coefficients read the responses

— label-tuned regularisation or stopping, feature learning. Pointwise-clipped affine rules are *not automatically* covered by $\mathcal{L}_{\text{aff}}$ either; note this is not a disjointness claim, since e.g. the zero rule lies in both $\mathcal{L}_{\text{aff}}$ and its clipped image. Extending Theorem D.2 to clipped affine rules is an open problem and is not claimed here.

Remark B.5 (Convention for constants)**.** A constant is called *structural* if it depends only on $(k, s, p, q, F, F_\infty, \sigma, c_+)$ for the upper half and additionally on c_- and on the atom-family parameters of Section F for the lower half. "Fixed k" means all of k, s (hence $s_{\max}$ and $s_{\min}$), p, q, F, F_∞, σ, $c_\pm$ are held fixed; no statement below is uniform in k. **No structural constant may depend on D, n, the resolution N, the unknown support S, or the target g.** Three constants entering the upper bound — $c_{(k,\mathsf{m})}$ of Lemma E.1, the norm-equivalence constant C_α of Lemma E.2, and the truncation constant C_{tr} of the same lemma — are named nowhere in their sources. The structural constants in the theoretical bounds remain symbolic; no numerical values or sample-size thresholds are inferred from them. The finite-prior illustration in Appendix H is specified separately and does not evaluate these constants.

C The estimator

C.1 The positional codebook

Definition C.1 (Codebook)**.** Let $m := \lceil 64k^2 \log(2D) \rceil$, let $s_1, \dots, s_D \in \{-1, +1\}^m$ be sign vectors and put

$$c_i := s_i/\sqrt{m} \in \mathbb{R}^m, \qquad \|c_i\|_2 = 1, \qquad \mu := \max_{i \neq j} \big|\langle c_i, c_j \rangle\big|. \tag{14}$$

We fix, once and for all and *before any data are seen*, a sign family with

$$\mu \;\leq\; \frac{1}{4k}. \tag{15}$$

Such a family exists: for i.i.d. Rademacher entries, $m\langle c_i, c_j\rangle$ is a sum of m i.i.d. signs, so Hoeffding gives $\mathbb{P}[|\langle c_i, c_j\rangle| > 1/(4k)] \leq 2\exp(-m/(32k^2))$, and a union bound over the $\binom{D}{2}$ pairs gives $\mathbb{P}[\mu > 1/(4k)] \leq D(D-1)\exp(-m/(32k^2)) \leq 1/4$ at $m = \lceil 64k^2 \log 2D \rceil$. One accepted family is then fixed; the estimator is deterministic given it, and the draw is neither label- nor support-dependent. Because μ is a maximum over *all* pairs, one codebook serves all $\binom{D}{k}$ supports at once; no union bound over supports is taken, which is why m carries $\log(2D)$ and not $k\log(2D/k)$.

Remark C.2 (What is and is not claimed about m)**.** Definition C.1 is a *sufficiency* statement. No matching necessity is claimed. In particular the Welch bound for D unit vectors in $\mathbb{R}^m$ with coherence μ reads $m \geq D/[1 + (D-1)\mu^2]$, which is *strictly smaller* than μ^{-2} for every finite D; the unconditional form $m \geq \mu^{-2}$ is only the $D \to \infty$ limit and is not a valid finite-D necessary condition. Nothing below uses any lower bound on m.

C.2 The sparse ReLU class and the one architecture schedule

Definition C.3 (Sparse ReLU networks)**.** For $L \geq 2$, $W \geq 1$, $S \geq 1$, $B > 0$ and input dimension d, let $\Phi_d(L, W, S, B)$ be the set of maps $f_\theta = (W^{(L)} \cdot + b^{(L)}) \circ \eta \circ (W^{(L-1)} \cdot + b^{(L-1)}) \circ \cdots \circ \eta \circ (W^{(1)} \cdot + b^{(1)})$ with $W^{(1)} \in \mathbb{R}^{W \times d}$, $W^{(l)} \in \mathbb{R}^{W \times W}$ for $1 < l < L$, $W^{(L)} \in \mathbb{R}^{1 \times W}$, $\eta(t) = \max(t, 0)$ applied coordinatewise, $\sum_{l=1}^{L} \big(\|W^{(l)}\|_0 + \|b^{(l)}\|_0\big) \leq S$ and $\max_{l \leq L} \|W^{(l)}\|_\infty \vee \|b^{(l)}\|_\infty \leq B$. *The final scalar affine readout $(W^{(L)}, b^{(L)})$ belongs to this parameter list:* it is counted in S and bounded by B. There is no trainable scale after it.

Lemma C.4 (Nominal versus source-admissible resolution)**.** *For $K \in \mathbb{Z}_{\geq 0}$ put*

$$\|K\| := \sum_{a=1}^{k} \Big\lfloor \frac{K s_{\min}}{s_a} \Big\rfloor, \qquad \mathcal{R}(s) := \big\{ 2^{\|K\|} : K \in \mathbb{Z}_{\geq 0} \big\}.$$

For every $N \in \mathbb{N}$ the number $N^\dagger := \min\{N' \in \mathcal{R}(s) : N' \geq N\}$ exists, and

$$N \;\leq\; N^\dagger \;<\; 2^k N.$$

Proof. $K \mapsto \|K\|$ is nondecreasing and unbounded, so the minimum exists. If $N^\dagger > 1$ then $2^{\|K-1\|} < N$ by minimality, and for each a, writing $c_a := s_{\min}/s_a \in (0,1]$, the integer $\lfloor Kc_a \rfloor - \lfloor (K-1)c_a \rfloor$ is nonnegative and strictly below $Kc_a - \big((K-1)c_a - 1\big) = c_a + 1 \le 2$, hence lies in $\{0,1\}$. Summing over a gives $\|K\| - \|K-1\| \le k$, so $N^\dagger \le 2^k 2^{\|K-1\|} < 2^k N$. □

Remark C.5 (Exact arithmetic)**.** $\|K\|$ must be evaluated exactly. In binary floating point, at $s = (0.7, 1.4, 2.8)$ and $K = 3$, the quantity $\lfloor 3 \cdot 0.7/0.7 \rfloor$ evaluates to 2.

Definition C.6 (The architecture schedule)**.** Given a nominal resolution $N \in \mathbb{N}$, $N \ge 2$, let $N^\dagger$ be as in Lemma C.4 and set, in this order,

$$\mathsf{m} := \min\{M \in \mathbb{N} :\ M > s_{\max} \text{ and } M > s_{\max} + 1 - 1/p\}, \tag{16}$$

$$W_0 := 6k\,\mathsf{m}(\mathsf{m}+2) + 2k\ \ (\ge 20k), \qquad Q := \lceil \log_2(k \vee \mathsf{m}) \rceil\ (\ge 1), \qquad B_0 := 2(\mathsf{m}+1)^{\mathsf{m}}, \tag{17}$$

$$A_0 := (k \vee \mathsf{m}) \log_2 3 + \log_2(1/c_k) + 5\ \ (\ge 5), \qquad c_k := \min\{1, c_{(k,\mathsf{m})}\}, \tag{18}$$

$$L_{\text{net}} := \max\Big\{3,\ 3 + 2Q\big\lceil A_0 + \bar{s}\log_2 N^\dagger + \log_2\log(2N^\dagger)\big\rceil\Big\}, \tag{19}$$

$$\mathsf{E} := \Big\lceil (\mathsf{m}+2)^k 2^{k+1} + (2^\nu - 1)^{-1} + (1+k)/\nu + 1 \Big\rceil, \tag{20}$$

$$W_{\text{net}} := \mathsf{E}\, N^\dagger W_0, \qquad S_{\text{net}} := \mathsf{E}\, N^\dagger \big(L_{\text{net}} W_0^2 + 1\big) + 2W_{\text{net}} + 1, \tag{21}$$

$$b := \max\Big\{1,\ (1+\nu^{-1}) \max\{\bar{s}/s_{\min},\ (1/p - \bar{s})_+\}\Big\}, \qquad B_{\text{ffn}} := \max\{1,\ C_B (2N^\dagger)^b\}, \tag{22}$$

where $\delta_{\text{ex}} := (1/p - 1/2)_+$, $\nu := (\bar{s} - \delta_{\text{ex}})/(2\delta_{\text{ex}})$, with $\nu := \infty$, $\nu^{-1} := 0$ and $(2^\nu - 1)^{-1} := 0$ when $\delta_{\text{ex}} = 0$, and C_B is the structural constant fixed in (54) below. Finally set, *once*,

$$\bar{L} := \big(W_{\text{net}}(B_{\text{ffn}} \vee 1)\big)^{L_{\text{net}}}, \qquad \tau_{N,D} := 2\log\big(e + 6kD\,\bar{L}\,N^{\dagger\bar{s}}\big), \qquad B_{\text{sel}} := 2\sqrt{k}\,\max\{1, \tau_{N,D}\}. \tag{23}$$

$\mathsf{m}, W_0, Q, \mathsf{E}, L_{\text{net}}, W_{\text{net}}, S_{\text{net}}$ are integers. **This is the only architecture schedule used anywhere in this document.**

Remark C.7 (The depth is well defined without knowing an unnamed constant)**.** A_0 involves $\log_2(1/c_{(k,\mathsf{m})})$ with $c_{(k,\mathsf{m})}$ named nowhere in Lemma E.1, so its sign is not known. Decreasing that constant only *increases* the depth L_0 that Lemma E.1 requires, and $\Phi_d(L, W, S, B) \subseteq \Phi_d(L', W, S + (L'-L)W, B \vee 1)$ for $L' \ge L$ by Lemma C.8. Replacing $c_{(k,\mathsf{m})}$ by $c_k = \min\{1, c_{(k,\mathsf{m})}\}$ in (18) is therefore legitimate and gives $A_0 \ge (k \vee \mathsf{m}) \log_2 3 + 5 \ge 5$. Hence the ceiling in (19) has a strictly positive argument for every $N^\dagger \ge 1$: *no depth expression is ever evaluated at a negative or undefined value, and no network is defined through one.*

Lemma C.8 (Depth padding)**.** *If $f \in \Phi_d(L, W, S, B)$ and $L' \ge L$ then $f \in \Phi_d(L', W, S + (L'-L)W, B \vee 1)$.*

Proof. Insert affine layers $(I_W, 0)$ immediately after a ReLU layer. Since $\eta(I_W \eta(u)) = \eta(u)$ exactly, the represented function is unchanged; each inserted layer has magnitudes $1 \le B \vee 1$ and contributes W nonzeros. □

Remark C.9 (Orders, with their structural constants)**.** Expressed through the nominal N using Lemma C.4,

$$W_{\text{net}} \le 2^k \mathsf{E} N W_0, \qquad B_{\text{ffn}} \le \max\{1, C_B 2^{kb} (2N)^b\}, \tag{24}$$

$$L_{\text{net}} \le 3 + 2Q\big\lceil A_0 + \bar{s}(k + \log_2 N) + \log_2\log(2^{k+1}N)\big\rceil, \tag{25}$$

and $S_{\text{net}} \le 2^k \mathsf{E} N (L_{\text{net}} W_0^2 + 1) + 2W_{\text{net}} + 1$. All inflations are structural. Hence

$$L_{\text{net}} = O(\log 2N), \quad W_{\text{net}} = O(N), \quad S_{\text{net}} = O(N \log 2N), \quad \log B_{\text{ffn}} = O(\log 2N),$$

with structural constants $2Q\bar{s}/\log 2$, $2^k \mathsf{E} W_0$, $2^k \mathsf{E} W_0^2 \cdot 2Q\bar{s}/\log 2$ and b respectively. These symbolic summaries are stated only *after* the integer schedule (16)–(22), and every proof below uses the integer forms.

C.3 THE CANDIDATE CLASS AND THE ESTIMATOR

Definition C.10 (Candidate class)**.** Fix $N \geq 2$ and the schedule of Definition C.6. A candidate is the composition of five stages, numbered (0)–(4), in order.

(0) Value tagger (fixed; no free parameters). Tokens are the pairs $e_i := (x_i, s_i) \in \mathbb{R}^{1+m}$. One shared affine-plus-ReLU block computes, coordinatewise in $r \in [m]$,

$$\big(\eta(x_i + s_{i,r} - 1) - \eta(x_i - s_{i,r} - 1)\big)/\sqrt{m} \;=\; x_i\, s_{i,r}/\sqrt{m}, \tag{26}$$

an *identity* for every $x_i \in [0, 2]$ and every $s_{i,r} \in \{\pm 1\}$ (check the two cases: at $s = +1$, $\eta(x) - \eta(x - 2) = x$; at $s = -1$, $\eta(x - 2) - \eta(x) = -x$), so $\tilde{v}_i = x_i c_i$ on the whole cube $[0, 1]^D$ with a factor-two margin. *The same block passes the positional part through unchanged*, so stage (0) emits the pair $(\tilde{v}_i, c_i)$: the key path of stage (1) needs c_i, which $\tilde{v}_i$ alone does not determine (it vanishes at $x_i = 0$). The pass-through is part of the definition, not a reading of it. Stage (0) has *zero* trainable parameters and is the same map for every candidate, so (26) holds on the whole class as a theorem, not as a quantifier to be checked; the code features are fixed inputs, not trainable biases.

(1) Selector (trainable, parameters θ_s). One head, with

$$\theta_s = (q, K, V, R), \qquad q \in \mathbb{R}^m, \quad K, V \in \mathbb{R}^{m\times m}, \quad R \in \mathbb{R}^{k\times m}, \tag{27}$$

positional logits, softmax pooling *at temperature one*, and a fixed deterministic gain k:

$$z_i := \langle q,\, Kc_i\rangle, \qquad \alpha := \mathrm{softmax}(z) \in \mathbb{R}^D, \qquad h_{\theta_s}(x) := k\, R \sum_{i=1}^{D} \alpha_i\, V(x_i c_i) \;\in \mathbb{R}^k. \tag{28}$$

The logit convention is *unscaled* (no $1/\sqrt{m}$), which is legitimate because $\|c_i\|_2 = 1$ deterministically by (14). The gain k is a fixed constant of the architecture, not a trainable scalar, and there is no post-decoder multiplier. All four groups live in one box $\|\theta_s\|_\infty \leq B_{\mathrm{sel}}$, and the free-parameter count is *exactly*

$$P_{\mathrm{sel}} \;=\; \underbrace{m}_{q} + \underbrace{m^2}_{K} + \underbrace{m^2}_{V} + \underbrace{km}_{R} \;=\; 2m^2 + km + m. \tag{29}$$

(2) Cube projection (fixed). $\Pi(t) := \big(\min(1, \max(0, t_a))\big)_{a\leq k}$.

(3) Feed-forward network (trainable, parameters θ_f). $\Phi_{\theta_f} \in \Phi_k(L_{\mathrm{net}}, W_{\mathrm{net}}, S_{\mathrm{net}}, B_{\mathrm{ffn}})$, its scalar readout included (Definition C.3).

(4) Output clip (fixed). $\mathcal{C}_{F_\infty}(t) := \max(-F_\infty, \min(F_\infty, t))$, the clipping level being F_∞, the sup-norm radius of Definition B.1.

The class is $\mathcal{H}_N := \{\, \mathcal{C}_{F_\infty} \circ \Phi_{\theta_f} \circ \Pi \circ h_{\theta_s} \,\}$, the union over all admissible (θ_s, θ_f).

Remark C.11 (One selector scale)**.** The quantity $\tau_{N,D}$ of (23) is a deterministic schedule quantity. It enters through the *magnitude* of q, which the witness of Lemma E.9 takes of order $\tau_{N,D}$, and correspondingly through the parameter box B_{sel}. The softmax itself is at temperature one; writing "softmax at temperature $\tau_{N,D}$" *in addition* would scale the logits twice. $\tau_{N,D}$ is not a trainable parameter and contributes no slot to (29).

Proposition C.12 (The ERM exists and is measurable)**.** *For every $N \geq 2$ the empirical risk $T \mapsto \sum_{i=1}^n \big(y_i - T(x_i)\big)^2$ attains its infimum on $\mathcal{H}_N$, and a measurable selection*

$$\hat{f}_n \in \operatorname*{arg\,min}_{T\in\mathcal{H}_N} \sum_{i=1}^{n} \big(y_i - T(x_i)\big)^2 \tag{30}$$

exists.

Proof. The parameter set is $\Theta := \{\|\theta_s\|_\infty \leq B_{\mathrm{sel}}\} \times \bigcup_P \{\theta_f \in [-B_{\mathrm{ffn}}, B_{\mathrm{ffn}}]^P\}$, a finite union (over the finitely many sparsity patterns P with $|P| \leq S_{\mathrm{net}}$) of compact boxes, hence compact. For fixed x

the map $\theta \mapsto T_\theta(x)$ is continuous ($\mathrm{softmax}$, Π, η and $\mathcal{C}_{F_\infty}$ are continuous, and finite compositions and products of continuous maps are continuous), so the empirical risk is continuous on a compact set and attains its minimum. Joint measurability in (θ, x, y) and compactness give a measurable minimiser by the Kuratowski–Ryll-Nardzewski selection theorem. Nothing here asserts that any optimisation algorithm finds $\hat{f}_n$; see Section G. □

Definition C.13 (Resolution schedule of the estimator)**.**

$$N_n := \max\Big\{2,\ \min\big(n,\ \lfloor (n/\log^3(2n))^{1/(2\bar{s}+1)} \rfloor\big)\Big\}, \tag{31}$$

and $\hat{f}_n$ is (30) at $N = N_n$. This reads $(n, \bar{s})$ only, so it is computable by an estimator that must know its own budgets. Three resolutions must be kept apart: the *estimator's* N_n, the *nominal* N of Definition C.6, and the *source-admissible* $N^\dagger$ of Lemma C.4.

D Results

Write $\alpha := \dfrac{2\bar{s}}{2\bar{s}+1}$ and $M := \binom{D}{k}$, and define the two selector envelopes

$$\bar{\Xi}_{n,D} := \sup_{2 \le N \le n} \Xi_{N,n,D}, \qquad \Xi_{N,n,D} := \log\Big(1 + 2B_{\mathrm{sel}}\, L_{\mathrm{sel}}\, \bar{L}\, n\Big), \tag{32}$$

$$\Xi_{\mathrm{safe}}(n, D) := C_\Xi\Big[\log^2(2n) + \log(2k) + \log\log(e^e D)\Big], \quad C_\Xi \text{ structural.} \tag{33}$$

Here B_{sel}, $\bar{L}$ are the schedule quantities (23) at resolution N and L_{sel} is the selector's parameter-to-output Lipschitz constant of Lemma E.15. Proposition E.16 proves $\bar{\Xi}_{n,D} \le \Xi_{\mathrm{safe}}(n, D)$. **The two are different objects and are never interchanged silently:** $\bar{\Xi}$ is an exact supremum over an integer schedule and is a staircase in n; Ξ_{safe} is a smooth majorant.

Theorem D.1 (Upper bound)**.** *Adopt Definitions B.1–B.3 and let $\hat{f}_n$ be the estimator of Definition C.13. There are structural constants $C_0, C_1, C_2 < \infty$, depending on $(k, s, p, q, F, F_\infty, \sigma, c_+)$ and on nothing else — in particular not on D, n, N, S, $f^\star$ or g — such that for every $n \ge 2$ and every $D \ge 2k$,*

$$\sup_{f \in \mathcal{F}} \mathbb{E}\big\|\hat{f}_n - f\big\|^2_{L^2(P_X)} \ \le\ U_n(D) \ :=\ C_0\, n^{-\alpha} \log^{3\alpha}(2n) \ +\ C_1\, \frac{P_{\mathrm{sel}}\, \bar{\Xi}_{n,D}}{n} \ +\ \frac{C_2}{n}, \tag{34}$$

with $P_{\mathrm{sel}} = 2m^2 + km + m = \Theta\big(k^4 \log^2(2D)\big)$ and $m = \lceil 64k^2 \log 2D \rceil$. Replacing $\bar{\Xi}_{n,D}$ by $\Xi_{\mathrm{safe}}(n, D)$ gives the weaker but smooth bound $U_n^{\mathrm{safe}}(D)$, which also holds.

The logarithmic exponent $3\alpha = 6\bar{s}/(2\bar{s}+1) < 3$ is *derived* in Section E.6, not prescribed; it is the exponent of this majorant and no claim is made that the minimax rate carries it.

Theorem D.2 (Lower bound for response-affine estimators)**.** *Adopt Definitions B.1–B.4. There are structural constants $c_1, c_2 > 0$, given explicitly in (78), such that for* every $\hat{f} \in \mathcal{L}_{\mathrm{aff}}$, *every $n \ge 1$ and every $D \ge 2k$,*

$$p \ge 2: \qquad \sup_{f \in \mathcal{F}} R(\hat{f}, f) \ \ge\ c_1 \min\Big\{1, \big(M/n\big)^{\alpha}\Big\}, \tag{35}$$

$$1 \le p < 2: \qquad \sup_{f \in \mathcal{F}} R(\hat{f}, f) \ \ge\ c_2 \min\Big\{1, \big(M/n\big)^{\beta}\Big\},$$

$$\beta := \frac{2s_*}{2s_*+1}, \quad s_* := \bar{s} - \Big(\frac{1}{p} - \frac{1}{2}\Big) > \frac{1}{2}. \tag{36}$$

No constraint ties n to the number of atoms. We write $L_n(D)$ for the right-hand side and (c, β) for (c_1, α) when $p \ge 2$ and (c_2, β) when $p < 2$.

Corollary D.3 (Comparison)**.** *Assume the two* printed *conditions*

$$\textit{(S10)} \qquad C_1\, P_{\mathrm{sel}}\, \bar{\Xi}_{n,D} \ \le\ C_0\, n^{1/(2\bar{s}+1)} \log^{3\alpha}(2n), \tag{37}$$

$$\textit{(R)} \qquad C_2 \ \le\ C_0\, n^{1-\alpha} \log^{3\alpha}(2n). \tag{38}$$

Then $U_n(D) \le 3C_0 n^{-\alpha}\log^{3\alpha}(2n)$ *and, for every* M *and every* D,

$$\frac{L_n(D)}{U_n(D)} \;\ge\; \frac{c}{3C_0}\;\cdot\;\frac{n^\alpha}{\log^{3\alpha}(2n)}\;\cdot\;\min\Big\{1,(M/n)^\beta\Big\}, \tag{39}$$

which splits into

$$M \le n:\;\; \frac{L_n}{U_n} \;\ge\; \frac{c}{3C_0}\,\frac{M^\beta n^{\alpha-\beta}}{\log^{3\alpha}(2n)}, \qquad\qquad M \ge n:\;\; \frac{L_n}{U_n} \;\ge\; \frac{c}{3C_0}\,\frac{n^\alpha}{\log^{3\alpha}(2n)}. \tag{40}$$

Proof. Under (37) the second term of (34) is at most $C_0 n^{-\alpha}\log^{3\alpha}(2n)$ after dividing by n, and under (38) so is the third; adding the first gives $U_n \le 3C_0 n^{-\alpha}\log^{3\alpha}(2n)$. Divide L_n by it. The split is the definition of $\min\{1,\cdot\}$. □

Remark D.4 (The branch condition is M versus n, and (37) is only sufficient)**.** Two independent conditions appear above and must not be conflated.

1. **Saturation** is governed by M versus n. The non-saturated display $M^\beta n^{\alpha-\beta}$ is valid *only* for $M \le n$. Applying it where $M \ge n$ overstates the certificate: at $k=1$, $p=1$, $s=3/2$ and $D=n^2$ one has $\alpha=3/4$, $\beta=2/3$, $M=D=n^2$, and the misapplied display gives the exponent $2\beta+(\alpha-\beta)=17/12$ against the correct $\alpha=9/12$, a spurious factor $n^{2/3}$.

2. **Condition** (37) caps D through $\log^2(2D)$. It is *sufficient* for the simplified certificate (39) and is *not necessary for a separation*. When it fails one compares the full $U_n(D)$ with the full $L_n(D)$; it is *not* the case that the upper bound becomes vacuous. For instance at $k=1$, $p=1$, $s=3/2$ and $\log D = n^{1/5}$ the selection term is $\Theta(n^{-3/5}\log^2 n)$ while the first term is $\Theta(n^{-3/4}\log^{9/4} n)$, so (37) fails for all large n — yet $U_n(D)\to 0$ while $M=D\gg n$ keeps $L_n = c_2$ constant, so the full comparison still certifies a constant-versus-vanishing gap.

Corollary D.5 (Certified sample-size separation)**.** *Fix a target risk* $r>0$ *and use the smooth envelope throughout. Let* $n_{\mathrm{att}}^{\mathrm{safe}}(r,D)$ *be the least integer* $n \ge n_* := \max\{n_0, \lceil e^3/2\rceil + 1\}$ *satisfying all three of*

$$C_0 n^{-\alpha}\log^{3\alpha}(2n) \le \tfrac{r}{3}, \qquad C_1 P_{\mathrm{sel}}\,\Xi_{\mathrm{safe}}(n,D)/n \le \tfrac{r}{3}, \qquad C_2/n \le \tfrac{r}{3}, \tag{41}$$

n_0 *being the threshold of Theorem D.1. Each of the three left-hand sides is strictly decreasing in* n *past* n_**, so the admissible set is upward closed and the least element is well defined; and* $\sup_{f\in\mathcal{F}}\mathbb{E}\|\hat f_n - f\|^2 \le r$ *for every such* n*. On the other side, for* $r < c$ *and on the non-saturated branch, Theorem D.2 forces every* $\hat f \in \mathcal{L}_{\mathrm{aff}}$ *to use* $n \ge M(c/r)^{1/\beta}$*; on the saturated branch no sample size achieves* $r < c$ *at all. Consequently a separation at risk level* r *is* ***certified whenever***

$$M\,(c/r)^{1/\beta} \;>\; n_{\mathrm{att}}^{\mathrm{safe}}(r,D). \tag{42}$$

Proof. Sufficiency of (41) is Theorem D.1 with $\bar\Xi \le \Xi_{\mathrm{safe}}$ (Proposition E.16). Monotonicity: for $h(n) = n^{-\alpha}\log^{3\alpha}(2n)$ one has $(\log h)'(n) = \frac{\alpha}{n}\big[\frac{3}{\log(2n)} - 1\big] < 0$ once $\log(2n) > 3$, i.e. $n > e^3/2$; for $v(n) = \big[\log^2(2n) + a_D\big]/n$ with $a_D := \log(2k) + \log\log(e^e D) > 0$ one has $v'(n) = \big[2\log(2n) - \log^2(2n) - a_D\big]/n^2 < 0$ once $\log(2n) > 2$; and C_2/n is always decreasing. The necessary side is (35)–(36) inverted. □

Remark D.6 (Why the smooth envelope is mandatory here)**.** The derivative computation above applies to Ξ_{safe}, *not* to $\bar\Xi$. The exact envelope $\bar\Xi_{n,D}$ is a staircase: at $k=1$ the admissible set $\mathcal{R}(s)$ of Lemma C.4 is the powers of two, so as n passes 2^j the supremum in (32) first admits an N whose $N^\dagger$ has doubled, and $\bar\Xi_{n,D}$ jumps upward while n has increased by one. Hence $\bar\Xi_{n,D}/n$ increases at infinitely many n and *no* eventual-monotonicity claim may be made for it. A reader who prefers to avoid the issue may drop the word "least" from Corollary D.5 and read it as: any n satisfying the three inequalities suffices. Both readings are used consistently; they are not mixed.

Remark D.7 (What Corollary D.5 is not)**.** The left-hand side of (42) comes from a *lower* bound and is therefore *necessary* for $\mathcal{L}_{\mathrm{aff}}$; the right-hand side comes from an *upper* bound and is therefore *sufficient* for $\hat f_n$. Neither is an exact sample complexity, and (42) is *not* an equivalence: its failure

does not show that no separation exists, only that this certificate does not reach it. At $p < 2$ the affine necessary budget does contain the factor M; what does not follow is that the whole gap is a factor M, since $\alpha \neq \beta$ places an exponent gap on top of it. Because C_0, C_1, C_2, c_1, c_2 are symbols (Remark B.5), the thresholds implicit in (41) and (42) are symbolic; **no numerical sample size, multiplier or onset is asserted**.

Remark D.8 (Interpretation)**.** The selector's routing depends only on positions at prediction time, while its parameters are estimated from the training sample. Theorems D.1 and D.2 compare the complete estimator with the response-affine class $\mathcal{L}_{\mathrm{aff}}$ and *do not identify a unique module responsible for any separation*. A separation is certified only in regimes where the upper guarantee is strictly below the response-affine lower bound; that inequality, and not any membership statement, is the evidence. Pointwise-clipped affine rules are *not* covered by Theorem D.2.

E Proof of Theorem D.1

E.1 The two imported statements

We import exactly two statements from Suzuki & Nitanda (2021), and one oracle inequality. Each is used as stated, with the hypotheses and the decay-parameter notation of Lemma E.2 made explicit below; nothing here asserts that their own upstream proofs have been re-audited.

Lemma E.1 (Spline emulation; Suzuki & Nitanda, 2021, Lemma 1)**.** *Let $M^k_{0,0}$ be the tensor-product cardinal B-spline of order m on $\mathbb{R}^k$. There is a constant $c_{(k,\mathsf{m})} > 0$, depending only on k and m, such that for every $\epsilon > 0$ there is $\check{M} \in \Phi_k\big(L_0(\epsilon), W_0, L_0(\epsilon)W_0^2, B_0\big)$ with $L_0(\epsilon) = 3 + 2\lceil \log_2\big(3^{k\vee\mathsf{m}}/(\epsilon\, c_{(k,\mathsf{m})})\big) + 5\rceil Q$, W_0, Q, B_0 as in (17), satisfying*

$$\big\|M^k_{0,0} - \check{M}\big\|_{L^\infty(\mathbb{R}^k)} \le \epsilon \qquad \text{and} \qquad \check{M}(t) = 0 \ \text{for all } t \notin [0, \mathsf{m}+1]^k. \tag{43}$$

The *exact compact support* in (43) is used in an essential way in Proposition E.7; it is what prevents the replacement error from acquiring a factor equal to the number of atoms.

At output norm L^2 the cited lemma carries two hypotheses, namely $\bar{s} > (1/p - 1/2)_+$ and $s_{\max} < \min\{\mathsf{m},\ \mathsf{m} - 1 + 1/p\}$. The first is implied by the standing assumption $\bar{s} > 1/p$ (the two coincide at $p = \infty$); the second is exactly the pair of inequalities defining m in (16), both of which are needed.

Lemma E.2 (Decomposition and coefficient norm; Suzuki & Nitanda, 2021, Lemma 2)**.** *Let $J_a(\kappa) := \{-\mathsf{m}, \dots, 2^{\lfloor \kappa s_{\min}/s_a \rfloor}\}$ and $J(\kappa) := J_1(\kappa) \times \cdots \times J_k(\kappa)$. Every u with $\|u\|_{B^s_{p,q}([0,1]^k)} \le 1$ admits $u = \sum_{\kappa \ge 0} \sum_{j \in J(\kappa)} a_{\kappa,j} M^k_{\kappa,j}$ with $M^k_{\kappa,j}(t) := M^k_{0,0}\big((2^{\lfloor \kappa s_{\min}/s_a \rfloor} t_a - j_a)_a\big)$, and the coefficient quasi-norm*

$$\big\|(a_{\kappa,j})\big\|_b := \Big\| \Big(w_\kappa \big(\textstyle\sum_{j \in J(\kappa)} |a_{\kappa,j}|^p \big)^{1/p} \Big)_{\kappa \ge 0} \Big\|_{\ell^q}, \qquad w_\kappa := 2^{\kappa s_{\min} - \|\kappa\|/p},$$

is equivalent to $\|u\|_{B^s_{p,q}}$. Moreover, for $K \in \mathbb{Z}_{\ge 0}$ and $N = 2^{\|K\|}$ there is a truncation

$$u_N = \sum_{\kappa=0}^{K} \sum_{j \in J(\kappa)} a_{\kappa,j} M^k_{\kappa,j} + \sum_{\kappa=K+1}^{K^*} \sum_{i=1}^{n_\kappa} a_{\kappa,j_i} M^k_{\kappa,j_i}, \qquad \|u - u_N\|_{L^2} \le C_{\mathrm{tr}} N^{-\bar{s}}, \tag{44}$$

with $K^* = \lceil K(1+\nu^{-1}) \rceil$, $n_\kappa = \lceil 2^{\|K\| - \nu(\|\kappa\| - \|K\|)} \rceil$ and $(j_i) \subseteq J(\kappa)$. We write $\mathcal{E}_N$ for the index set of (44) and C_α for the constant of the norm equivalence in the direction $\|(a)\|_b \le C_\alpha \|u\|_{B^s_{p,q}}$.

The display for n_κ in the cited Lemma 2 writes the decay parameter as ϵ, a symbol that statement does not define; we write ν, which is the parameter its own proof uses in the retained-level exponent (arXiv:1910.12799v2, p. 17). For $p < 2$ that exponent is $\bar{s} - \delta_{\mathrm{ex}} - \delta_{\mathrm{ex}}\nu = (\bar{s} - \delta_{\mathrm{ex}})/2 > 0$, and K^* satisfies $(\bar{s} - \delta_{\mathrm{ex}})(1 + \nu^{-1}) = \bar{s} + \delta_{\mathrm{ex}}$; for $p \ge 2$ the conventions of Definition C.6 make the second sum of (44) empty, so n_κ is not evaluated.

Lemma E.3 (Bounded-class oracle inequality; Suzuki & Nitanda, 2021, Prop. 4, cf. Schmidt-Hieber, 2020, Lemma 4)**.** *Let $\mathcal{T}$ be a class of functions $[0,1]^D \to \mathbb{R}$ with $\|T\|_\infty \le F'$ for every $T \in \mathcal{T}$ and*

$\|f^\star\|_\infty \le F'$ *for some* $F' \ge 1$*, and let* $\hat{f}$ *be the exact least-squares minimiser over* $\mathcal{T}$ *in the model of Definition B.3. Then for every* $\delta \in (0,1]$ *with* $\mathcal{N}(\delta, \mathcal{T}, \|\cdot\|_\infty) \ge 3$*,*

$$\mathbb{E}\|\hat{f} - f^\star\|^2_{L^2(P_X)} \;\le\; C_{\mathrm{o}}\Big[\inf_{T\in\mathcal{T}} \|T - f^\star\|^2_{L^2(P_X)} + \frac{(F'^2 + \sigma^2)\log \mathcal{N}(\delta, \mathcal{T}, \|\cdot\|_\infty)}{n} + \delta(F' + \sigma)\Big], \tag{45}$$

C_{o} *universal. The remainder is* first order *in* δ*, which is what forces the choice* $\delta = 1/n$ *below and the resulting logarithms.*

Remark E.4 (Hypotheses of Lemma E.3 and where they are met)**.** The envelope is met with $F' := \max(1, F_\infty)$ by construction of stage (4) of Definition C.10: $|T(x)| \le F_\infty$ for *every* parameter and *every* real x, needing no property of Φ_{θ_f}, h_{θ_s} or x; and $\|f^\star\|_\infty \le F_\infty$ is a defining hypothesis of $\mathcal{F}$. The covering-number floor $\mathcal{N}(1/n, \mathcal{H}_N, \|\cdot\|_\infty) \ge 3$ holds once $n \ge n_{\mathrm{cov}} := \lceil 5/\min(1, F_\infty)\rceil$, since $\mathcal{H}_N$ contains three functions pairwise $2/n$-separated in sup-norm for such n. Exact minimisation is Proposition C.12. If instead one uses Schmidt-Hieber (2020)'s Lemma 4, whose numerals bake in $\sigma = 1$, the rescaled version requires $\delta = \min(1/n, \sigma)$ — whence $32\delta F''' \le 32F'''/n$ with $F''' = \max(1, F_\infty, \sigma)$, and the extra threshold $n \ge \lceil 1/\sigma\rceil$. Only the first branch is used below.

E.2 The realised atom count

Lemma E.5 (Atom count)**.** *With* $\mathcal{E}_N$ *as in Lemma E.2 and* E *as in* (20)*,*

$$N \;<\; |\mathcal{E}_N| \;\le\; \mathsf{E}\,N.$$

E *depends on* (k, s, p) *only.*

Proof. Since $|J(\kappa)| = \prod_a \big(2^{\lfloor \kappa s_{\min}/s_a\rfloor} + \mathsf{m} + 1\big) > 2^{\|\kappa\|}$, the level $\kappa = K$ alone contributes more than $2^{\|K\|} = N$, giving the strict lower bound. For the upper bound put $c := s_{\min}/\bar{s} = s_{\min}\sum_a s_a^{-1} \in [1, k]$ and estimate the four pieces. *(i)* $2^A + \mathsf{m} + 1 \le 2^A(\mathsf{m}+2)$ for $A \ge 0$, so $|J(\kappa)| \le (\mathsf{m}+2)^k 2^{\|\kappa\|}$. *(ii)* $\|K\| \ge Kc - k$ and $\|\kappa\| \le \kappa c$ give $\|K\| - \|\kappa\| \ge (K-\kappa)c - k$, so $\sum_{\kappa\le K} 2^{\|\kappa\|} \le 2^{\|K\|}2^k \sum_{i\ge 0} 2^{-ic} \le 2^{k+1}N$ using $c \ge 1$. *(iii)* The coordinate with $s_a = s_{\min}$ contributes κ exactly while the others are nondecreasing, so $\|\kappa\| - \|K\| \ge \kappa - K$ and $n_\kappa \le N2^{-\nu(\kappa-K)} + 1$, whence $\sum_{\kappa=K+1}^{K^*} n_\kappa \le N(2^\nu - 1)^{-1} + (K^* - K)$. *(iv)* $K^* - K \le K/\nu + 1$ and $K \le \|K\| + k = \log_2 N + k \le N(1+k)$, so $K^* - K \le N(1+k)/\nu + 1$. Adding (i)–(iv) and taking the ceiling gives E. When $\delta_{\mathrm{ex}} = 0$ one has $K^* = K$ and the sums in (iii)–(iv) are *empty*: the two ν-terms are the value of an empty sum, not a convention. □

Remark E.6 (Scope of Lemma E.5)**.** The realised atom count exceeds N at every resolution. Consequently a width and sparsity budget of the form $W = NW_0$, $S = [(L-1)W_0^2 + 1]N$ does *not* host the network that Proposition E.7 constructs. This is a statement about which budgets the construction needs; it says nothing about whether some *other* construction could achieve the same approximation rate with smaller budgets, and it does not affect the asymptotic approximation rate $N^{-\bar{s}}$ itself, since (21) inflates width and sparsity by a structural factor only.

E.3 The approximation, normalised once

Proposition E.7 (Constructive approximation on the cube)**.** *There is a structural constant* $C^{\mathrm{syn}}_{\mathrm{app}} = C^{\mathrm{syn}}_{\mathrm{app}}(k, s, p, q)$ *such that for* every g *with* $\|g\|_{B^s_{p,q}([0,1]^k)} \le F$ *there is*

$$\Phi_g \;\in\; \Phi_k\big(L_{\mathrm{net}}, W_{\mathrm{net}}, S_{\mathrm{net}}, B_{\mathrm{ffn}}\big) \qquad \textit{with} \qquad \|g - \Phi_g\|_{L^2([0,1]^k, \mathrm{Leb})} \;\le\; C^{\mathrm{syn}}_{\mathrm{app}}\, F\, N^{-\bar{s}},$$

the budgets being exactly those of Definition C.6. The architecture is quantified before *the supremum over* g*: one quadruple serves the whole class, and only the element* Φ_g *depends on* g*. The budgets depend on neither* g *nor* S*.*

Proof. **Step 1 (normalise once).** Put $u := g/F$, so $\|u\|_{B^s_{p,q}} \le 1$. Everything below is done for u, and F is reinstated exactly once, in Step 6.

Step 2 (truncation, at the admissible resolution). Apply Lemma E.2 at $N^\dagger = 2^{\|K\|}$, giving $u_{N^\dagger}$ with $\|u - u_{N^\dagger}\|_{L^2} \le C_{\mathrm{tr}} N^{\dagger\,-\bar{s}} \le C_{\mathrm{tr}} N^{-\bar{s}}$ by Lemma C.4. *The rounding* $N \mapsto N^\dagger$ *therefore*

improves the error and costs no constant; what it inflates is the budgets, in (24), not the error. By Lemma E.5, $|\mathcal{E}_{N^\dagger}| \le \mathsf{E} N^\dagger$.

Step 3 (unit-ball coefficient bound, and why it is legitimate here). For every κ,

$$\|\kappa\| = \sum_a \left\lfloor \frac{\kappa s_{\min}}{s_a} \right\rfloor \le \kappa s_{\min} \sum_a \frac{1}{s_a} = \frac{\kappa s_{\min}}{\bar{s}},$$

$$\text{hence} \qquad w_\kappa = 2^{\kappa s_{\min} - \|\kappa\|/p} \;\ge\; 2^{\kappa s_{\min}\left(1-\frac{1}{p\bar{s}}\right)} \;\ge\; 1,$$

the last step because (12) says $p\bar{s} > 1$ (at $p=\infty$, $1/p=0$ and $w_\kappa \ge 1$ trivially). Since $\max_j |a_{\kappa,j}| \le (\sum_j |a_{\kappa,j}|^p)^{1/p} \le w_\kappa^{-1}\|(a)\|_b$, we get

$$\max_{(\kappa,j)\in\mathcal{E}_{N^\dagger}} |a_{\kappa,j}| \;\le\; \|(a)\|_b \;\le\; C_\alpha \|u\|_{B^s_{p,q}} \;\le\; C_\alpha. \tag{46}$$

It is essential that this is applied to u and not to g: (46) is a *unit-ball* bound, and mixing it with a radius-F target is exactly the double-normalisation this proof avoids.

Step 4 (replace each atom, and bound the error of the whole sum). Let $\check{M}$ be as in Lemma E.1 at accuracy

$$\epsilon_N := \frac{{N^\dagger}^{-\bar{s}}}{\log(2N^\dagger)} \tag{47}$$

and set $\check{M}^k_{\kappa,j}(t) := \check{M}\big((2^{\lfloor \kappa s_{\min}/s_a \rfloor} t_a - j_a)_a\big)$, the *same* affine map that defines $M^k_{\kappa,j}$, and $\check{u}_{N^\dagger} := \sum_{(\kappa,j)\in\mathcal{E}_{N^\dagger}} a_{\kappa,j}\check{M}^k_{\kappa,j}$. Both $M^k_{0,0}$ and $\check{M}$ vanish outside $[0,\mathsf{m}+1]^k$ (the former because a cardinal B-spline of order m does, the latter by (43)), so for fixed t and fixed κ the terms that can be nonzero are those with $t_a \in \big[j_a 2^{-\lambda_a}, (j_a+\mathsf{m}+1)2^{-\lambda_a}\big]$ for every a, where $\lambda_a := \lfloor \kappa s_{\min}/s_a \rfloor$. In each coordinate the admissible j_a form a set of at most $\mathsf{m}+2$ consecutive integers, so at most $(\mathsf{m}+2)^k$ pairs (κ, j) contribute at level κ. Summing over the K^*+1 levels present in $\mathcal{E}_{N^\dagger}$ and using (46) and (43),

$$\big\|u_{N^\dagger} - \check{u}_{N^\dagger}\big\|_{L^\infty} \;\le\; C_\alpha\,(\mathsf{m}+2)^k\,(K^*+1)\,\epsilon_N. \tag{48}$$

The number of atoms $|\mathcal{E}_{N^\dagger}|$ does not appear here; only the finite overlap does.

Step 5 (the level count, and the exact cancellation of the logarithm). Put $a := (1+\nu^{-1})\bar{s}/s_{\min}$. From $\|K\| \ge K s_{\min}/\bar{s} - k$ we get $K \le (\log_2 N^\dagger + k)\bar{s}/s_{\min}$, hence $K^*+1 \le \lceil K(1+\nu^{-1})\rceil + 1 \le a\log_2 N^\dagger + ak + 2$. With

$$C_K := \big[a(k+1)+2\big]/\log 2 \qquad \text{one has} \qquad K^*+1 \;\le\; C_K \log(2N^\dagger), \tag{49}$$

since $C_K \log(2N^\dagger) = [a(k+1)+2](1+\log_2 N^\dagger)$ and $[a(k+1)+2](1+\log_2 N^\dagger) - (a\log_2 N^\dagger + ak + 2) = (ak+2)\log_2 N^\dagger + a \ge 0$. Substituting (49) and (47) into (48),

$$\big\|u_{N^\dagger} - \check{u}_{N^\dagger}\big\|_{L^\infty} \;\le\; C_\alpha(\mathsf{m}+2)^k C_K \,\frac{\log(2N^\dagger)}{\log(2N^\dagger)}\, {N^\dagger}^{-\bar{s}} \;=\; C_\alpha(\mathsf{m}+2)^k C_K\, {N^\dagger}^{-\bar{s}}. \tag{50}$$

The $\log(2N^\dagger)$ built into ϵ_N cancels the level count *exactly*: no residual logarithm and no positive power of N survives.

Step 6 (reinstate F, once). Define $\Phi_g := F\,\check{u}_{N^\dagger}$. Since $\mathrm{Leb}([0,1]^k) = 1$ gives $\|\cdot\|_{L^2} \le \|\cdot\|_{L^\infty}$ there, Steps 2 and 5 combine to

$$\|g - \Phi_g\|_{L^2} = F\,\|u - \check{u}_{N^\dagger}\|_{L^2} \le F\Big[C_{\mathrm{tr}} + C_\alpha(\mathsf{m}+2)^k C_K\Big] {N^\dagger}^{-\bar{s}} \le C^{\mathrm{syn}}_{\mathrm{app}} F\, N^{-\bar{s}},$$

with $C^{\mathrm{syn}}_{\mathrm{app}} := C_{\mathrm{tr}} + C_\alpha(\mathsf{m}+2)^k C_K$. *$F$ multiplies exactly once*, in this step.

Step 7 (the constructed network lies in $\Phi_k(L_{\mathrm{net}}, W_{\mathrm{net}}, S_{\mathrm{net}}, B_{\mathrm{ffn}})$). Each $\check{M}^k_{\kappa,j}$ is $\check{M}$ precomposed with $t \mapsto D_\kappa t - j$, $D_\kappa := \mathrm{diag}(2^{\lambda_a})$. Absorbing this into the first layer turns $W^{(1)}t + b^{(1)}$ into $W^{(1)}D_\kappa t + (b^{(1)} - W^{(1)}j)$, so with $|j_a| \le 2^{\lambda_a} + \mathsf{m} \le 2^{K^*} + \mathsf{m}$,

$$\|W^{(1)}D_\kappa\|_{\max} \le B_0 2^{K^*}, \qquad \|b^{(1)} - W^{(1)}j\|_\infty \le B_0\big[1 + k\big(2^{K^*} + \mathsf{m}\big)\big]. \tag{51}$$

From $\|K\| \geq K s_{\min}/\bar{s} - k$, $2^K \leq (2^k N^\dagger)^{\bar{s}/s_{\min}}$ and $K^* \leq K(1+\nu^{-1})+1$, so

$$2^{K^*} \;\leq\; 2\left(2^k N^\dagger\right)^{(1+\nu^{-1})\bar{s}/s_{\min}} \;\leq\; 2\cdot 2^{kb} {N^\dagger}^b, \tag{52}$$

which is where the exponent b of (22) comes from. *Diagonal rescaling does not enlarge the weight support, but a bias may pass from zero to nonzero, so the sparsity pattern is not unchanged; the increase is at most W_0 nonzeros per block.* Padding each block to the common depth $L_{\text{net}} \geq L_0(\epsilon_N)$ (Lemma C.8, using Remark C.7) costs at most $(L_{\text{net}} - L_0)W_0 \leq (L_{\text{net}} - L_0)W_0^2$ further nonzeros, so each block has at most $L_0 W_0^2 + W_0 + (L_{\text{net}} - L_0)W_0^2 \leq L_{\text{net}} W_0^2 + W_0$ nonzeros and width W_0. Stacking the $|\mathcal{E}_{N^\dagger}| \leq \mathsf{E}N^\dagger$ blocks in parallel and joining them by a single linear readout with weights $a_{\kappa,j} W_{\check{M}}^{(L)}$ and bias $\big(\sum_{(\kappa,j)} a_{\kappa,j}\big) b_{\check{M}}^{(L)}$, and then applying Step 6's scaling by F to that last layer only,

$$\begin{aligned} \text{width} \leq \mathsf{E}N^\dagger W_0 = W_{\text{net}}, \qquad & \text{nnz} \leq \mathsf{E}N^\dagger\big(L_{\text{net}} W_0^2 + W_0\big) + \mathsf{E}N^\dagger + 1 \\ & = \mathsf{E}N^\dagger\big(L_{\text{net}} W_0^2 + 1\big) + W_{\text{net}} + 1 \;\leq\; S_{\text{net}}, \end{aligned} \tag{53}$$

the slack in (53) being exactly W_{net}. For the magnitudes, collect (51), (52), (46), $|b_{\check{M}}^{(L)}| \leq B_0$ and $\|W_{\check{M}}^{(L)}\|_\infty \leq B_0$: the classes to cover are the first-layer weights $\leq B_0 2^{K^*}$, the first-layer biases $\leq B_0[1 + k(2^{K^*} + \mathsf{m})]$, the interior parameters $\leq B_0$, the readout weights $\leq \max(1,F) C_\alpha B_0$, and the composite output bias $\leq \max(1,F)\mathsf{E}N^\dagger C_\alpha B_0$. All five are at most $C_B(2N^\dagger)^b \leq B_{\text{ffn}}$ provided

$$C_B \;\geq\; \max\Big\{1,\;\; B_0\big[1 + k\mathsf{m} + 2k\,2^{kb}\big],\;\; \max(1,F)\, C_\alpha B_0\,(1+\mathsf{E})\Big\}, \tag{54}$$

which we take as the definition of C_B (any larger value is admissible). Indeed $b \geq 1$ gives $(2N^\dagger)^b \geq 2{N^\dagger}^b$ and $(2N^\dagger)^b \geq 2N^\dagger$, so the first-layer terms are covered by the middle entry of (54) through (52), and the readout and the output bias — the latter growing *linearly* in $N^\dagger$, which is precisely why $b \geq 1$ is required and is not slack — by the last entry. Every quantity in (54) is structural, and none reads D, n, N, S or g. □

Lemma E.8 (Transfer to $L^2(P_X)$)**.** *For every $\psi, g : [0,1]^k \to \mathbb{R}$ and every S with $|S| = k$, $\|\psi((\cdot)_S) - g((\cdot)_S)\|^2_{L^2(P_X)} \leq c_+ \|\psi - g\|^2_{L^2([0,1]^k,\text{Leb})}$.*

Proof. Write $x = (v,z)$ with $v = x_S$. By Tonelli the law of X_S has Lebesgue density $p_{X_S}(v) = \int_{[0,1]^{D-k}} p_X(v,z)\,dz \leq c_+ \,\text{Leb}([0,1]^{D-k}) = c_+$, the last equality being exact and the reason the bound is free of D. No product structure or independence of the coordinates of P_X is used, and the lower bound c_- is not used at all. □

E.4 The finite-temperature witness

Lemma E.9 (Whitened witness and recovery)**.** *Fix $S = \{i_1 < \cdots < i_k\}$, write $C_S := [c_{i_1}, \ldots, c_{i_k}] \in \mathbb{R}^{m\times k}$ and $G_S := C_S^\top C_S$, and set*

$$q_S := \tau_{N,D}\, C_S G_S^{-1} \mathbf{1}_k, \qquad K = V = I_m, \qquad R_S := G_S^{-1} C_S^\top. \tag{55}$$

Then $\theta_s^{\text{wit}} := (q_S, I_m, I_m, R_S)$ satisfies $\|\theta_s^{\text{wit}}\|_\infty \leq B_{\text{sel}}$ for every S, and

$$\begin{aligned} \sup_{x\in[0,1]^D} \big\|h_{\theta_s^{\text{wit}}}(x) - x_S\big\|_2 \;&\leq\; \Big(\tfrac{1}{\sqrt{k}} + \tfrac{2}{\sqrt{3}}\Big)(D-k) e^{-2\tau_{N,D}/3} \\ &\leq\; C_{\text{w}}\,(D-k)\,e^{-\tau_{N,D}/2}, \qquad C_{\text{w}} := 1 + \tfrac{2}{\sqrt{3}}. \end{aligned} \tag{56}$$

Consequently, for every $g \in \mathcal{G}$ and every S,

$$\inf_{T\in\mathcal{H}_N} \big\|T - g((\cdot)_S)\big\|_{L^2(P_X)} \;\leq\; \Big(\sqrt{c_+}\, C_{\text{app}}^{\text{syn}} F + \frac{C_{\text{w}}}{6k}\Big) N^{-\bar{s}}. \tag{57}$$

Proof. **(a) One S-independent budget.** G_S has unit diagonal and off-diagonal entries of modulus at most $\mu \leq 1/(4k)$, so by Gershgorin every eigenvalue lies within $(k-1)\mu \leq 1/4$ of 1, i.e. $\text{spec}(G_S) \subseteq [3/4, 5/4]$ and $\|G_S^{-1}\|_{\text{op}} \leq 4/3$. Hence

$$\begin{aligned} \|q_S\|_\infty &\leq \tau \|C_S G_S^{-1}\mathbf{1}_k\|_2 \\ &= \tau\sqrt{\mathbf{1}^\top G_S^{-1}\mathbf{1}} \leq \tau\sqrt{4k/3} \leq 2\sqrt{k}\,\tau. \end{aligned}$$

For the decoder, use its pseudo-inverse structure rather than multiplying separate operator-norm bounds:

$$R_S R_S^\top = G_S^{-1} C_S^\top C_S G_S^{-1} = G_S^{-1}.$$

Since $\|A\|_{\mathrm{op}}^2 = \|AA^\top\|_{\mathrm{op}}$ for any real matrix A,

$$\|R_S\|_{\max} \le \|R_S\|_{\mathrm{op}} = \sqrt{\|G_S^{-1}\|_{\mathrm{op}}} \le \frac{2}{\sqrt{3}},$$

and $\|K\|_{\max} = \|V\|_{\max} = 1$. All four are at most $B_{\mathrm{sel}} = 2\sqrt{k}\max\{1, \tau_{N,D}\}$, *for every* S, because every estimate descends from the single scalar μ of (15), which is a maximum over *all* pairs in $[D]$. There is no S-dependent budget anywhere.

(b) The active logits are exactly equal — this is what the whitening buys. Since $C_S^\top c_{i_a} = G_S e_a$,

$$z_{i_a} = \langle q_S, c_{i_a}\rangle = \tau\, \mathbf{1}_k^\top G_S^{-1} G_S e_a = \tau\, \mathbf{1}_k^\top e_a = \tau \qquad \textit{exactly}, \text{ for every } a \in [k], \tag{58}$$

so the intra-S spread is *zero*. Off support, writing $w_i := C_S^\top c_i$ with $\|w_i\|_2 \le \sqrt{k}\,\mu$, $|z_i| = \tau|\mathbf{1}^\top G_S^{-1} w_i| \le \tau \|G_S^{-1}\mathbf{1}\|_2 \|w_i\|_2 \le \tau \cdot \frac{4}{3}\sqrt{k} \cdot \sqrt{k}\mu \le \tau \cdot \frac{4}{3}k \cdot \frac{1}{4k} = \tau/3$. The margin is therefore at least $2\tau/3 \ge \tau/2$.

(c) Recovery. By (58) the softmax assigns *the same* weight to every $i \in S$: with $\epsilon_{\mathrm{off}} := \sum_{i \notin S} \alpha_i$,

$$\alpha_{i_a} = \frac{1 - \epsilon_{\mathrm{off}}}{k} \quad (a \in [k]), \qquad \epsilon_{\mathrm{off}} \le \frac{(D-k)e^{\tau/3}}{k e^{\tau}} = \frac{D-k}{k} e^{-2\tau/3}.$$

At $K = V = I_m$ the forward map (28) is $h = kR_S \sum_i \alpha_i x_i c_i$, and $R_S C_S = G_S^{-1} C_S^\top C_S = I_k$ *exactly*, so

$$h(x) = (1 - \epsilon_{\mathrm{off}})\, x_S \;+\; kR_S \sum_{i \notin S} \alpha_i x_i c_i. \tag{59}$$

With $x \in [0,1]^D$ and $\|c_i\|_2 = 1$, the first term differs from x_S by at most $\epsilon_{\mathrm{off}}\|x_S\|_2 \le \sqrt{k}\,\epsilon_{\mathrm{off}}$ and the second has norm at most $k\|R_S\|_{\mathrm{op}}\epsilon_{\mathrm{off}} \le (2k/\sqrt{3})\epsilon_{\mathrm{off}}$, whence $\|h - x_S\|_2 \le (\sqrt{k} + 2k/\sqrt{3})\epsilon_{\mathrm{off}}$, which is (56) after substituting the bound on ϵ_{off} and $e^{-2\tau/3} \le e^{-\tau/2}$.

(d) The approximation statement, and why it decays. Take Φ_g from Proposition E.7 and $T := \mathcal{C}_{F_\infty} \circ \Phi_g \circ \Pi \circ h_{\theta_s^{\mathrm{wit}}} \in \mathcal{H}_N$. Since $\mathcal{C}_{F_\infty}$ is 1-Lipschitz and $\|g((\cdot)_S)\|_\infty \le F_\infty$, and since Π is 1-Lipschitz and does not increase the distance to $x_S \in [0,1]^k$,

$$\|T - g((\cdot)_S)\|_{L^2(P_X)} \;\le\; \underbrace{\|\Phi_g((\cdot)_S) - g((\cdot)_S)\|_{L^2(P_X)}}_{\le \sqrt{c_+} C_{\mathrm{app}}^{\mathrm{syn}} F N^{-\bar{s}} \text{ by Lem. E.8, Prop. E.7}} \;+\; \underbrace{\bar{L} \sup_x \|\Pi(h_{\theta_s^{\mathrm{wit}}}(x)) - x_S\|_2}_{\le \bar{L}\, C_{\mathrm{w}}(D-k) e^{-\tau_{N,D}/2}},$$

$\bar{L} \le \bar{L}$ being the Lipschitz constant of Φ_g on $[0,1]^k$ by Lemma E.11. The schedule (23) gives $e^{-\tau_{N,D}/2} \le \left(e + 6kD\bar{L}N^{\dagger\bar{s}}\right)^{-1} \le \left(6kD\bar{L}N^{\dagger\bar{s}}\right)^{-1}$, so the second term is at most

$$\bar{L} C_{\mathrm{w}}(D-k) \cdot \frac{1}{6kD\bar{L}N^{\dagger\bar{s}}} \;=\; \frac{C_{\mathrm{w}}}{6k} \cdot \frac{D-k}{D} \cdot N^{\dagger -\bar{s}} \;\le\; \frac{C_{\mathrm{w}}}{6k} N^{-\bar{s}},$$

using $N^\dagger \ge N$ and $(D-k)/D < 1$. *The factors $\bar{L}$ and $6kD$ cancel exactly, which is why no constant here carries D and why the second term carries the same power $N^{-\bar{s}}$ as the first.* Adding the two gives (57). □

Remark E.10 (Whitening is not cosmetic)**.** Replacing q_S by the raw code sum $\frac{\tau}{2}\sum_{i \in S} c_i$ destroys (58): the active logits become $\frac{\tau}{2}$ times the row sums of G_S, which differ in general, and the softmax then concentrates on the largest of them as τ grows. The following unit-vector family exhibits the mechanism. *It satisfies the coherence constraint* (15) *only, and is not an instance of the sign codebook of Definition C.1*: take $k = 3$, $D = 6$, $c_1 = e_1$, $c_2 = \frac{1}{12}e_1 + \sqrt{143}/12\, e_2$, $c_3 = e_3$ and three further orthonormal codes, all padded to $\mathbb{R}^m$. Every vector has unit norm and $\mu = 1/12 = 1/(4k)$ exactly, so (15) holds; but their coordinates are not all of modulus $m^{-1/2}$, so (14) fails and nothing here asserts that the same failure occurs for an admissible codebook. *Only the query is replaced; the decoder is held at R_S of* (55). With that query the active logits are $(\frac{13\tau}{24}, \frac{13\tau}{24}, \frac{\tau}{2})$; at $\tau = 40$ the inactive mass is 5.3×10^{-10} while $\alpha_3 = 0.086$, and taking $x_3 = 1$ with the other coordinates zero

gives recovery error 0.741, against the value $C_{\mathrm{w}}(D-k)e^{-\tau/2} = 1.3 \times 10^{-8}$ that the bound (56) takes at this τ. *A small inactive mass does not by itself give recovery; equality of the active weights does, and that is exactly what G_S^{-1} supplies.* Raising τ makes this failure worse, not better. What the example refutes is the raw code-sum query *with the decoder held at R_S*; it does not show that every unwhitened query fails, nor that a decoder redesigned for a given unwhitened query fails. It is an illustration only: no step in the proof of Lemma E.9 uses it.

E.5 Metric entropy of $\mathcal{H}_N$

Lemma E.11 (Parameter-to-function Lipschitz constant)**.** *Let $L \geq 2$, $W \geq 1$, $B \geq 1$, $d \leq W$ and $A \subseteq \{t \in \mathbb{R}^d : \|t\|_\infty \leq 1\}$. For any two parameter vectors of $\Phi_d(L, W, S, B)$ with $\|\theta - \theta'\|_\infty \leq \rho$,*

$$\sup_{x \in A} \big| f_\theta(x) - f_{\theta'}(x) \big| \;\leq\; K_{\mathrm{par}}\, \rho, \qquad K_{\mathrm{par}} := L\,(B \vee 1)^{L-1}(W+1)^L. \tag{60}$$

In particular $\sup_{x \in A} |f_\theta(x) - f_\theta(x')| \leq (W(B \vee 1))^L \|x - x'\|_\infty$, so the Lipschitz constant of any member of $\Phi_k(L_{\mathrm{net}}, W_{\mathrm{net}}, S_{\mathrm{net}}, B_{\mathrm{ffn}})$ on $[0,1]^k$ is at most $\bar{L}$ of (23).

Proof. Write $h^{(0)} = x$, $h^{(l)} = \eta(W^{(l)} h^{(l-1)} + b^{(l)})$ for $l \leq L-1$, and $f_\theta = W^{(L)} h^{(L-1)} + b^{(L)}$. Every row of every $W^{(l)}$ has at most W entries — at most $d \leq W$ in the first layer, which is where $d \leq W$ is used — each of modulus at most B, so $\|W^{(l)}\|_{\infty \to \infty} \leq WB$. Put $\gamma := B(W+1)$ and $M_l := \max\{1, \|h^{(l)}\|_{\infty, A}\}$. Then $\|h^{(l)}\|_\infty \leq WB \|h^{(l-1)}\|_\infty + B \leq \gamma M_{l-1}$ and $\gamma \geq 1$, so $M_l \leq \gamma M_{l-1}$ and hence $M_l \leq \gamma^l$ (as $M_0 \leq 1$). With $\Delta_l := \|h^{(l)}(\theta) - h^{(l)}(\theta')\|_{\infty, A}$ and $\Delta_0 = 0$, and using that η is 1-Lipschitz,

$$\begin{aligned} \Delta_l \;&\leq\; \underbrace{\|W^{(l)}(\theta)\|_{\infty \to \infty}}_{\leq WB} \Delta_{l-1} \;+\; \underbrace{\big\| \big(W^{(l)}(\theta) - W^{(l)}(\theta')\big) h^{(l-1)}(\theta') \big\|_\infty}_{\leq W \rho M_{l-1}} \\ &\quad + \underbrace{\big\| b^{(l)}(\theta) - b^{(l)}(\theta') \big\|_\infty}_{\leq \rho} \;\leq\; WB\, \Delta_{l-1} + \rho(W+1)\gamma^{l-1}, \end{aligned}$$

where $W\rho M_{l-1} + \rho \leq \rho(W+1)M_{l-1}$ was used. Since $WB \leq \gamma$, unrolling gives $\Delta_l \leq \rho(W+1) \sum_{i=0}^{l-1} (WB)^{l-1-i} \gamma^i \leq \rho(W+1)\, l\, \gamma^{l-1}$, and at $l = L$ (the last layer carries no η and obeys the same recursion) $\Delta_L \leq \rho L(W+1)\big(B(W+1)\big)^{L-1} = \rho L B^{L-1}(W+1)^L$. The statement about $x \mapsto f_\theta(x)$ follows from the same recursion with $\rho = 0$ and $\Delta_0 = \|x - x'\|_\infty$. □

Remark E.12 (The exponent L is not slack)**.** Replacing $(W+1)^L$ by $(W+1)^{L-1}$ in (60) makes the statement false. Take $d = 1$, $L = 3$, $W = 20$, $B = 1$ and let all 481 weights and biases equal $t \in [0,1]$; at $x = 1$ every η is active, so $f_t(1) = 800t^3 + 20t^2 + t$. At $t = 1$ and $t = 999/1000$ the parameter distance is $1/1000$ while $\big(f_1(1) - f_{999/1000}(1)\big)/(1/1000) = 1524113/625 > 1323 = 3 \cdot 21^2$, whereas (60) gives $3 \cdot 21^3 = 27783$.

Lemma E.13 (Conservative sup-norm cover of Φ_d)**.** *Let $L \geq 3$, $W \geq 20$, $B \geq 1$, $d \leq W$, $A \subseteq \{t \in \mathbb{R}^d : \|t\|_\infty \leq 1\}$, and let*

$$T := Wd + W + (L-2)(W^2 + W) + (W+1) \;\leq\; 2LW^2 \tag{61}$$

be the total number of parameter slots, every bias included and the scalar output bias among them. *Then for every $0 < \epsilon \leq BK_{\mathrm{par}}$,*

$$\begin{aligned} &\log \mathcal{N}\big(\epsilon, \Phi_d(L, W, S, B), \|\cdot\|_{L^\infty(A)}\big) \\ &\leq\; \log \sum_{i=0}^{\min(S,T)} \binom{T}{i} \;+\; S \log\Big(1 + \frac{BK_{\mathrm{par}}}{\epsilon}\Big) \\ &\leq\; S \log \frac{6LW^2 BK_{\mathrm{par}}}{\epsilon}. \end{aligned} \tag{62}$$

Proof. Fix a sparsity pattern P with $|P| \leq S$; on it the free parameters range over the box $[-B, B]^{|P|}$. An ℓ^∞ net of $[-B, B]$ of radius $\rho := \epsilon / K_{\mathrm{par}}$ has at most $\lceil B/\rho \rceil \leq 1 + BK_{\mathrm{par}}/\epsilon$ points, so the

product grid over P has at most $(1 + BK_{\mathrm{par}}/\epsilon)^S$ points, and by Lemma E.11 every f_θ with pattern P lies within ϵ of a grid centre in $\|\cdot\|_{L^\infty(A)}$. Summing over patterns and taking logarithms gives the first bound. For the second, $\sum_{i\le\min(S,T)} \binom{T}{i} \le (T+1)^S \le (2LW^2+1)^S \le (3LW^2)^S$ and $1 + BK_{\mathrm{par}}/\epsilon \le 2BK_{\mathrm{par}}/\epsilon$ for $\epsilon \le BK_{\mathrm{par}}$. Finally $T \le (L-1)(W^2+W) + W + 1 \le 2LW^2$ for $W \ge 2$, using $d \le W$. □

Remark E.14 (Attribution)**.** The *argument* of Lemma E.13 — sparsity patterns times a grid on the nonzero parameters, with a layerwise propagation constant, on inputs of sup-norm at most one — is that of Suzuki & Nitanda (2021, App. G, Lemma 6), and the input-domain restriction is forced by their proof. The *statement* above is not theirs: their printed form counts $1/\rho$ grid points per free parameter although those parameters range over $[-B, B]$, which needs $\lceil B/\rho \rceil$. We therefore state Lemma E.13 as our own conservative lemma and credit the source for the argument it supplies. It may not be attributed to Schmidt-Hieber (2020)'s Lemma 5, whose class hard-codes all parameters at most one *in its definition*, so that a class with $B > 1$ is not a subset of it and covering numbers are not monotone in that direction.

Lemma E.15 (Selector Lipschitz constant)**.** *On the box $\|\theta_s\|_\infty \le B_{\mathrm{sel}}$, the map $\theta_s \mapsto h_{\theta_s}$ satisfies $\sup_{x\in[0,1]^D} \|h_{\theta_s}(x) - h_{\theta'_s}(x)\|_\infty \le L_{\mathrm{sel}} \|\theta_s - \theta'_s\|_\infty$ with $L_{\mathrm{sel}} := 6k^{3/2}m^3 B_{\mathrm{sel}}^3$.*

Proof. Write $\rho := \|\theta_s - \theta'_s\|_\infty$, $B := B_{\mathrm{sel}}$, and recall $\|c_i\|_2 = 1$.

(i) Logits. Entrywise bounds give $\|q\|_2 \le \sqrt{m}B$ and $\|K\|_{\mathrm{op}} \le \|K\|_F \le mB$, and likewise $\|q - q'\|_2 \le \sqrt{m}\rho$, $\|K - K'\|_{\mathrm{op}} \le m\rho$. Hence

$$|z_i - z'_i| \le |\langle q - q', Kc_i\rangle| + |\langle q', (K - K')c_i\rangle| \le \sqrt{m}\rho \cdot mB + \sqrt{m}B \cdot m\rho = 2m^{3/2}B\rho,$$

so $\|z - z'\|_\infty \le 2m^{3/2}B\rho$. *The factor $\sqrt{m}$ is not optional: it comes from $\|q\|_2 \le \sqrt{m}B$ under an entrywise budget.*

(ii) Pooling. softmax is 2-Lipschitz from $(\mathbb{R}^D, \|\cdot\|_\infty)$ to $(\mathbb{R}^D, \|\cdot\|_1)$, so $\|\alpha - \alpha'\|_1 \le 4m^{3/2}B\rho$. Put $b(\theta_s) := \sum_i \alpha_i x_i c_i$. Since $x_i \in [0,1]$ and $\|c_i\|_2 = 1$, b is a sub-convex combination of unit vectors, so $\|b\|_2 \le 1$ and $\|b - b'\|_2 \le \sum_i |\alpha_i - \alpha'_i|\,|x_i|\,\|c_i\|_2 \le \|\alpha - \alpha'\|_1 \le 4m^{3/2}B\rho$.

(iii) Decoder, all three variations. By (28), $h = kRVb$, so

$$h - h' = k\big[(R - R')Vb + R'(V - V')b + R'V'(b - b')\big],$$

and none of the three may be dropped. With $\|R\|_{\mathrm{op}} \le \|R\|_F \le \sqrt{km}B$, $\|V\|_{\mathrm{op}} \le mB$, $\|R - R'\|_{\mathrm{op}} \le \sqrt{km}\rho$ and $\|V - V'\|_{\mathrm{op}} \le m\rho$,

$$\begin{aligned}\|h - h'\|_2 &\le k\Big[\sqrt{km}\,\rho \cdot mB \;+\; \sqrt{km}\,B \cdot m\rho \;+\; \sqrt{km}\,B \cdot mB \cdot 4m^{3/2}B\rho\Big] \\ &= \big[2k^{3/2}m^{3/2}B + 4k^{3/2}m^3B^3\big]\rho \;\le\; 6k^{3/2}m^3B^3\rho,\end{aligned}$$

using $m \ge 1$ and $B \ge 1$. Finally $\|\cdot\|_\infty \le \|\cdot\|_2$ on $\mathbb{R}^k$. The bound holds on the *entire* parameter box, not only at the witness, and is not claimed to be sharp. □

Proposition E.16 (Class entropy and the two envelopes)**.** *For $0 < \epsilon \le 1$,*

$$\begin{aligned}\log \mathcal{N}\big(\epsilon, \mathcal{H}_N, \|\cdot\|_{L^\infty([0,1]^D)}\big) \;&\le P_{\mathrm{sel}} \log\Big(1 + \frac{2B_{\mathrm{sel}}L_{\mathrm{sel}}\bar{L}}{\epsilon}\Big) \\ &\quad + \; S_{\mathrm{net}} \log \frac{12\,L_{\mathrm{net}}W_{\mathrm{net}}^2 B_{\mathrm{ffn}}\,K_{\mathrm{par}}(L_{\mathrm{net}}, W_{\mathrm{net}}, B_{\mathrm{ffn}})}{\epsilon}.\end{aligned} \tag{63}$$

At $\epsilon = 1/n$ and $2 \le N \le n$ the second summand is at most $C_{\mathrm{ffn}} N \log^3(2n)$ with C_{ffn} structural, and the first is $P_{\mathrm{sel}}\,\Xi_{N,n,D}$ with Ξ as in (32). Moreover $\bar{\Xi}_{n,D} \le \Xi_{\mathrm{safe}}(n, D)$ for a structural C_Ξ.

Proof. Composition. Cover the selector box at radius $\epsilon/(2L_{\mathrm{sel}}\bar{L})$ (a product grid, at most $\big(1 + 2B_{\mathrm{sel}}L_{\mathrm{sel}}\bar{L}/\epsilon\big)^{P_{\mathrm{sel}}}$ points) and the FFN at radius $\epsilon/2$ on $[0,1]^k$ by Lemma E.13 with $d = k \le W_{\mathrm{net}}$ — legitimate because Π places the FFN's argument in $[0,1]^k \subseteq \{\|t\|_\infty \le 1\}$. For T, T' with selector parameters within $\epsilon/(2L_{\mathrm{sel}}\bar{L})$ and FFNs within $\epsilon/2$, $\|T - T'\|_\infty \le \bar{L}L_{\mathrm{sel}} \cdot \epsilon/(2L_{\mathrm{sel}}\bar{L}) + \epsilon/2 = \epsilon$,

using that $\mathcal{C}_{F_\infty}$ and Π are 1-Lipschitz and Lemma E.11 for the Lipschitz constant of the FFN in its argument. *Two different domains appear in one chain and both are written: the FFN is covered on* $[0,1]^k$*, the class on* $[0,1]^D$.

The exponent 3. Insert the schedule of Remark C.9 with $\ell := \log(2N) \le \Lambda := \log(2n)$. The bracket of the second summand is $\log 12 + 2\log L_{\text{net}} + 2\log W_{\text{net}} + L_{\text{net}} \log(B_{\text{ffn}} \vee 1) + L_{\text{net}} \log(W_{\text{net}}+1) + \log n + O(1)$; the two products are $O(\ell^2)$ with structural coefficients $C_L b$ and C_L respectively, and everything else is $O(\ell+\Lambda)$. Hence the second summand is at most $C_S N \ell \big[C_L(b+1)\ell^2 + O(\ell+\Lambda)\big] \le C_{\text{ffn}} N \Lambda^3$. The one extra power of $(B_{\text{ffn}} \vee 1)$ that Lemma E.11 carries relative to the source's printed form enters only as one more $L_{\text{net}} \log B_{\text{ffn}} = O(\ell^2)$ term and does not change the exponent.

The envelope. By (23), $\log \bar{L} = L_{\text{net}} \log(W_{\text{net}}(B_{\text{ffn}} \vee 1)) = O(\ell^2)$ and $\log B_{\text{sel}} \le \log(2\sqrt{k}) + \log \max(1, \tau_{N,D})$ with $\tau_{N,D} = 2\log(e + 6kD\bar{L}N^{\dagger^{\bar{s}}}) = O(\log(2D) + \ell^2)$; also $\log L_{\text{sel}} = O(\log(2k) + \log m + \log B_{\text{sel}})$ and $\log m = O(\log k + \log\log(2D))$. Therefore $\Xi_{N,n,D} = \log(1 + 2B_{\text{sel}} L_{\text{sel}} \bar{L} n) = O\big(\ell^2 + \Lambda + \log\log(e^e D) + \log(2k)\big)$ uniformly in $2 \le N \le n$, because $\ell \le \Lambda$; taking the supremum over N and collecting the structural coefficients into C_Ξ gives $\bar{\Xi}_{n,D} \le \Xi_{\text{safe}}(n, D)$. *This is the step that makes the second term of* (34) *free of* N*, and it must be taken before* N *is optimised.* Note that D enters Ξ only through $\log(2D)$ inside τ, i.e. through $\log\log(e^e D)$ after the outer logarithm, and through P_{sel}; *no* $N^a \log D$ *term with* $a > 0$ *appears anywhere*. □

E.6 The rate

Proof of Theorem D.1. Let $n \ge n_0 := \max\{2, n_{\text{cov}}, n_\dagger\}$, where

$$n_\dagger := \min\big\{n \ge 11 :\ n \ge 2^{2\bar{s}+1} \log^3(2n)\big\}$$

is well defined, and the condition is upward closed because $\frac{d}{dn}\big[n/\log^3(2n)\big] > 0$ for $n > e^3/2$. Apply Lemma E.3 to $\mathcal{T} = \mathcal{H}_{N_n}$ with $F' = \max(1, F_\infty)$ and $\delta = 1/n$ (Remark E.4). For the approximation term, take any $f = g((\cdot)_S) \in \mathcal{F}$ and use *exactly* the display (57) of Lemma E.9, whose right-hand side is $C_{\text{U}} N^{-\bar{s}}$ with

$$C_{\text{U}} := \sqrt{c_+}\, C_{\text{app}}^{\text{syn}} F + \frac{C_{\text{w}}}{6k}; \tag{64}$$

squaring gives $C_{\text{U}}^2 N^{-2\bar{s}}$, *both* of whose summands decay — no non-decaying term is absorbed into a decaying constant. For the entropy term use Proposition E.16. This yields

$$\sup_{f \in \mathcal{F}} \mathbb{E}\|\hat{f}_n - f\|^2_{L^2(P_X)} \ \le\ C_{\text{o}} \Big[C_{\text{U}}^2 N_n^{-2\bar{s}} + \frac{(F'^2 + \sigma^2)\big(P_{\text{sel}} \bar{\Xi}_{n,D} + C_{\text{ffn}} N_n \log^3(2n)\big)}{n} + \frac{F' + \sigma}{n} \Big]. \tag{65}$$

Because $P_{\text{sel}} \bar{\Xi}_{n,D}$ is *constant in* N, optimising (65) over N sees exactly two terms, $G(N) = AN^{-2\bar{s}} + \mathcal{B} N \log^3(2n)/n$ with $A = C_{\text{U}}^2$ and $\mathcal{B} = (F'^2 + \sigma^2) C_{\text{ffn}}$. G is strictly convex, and $G'(N_*) = 0$ gives $\mathcal{B} \log^3(2n) N_*/n = 2\bar{s} A N_*^{-2\bar{s}}$, whence

$$G(N_*) = A(1 + 2\bar{s}) N_*^{-2\bar{s}} = A(1+2\bar{s}) \Big(\tfrac{2\bar{s}A}{\mathcal{B}}\Big)^{-\alpha} n^{-\alpha} \log^{3\alpha}(2n), \tag{66}$$

so the logarithmic exponent is $3\alpha = 6\bar{s}/(2\bar{s}+1)$ — *derived*, and strictly below 3. At $n \ge n_0$ the choice (31) has $t_n := (n/\log^3 2n)^{1/(2\bar{s}+1)} \ge 2$ and $N_n = \lfloor t_n \rfloor \ge t_n/2$, so $N_n^{-2\bar{s}} \le 2^{2\bar{s}} n^{-\alpha} \log^{3\alpha}(2n)$ and $N_n \log^3(2n)/n \le n^{-\alpha} \log^{3\alpha}(2n)$; substituting into (65) gives (34) with

$$C_0^{\text{main}} := C_{\text{o}}\big[2^{2\bar{s}} C_{\text{U}}^2 + (F'^2 + \sigma^2) C_{\text{ffn}}\big], \qquad C_1 := C_{\text{o}}(F'^2 + \sigma^2), \qquad C_2 := C_{\text{o}}(F' + \sigma).$$

For $2 \le n < n_0$ the clip gives $\|\hat{f}_n - f\|_\infty \le 2F_\infty$ surely, so the risk is at most

$$4F_\infty^2 \ \le\ 4F_\infty^2\, n_0^\alpha\, \frac{n^{-\alpha} \log^{3\alpha}(2n)}{\log^{3\alpha} 4}, \qquad \text{and taking} \qquad C_0 := \max\Big\{C_0^{\text{main}},\ \frac{4F_\infty^2 n_0^\alpha}{\log^{3\alpha} 4}\Big\}$$

covers both regimes. *On* $2 \le n < n_0$ *the bound therefore carries no rate content.* Finally C_0 is free of D: every ingredient reads $(k, s, p, q, F, F_\infty, \sigma, c_+)$ only, and the D inside $\tau_{N,D}$ cancels against the $6kD$ of Lemma E.9. The D-carrying objects are the *counts and budgets* $m, P_{\text{sel}}, B_{\text{sel}}, \tau_{N,D}$, never the constants. □

F Proof of Theorem D.2

F.1 A finite-level orthonormal family inside the target class

Definition F.1 (Atoms). Fix a profile $\eta \in C^{r_{\mathrm{diff}}}(\mathbb{R})$ supported in $[0,1]$ with $\int \eta = 0$ and $\int \eta^2 = 1$, and set $A := \max_{0\le i\le r_{\mathrm{diff}}} \|\eta^{(i)}\|_\infty$, $C_\eta := \|\eta\|_\infty \ge 1$. For a real level $u \ge 0$ put

$$j_a := \lfloor u\,\bar{s}/s_a \rfloor \ (a \le k), \qquad Q_u := \prod_{a=1}^{k} 2^{j_a}, \qquad \Lambda_u := \prod_{a=1}^{k} \{0,1,\dots,2^{j_a}-1\}, \tag{67}$$

and for $S = \{i_1 < \cdots < i_k\}$ and $\ell = (\ell_1,\dots,\ell_k) \in \Lambda_u$,

$$\psi_{S,\ell}(x) := \prod_{a=1}^{k} 2^{j_a/2}\,\eta\big(2^{j_a}x_{i_a} - \ell_a\big). \tag{68}$$

Because $\sum_a u\bar{s}/s_a = u$ and each floor removes less than one,

$$2^{\,u-k} \;<\; Q_u \;\le\; 2^{\,u}. \tag{69}$$

The floors are never removed. Only the lower half of (69) is used in the level optimisation of Section F.4, and the upper half only in the sup-norm check (76); the identification $Q_u = 2^u$ is nowhere written.

Lemma F.2 (Orthonormality). *The family $\{\psi_{S,\ell}\}_{|S|=k,\ \ell\in\Lambda_u}$ is orthonormal in $L^2([0,1]^D, \mathrm{Leb})$, and $J_u := |\{(S,\ell)\}| = M\,Q_u$.*

Proof. For a fixed S, distinct ℓ give products of η's with disjoint supports in at least one coordinate, hence orthogonality; and $\int_{[0,1]} \big(2^{j/2}\eta(2^j t - \ell)\big)^2 dt = \int \eta^2 = 1$ gives normalisation, the remaining $D-k$ coordinates integrating to $\mathrm{Leb}([0,1]^{D-k}) = 1$. For $S \ne S'$ pick $i \in S \setminus S'$; the integrand factorises and the i-th factor is $\int_{[0,1]} 2^{j_a/2}\eta(2^{j_a}t - \ell_a)\,dt = 0$ because $\int \eta = 0$. □

Proposition F.3 (Membership in the bounded class). *Write $g_\theta := \sum_{\ell\in\Lambda_u} \theta_\ell \prod_a 2^{j_a/2}\eta(2^{j_a}\cdot_a - \ell_a)$ for the k-variate profile attached to one support. There are structural constants $C_{\mathrm{B}} = C_{\mathrm{B}}(k,s,\eta)$, free of p and q, such that*

$$\|g_\theta\|_{B^s_{p,q}([0,1]^k)} \;\le\; C_{\mathrm{B}}\,2^{u\bar{s}}\,Q_u^{1/2-1/p}\,\|\theta\|_{\ell^p}, \qquad \|g_\theta\|_{L^\infty} \;=\; C_\eta^k\,Q_u^{1/2}\,\|\theta\|_{\ell^\infty}. \tag{70}$$

Proof, and where the joint modulus is used. The L^p norm is immediate from disjoint supports:

$$\|g_\theta\|_{L^p} = Q_u^{1/2-1/p}\,\|\eta\|_{L^p}^k\,\|\theta\|_{\ell^p},$$

and likewise for the sup-norm, with equality because the supports are disjoint and η attains C_η. For the seminorm (11) fix $v \ge 0$ and $t = (2^{-v/s_1},\dots,2^{-v/s_k})$ and split.

Small scales, $v \le u\bar{s}$. Use $\|\Delta_h^r g\|_{L^p} \le 2^r \|g\|_{L^p}$, giving a geometric sum in v that contributes $O(2^{u\bar{s}})$ after the ℓ^q norm.

Large scales, $v > u\bar{s}$. Here the joint modulus is essential: Δ_h^r is taken along *one* vector h with $|h_a| \le t_a$, and by Taylor $\|\Delta_h^r g_\theta\|_{L^p} \le \|D_h^r g_\theta\|_{L^p}$ with $D_h = \sum_a h_a\partial_a$. Expanding the multinomial and using $\|\partial^\gamma g_\theta\|_{L^p} \le A^k \prod_a 2^{j_a\gamma_a}\cdot Q_u^{1/2-1/p}\|\theta\|_{\ell^p}$, all cross terms collapse into

$$\|D_h^r g_\theta\|_{L^p} \;\le\; A^k \Big(\sum_a |h_a| 2^{j_a}\Big)^r Q_u^{1/2-1/p}\|\theta\|_{\ell^p}.$$

Here $|h_a|2^{j_a} \le 2^{-v/s_a}2^{u\bar{s}/s_a} = 2^{-(v-u\bar{s})/s_a} \le 2^{-(v-u\bar{s})/s_{\max}}$, because $v > u\bar{s}$ makes the exponent negative and the largest s_a is then the binding one, so $\sum_a |h_a|2^{j_a} \le k\,2^{-(v-u\bar{s})/s_{\max}}$. Hence $2^v\|\Delta_h^r g_\theta\|_{L^p} \le A^k k^r 2^v 2^{-r(v-u\bar{s})/s_{\max}} Q_u^{1/2-1/p}\|\theta\|_{\ell^p}$, and with $\rho := r_{\mathrm{diff}}/s_{\max} - 1 > 0$ the ℓ^q sum over $v > u\bar{s}$ is geometric with ratio $2^{-\rho}$ and total $2^{u\bar{s}}(1-2^{-\rho})^{-1}$. Collecting, $C_{\mathrm{B}} := A^k\big[1 + 2^{r_{\mathrm{diff}}}k^{r_{\mathrm{diff}}}\big(2 + (1-2^{-\rho})^{-1}\big)\big]$ works, and depends on (k,s,η) only. *It is the joint modulus that produces the single factor $(\sum_a |h_a|2^{j_a})^r$; a decoupled per-direction modulus would give a different — and here inadmissible — functional.* □

Remark F.4 (Scope of Proposition F.3)**.** (70) is a two-sided-in-spirit but one-sided-in-use estimate for *this finite-level family*. It is not, and is not used as, a characterisation of the whole Besov ball by tensor-wavelet coefficient sequences.

F.2 The direct random-design bound for response-affine rules

Lemma F.5 (Affine risk lower bound)**.** *Let $\psi_1, \dots, \psi_J$ be orthonormal in $L^2([0,1]^D, \mathrm{Leb})$ — no other property is used, in particular no boundedness, no support structure and no P_X-orthogonality — and let $f_\theta := \sum_j \theta_j \psi_j$. Let π be a law on $\mathbb{R}^J$, chosen* before *the design is drawn and independent of (X, ξ, ω), with*

$$\textit{(H1) } f_\theta \in \mathcal{F} \textit{ for } \pi\textit{-a.e. } \theta, \qquad \textit{(H2) } \mathbb{E}_\pi \theta = 0, \qquad \textit{(H3) } \mathrm{Cov}_\pi(\theta) = v\, I_J,\ v > 0. \tag{71}$$

Then every $\hat{f} \in \mathcal{L}_{\mathrm{aff}}$ satisfies

$$\sup_{f \in \mathcal{F}} R(\hat{f}, f) \;\geq\; \frac{c_-\, J\, v\, \sigma^2}{\sigma^2 + n\, c_+\, v} \;\geq\; \frac{c_-}{2}\, \min\Big\{ Jv,\ \frac{J\sigma^2}{n\, c_+} \Big\}. \tag{72}$$

Proof. Since π is supported in $\mathcal{F}$, $\sup_{f \in \mathcal{F}} R(\hat{f}, f) \geq \mathbb{E}_{\theta \sim \pi} R(\hat{f}, f_\theta)$. Let P_V be the $L^2(\mathrm{Leb})$-orthogonal projection onto $V := \mathrm{span}\{\psi_j\}$ and let $\hat{\theta}$ be the coefficient vector of $P_V \hat{f}$. By the density lower bound, $\|\hat{f} - f_\theta\|^2_{L^2(P_X)} \geq c_- \|\hat{f} - f_\theta\|^2_{L^2(\mathrm{Leb})} \geq c_- \|\hat{\theta} - \theta\|_2^2$, using Parseval and that projection is a contraction. Because $\hat{f} \in \mathcal{L}_{\mathrm{aff}}$, $\hat{\theta} = b + A\, y$ with $b \in \mathbb{R}^J$ and $A \in \mathbb{R}^{J \times n}$ measurable in (X, ω) and *not* in y; and $y = Z\theta + \xi$ with $Z_{ij} := \psi_j(x_i)$. Condition on (X, ω) and average over $\theta \sim \pi$ and ξ, which are independent of each other and of (X, ω); using (71),

$$\mathbb{E} \big\| b + A(Z\theta + \xi) - \theta \big\|_2^2 \;=\; \|b\|_2^2 \;+\; v\, \|AZ - I_J\|_F^2 \;+\; \sigma^2 \|A\|_F^2, \tag{73}$$

an *exact* identity requiring only the first two moments of π and of ξ — in particular π is not assumed Gaussian, and no interchange of $\sup_f$ with $\mathbb{E}_X$ is performed. Minimising (73) over (b, A) gives $b = 0$ and, by the normal equations,

$$\min_{b, A} (73) \;=\; \mathrm{tr}\Big[\big(v^{-1} I_J + \sigma^{-2} Z^\top Z \big)^{-1} \Big], \tag{74}$$

which needs neither $n \geq J$ nor invertibility of $Z^\top Z$. For a positive definite H one has $\mathrm{tr}(H^{-1}) \geq J^2 / \mathrm{tr}(H)$ by Cauchy–Schwarz, so with $H := v^{-1} I_J + \sigma^{-2} Z^\top Z$,

$$\mathrm{tr}(H^{-1}) \;\geq\; \frac{J^2}{J/v + \sigma^{-2}\, \mathrm{tr}(Z^\top Z)}.$$

Now average over X: $t \mapsto J^2/(J/v + \sigma^{-2} t)$ is convex, so Jensen gives $\mathbb{E}_X \mathrm{tr}(H^{-1}) \geq J^2 / \big(J/v + \sigma^{-2} \mathbb{E}_X \mathrm{tr}(Z^\top Z) \big)$, and $\mathbb{E}_X \mathrm{tr}(Z^\top Z) = n \sum_j \mathbb{E} \psi_j(X)^2 \leq n\, c_+ \sum_j \int \psi_j^2 = n\, c_+ J$ by the density upper bound and orthonormality. Substituting gives the first inequality of (72); the second follows from $\sigma^2 + n c_+ v \leq 2 \max\{\sigma^2, n c_+ v\}$. □

Remark F.6 (What Lemma F.5 does and does not give)**.** It is a finite-dimensional statement and carries no information about Besov membership: it is vacuous until (H1) is discharged, which is what Section F.3 does. It bounds the $L^2(P_X)$ integrated risk only. The coefficients b, A are label-independent by hypothesis; *the lemma says nothing about rules whose coefficients read y, nor about pointwise-clipped affine rules*. The bound is explicit but not exact: Cauchy–Schwarz and Jensen are strict unless H's eigenvalues coincide and $\mathrm{tr}(Z^\top Z)$ is a.s. constant.

F.3 Two priors inside the bounded class

Lemma F.7 (The prior machine)**.** *Fix $u \geq 0$ and a magnitude profile $c \in \mathbb{R}^{Q_u}_{\geq 0}$. Draw independently S^* uniform on the M supports, a uniform random bijection $\varpi : \Lambda_u \to [Q_u]$, and i.i.d. Rademacher (ε_ℓ), and set $\theta_{(S,\ell)} := \mathbf{1}\{S = S^*\} \varepsilon_\ell c_{\varpi(\ell)}$. Then (i) every realisation has the same magnitude multiset, so membership is a* deterministic *check valid at every point of the support, not on average; (ii) $\mathbb{E}\theta = 0$; (iii) $\mathrm{Cov}(\theta) = \frac{\mathcal{E}}{M Q_u} I_{J_u}$ with $\mathcal{E} := \sum_i c_i^2$; (iv) $\mathcal{E} = \|f_\theta\|^2_{L^2(\mathrm{Leb})}$; and (v) the randomisation uses*

only $(S^*, \varpi, \varepsilon)$*, none of which sees* X *or* y*. The conjunction of* (75) *and* (76) *below is* sufficient *for membership* (71)*(H1) — Proposition F.3 is an upper estimate of the two norms, so no necessity is claimed, and none is needed: both profiles of Lemma F.8 satisfy them. The conditions are*

$$\|c\|_{\ell^p} \;\le\; \rho_u := \big(F/C_{\mathrm{B}}\big)2^{-u\bar{s}}Q_u^{1/p-1/2} \qquad \textit{and} \tag{75}$$

$$\|c\|_{\ell^\infty} \;\le\; h_u := \big(F_\infty/C_\eta^k\big)Q_u^{-1/2}, \tag{76}$$

by Proposition F.3.

Proof. (i) is immediate. (ii) follows from $\mathbb{E}\varepsilon_\ell = 0$. For (iii): if $S \neq S'$ then $\mathbf{1}\{S^* = S\}\mathbf{1}\{S^* = S'\} = 0$ *pointwise* — disjointness of events, not independence; if $S = S'$ and $\ell \neq \ell'$ the sign independence gives 0; and the diagonal is $\frac{1}{M} \cdot \frac{1}{Q_u}\sum_i c_i^2$ because $\varpi(\ell)$ is marginally uniform on $[Q_u]$. (iv) is Parseval and Lemma F.2. (v) is by construction. *The prior is not a product across blocks — only one block is ever active — yet its covariance is exactly isotropic, and Lemma F.5 consumes only the first two moments.* □

Lemma F.8 (The two profiles)**.** *Put* $a_0 := \min\{F/C_{\mathrm{B}},\ F_\infty/C_\eta^k\} > 0$ — both *radii are consumed — and* $c_{p,k} := 2^{-k(2/p-1)}$.

(a) $p \ge 2$, dense signs. $c_i \equiv a_u := a_0 2^{-u\bar{s}}Q_u^{-1/2}$ *satisfies* (75)–(76) *and gives* $\mathcal{E}_u = a_0^2 2^{-2u\bar{s}}$ *exactly.*

(b) $1 \le p < 2$ and $\bar{s} > 1/p$, single spike. $c_1 = a_0 2^{-u\bar{s}}Q_u^{1/p-1/2}$ *and* $c_i = 0$ *otherwise satisfies* (75)–(76) *— the sup-norm check* (76) *is where* $\bar{s} > 1/p$ *and the* upper *half of* (69) *are used, since the spike height is controlled by* $2^{-u(\bar{s}-1/p)}$ *— and gives* $\mathcal{E}_u > a_0^2 c_{p,k} 2^{-2us_*}$ *with* $s_* = \bar{s} - (1/p - 1/2)$. On this branch, and only on it, $2/p - 1 \in (0, 1]$ *so* $c_{p,k} = 2^{-k(2/p-1)} \in [2^{-k}, 1)$*; at* $p = 2$ *it equals* 1 *and at* $p > 2$ *it exceeds* 1*, so that interval may not be printed universally.*

F.4 The level optimisation

Proof of Theorem D.2. For each $u \ge 0$, Lemmas F.2, F.7 and F.8 supply hypotheses (H1)–(H3) of Lemma F.5 with $J = J_u = MQ_u$ and $v = v_u = \mathcal{E}_u/(MQ_u) > 0$; neither $n \ge J_u$ nor an invertible Gram is needed, so J_u may vastly exceed n. Hence for every $\hat{f} \in \mathcal{L}_{\mathrm{aff}}$,

$$\sup_{f \in \mathcal{F}} R(\hat{f}, f) \;\ge\; L(u) := \frac{c_- \sigma^2 \mathcal{E}_u}{\sigma^2 + n\, c_+ \mathcal{E}_u / J_u}. \tag{77}$$

The estimator is fixed *before* u is chosen and (77) holds for every u simultaneously, so $\sup_f R \ge \sup_{u \ge 0} L(u)$; each u indexes one prior used inside one instance of Lemma F.5, and no interchange of $\sup_f$ with $\mathbb{E}_X$ occurs. Since $z \mapsto \sigma^2 z/(\sigma^2 + nc_+ z/J)$ is increasing in z and L is increasing in J, we may replace $\mathcal{E}_u$ and J_u by *lower* bounds, which is the only way (69) is used here — always through its lower half $Q_u > 2^{u-k}$.

Case $p \ge 2$. With $t := 2^{-u}$, $\mathcal{E}_u = a_0^2 t^{2\bar{s}}$ and $J_u > M2^{-k}t^{-1}$, so $L(u) > c_- a_0^2 h(t)$ with $h(t) = t^{2\bar{s}}/(1 + b_{\mathrm{opt}} t^{2\bar{s}+1})$ and the *local balance coefficient*

$$b_{\mathrm{opt}} \;:=\; \frac{n\, c_+ A\, 2^k}{M\sigma^2}, \qquad A := a_0^2, \qquad r := \bar{s}.$$

This is a coefficient, not an exponent: it is dimensionless but depends on n *and* M*, whereas the rate exponents* α *and* β *of* (35)–(36) *are fixed by* $\bar{s}, s_*$ *alone. The two are distinct objects and are never equated.* The numerator of h' is $t^{2r-1}[2r - b_{\mathrm{opt}} t^{2r+1}]$, so if $b_{\mathrm{opt}} \ge 2r$ the maximiser is $t_* = (2r/b_{\mathrm{opt}})^{1/(2r+1)} \le 1$ and $h(t_*) = t_*^{2r}/(1 + 2r)$; otherwise $b_{\mathrm{opt}} t^{2r+1} \le b_{\mathrm{opt}} < 2r$ on $(0, 1]$, so h increases there and $h(1) = 1/(1 + b_{\mathrm{opt}}) > 1/(1 + 2r)$. Both branches give

$$\sup_{u \ge 0} L(u) \;\ge\; \frac{c_- A}{1 + 2r} \min\Big\{1, (2r/b_{\mathrm{opt}})^{\alpha}\Big\},$$

and factoring $(2r/b_{\mathrm{opt}})^{\alpha} = \left[2r\sigma^2/(2^k c_+ A)\right]^{\alpha}(M/n)^{\alpha}$ together with $\min\{1, xy\} \geq \min\{1, x\}\min\{1, y\}$ yields (35) with

$$c_1 := \frac{c_- a_0^2}{1+2\bar{s}} \min\Big\{1, \Big(\tfrac{2\bar{s}\sigma^2}{2^k c_+ a_0^2}\Big)^{\alpha}\Big\}, \qquad c_2 := \frac{c_- a_0^2 c_{p,k}}{1+2s_*} \min\Big\{1, \Big(\tfrac{2s_*\sigma^2}{2^k c_+ a_0^2 c_{p,k}}\Big)^{\beta}\Big\}. \tag{78}$$

Case $1 \leq p < 2$. The display above was written in terms of (r, A), so it is reused with $(r, A) = (s_*, a_0^2 c_{p,k})$, the energy lower bound of Lemma F.8(b) replacing that of (a). *The local balance coefficient changes with it*: it is $b_{\mathrm{opt}} = nc_+ a_0^2 c_{p,k} 2^k/(M\sigma^2)$ on this branch, a different number from the $p \geq 2$ one, since $c_{p,k} = 2^{-k(2/p-1)} < 1$ there; the $p \geq 2$ value is not carried over. The exponent produced is $2s_*/(2s_*+1) = \beta$, giving (36) with c_2 as in (78).

Rounding. The level $u \geq 0$ is continuous; the calculation above optimises the smooth lower envelope $c_- A\, h(2^{-u})$, not the exact expression $L(u)$ containing the floors. No integer rounding of u is needed. The lower bound $Q_u > 2^{u-k}$ contributes the factor 2^k to b_{opt} through $1/J_u$ — as 2^{+k}, not 2^{-k}, because J_u sits in the *denominator* of (77). On the $p < 2$ branch it also contributes $c_{p,k} = 2^{-k(2/p-1)}$ through the spike energy, so both factors are already included in b_{opt} with $A = a_0^2 c_{p,k}$. These factors depend only on structural parameters; the resulting constants are those in (78), and the rate exponents are unchanged. No assertion about a precise shift of the maximiser of $L(u)$ is needed. □

Remark F.9 (Consistency check, and the exponent ordering)**.** The zero estimator lies in $\mathcal{L}_{\mathrm{aff}}$, so any valid bound must be at most $\sup_{f\in\mathcal{F}} \|f\|^2_{L^2(P_X)} \leq c_+ F_\infty^2$; indeed $c_1 \leq c_- a_0^2 \leq c_- F_\infty^2/C_\eta^{2k} \leq F_\infty^2$, using $c_- \leq 1 \leq c_+$ (forced, both being densities on a unit-volume domain) and $C_\eta \geq 1$. Since $s_* < \bar{s}$ for $p < 2$ and $x \mapsto 2x/(2x+1)$ is increasing, the $p < 2$ exponent β is the smaller one. For $p > 2$ the spike prior is admissible too but yields the weaker exponent attached to $\bar{s} + (1/2 - 1/p)$; at $p = 2$ the two coincide.

G Scope

Remark G.1 (Everything the two theorems carry)**. Named conditions of Theorem D.1**, all stated mathematically above and none hidden: the codebook coherence $\mu \leq 1/(4k)$ at $m = \lceil 64k^2 \log 2D \rceil$ (Definition C.1); $k \leq W_{\mathrm{net}}$ and $T \leq 2L_{\mathrm{net}} W_{\mathrm{net}}^2$, used at both of their occurrences (Lemmas E.11 and E.13); $\mathcal{N}(1/n, \mathcal{H}_N, \|\cdot\|_\infty) \geq 3$, i.e. $n \geq n_{\mathrm{cov}}$ (Remark E.4); source-admissibility of the resolution (Lemma C.4) and the requirement that $\|K\|$ be computed exactly (Remark C.5); the existence of $\mathsf{m} \geq 2$ and hence the dependence of every architecture constant on $s_{\max}$ and not on $\bar{s}$ alone (16); the regularisation $\epsilon_N = {N^\dagger}^{-\bar{s}}/\log(2N^\dagger)$ (47); $n \geq 2$ and $2 \leq N_n \leq n$; and, on the Schmidt-Hieber branch only, $\delta = \min(1/n, \sigma)$ with $n \geq \lceil 1/\sigma \rceil$. For $2 \leq n < n_0$ the bound reduces to the clipped cap $4F_\infty^2$ and *carries no rate content there*.

Unnamed inherited constants. $c_{(k,\mathsf{m})}$, C_α and C_{tr} are named nowhere in Suzuki & Nitanda (2021); the first is proved in Suzuki (2019) and the second rests on a norm equivalence attributed there to Leisner (2003), neither of whose proofs is re-derived here. Both are used exactly as stated. The inherited structural constants and the thresholds derived from them remain symbolic. Appendix H reports a separately specified finite-prior illustration, not numerical values for those constants or a calibration of the theorem bounds.

Not proved here, and not to be inferred. (i) *No minimax optimality.* Theorem D.2 is a statement about $\mathcal{L}_{\mathrm{aff}}$; whether $n^{-\alpha}$ is minimax over $\mathcal{F}$ against *all* estimators is a different question, not addressed. (ii) *No trainability.* Proposition C.12 gives existence and measurability of an exact global minimiser; nothing asserts that any optimisation algorithm finds it. The imported approximation statement carries its source's own rider: it is an approximation error in an oracle setting and no sample complexity appears in it. (iii) *No fixed design and no sub-Gaussian noise.* Lemma E.3 inherits i.i.d. random design and Gaussian noise. (iv) *No growing* k, no $p < 1$, no $q < 1$, no $\bar{s} \leq 1/p$. (v) *No claim about clipped affine rules.* Theorem D.2 covers $\mathcal{L}_{\mathrm{aff}}$ as defined; extending it to $\{\mathcal{C}_B \circ T : T \in \mathcal{L}_{\mathrm{aff}}\}$ is an open problem, and clipping being 1-Lipschitz bounds a clipped rule's risk from *above* and therefore cannot be run backwards into a lower bound. (vi) *No module attribution.* See Remark D.8. In particular, that h_{θ_s} is linear in x at fixed parameters does *not* imply that the fitted estimator is affine in the responses, since $\hat{\theta}$ depends on y; that a rule is clipped does not by itself place it outside $\mathcal{L}_{\mathrm{aff}}$ (clipping the zero rule returns the zero rule, which is affine); and non-membership in $\mathcal{L}_{\mathrm{aff}}$ is in

any case not evidence of a separation, since a rule outside $\mathcal{L}_{\text{aff}}$ may be far worse. *The evidence for a separation is the inequality $U_n(D) < L_n(D)$ in the same model, nothing else.* (vii) *No sharpness.* 3α is the exact optimum of *this* majorant; L_{sel} and $\bar{L}$ are majorants; $m = O(k^2 \log 2D)$ is sufficiency only (Remark C.2).

A support-search reference class. The upper bound of Theorem D.1 is not the only way to obtain a vanishing guarantee on $\mathcal{F}$, and it is worth recording what a direct search over supports gives, so that the comparison with $\mathcal{L}_{\text{aff}}$ is not read as an architecture-specific effect. Keep N and every structural parameter, and keep the feed-forward budget $(L_{\text{net}}, W_{\text{net}}, S_{\text{net}}, B_{\text{ffn}})$ of Definition C.6; these do not depend on D or on the true support. Fix, before any data are seen,

$$\mathcal{T}_N \;:=\; \bigcup_{|S|=k} \big\{\, x \mapsto \mathcal{C}_{F_\infty}\big(\Phi(x_S)\big) \;:\; \Phi \in \Phi_k(L_{\text{net}}, W_{\text{net}}, S_{\text{net}}, B_{\text{ffn}}) \,\big\},$$

and let $\hat{f}_n^{\text{sea}}$ be an exact least-squares minimiser over $\mathcal{T}_N$ at $N = N_n$. The true S is not supplied to the estimator; the union is over all M of them. Each member is an ordinary D-input network: replacing $W^{(1)}$ by $W^{(1)}P_S$ places the k columns on S and leaves every other column zero, so by Definition C.3 the number of nonzeros, the magnitude bound, the width and the depth are unchanged, and $\mathcal{T}_N \subseteq \mathcal{C}_{F_\infty} \circ \Phi_D(L_{\text{net}}, W_{\text{net}}, S_{\text{net}}, B_{\text{ffn}})$ with the same quadruple.

Three estimates then run in the present model. For $f = g((\cdot)_S) \in \mathcal{F}$, Proposition E.7 supplies Φ_g and, since $\mathcal{C}_{F_\infty}$ contracts towards a target with $\|g\|_\infty \le F_\infty$, since $p_X \le c_+$, and since the $D-k$ unselected coordinates integrate to one, $\|\mathcal{C}_{F_\infty}(\Phi_g((\cdot)_S)) - f\|^2_{L^2(P_X)} \le c_+ (C_{\text{app}}^{\text{syn}} F)^2 N^{-2\bar{s}}$, with no constant carrying D. For the entropy, $\Phi \mapsto \Phi((\cdot)_S)$ is a sup-norm isometry from $[0,1]^k$ onto the S-block and $\mathcal{C}_{F_\infty}$ is 1-Lipschitz, so a net for each block and a union over the M blocks give

$$\log \mathcal{N}\big(\epsilon, \mathcal{T}_N, \|\cdot\|_\infty\big) \;\le\; \log M \;+\; S_{\text{net}} \log \frac{6 L_{\text{net}} W_{\text{net}}^2 B_{\text{ffn}} K_{\text{par}}}{\epsilon},$$

by Lemma E.13 applied with $d = k \le W_{\text{net}}$. *This covers all of $\mathcal{T}_N$ — every support, every sparsity pattern, the whole magnitude box — and it is not a covering bound for Φ_D, which is a larger class and is never covered here.* The hypothesis $d \le W_{\text{net}}$ of Lemma E.13 is a condition of that covering argument, not a statement about what a network with D inputs can represent. Finally $\mathcal{T}_N$ contains the constants $\pm\min(1, F_\infty)$ and 0, so the covering floor of Lemma E.3 holds at the same n_{cov}, and its parameter set is a finite union over S of the sets of Proposition C.12, so an exact measurable minimiser exists. For $n \ge n_0$ from Section E.6, Lemma E.3 at $\delta = 1/n$, with $\log M$ constant in N so that the optimisation of Section E.6 is unchanged, gives

$$\sup_{f \in \mathcal{F}} \mathbb{E}\big\|\hat{f}_n^{\text{sea}} - f\big\|^2_{L^2(P_X)} \;\le\; C_0'\, n^{-\alpha} \log^{3\alpha}(2n) \;+\; C_1' \frac{\log \binom{D}{k}}{n} \;+\; \frac{C_2'}{n},$$

with C_0', C_1', C_2' depending only on the fixed structural parameters. The same clipping argument as in Section E.6, enlarging C_0', covers $2 \le n < n_0$. At $D = n$ with k fixed, $\log \binom{n}{k} \le k \log(en/k)$, so this bound vanishes as well. *Consequently the separation against $\mathcal{L}_{\text{aff}}$ at $D = n$ is a statement about response-affine versus nonlinear estimation, and is not evidence of an advantage specific to attention.* The reference complexity factor is $\log M = O(k \log(D/k))$, rather than $P_{\text{sel}}\Xi_{n,D}$ in (34); this is not a numerical ranking of the full bounds with unspecified structural constants, or of actual risks. The positional parameterisation supplies a single continuous parameter box with a complete construction, budget and entropy certificate, in place of an explicit enumeration of M supports. Neither estimator comes with a search or optimisation guarantee: both are statements about exact minimisers, and $\mathcal{T}_N$ in particular is not claimed to be searchable efficiently. This paragraph is an upper bound for one specified estimator; it is not a minimax statement and not a guarantee for Φ_D.

The affine comparison and the convex hull. Theorem D.2 is a statement about $\mathcal{L}_{\text{aff}}$, and it is worth recording how it sits inside the classical picture for affine procedures, as used by Hayakawa & Suzuki (2020, Thm. 3.3). Fix $T \in \mathcal{L}_{\text{aff}}$ with offset and coefficients measurable in (X, ω) and independent of the responses, take $f_1, \dots, f_r \in \mathcal{F}$, weights $\lambda_j \ge 0$ summing to one, and $h = \sum_j \lambda_j f_j$. Using *one* draw of (X, ξ, ω) for all of them and writing $e_f := T(X, f(X) + \xi, \omega) - f$, response-affinity and $\sum_j \lambda_j = 1$ give the pointwise identity $e_h = \sum_j \lambda_j e_{f_j}$, whence by convexity of the squared norm $R(T,h) \le \sum_j \lambda_j R(T, f_j) \le \sup_{f \in \mathcal{F}} R(T, f)$. It is here that ω must not read the responses: the

same ω is used for every f_j. Nothing in this step uses a uniform design. Passing from finite mixtures to the L^2-closed hull is a Fatou argument along an almost-everywhere convergent subsequence, available because $c_- \le p_X \le c_+$ makes the $L^2(P_X)$ and Lebesgue closures agree. So the convex-hull reduction is available in the present model, and the design used by Hayakawa & Suzuki (2020) is not an obstacle to it.

What that reduction does not do is produce the factor M. It says that no affine rule can separate $\mathcal{F}$ from its closed convex hull — which is the qualitative reason a union of M balls is hard for such rules — while the quantitative statement is the computation carried out above: the prior of Lemma F.7 is supported in $\mathcal{F}$ pointwise yet spreads its covariance over all $J_u = MQ_u$ coordinates, and the two-moment bound of Lemma F.5 turns that dimension into $\min\{1, (M/n)^\beta\}$ through the level optimisation. Because that prior already lives in $\mathcal{F}$, the argument is run directly there and the hull is not needed as a step; it is recorded here only to locate the result, not to claim a new principle and not to claim that Hayakawa & Suzuki (2020) state this rate. The hull equality gives an equivalent interpretation of the same prior–moment calculation, not an independent alternative lower-bound proof.

H THE FINITE-SAMPLE ILLUSTRATION IN FULL

This appendix records the experiment summarised in Section 5.5. It is an illustration of one mechanism on one fixed resolution of the lower-bound template. It does *not* verify the rate of Theorem 5.1, the exponent β of Theorem 5.2, the $p < 2$ shift, any logarithmic factor, or any worst-case risk; and it involves no attention training and no learned dictionary.

H.1 TARGET, DESIGN AND NOISE

Fix $k = 2$, $p = q = 2$, $s = (3, 3)$ and the level $u = 0$, so that $j_a = 0$, $Q_u = 1$ and each support carries one atom. Let

$$q_0(t) = t^5(1-t)^5(2t-1), \qquad \eta = C\, q_0 \text{ on } [0,1], \quad \eta = 0 \text{ outside}, \tag{79}$$

so η is odd about $t = 1/2$, hence $\int \eta = 0$, and its derivatives of order $0, \dots, 4$ vanish at both endpoints, so the zero extension lies in C^4 as $r_{\text{diff}} = 4$ requires. Writing $z = 2t - 1$ gives $q_0 = 4^{-5} z (1 - z^2)^5$, whose modulus is maximal at $z^2 = 1/11$, and

$$\int_0^1 q_0^2 = \frac{(10!)^2}{23 \cdot 21!}, \qquad C = \left(\tfrac{23 \cdot 21!}{(10!)^2}\right)^{1/2}, \qquad C_\eta = \|\eta\|_\infty = C\, 4^{-5} (10/11)^5 / \sqrt{11}, \tag{80}$$

so $\int \eta^2 = 1$ exactly. Put $a = C_\eta^{-2}$ and $\sigma = a$. With $\psi_S(x) = \eta(x_i)\eta(x_j)$ for $S = \{i < j\}$, Lemma F.2 gives orthonormality in $L^2([0,1]^D)$, and Proposition F.3 gives $\|a\psi_S\|_\infty = a\, C_\eta^k = 1 = F_\infty$, so the amplitude a saturates the sup-norm envelope and $F = a\, C_{\mathrm{B}}$ is a fixed sufficient radius, held constant across the grid. The design is $X_i \sim \mathrm{Unif}([0,1]^D)$ i.i.d. and $y_i = f^*(X_i) + \sigma z_i$ with $z_i \sim N(0,1)$; the support S^*, the sign ε, the design and the noise are drawn from separate independent streams, and a fresh draw of all four is made for every dataset. By orthonormality the integrated squared error of any rule whose output lies in $\mathrm{span}\{\psi_S\}$ equals $\|\hat\theta - \theta^*\|_2^2$ exactly, with $\theta^* = \varepsilon a\, e_{S^*}$; the null rule $\hat f \equiv 0$ has risk exactly a^2.

H.2 THE SEVEN RULES AND WHAT EACH IS ALLOWED TO KNOW

Let $Z \in \mathbb{R}^{n \times M}$ have columns $z_S = (\psi_S(X_i))_i$, and let $v = a^2/M$, which is the prior covariance scale of Lemma F.7 at $u = 0$. All seven rules know the profile η, the value of k, the dictionary $\{\psi_S\}$ and the amplitude bound a. None of the first four receives S^* or ε; the three oracles receive S^* only, never ε.

A the element of $\mathcal{L}_{\text{aff}}$ minimising the prior-average risk, $\hat\theta_A = vZ^\top(vZZ^\top + \sigma^2 I_n)^{-1} y$, equivalently ridge at $\lambda = \sigma^2/v = M\sigma^2/a^2$. The level is fixed by the prior and no label-based tuning is performed, so A is the strongest member of $\mathcal{L}_{\text{aff}}$ for this prior-average risk. Its conditional risk is $\mathrm{tr}\{(v^{-1} I_M + \sigma^{-2} Z^\top Z)^{-1}\}$, and $\mathrm{tr}(H^{-1}) \ge M^2/\mathrm{tr}(H)$ with $\mathbb{E}\,\mathrm{tr}(Z^\top Z) = nM$ gives $\mathbb{E}_X$ of it at least $a^2 M/(M+n)$ when $\sigma = a$: a lower bound on the *mean*, not on an individual dataset.

AC the pointwise clipping $x \mapsto \max\{-1, \min\{1, \hat{f}_A(x)\}\}$. Since $\|f^*\|_\infty = 1$ this cannot increase the pointwise error. AC leaves the span, so for it only the test-point estimate is available.

B the exact posterior mean over all $2M$ hypotheses $(S, \pm)$ under the same prior: with $r_S = a\, z_S^\top y/\sigma^2$ and $b_S = -a^2\|z_S\|^2/(2\sigma^2)$, $\hat{\theta}_{B,S} = a\, p_S \tanh r_S$ where $p_S = \mathrm{softmax}_S\{b_S + \log\cosh r_S\}$, evaluated by log-sum-exp. No optimiser and no approximation is involved.

E support search with bounded amplitude, and *not* a Bayes rule: it uses neither σ nor the prior probabilities. For each S the inner problem $\min_{|t|\le a} \|y - t z_S\|^2$ is a one-dimensional convex quadratic on an interval, so its solution is the projection $t_S = \mathrm{clip}(z_S^\top y/\|z_S\|^2, -a, a)$, with $t_S = 0$ when $\|z_S\|^2 = 0$; E maximises $2t_S z_S^\top y - t_S^2\|z_S\|^2$ over *all* M supports, exact ties going to the first in a fixed enumeration, and returns the single coefficient $t_{\hat{S}}$. If $\hat{S} \neq S^*$ its error is $t^2 + a^2 \ge a^2$, at most $2a^2$; that is a bound for the wrong-support case only, not for the experiment.

OA, OB, OE the same three constructions restricted to the true support, the sign still fitted from the labels.

H.3 Grid, statistics and what was run

Twelve cells: $n \in \{32, 128\}$ at $D \in \{8, 16, 32, 64\}$, together with $n \in \{8, 16, 64, 256\}$ at $D = 32$; $M = \binom{D}{2}$ is $28, 120, 496, 2016$. Each cell uses 100 independent datasets, all seven rules paired within a dataset, 1200 datasets and 8400 rows in total, with no failures. Streams are derived from master seed 20260922 as $\mathrm{PCG64}(\mathrm{SeedSequence}([\mathrm{seed}, D, n, \mathrm{replicate}, \mathrm{stream}]))$. An earlier pilot at master seed 20260920 with 20 datasets per cell is *not* pooled with these data. Test-point risks use 8192 independent uniform points per dataset in blocks of 512, shared by all seven rules so that every contrast is paired. Intervals are 95% percentile bootstrap over datasets with 2000 resamples, one shared resample index matrix per cell; per-rule intervals and paired-difference intervals are distinct objects and are reported separately. The pre-specified primary comparison is $(\mathrm{ISE}_A - \mathrm{ISE}_E)/a^2$ at $D = 32$, $n = 128$ on the exact basis, and the pre-specified confirmation is the same contrast against B.

Table 1: Mean ISE/a^2 on the shared test points; the null rule is 1. Oracles are marked $*$. Entries shown as 0.000 are numerically negligible, not exactly zero; the raw values are in the shipped per-dataset file.

D	n	M	A	AC	B	E	OA*	OB*	OE*	$M/(M+n)$
8	32	28	0.638	0.638	0.021	0.063	0.029	0.000	0.015	0.467
8	128	28	0.227	0.225	0.000	0.004	0.008	0.000	0.004	0.179
16	32	120	0.892	0.892	0.085	0.136	0.035	0.000	0.016	0.789
16	128	120	0.629	0.629	0.000	0.005	0.008	0.000	0.005	0.484
32	8	496	0.991	0.991	0.956	1.620	0.170	0.051	0.122	0.984
32	16	496	0.986	0.986	0.683	1.059	0.088	0.008	0.051	0.969
32	32	496	0.967	0.967	0.176	0.255	0.031	0.000	0.016	0.939
32	64	496	0.942	0.942	0.000	0.020	0.017	0.000	0.008	0.886
32	128	496	0.882	0.882	0.000	0.003	0.008	0.000	0.003	0.795
32	256	496	0.780	0.780	0.000	0.002	0.003	0.000	0.002	0.660
64	32	2016	0.994	0.994	0.215	0.323	0.033	0.000	0.015	0.984
64	128	2016	0.969	0.969	0.000	0.004	0.009	0.000	0.004	0.940

H.4 What the grid shows, including where the label-driven rule fails

Table 1 gives the full grid. At fixed n the affine risk rises towards the null rule as M grows (Figures 3 and 4), while B and E stay near zero once n suffices to identify the support. At $D = 32$ the identification threshold is visible directly: E selects the true support in 11%, 41%, 85%, 99%, 100%, 100% of datasets at $n = 8, 16, 32, 64, 128, 256$, and its risk is 1.620 and 1.059 at the first two, i.e. above the null rule, because a wrong support places all the mass on a wrong coordinate. The affine rule fails in the opposite direction: it shrinks towards zero and approaches the null rule from below. Figure 5 plots both identification curves.

Clipping is not a remedy here. On the shared test points the paired difference between A and AC is identically 0 in all one hundred datasets of every $D = 32$ cell except $n = 256$, where a single dataset contributes 3.8×10^{-5}, and in every $D = 64$ cell; at $D = 8$, the cell-mean reduction is at most 1.4×10^{-3}. The two curves are therefore visually indistinguishable in the large-M cells without being exactly equal everywhere. Where the difference vanishes, the affine predictor has already shrunk inside the envelope, so the clip does not bind at the points evaluated. This is a statement about those test points, not about the whole cube: a sample maximum is not a supremum, and feasible points with $|\hat{f}_A(x)| > 1$ do exist in cells where no test point triggered the clip. Nothing here bears on whether the lower bound extends to pointwise-clipped affine rules, which remains open.

H.5 Reproduction

Paths below are relative to the root of the supplement's reproduction directory. It contains the per-dataset file `results/per_trial.csv` (8400 rows), the cell summaries and paired contrasts, the fixed configuration `config.json`, the estimator, runner and analysis sources under `src/`, the recorded software versions, a data dictionary describing every column, and three fixed raw datasets for independent recomputation. A single command, `bash run.sh`, reproduces every number and figure from the configuration; the run takes about half a minute on two CPU threads with no GPU. The figures in this appendix are generated from that summary file by `src/make_figures.py`; panel (a) of Figure 2 is drawn from the same file by the script that composes the main-text figure.

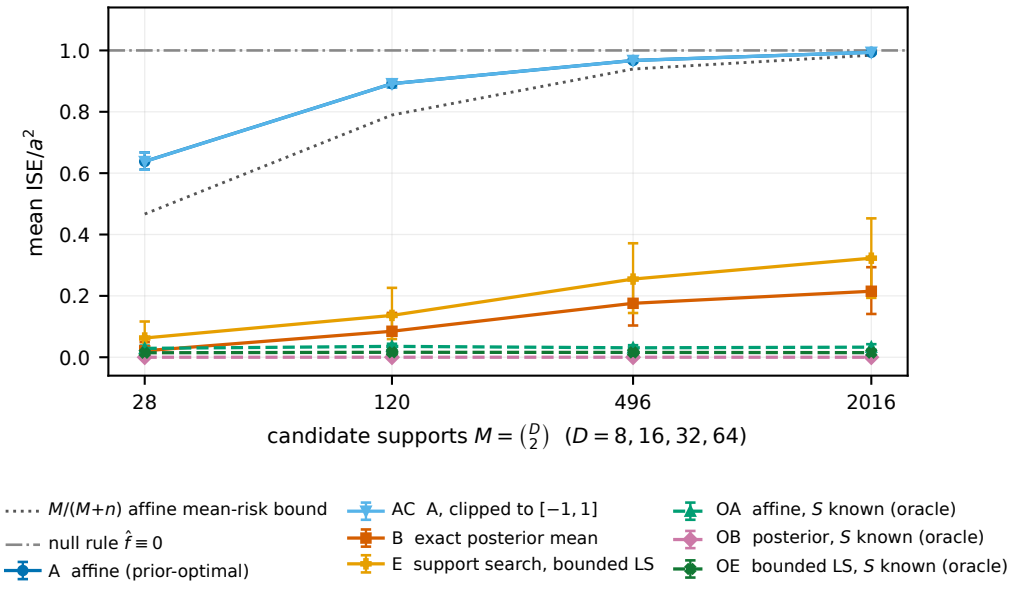


Figure 3: Risk against the number of candidate supports at fixed $n = 32$; axes, bars and conventions as in Figure 2(a).

I The whitening step of the witness, in full

This appendix records the mechanism check summarised in Section 5.5. It is a *known-support* diagnostic of one construction inside the proof of Lemma E.9. Nothing is trained, no support is estimated, and no statement here bears on Theorem 5.1, Theorem 5.2 or any risk constant.

I.1 The two queries and the fixed decoder

Fix an admissible sign codebook of Definition C.1, a support $S = \{i_1 < \cdots < i_k\}$, and write $C_S = [c_{i_1}, \ldots, c_{i_k}]$, $G = C_S^\top C_S$. Both arms use the forward map (2) at $K = V = I_m$ and *the same*

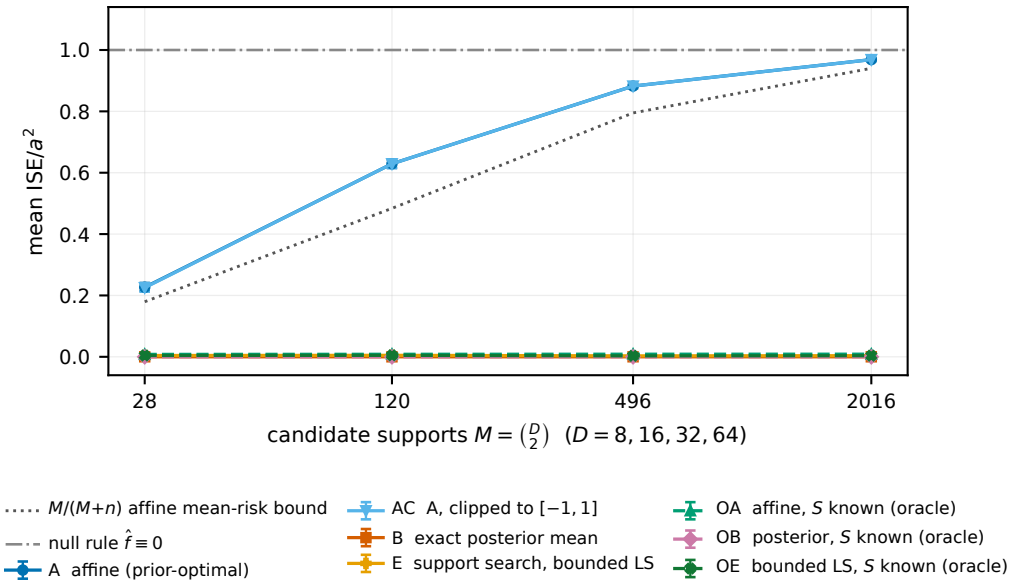


Figure 4: Risk against the number of candidate supports at fixed $n = 128$. The affine rule rises from 0.227 to 0.969 of the null risk as M grows from 28 to 2016, while the two label-driven rules and the three oracles stay near zero.

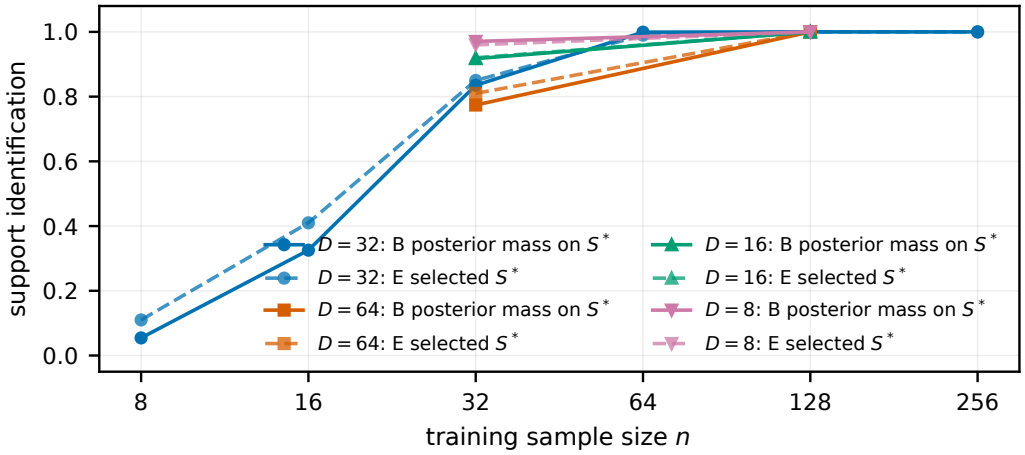


Figure 5: Support identification: posterior mass placed on the true support by B (solid) and the fraction of datasets in which E selected it (dashed). This is a diagnostic for B and E only; the affine rule makes no support-recovery claim.

decoder $R_S = G^{-1}C_S^\top$ of (55); only the query differs:

$$\text{(W)} \quad q_W = \tau\, C_S G^{-1}\mathbf{1}_k, \qquad \text{(U)} \quad q_U = \frac{\tau}{d}\, C_S\mathbf{1}_k, \qquad d := \frac{\mathbf{1}^\top G\mathbf{1}}{k}.$$

By (58) the active logits of W are all exactly τ; those of U are $(\tau/d)(G\mathbf{1})_a$, whose *mean* over $a \in [k]$ is τ as well. The scaling by d is what makes the comparison a test of the intra-support spread rather than of overall query strength: the raw code sum $\frac{\tau}{2}\sum_{i\in S} c_i$ of Remark E.10 carries a different mean active logit, and replacing it by q_U removes that confound. Since $\mathrm{spec}(G) \subseteq [3/4, 5/4]$ gives $d \in [3/4, 5/4]$, we have $\|q_U\|_2^2 = \tau^2 k/d \le 4k\tau^2/3$, the same certificate the proof uses for q_W; both queries sit inside the reference box introduced below.

Off-support mass is *measured*, not held equal: the two arms are matched on the mean active logit only.

Throughout, τ is a *logit scale*, i.e. the magnitude of the query. The softmax temperature of (2) is one and is never varied. Sweeping τ does not evaluate the integer schedule of Definition C.6, and no τ below corresponds to a computed N; the budget reference used for bookkeeping is

$$B(\tau) := 2\sqrt{k}\max\{1, \tau\},$$

the S-independent bound that part (a) of the proof of Lemma E.9 establishes for θ_s^{wit} at logit scale τ.

I.2 What is measured

At fixed parameters the selector is the linear map $h(x) = k\,(R\,C)\,\mathrm{diag}(\alpha)\,x =: B_{\mathrm{eff}}x$ with $\alpha = \mathrm{softmax}(C^\top q)$. Writing P_S for the $k \times D$ selection matrix and $H = B_{\mathrm{eff}} - P_S$, and taking $X \sim \mathrm{Unif}([0,1]^D)$, so that $\mathbb{E}[XX^\top] = \frac{1}{12}I + \frac{1}{4}\mathbf{1}\mathbf{1}^\top$,

$$\frac{1}{k}\,\mathbb{E}\big\|h(X) - X_S\big\|_2^2 \;=\; \frac{1}{k}\Big(\tfrac{1}{12}\|H\|_F^2 + \tfrac{1}{4}\|H\mathbf{1}\|_2^2\Big), \tag{81}$$

a closed form with no Monte-Carlo error. This is a *mean per-coordinate* L^2 quantity before the cube projection. It is not the supremum that (56) bounds, and it calibrates no constant of either theorem. Alongside it we record the off-support mass $\epsilon_{\mathrm{off}} = \sum_{i\notin S}\alpha_i$ and the conditional within-support imbalance $\frac{1}{2}\sum_a |p_a - 1/k|$, $p_a = \alpha_{i_a}/\sum_b \alpha_{i_b}$, which is identically zero for W in exact arithmetic by (58).

I.3 Grid and statistics

$k \in \{3,4\}$, $D \in \{32,64\}$, with the code length $m = \lceil 64k^2 \log 2D\rceil$ of Definition C.1 giving $m = 2396, 2795, 4259, 4969$. Ten independent codebooks per cell, each drawn until the integer coherence test $4k|\langle s_i, s_j\rangle| \le m$ passes for all pairs; all forty were accepted on the first draw, so no codebook was selected, even implicitly. Twenty distinct supports are drawn per codebook before any τ is evaluated, and eight logit scales $\tau \in \{1,2,4,8,16,32,64,128\}$ are applied to all three arms: 19,200 conditions. The statistical unit is the *codebook*: the twenty supports of a codebook are averaged first, and intervals are 95% paired-codebook percentile bootstrap with 2000 resamples and one shared resample index array. The 6400 support–τ conditions of an arm are not independent repetitions.

I.4 The grid

Table 2 gives the codebook-mean of (81) for all four cells. U-comp is the diagnostic arm of Section I.6.

The pre-specified contrast is $k = 3$, $D = 32$, $\tau = 64$, where U − W has codebook mean 0.2166 with 95% paired-codebook interval $[0.1958, 0.2409]$ and all ten codebooks of one sign. The same contrast along the whole sweep is $-1.07{\cdot}10^{-5}$, $+1.53{\cdot}10^{-5}$, $1.15{\cdot}10^{-3}$, $6.47{\cdot}10^{-3}$, $2.49{\cdot}10^{-2}$, $8.47{\cdot}10^{-2}$, $2.17{\cdot}10^{-1}$, $3.81{\cdot}10^{-1}$ at $\tau = 1, \ldots, 128$: *negative* at $\tau = 1$, where the unwhitened query is very slightly the better of the two, and positive thereafter. The U curve itself is not monotone — in every cell it falls to a minimum near $\tau = 4$ or 8 before rising — so nothing here describes a threshold law in m or a monotone degradation. Entries below 10^{-24} are reported but are not ranked against one another: at that size the value is dominated by the rounding of an identity that holds exactly in exact arithmetic, and this experiment certifies no accuracy law for it.

I.5 Two controls that must show no effect

If $k = 2$ then $G\mathbf{1} = (1+r)\mathbf{1}$ with $r = \langle c_{i_1}, c_{i_2}\rangle$, so $G^{-1}\mathbf{1} = (1+r)^{-1}\mathbf{1}$, $d = 1 + r$ and $q_W = q_U$ *identically*. Computed along the two separate code paths over 270 conditions the largest discrepancy in any recorded quantity is $1.4 \cdot 10^{-14}$, against a scale of 0.24. Orthonormal unit codes give $G = I$ and again $q_W = q_U$; over 48 conditions the discrepancy is exactly 0. Those codes satisfy the coherence constraint (15) but not (14), so they are an algebraic control on the implementation and are never pooled with the forty sign codebooks.

Table 2: Codebook-mean per-coordinate recovery error (81). Entries below 10^{-24} are shipped and plotted, but are not used to rank arms; that cutoff is a reading convention adopted here, not a bound on double-precision accuracy.

		$D = 32$			$D = 64$		
k	τ	W	U	U-comp	W	U	U-comp
3	1	$2.03{\cdot}10^{-1}$	$2.03{\cdot}10^{-1}$	$5.72{\cdot}10^{-4}$	$2.59{\cdot}10^{-1}$	$2.59{\cdot}10^{-1}$	$1.06{\cdot}10^{-3}$
3	2	$1.07{\cdot}10^{-1}$	$1.07{\cdot}10^{-1}$	$8.00{\cdot}10^{-5}$	$1.79{\cdot}10^{-1}$	$1.79{\cdot}10^{-1}$	$1.56{\cdot}10^{-4}$
3	4	$7.60{\cdot}10^{-3}$	$8.75{\cdot}10^{-3}$	$1.69{\cdot}10^{-6}$	$2.48{\cdot}10^{-2}$	$2.54{\cdot}10^{-2}$	$3.62{\cdot}10^{-6}$
3	8	$3.70{\cdot}10^{-6}$	$6.47{\cdot}10^{-3}$	$9.11{\cdot}10^{-10}$	$1.64{\cdot}10^{-5}$	$5.02{\cdot}10^{-3}$	$2.24{\cdot}10^{-9}$
3	16	$5.30{\cdot}10^{-13}$	$2.49{\cdot}10^{-2}$	$3.56{\cdot}10^{-16}$	$2.30{\cdot}10^{-12}$	$1.95{\cdot}10^{-2}$	$8.94{\cdot}10^{-16}$
3	32	$1.98{\cdot}10^{-26}$	$8.47{\cdot}10^{-2}$	$1.90{\cdot}10^{-28}$	$6.89{\cdot}10^{-26}$	$6.82{\cdot}10^{-2}$	$3.79{\cdot}10^{-28}$
3	64	$5.90{\cdot}10^{-27}$	$2.17{\cdot}10^{-1}$	$1.07{\cdot}10^{-28}$	$8.56{\cdot}10^{-27}$	$1.85{\cdot}10^{-1}$	$1.48{\cdot}10^{-28}$
3	128	$2.33{\cdot}10^{-26}$	$3.81{\cdot}10^{-1}$	$2.35{\cdot}10^{-24}$	$3.42{\cdot}10^{-26}$	$3.48{\cdot}10^{-1}$	$5.93{\cdot}10^{-26}$
4	1	$1.73{\cdot}10^{-1}$	$1.73{\cdot}10^{-1}$	$3.26{\cdot}10^{-4}$	$2.38{\cdot}10^{-1}$	$2.38{\cdot}10^{-1}$	$5.74{\cdot}10^{-4}$
4	2	$7.88{\cdot}10^{-2}$	$7.89{\cdot}10^{-2}$	$4.54{\cdot}10^{-5}$	$1.49{\cdot}10^{-1}$	$1.49{\cdot}10^{-1}$	$8.13{\cdot}10^{-5}$
4	4	$4.35{\cdot}10^{-3}$	$5.74{\cdot}10^{-3}$	$9.18{\cdot}10^{-7}$	$1.56{\cdot}10^{-2}$	$1.66{\cdot}10^{-2}$	$1.73{\cdot}10^{-6}$
4	8	$1.93{\cdot}10^{-6}$	$7.15{\cdot}10^{-3}$	$4.28{\cdot}10^{-10}$	$8.80{\cdot}10^{-6}$	$7.06{\cdot}10^{-3}$	$9.00{\cdot}10^{-10}$
4	16	$2.62{\cdot}10^{-13}$	$2.83{\cdot}10^{-2}$	$1.26{\cdot}10^{-16}$	$1.17{\cdot}10^{-12}$	$2.80{\cdot}10^{-2}$	$2.98{\cdot}10^{-16}$
4	32	$9.32{\cdot}10^{-27}$	$1.03{\cdot}10^{-1}$	$1.54{\cdot}10^{-28}$	$3.09{\cdot}10^{-26}$	$1.03{\cdot}10^{-1}$	$2.55{\cdot}10^{-29}$
4	64	$5.62{\cdot}10^{-27}$	$2.87{\cdot}10^{-1}$	$2.10{\cdot}10^{-28}$	$7.18{\cdot}10^{-27}$	$2.95{\cdot}10^{-1}$	$2.03{\cdot}10^{-29}$
4	128	$2.19{\cdot}10^{-26}$	$5.41{\cdot}10^{-1}$	$5.86{\cdot}10^{-24}$	$2.87{\cdot}10^{-26}$	$5.61{\cdot}10^{-1}$	$1.05{\cdot}10^{-25}$

I.6 The decoder rebuilt for the unwhitened query

Remark E.10 already notes that its example does not show that a decoder redesigned for a given unwhitened query fails. Keeping q_U and the weights it produces, and writing $A_S = \mathrm{diag}(\alpha_{i_1}, \dots, \alpha_{i_k})$, the choice $R_{\mathrm{comp}} := \frac{1}{k} A_S^{-1} R_S$ satisfies $k\, R_{\mathrm{comp}} C_S A_S = I_k$ exactly, so the support block of B_{eff} is exact and only the off-support leakage survives. This arm is the third column of each block in Table 2. It recovers at every τ, and below $\tau = 32$ it is the more accurate of the two by two to six orders of magnitude — at $k = 3$, $D = 32$, $\tau = 1$ its error is $5.7 \cdot 10^{-4}$ against $2.0 \cdot 10^{-1}$ for W — because it also compensates the deficit of total active mass that both queries suffer at small τ. At $\tau = 128$ we do not rank W against the rebuilt decoder using these very small numerical residuals; Table 2 retains the cell means without truncation.

What it costs is the budget, and this is a statement about the specific rebuilt decoder, not about unwhitened queries in general. Writing $z_i = \tau \ell_i$,

$$\big\| R_{\mathrm{comp}}[a,:] \big\|_{\max} = \frac{\| R_S[a,:] \|_{\max}}{k} \sum_{i=1}^{D} \exp\big(\tau(\ell_i - \ell_{i_a})\big), \tag{82}$$

an identity verified to a relative residual of $3.4 \cdot 10^{-16}$ on the shipped conditions. Put $\Delta_a = \max_i \ell_i - \ell_{i_a}$. The row grows like $e^{\tau \Delta_a}$ only when $\Delta_a > 0$; when $\Delta_a = 0$ — which is exactly what (58) arranges for W on the active rows — the sum is at most D and there is no growth in τ. So no general claim is made that a compensated decoder must leave any budget, and at finite τ a bound uniform in S is available from (15). Empirically, 6396 of the 6400 U-comp conditions have $\|R_{\mathrm{comp}}\|_{\max} \le B(\tau)$; the four exceptions all occur at $\tau = 128$, at $(k, D) = (3, 32)$ twice and $(3, 64)$, $(4, 32)$ once each, with ratios 10.44, 4.65, 4.45 and 1.007. They are reported, not filtered. By contrast R_S obeys the S- and τ-independent bound of the proof; its largest entry anywhere in the grid is 0.0227, consistent with $\|R_S\|_{\max} \le \|R_S\|_{\mathrm{op}} \le 2/\sqrt{3}$.

I.7 What this does and does not establish

Whitening supplies *one* sufficient construction for the witness, and the one whose budget certificate is uniform in S — which is what part (a) of the proof needs. It is not shown to be necessary, optimal, or the only decoder–query pair that recovers; the compensated arm recovers too, and usually inside $B(\tau)$. With the original decoder held fixed, a small off-support mass alone does not deliver recovery, while equality of the active weights together with that same small off-support mass does. Neither statement is about equal weights on their own, and neither asserts recovery to arbitrary accuracy at any fixed τ. All of this is at known support, on one design, for a linear selector, in one L^2 norm.

I.8 Reproduction

The shipped module contains the per-condition file (all 19,200 rows), the two control files, the codebook-level summaries and every pairwise contrast, the forty codebooks as packed sign bits with their convention, the fixed configuration and the recorded software versions. The default entry point analyses and redraws from those files and generates no data; panels (b) and (c) of Figure 2 are produced from the per-condition file alone.

J Exploratory training of a reduced parameterisation

This pilot concerns a reduced, unconstrained parameterisation trained for a fixed number of Adam steps; it is not the exact empirical minimiser of Definition 4.2 in Theorem 5.1. We report all six target–sample-size conditions and all four methods. The results are exploratory and mixed.

J.1 What is trained, and how it differs from the object of the theorem

The forward order of Section 4 is kept — tagger, selector, cube projection, network, output clip — with $K = V = I_m$ *fixed* and only q, R and a small ReLU network trained. We call this `PA-QR`. At $k = 2$, $D = 32$ the code length of Definition C.1 is $m = 1065$, so the trained selector has $m + km = 1065 + 2130 = 3195$ parameters and the fixed $2 \to 32 \to 32 \to 1$ network has 1185, for 4380 in total. The selector parameter count of (3) is $P_{\mathrm{sel}} = 2m^2 + km + m = 2{,}271{,}645$; it counts q, K, V, R and contains no network parameters, so the like-for-like comparison is 3195 against 2,271,645 and mixing the two conventions would misstate it.

Four further differences from the object of Theorem 5.1 are load-bearing. The network is a fixed $2 \to 32 \to 32 \to 1$ net, not the integer schedule of Definition C.6. The box $\|\theta_s\|_\infty \le B_{\mathrm{sel}}$ is *not* enforced. Six hundred Adam steps are not the exact global minimiser that Definition 4.2 requires. And `RF-ridge` is the one rule here built to satisfy the response-affine restriction of Definition 3.4 explicitly; the three feature-learning procedures are not constrained to it. Nothing below tests Theorem 5.2.

One algebraic caveat bounds what a comparison against the free bottleneck can mean. With $m \ge D$ and C of full column rank, $C^+ = (C^\top C)^{-1} C^\top$ satisfies $C^+ C = I$, so uniform weights and $R = (D/k) B C^+$ realise *any* $B \in \mathbb{R}^{k \times D}$. Without the box, the two parameterisations therefore have the same achievable linear input maps here, so a gap in either direction cannot be attributed to that difference. What remains — fitting dynamics, the implicit bias of the parameterisation, finite-sample generalisation — this pilot does not separate. None of this says anything about the candidate classes inside the original budget.

J.2 Data, targets and methods

$D = 32$, $k = 2$; the support is drawn uniformly from the $\binom{32}{2} = 496$ two-element subsets, its coordinates listed in increasing order, and a sign $\varepsilon \in \{\pm 1\}$ is drawn independently, per dataset. With $t = x_S$, $u = 2t_1 - 1$, $v = 2t_2 - 1$, $a = \sin(2\pi t_1)$, $b = \sin(4\pi t_2)$, the three fixed targets are

$$\mathrm{G1} = \tfrac{1}{3}(u + v + uv), \qquad \mathrm{G2} = uv, \qquad \mathrm{G3} = \tfrac{1}{3}(a + b + ab),$$

with $\|g\|_\infty = 1 = F_\infty$ and exact zero-rule risks $\mathbb{E}g^2 = 7/81, 1/9, 5/36$. Labels are $y = \varepsilon g(x_S) + 0.1\, z$, $z \sim N(0,1)$, $X \sim \mathrm{Unif}([0,1]^{32})$. At $p = q = 2$, $s = (3,3)$, so $r_{\mathrm{diff}} = 4$ and $\bar{s} = 3/2 > 1/p$, every fourth-order mixed partial of the three targets is dominated by $A_4 = (4\pi)^4$ — the binding one is $2(4\pi)^4/3$ for G3, and G1, G2 have vanishing fourth differences — whence $F = 1 + 16A_4/\sqrt{1 - 2^{-2/3}}$ is a common *sufficient* radius. It is not an estimate of their minimal Besov radius, and six conditions say nothing about the worst case over the ball. The four fitters are listed in Table 3.

`RF-ridge` uses $r = 128$ fixed random ReLU features $\Phi(X) = \max\{0, (2X - 1)W^\top + b\}/\sqrt{r}$ drawn from a label-independent stream, and solves the ridge problem at the fixed level $\lambda = 10^{-3}$ with an unpenalised intercept. Both the coefficients and the intercept are linear functionals of y with X-only coefficients, so the rule lies in $\mathcal{L}_{\mathrm{aff}}$; its output is not clipped. It is a *designated legal*

Table 3: The four fitters. Only the oracle receives the support, as an ordered pair of column indices and nothing else; no fitter receives the sign, the target name or its formula.

method	input	trained	params
`PA-QR`	X, codebook	q, R, network; $K = V = I$ fixed, no box	4380
`Linear-bottleneck`	X	free $B \in \mathbb{R}^{k \times D}$, same network	1249
`RF-ridge`	X	ridge coefficients only; features fixed	129
`Support-oracle`*	X_S	same network	1185

affine baseline, not the optimal element of $\mathcal{L}_{\text{aff}}$ for this target distribution, and no tuning on labels is performed.

`PA-QR` and `Linear-bottleneck` are initialised to the *same function*: B_0 has independent Dirichlet$(1, \ldots, 1)$ rows, $q_0 = 0$ gives uniform weights, and $R_0 = (D/k)B_0C^+$ makes $B_{\text{eff}}(0) = B_0$; the three networks share one set of initial values. B_0 is a dense random convex combination of all D coordinates drawn from a stream disjoint from the data, not a warm start towards S.

J.3 Statistics

Three targets $\times$ two sample sizes $n \in \{128, 512\}$ $\times$ five independent datasets $\times$ two initialisations gives 60 data–initialisation combinations and 240 method terminal states, of which 180 are network fits. Every network ran exactly 600 full-batch Adam steps at learning rate 0.003; there is no early stopping, no restart selection and no tuning on the test set, and no fit produced a non-finite value. The statistical unit is the *dataset*: the two initialisations are averaged first, leaving five independent units per cell. Intervals are 95% percentile bootstrap over datasets with 2000 resamples and one shared index array. *Five units per cell is exploratory*; these intervals are not a stable benchmark and carry no multiple-comparison correction. Risk is the mean squared error against the *noiseless* target on 8192 independent points, reported in units of the exact zero-rule risk of that target, so 1.000 is the null rule.

J.4 All six conditions

Table 4 reports every condition.

Table 4: Normalised test risk; 1.000 is the zero rule. Mean over five datasets after averaging the two initialisations, with the 95% dataset bootstrap interval. The four column headings abbreviate the methods of Table 3; * marks the oracle, which is given the true coordinates and is not a competitor.

target	n	`PA-QR`	`Bottleneck`	`RF-ridge`	`Oracle`*
G1	128	0.217 [0.207, 0.233]	0.269 [0.205, 0.370]	0.336 [0.288, 0.375]	0.029 [0.016, 0.041]
G1	512	0.046 [0.036, 0.056]	0.081 [0.054, 0.124]	0.256 [0.247, 0.265]	0.008 [0.006, 0.010]
G2	128	2.318 [2.138, 2.504]	2.313 [2.048, 2.525]	1.167 [1.139, 1.195]	0.031 [0.027, 0.034]
G2	512	0.576 [0.203, 0.953]	0.348 [0.330, 0.366]	1.035 [1.022, 1.046]	0.012 [0.010, 0.016]
G3	128	1.463 [1.400, 1.523]	1.672 [1.542, 1.803]	0.883 [0.874, 0.892]	0.184 [0.168, 0.198]
G3	512	0.870 [0.806, 0.961]	0.983 [0.910, 1.060]	0.776 [0.764, 0.790]	0.069 [0.053, 0.086]

`PA-QR` has the smallest mean among the three non-oracle rules in two of the six conditions, both of them G1. In G2 at $n = 512$ the free bottleneck is ahead on the mean; in G2 at $n = 128$ and in both G3 conditions the fixed affine baseline is. Three conditions contain rules *worse* than predicting zero: `PA-QR` and the bottleneck at 2.3 and 1.5–1.7 in G2 and G3 at $n = 128$, and `RF-ridge` at 1.17 and 1.03 in G2. Being first on the mean in a cell is not a significant win; Table 5 gives the paired intervals. The primary pre-specified readout was `PA-QR` on G1 at $n = 512$, 0.046 [0.036, 0.056]: a feasibility observation, not a rate and not a benchmark. The worst single initialisation of `PA-QR` in each cell is 0.366, 0.067, 2.886, 1.439, 1.658 and 1.057.

J.5 Fit versus generalisation, and what the oracle does not settle

In the two hardest conditions the training loss falls close to the noise floor $\sigma^2 = 0.01$ while the noiseless test risk does not follow: for `PA-QR` the final training MSE and test MSE are 0.0289 and

Table 5: Paired differences of normalised risk, by dataset; negative favours the first method. † marks intervals excluding 0 at five datasets per cell — exploratory, with no correction for the eighteen comparisons shown.

target	n	PA−QR − bottleneck	RF−ridge − PA−QR	PA−QR − oracle
G1	128	$-0.052\,[-0.137, 0.003]$	$+0.118\,[0.079, 0.157]^{\dagger}$	$+0.188\,[0.177, 0.199]^{\dagger}$
G1	512	$-0.035\,[-0.072, -0.010]^{\dagger}$	$+0.210\,[0.201, 0.217]^{\dagger}$	$+0.038\,[0.027, 0.050]^{\dagger}$
G2	128	$+0.006\,[-0.117, 0.143]$	$-1.152\,[-1.330, -0.980]^{\dagger}$	$+2.288\,[2.107, 2.471]^{\dagger}$
G2	512	$+0.227\,[-0.131, 0.591]$	$+0.459\,[0.080, 0.835]^{\dagger}$	$+0.563\,[0.190, 0.942]^{\dagger}$
G3	128	$-0.209\,[-0.349, -0.069]^{\dagger}$	$-0.581\,[-0.638, -0.519]^{\dagger}$	$+1.280\,[1.209, 1.350]^{\dagger}$
G3	512	$-0.113\,[-0.139, -0.091]^{\dagger}$	$-0.094\,[-0.175, -0.039]^{\dagger}$	$+0.801\,[0.722, 0.904]^{\dagger}$

0.2576 at G2, $n = 128$, and 0.0134 and 0.2032 at G3, $n = 128$. The oracle network, given the two coordinates, reaches 0.0034 and 0.0255 on the same targets. That the oracle does well shows the *same* network can fit these targets once the support is supplied; it does not separate support estimation, finite-sample overfitting, the implicit bias of the parameterisation and non-convex optimisation. The accurate statement is that learning with an unknown support still faces a substantial obstacle here, and this pilot does not identify its cause. No run was extended, reseeded or restarted to look for a better outcome.

J.6 Diagnostics

Output dependence is probed by permuting one input coordinate at a time on a fixed subset of 2048 test points and reading the change in the final prediction; attention weights are not used as a recovery certificate. The four methods draw from one random stream in a fixed order, so they consume *different* permutation indices; the probe is therefore not matched coordinate-for-coordinate across methods, while every risk comparison remains paired by dataset. Both true coordinates appear in the top two of 32 in 63.3% of PA−QR fits, 65.0% for the bottleneck and 60.0% for RF−ridge, against a chance level of $1/496$; the oracle is at 100% by construction. Recovery tracks the risks: the top two columns of B_{eff} hit both coordinates in 2.00 of 2 on G1 at both n and on G3 at $n = 512$, but in 0.10 of 2 on G2 at $n = 128$. The cube projection clips a mean of 11.8% of the $8192k$ coordinate entries for PA−QR (maximum 58.2%) and the output clip a mean of 1.4% of the 8192 scalar outputs (maximum 9.7%); these are fractions on the evaluated test points, not over the cube and not suprema.

J.7 Reproduction, and what is reproducible

The module ships the per-fit file, every learning curve, all 180 network terminal states, the 60 ridge solutions and three pre-specified raw cases. Re-scoring the stored terminal states reproduces the numbers above to float32 tolerance. Re-*training* is a different matter: an independent run of three pre-specified cases in a different software environment completed normally but moved the normalised risk of PA−QR by up to 0.073, so optimisation trajectories are not reproduced bit-for-bit across environments, and only checkpoint evaluation is. The recorded versions are NumPy 2.5.3 with PyTorch 2.14.0+cpu.