Sharp limits for high-dimensional learning with spectral structure

Jae Jun Han[1,*]

[1]Department of Physics, Chungnam National University, Daejeon 34134, Republic of Korea

[*]Correspondence: 202550282@o.cnu.ac.kr

High-dimensional learning is often limited by the number of variables that must be explored, even when only a small set determines the response[1,2]. Whether that statistical cost can be separated from the number of predictive degrees of freedom depends on how the representation is identified and how rapidly its spectral coefficients decay. Here we show that a certified three-dimensional flat quotient with Gaussian heat regularity admits a sharp, nearly parametric prediction law. A freely acting reflected half-translation reduces the leading spectral mode count by half, while preserving the intrinsic dimension. For ordinary regression with unknown active coordinates, we obtain a distinct, explicitly quantified support-search cost at a critical growth rate of the ambient dimension. We also give finite-resolution criteria for recovering the geometry and establish why response labels alone do not force a torus. These results specify when geometry and smoothness permit efficient statistical prediction and when additional representation information is indispensable; they do not remove the curse of dimensionality for arbitrary data or guarantee a polynomial-time exhaustive search.

Two different costs of dimension

Many recorded coordinates need not be equally relevant to prediction. Statistical rates can depend on a lower-dimensional structural representation, but discovering that representation is a separate inference problem. Existing theory establishes strong sparsity-adaptive procedures and lower bounds[1,3], yet a geometric representation can impose additional spectral constraints not captured by a simple count of active variables. We study those two costs separately: approximation

and estimation *within* a representation, and the information required to *identify* the relevant representation.

For a feature vector $X \in \mathbb{R}^D$ and response Y, the minimal predictive quotient identifies two feature values only if their conditional response laws agree. Such a quotient is defined by the experiment; it is not automatically a smooth manifold, a flat space, or a torus. Our geometric conclusions begin only after additional heat-kernel and orbit-recovery evidence has been supplied, as summarized in Fig. 1. This distinction prevents a structural assumption from being mistaken for a universal data-driven discovery.

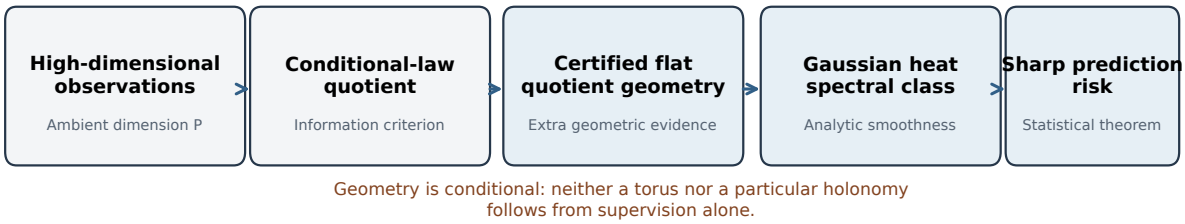


Figure 1: **Conditional route from observations to sharp prediction risk.** The minimal conditional-law quotient is defined by predictive equivalence. Identification as a specific flat quotient then requires separate population or finite-resolution geometric certificates; only after a Gaussian heat class is specified does the corresponding spectral prediction law apply. Arrows indicate logical requirements, not an empirically observed reconstruction.

Geometry determines the spectral budget

Suppose the quotient has a closed connected Riemannian realization with a compatible heat semigroup. If its universal-cover heat kernel is exactly the Euclidean three-dimensional Gaussian and the lifted orbit kernel is available, rigidity yields a Euclidean cover and identifies the crystallographic deck group[4,5]. The translation-lattice Gram matrix and affine holonomy can then be tested. Passing the cubic, reflected half-translation, and metric-normal-form certificates singles out a torus cover with a freely acting involution,

$$M = \mathbb{T}^3_{2\pi}/\langle \iota \rangle, \quad \iota(\theta, u, v) = (\theta + \pi, -u, v), \quad M \cong K^2 \times S^1. \tag{1}$$

This identification is conditional: the statistical quotient alone does not provide the certificates, and flat quotients with other Gram matrices or holonomies remain possible. The non-orientable factor describes a global identification and does not alter the local Euclidean dimension.

A torus Fourier mode $e^{i(m\theta+ku+pv)}$ is transformed to $(-1)^m e^{i(m\theta-ku+pv)}$. Thus each $k > 0$ pair contributes one invariant mode and, for $k = 0$, only even m survives. In the normalized cubic case,

$$N_M(\Lambda) = \frac{1}{2} N_{\mathbb{T}^3}(\Lambda) + O(\Lambda^{1/2}) \sim \frac{2\pi}{3}\Lambda^{3/2}. \tag{2}$$

The quotient changes the *coefficient*, not the $3/2$ Weyl exponent[6]. Figure 2 compares exact finite lattice enumerations with the leading spectral approximations. No measurement or fitted parameter enters this figure.

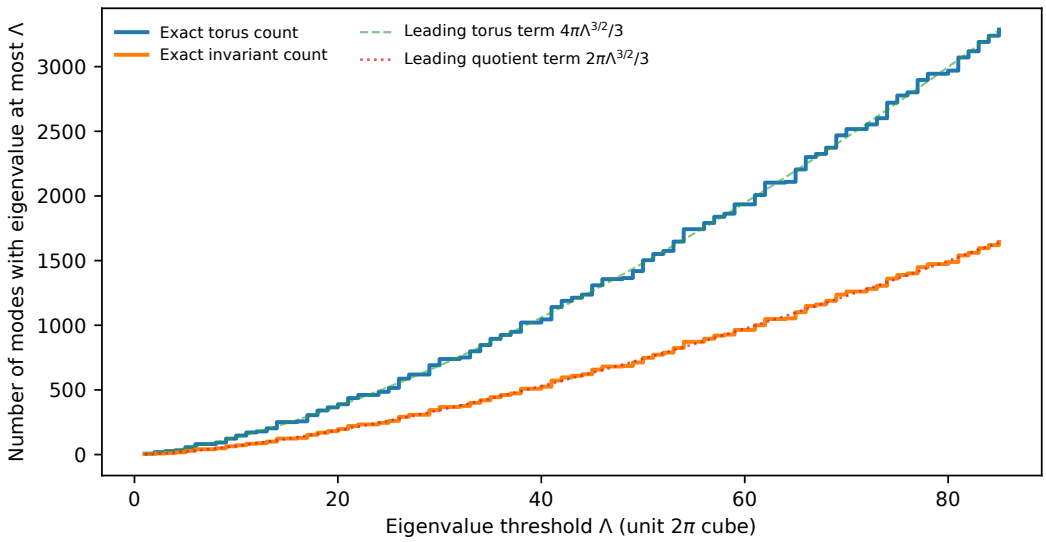


Figure 2: **An invariant mode count has half the leading density.** Exact integer-lattice enumeration for the 2π-period cubic convention is contrasted with the two Weyl leading terms. The zero mode is included. The asymptotic ratio tends to one half; this does not imply a change in intrinsic dimension or apply to an uncertified quotient.

A nearly parametric risk law

Let $-\Delta_M \phi_j = \lambda_j \phi_j$ and define the Gaussian heat ellipsoid

$$\mathcal{H}_\tau(B) = \left\{ f = \sum_j \theta_j \phi_j : \sum_j e^{\tau\lambda_j} |\theta_j|^2 \le B^2 \right\}, \qquad \tau, B > 0. \tag{3}$$

Equivalently, $f = e^{\tau\Delta_M/2}h$ with $\|h\|_2 \le B$. Its spectral tail obeys $\|f - P_\Lambda f\|_2^2 \le B^2 e^{-\tau\Lambda}$. Because only $O(\Lambda^{3/2})$ modes lie below the cutoff, selecting $\Lambda \simeq \tau^{-1}\log(nB^2/\sigma^2)$ balances squared bias and Gaussian estimation variance. In the Gaussian sequence experiment, and in the specified observed-coordinate regression experiment, the squared minimax prediction risk is

$$\mathcal{R}^*_{n,M} \sim \frac{2\pi}{3\tau^{3/2}}\frac{\sigma^2}{n}\left[\log\left(\frac{nB^2}{\sigma^2}\right)\right]^{3/2}, \quad M = \mathbb{T}^3_{2\pi}/\langle\iota\rangle. \tag{4}$$

The constant follows from the exact asymptotic eigenvalue count, while a compactly supported coefficient prior yields a matching information-theoretic lower bound. This is close to the parametric $1/n$ rate apart from a spectral logarithm. It holds for *fixed* B, τ, σ and the stated geometry; an arbitrary function in a high-dimensional ambient space need not satisfy (3). Smoothness is part of the result, not a consequence of dimensional reduction.

The topological effect is quantitative: at matched scale and normalization the quotient has one half of the torus's leading prediction constant. It is not a claim that a topological identification by itself universally halves data requirements. Under non-Haar or singular designs the relevant operator is instead the heat reproducing-kernel Hilbert space restricted to the observed design[7,8]; the original manuscript gives design-specific bounds and counterexamples to a distribution-free exact constant.

The cost of finding three coordinates

Now suppose $Y_i = f(X_i) + \varepsilon_i$, where X_i are independent product-Haar inputs on $P = P_n$ coordinates and $\varepsilon_i \sim N(0, \sigma^2)$. Only an unknown triple of coordinates is active; its ordering, reflected phase and heat function are also unknown. Consider the critical support-search regime

$$\frac{\log P_n}{(\log n)^{3/2}} \longrightarrow c \in (0, \infty), \qquad A_\tau = \frac{2\pi}{3\tau^{3/2}}. \tag{5}$$

The exact leading risk reported for this fully specified sparse class is

$$\mathcal{R}^*_{n,P_n} \sim A_\tau \max\{1, a(c)\}\frac{\sigma^2}{n}(\log n)^{3/2}, \qquad a(c) - \log(1 + a(c)) = \frac{6c}{A_\tau}. \tag{6}$$

Here $a(c)$ is the unique positive root. It measures the informational support-search burden, beyond the spectral estimation budget. The threshold $c_* = A_\tau(1 - \log 2)/6$ separates the region where the known-representation constant survives from the region where the search term dominates (Fig. 3). Equation (6) is *not* a polynomial-time algorithm guarantee: the constructive estimator performs an exhaustive candidate scan in its general formulation.

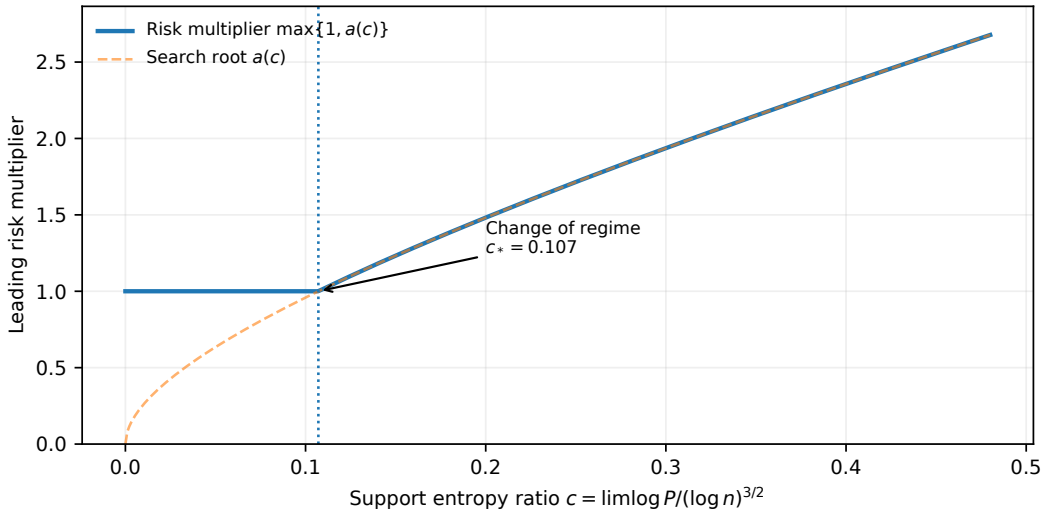


Figure 3: **Predicted search penalty from the exact root equation.** The leading risk multiplier $\max\{1, a(c)\}$ is plotted against the asymptotic support entropy ratio c for $\tau = 1$. The transition c_* is analytic, not estimated from experimental data. The graph applies to the specified sparse Gaussian heat/Haar experiment.

The lower bound uses a family of mutually orthogonal exact-three-way Fourier blocks indexed by supports and a Gaussian-mixture testing argument. A matching upper construction partitions each candidate into an $O((\log n)^{3/2})$ exact-three-way core and an $O(\log n)$ shared lower-order boundary. A screening sample chooses a candidate; a separate smaller sample refits the boundary; a guarded core flag prevents paying full-dimensional estimation variance when the wrong support has been chosen. This *selective refitting* distinction is what changes an earlier $1 + a(c)$ upper bracket into the reported $\max\{1, a(c)\}$ constant. The complete dependence on sample Gram concentration, shared-coordinate leakage, phase profiling and tail control is retained in Supplementary Notes, including an additional independent-check annex. Finite diagnostic tests are not substitutes for peer review of the asymptotic arguments.

In the explicitly independent, continuous-marginal model, empirical rank calibration can learn the coordinate transforms at a lower-order cost; it does not extend the same constant to arbitrary

correlated inputs. The general manuscript separately analyses dependent and singular designs, unknown interaction dimension, approximate regularity and structured nonlinear shear. Each extension has distinct hypotheses; they are not all collapsed into the equation above.

When the geometry can be recovered

Exact orbit recovery assumes exact population kernels. In a finite-resolution Gaussian-shift observation model, discrete candidate deck groups are distinguishable only when their template separation exceeds the deterministic error budget and the sampling noise. For continuous parameters, a locally bi-Lipschitz template map with sensitivity lower bound μ yields a localization scale of the form

$$\|\widehat{\vartheta} - \vartheta_*\| = O_{\mathbb{P}}\left(\frac{\delta}{\mu} + \frac{L_{\mathrm{tpl}}}{\mu\sqrt{N}} + \frac{\sigma}{\mu}\sqrt{\frac{q\log N}{N}}\right), \tag{7}$$

for a fixed-dimensional q-parameter family, deterministic template error δ, and a controlled upper sensitivity L_{tpl}. This expresses finite-resolution localization, not exact recovery of continuous geometric parameters from finite noisy data. A uniform L^2-to-geometry inference additionally requires regularity assumptions; without them narrow heat-kernel spikes defeat the upgrade. Local Euclideanity also does not determine a global deck group, which is why approximate equivariance is required to control the reconstructed quotient.

The results separate three independent obstructions: arbitrary regression functions may have no low-dimensional predictive quotient; even if one exists, its geometry may not be identifiable from available observations; and, even if the geometry and smoothness are known, searching a large ambient support space may carry a non-negligible information cost. Under the stated certificates, analytic spectral decay and a learned sparse representation make nearly parametric prediction possible. Outside those certificates there is no universal conclusion that the curse of dimensionality has disappeared.

References

[1] Yang, Y. & Tokdar, S. T. Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43**, 652–674 (2015).

[2] Stone, C. J. Consistent nonparametric regression. *Ann. Statist.* **5**, 595–620 (1977).

[3] Tsybakov, A. B. *Introduction to Nonparametric Estimation* (Springer, 2009).

[4] Carron, G. & Tewodrose, D. A rigidity result for metric measure spaces with Euclidean heat kernel. *J. Éc. polytech. Math.* **9**, 101–154 (2022).

[5] Charlap, L. S. *Bieberbach Groups and Flat Manifolds* (Springer, 1986).

[6] Chavel, I. *Eigenvalues in Riemannian Geometry* (Academic Press, 1984).

[7] Aronszajn, N. Theory of reproducing kernels. *Trans. Am. Math. Soc.* **68**, 337–404 (1950).

[8] Caponnetto, A. & De Vito, E. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7**, 331–368 (2007).

[9] Laurent, B. & Massart, P. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 1302–1338 (2000).

[10] Barron, A. R., Birgé, L. & Massart, P. Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301–413 (1999).

Methods

Predictive quotient and geometry. The conditional-law equivalence identifies features with indistinguishable conditional response distributions. The geometric branch is invoked only when the quotient admits the prescribed smooth metric-measure realization, its universal-cover heat kernel equals the Euclidean Gaussian, and its lifted orbit information determines the deck group. Gram and affine-holonomy tests distinguish a cubic torus cover from other flat

quotients; the reflected-half-translation normal form is used in equations (1)–(4). The 124-page supplementary source includes the exact-population statements, counterexamples, and finite-resolution replacements.

Spectrum and sharp oracle constant. An invariant Fourier projection $P_M = (I + U_\iota)/2$ gives equation (2) by counting positive and negative reflected-coordinate frequencies, with a lower-order contribution from the zero-frequency plane. Truncation at $\Lambda = \tau^{-1} \log(nB^2/\sigma^2)$ gives a variance $\sigma^2 N_M(\Lambda)/n$ plus a heat tail bounded by $B^2 e^{-\tau\Lambda}$. A matching lower bound uses a product compact prior on modes below a slightly reduced cutoff; the technical argument is given in the Supplementary Notes.

Unknown-support boundary. For equation (6), the active triple is unknown among P_n coordinates. The design is independent product Haar, the noise is Gaussian, and the target belongs to the specified invariant heat ball. The statistical proof decomposes low-frequency modes into exact-three-way core and lower-order boundary blocks. A joint Gaussian-block lower bound identifies $a(c)$; the matching selective-refitting upper bound uses sample splitting, simultaneous Gram control and Gaussian tail inequalities[9,10]. The procedure is statistical rather than uniformly polynomial-time computational.

Finite-resolution geometry. The population theorem is supplemented by a Gaussian template experiment with deterministic error, finite candidate separation and, where needed, bi-Lipschitz localization of continuous geometric parameters. Local heat-kernel almost-rigidity and global quotient reconstruction require distinct information. Tests that only check observed finite spectra do not prove population-level heat-kernel equality or exact metric cubicity.

Numerical diagnostics. Figure 2 is computed by integer enumeration of frequencies with $m^2 + k^2 + p^2 \le \Lambda$; Figure 3 plots the analytic positive root of $a - \log(1 + a) = 6c/A_\tau$. Figure 1 is a schematic. The supplied Python scripts reproduce these displays and the existing finite checks. The figures do not represent empirical or simulated validation of all theorem statements.

Use of generative AI tools. ChatGPT (OpenAI) was used to assist with restructuring the article for journal submission, editing the presentation, and producing reproducibility and figure-preparation code. The named author remains responsible for verifying all mathematical claims, bibliographic details and declarations, and for approving the submitted manuscript.

Data availability

No external empirical dataset is analysed. Figures and quantitative numerical checks are generated from exact finite mode enumerations or deterministic formulae; corresponding machine-readable diagnostics are included in the accompanying reproducibility materials. No external accessions are claimed.

Code availability

The reproducibility package supplied with the manuscript includes the figure-generation program and the available finite-check scripts. These scripts are diagnostic and do not constitute formal proof certificates. A permanent public code repository/identifier has not yet been provided; the author must confirm deposition arrangements before publication.

Acknowledgements

No acknowledgements are specified in the source manuscript.

Funding

Funding declaration requires confirmation by the corresponding author before formal submission; no grant information was provided with the source manuscript.

Author contributions

J.J.H. is the sole listed author and is responsible for the theoretical manuscript and approval of the submitted version.

Competing interests

The corresponding author must confirm the competing-interests declaration before formal submission.

Additional information

Supplementary Information is provided as a complete mathematical-source and verification package. Correspondence and requests for materials should be addressed to Jae Jun Han (202550282@o.cnu.ac.kr).