WHEN LINEAR PROBES GO BLIND: THE ORDER SPECTRUM OF LEARNED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear probes are widely used to assess properties encoded in neural network representations. However, poor linear recovery cannot distinguish absent information from information accessible through nonlinear readouts, while commonly used MLP heads do not characterize how information is distributed across orders. We formalize this distinction by introducing the order spectrum, a Hermite decomposition of the target regression function that, under Gaussian representations, resolves predictable variance by polynomial degree. Specifically, we show that optimal affine probes measure the first-degree energy of the spectrum, while increments between optimal nested polynomial probes recover successive energies. For the off-Gaussian case, we bound the discrepancy in terms of the Hermite-feature Gram deviation and a target-dependent approximation residual, and validate the bound in controlled settings. Leveraging this framework, we show that quadratic probes recover substantial board-color information in OthelloGPT and pendulum energy in a JEPA trained without energy labels, where linear probes fail. Further, by analyzing grokking in modular addition, we find that partial linear recovery can coexist with substantial quadratic accessibility, while quantitative Hermite-energy interpretations require explicit distributional and approximation conditions. More broadly, the framework provides a degree-resolved account of how target information becomes accessible in learned representations.

1 INTRODUCTION

Understanding how neural networks internally encode information is central to demystifying their black-box nature. A mechanistic understanding of internal circuits not only helps predict model behavior but also enables targeted interventions (Meng et al., 2022; 2023; Rimsky et al., 2024). Probing addresses this by predicting targets from frozen layer activations (Alain & Bengio, 2017; Belinkov, 2022). However, the performance depends heavily on both the chosen coordinate system and the expressive capacity of the probe family (Hewitt & Liang, 2019; Pimentel et al., 2020). Specifically, a low linear probe score cannot determine whether the target information is absent or accessible via a non-linear transformation. Although expressive probes, such as multilayer perceptrons (MLPs), overcome this bottleneck by modeling richer interactions, they offer little insight into the functional complexity required for the target. Consequently, comparing linear probes to expressive MLPs leaves the functional complexity of target accessibility fundamentally unresolved.

Recent studies have examined polynomial degree as a measure for readout complexity (Hewitt & Manning, 2019; Oldfield et al., 2026; Lee & Kondor, 2026). However, a principled theoretical framework for disentangling the contributions of different polynomial degrees to target predictability is missing. The key challenge stems from distributional variability since representations vary significantly across architectures and layers, confounding how individual polynomial components contribute to target predictability.

We address this challenge in two steps. First, we introduce the **order spectrum** and show that, under an isotropic Gaussian representation assumption, the total predictable variance of the target regression function decomposes into mutually orthogonal Hermite-degree components. Specifically, we prove that the optimal population linear probe isolates the degree-one energy within this spectrum. In contrast, population increments between optimal nested polynomial probes systematically recover higher-order degree energies.

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.

Second, in off-Gaussian regimes, where non-Gaussian laws break Hermite feature orthogonality, we establish an upper bound on the induced error. We show that the discrepancy between population probe increments and reference Gaussian degree energies is governed jointly by two factors: 1) the operator-norm deviation of the Hermite-feature Gram matrix from the identity, and 2) a target-dependent polynomial approximation residual.

Finally, we deploy our theoretical framework to investigate representation dynamics across diverse deep learning paradigms: OthelloGPT (Li et al., 2023), Joint-Embedding Predictive Architectures (JEPAs) (Assran et al., 2023), and modular arithmetic grokking (Power et al., 2022). Notably, in OthelloGPT, quadratic probes recover substantial board-color information at intermediate layers, without requiring the target-coordinate reformulations introduced by Nanda et al. (2023b). Similar quadratic accessibility appears when decoding physical pendulum energy from a SIGReg-regularized JEPA representation (Balestriero & LeCun, 2025). Furthermore, during modular-addition grokking, the quadratic accessibility of answer Fourier features closely tracks the onset of generalization. Together, these diverse applications demonstrate that our order spectrum framework serves as a principled, unifying lens for quantifying the emergence and functional complexity of latent representations across fundamentally distinct learning regimes.

The remainder of this paper is structured as follows: Sections 2 and 3 formalize the theoretical order spectrum framework and derive the off-Gaussian stability bound. Sections 4 through 8 present controlled validations and learned-representation case studies. Finally, Section 9 contextualizes our contributions within interpretability literature, and Section 10 discusses the findings and limitations.

2 The Order Spectrum

In this section, we introduce the order spectrum. We first briefly review probabilists' Hermite polynomials, outlining key concepts we then use to define the order spectrum. Then, we show that affine probes (linear probes with an intercept) correspond to first-order members of the order spectrum, and we present an algorithmic way to measure this spectrum.

Preliminaries: Let (Z, Y) be jointly distributed, where $Z \in \mathbb{R}^d$ denotes the internal representation of a neural network and $Y \in \mathbb{R}$ is a centered scalar target with finite, nonzero variance. Let γ denote the marginal distribution of Z. We define the target function $f \in L^2(\gamma)$ as $f(z) = \mathbb{E}[Y \mid Z = z]$.

Assuming γ to be an Isotropic Gaussian, $\gamma = \mathcal{N}(0, I_d)$, one can define probabilists' Hermite polynomials, $\{\mathrm{He}_\alpha\}_{\alpha \in \mathbb{N}^d}$, $\mathrm{He}_\alpha(z) = \prod_i \mathrm{He}_{\alpha_i}(z_i)$, which form an orthogonal basis of $L^2(\gamma)$ with $\mathbb{E}[\mathrm{He}_\alpha \mathrm{He}_\beta] = (\prod_i \alpha_i!)\, \delta_{\alpha\beta}$. Grouping the basis elements by their total degree $k = |\alpha|$, one can define the orthogonal subspaces $H_k = \mathrm{span}\{\mathrm{He}_\alpha : |\alpha| = k\}$. The orthogonal projection of f onto H_k is given as:

$$\mathrm{Proj}_{H_k} f = \sum_{|\alpha|=k} \frac{\mathbb{E}[f(Z)\mathrm{He}_\alpha(Z)]}{\prod_i \alpha_i!}\, \mathrm{He}_\alpha,$$

and consequently every $f \in L^2(\gamma)$ admits the following decomposition with convergence in $L^2(\gamma)$:

$$f = \sum_{k=0}^{\infty} \mathrm{Proj}_{H_k} f.$$

Definition 1 (Order spectrum)**.** *We define the order spectrum of f under γ as the sequence of normalized degree energies $\{s_k(Y; Z)\}_{k=0}^{\infty}$, where each $s_k(Y; Z)$ is defined as*

$$s_k(Y; Z) = \frac{\|\mathrm{Proj}_{H_k} f\|^2_{L^2(\gamma)}}{\mathrm{Var}(Y)} \tag{1}$$

If Y is measurable with respect to $\sigma(Z)$, then $f(Z) = Y$ almost surely and hence $\sum_{k=0}^{\infty} s_k(Y; Z) = 1$. However, in general, the law of total variance gives:

$$\sum_{k=0}^{\infty} s_k(Y; Z) = 1 - \frac{\mathbb{E}[\mathrm{Var}(Y \mid Z)]}{\mathrm{Var}(Y)} \leq 1$$

Proposition 1. *Assume $Z \sim \mathcal{N}(0, I_d)$. The population R^2 of the optimal affine predictor of Y from Z, $z \mapsto a^\top z + b$ is equal to $s_1(Y; Z)$.*

Proposition 1 is central to our argument and offers an alternative way to view statistics derived from affine probes. From this point of view, a near-zero R^2 from an optimal population affine probe only certifies that $s_1(Y; Z) \approx 0$ and says nothing about the higher-degree energies s_k for $k > 1$. Therefore, the failure of an affine probe is not necessarily due to missing information; it could be because f inherits a higher Hermite order under γ, which requires a more expressive probe. The proof of Proposition 1 is given in Appendix A.

To estimate the order spectrum, we use a nested probe ladder. Under an isotropic Gaussian representation, Hermite blocks of different degrees are orthogonal in population. Consequently, sequentially fitting ordinary least squares models of increasing degree $k \in \{1, 2, \dots\}$ mirrors, in the limit, the orthogonal projection onto $H_0 \oplus H_1 \oplus H_2 \oplus \cdots \oplus H_k$. We can therefore isolate the degree-k energy by measuring the incremental performance gain.

$$s_k = R^2_{\leq k} - R^2_{\leq k-1} \tag{2}$$

In practice, finite-sample fitted probe scores can differ from these population quantities due to sampling, regularization, and optimization. We therefore reserve direct spectral interpretation for settings where the population assumptions are justified; otherwise, fitted probe comparisons measure degree-limited predictive accessibility. Under the Gaussian population setting, the affine probe isolates s_1 (Proposition 1), while the quadratic increment extracts s_2. An MLP provides an additional nonlinear comparison, but its fitted performance does not establish the total recoverable variance or whether target information is absent.

3 Analyzing the Off-Gaussian Spectrum

A key limitation of the order spectrum is its assumption of an isotropic Gaussian representation, since learned activations are generally non-Gaussian. Consequently, population probe increments need not equal the Gaussian-reference degree energies s_k. We address this issue by demonstrating that, under appropriate conditions, this degree leakage admits an upper bound. Specifically, we show that the leakage depends directly on the operator norm of the Gram-deviation matrix and the target function's polynomial approximation residual. We formally state our results below.

Let μ denote the non-Gaussian law governing the learned activations $Z \in \mathbb{R}^d$, and let $\Phi_{\leq k}$ denote the Gaussian-orthonormal Hermite features up to degree k, including the intercept. We define the Hermite-Gram matrix G_k and its deviation ρ_k as follows:

$$G_k = \mathbb{E}_\mu[\Phi_{\leq k}\Phi_{\leq k}^\mathsf{T}] \qquad \rho_k = \|G_k - I\|_{\mathrm{op}} \tag{3}$$

The scalar ρ_k measures departure from Gaussian-orthonormal feature geometry through degree k; it is not a complete measure of non-Gaussianity. Under an isotropic Gaussian measure, $G_k = I$ and $\rho_k = 0$. In the off-Gaussian regime, ρ_k jointly captures shifts in feature scale and spurious correlations across Hermite degrees. Crucially, evaluating ρ_k requires only the representations and is completely independent of the target labels.

To bound the degree leakage, we first formalize our regression setting. Let the target variable Y satisfy $\mathbb{E}[Y] = 0$ and have bounded, non-zero variance $\sigma^2 = \mathrm{Var}(Y)$. Specify a measurable function $f \in L^2(\mu) \cap L^2(\gamma)$ on $\mathbb{R}^d$ satisfying $f(Z) = \mathbb{E}[Y \mid Z]$ almost surely, where $\gamma = \mathcal{N}(0, I_d)$. This specifies the function used for the Gaussian reference, including outside the support of μ. For singular or lower-dimensional μ, the Gaussian-reference energies can depend on the chosen off-support extension of f, so in such settings we interpret polynomial probes as measures of degree-limited predictive accessibility rather than intrinsic spectral energies.

Let $P^\gamma_{\leq k}$ denote the orthogonal projection onto polynomials of degree at most k under the Gaussian measure γ, and let $f_k = \mathrm{Proj}_{H_k} f$. We quantify the model misspecification via the normalized truncation error under the true law μ, defined as $\tau_k = \|f - P^\gamma_{\leq k} f\|_{L^2(\mu)}/\sigma$. Furthermore, let $\Delta_k = R^2_{\leq k} - R^2_{\leq k-1}$ denote the increment between optimal population polynomial-probe scores, and $s_k = \|f_k\|^2_{L^2(\gamma)}/\sigma^2$ be the Gaussian-reference degree energy.

Proposition 2. *Fix an integer $k \geq 1$ and assume $\mathbb{E}_\mu \|Z\|_2^{2k} < \infty$. If ρ_k (Gram deviation, equation 3) satisfies $\rho_k \leq \rho < 1$, then the deviation of the nested increment from the Gaussian degree energy is bounded by:*

$$|\Delta_k - s_k| \leq \rho s_k + 2\tau_k \sqrt{(1+\rho)s_k} + \tau_k^2. \tag{4}$$

Evaluating the truncation parameter τ_k directly is generally intractable because the target function f is unknown. However, when f is assumed or known to be a polynomial of degree at most k, the truncation error vanishes ($\tau_k = 0$). In this setting, equation 4 simplifies to a pure relative perturbation bound:

$$|\Delta_k - s_k| \leq \rho s_k. \tag{5}$$

For $s_k > 0$, the relative discrepancy $|\Delta_k - s_k|/s_k$ is therefore at most ρ. The absolute bound also depends on the target energy s_k. An empirical Gram estimate alone does not certify the population condition $\rho_k \leq \rho < 1$. We defer the proof to Appendix B.

For empirical validation, we evaluate our theoretical bound using the standardized `sinh-arcsinh` distribution family, systematically varying the skewness ϵ and tail deformation δ around the Gaussian reference $(\epsilon, \delta) = (0, 1)$. We define a target function of maximum degree two: $Y = 0.6\mathrm{He}_1(Z) + 0.4\mathrm{He}_2(Z)$. Because the target is quadratic, the truncation error vanishes ($\tau_2 = 0$). Consequently, following equation 5, our theoretical bound reduces to a strictly relative form:

$$E_2 := |\Delta_2 - s_2| \leq B_2 := \rho_2 s_2 \qquad \text{for all } \rho_2 < 1.$$

We evaluate Proposition 2 across over 2,000 distinct configurations of the skew and tail parameters. As depicted in Figure 1a, the condition $\rho_2 < 1$ delineates the regime covered by Proposition 2 around the Gaussian baseline. Within this region, comprising approximately 1,400 configurations, we observe zero bound violations (Figure 1c, blue circles). Furthermore, the scatter plot shows that the empirical error closely tracks the theoretical upper bound within this admissible region. Conversely, the red crosses represent settings where $\rho_2 \geq 1$. The sweep protocol is given in Appendix C.3. Additional checks with a nonzero approximation residual are reported in Appendix C.4. In the upcoming sections, we make our framework operational and analyze a diverse set of scenarios through the lens of the order spectrum.

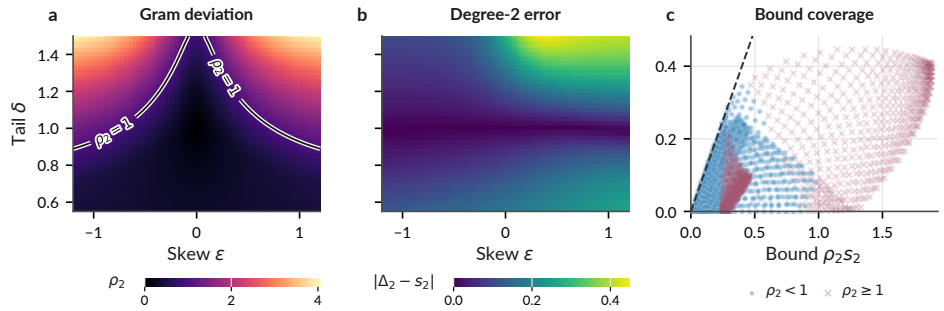


Figure 1: **(a)** Gram deviation ρ_2 as a function of skew ϵ and tail δ, highlighting the admissible boundary $\rho_2 = 1$. **(b)** Absolute empirical error $E_2 = |\Delta_2 - s_2|$. **(c)** Error E_2 versus the theoretical bound $B_2 = \rho_2 s_2$. Configurations satisfying $\rho_2 < 1$ (blue circles) lie below the dashed equality line, consistent with the bound.

4 Validating the Spectrum

The order spectrum, equation 1, explains how the total predictable variance of the target is distributed across different degrees in the Hermite polynomial space. To demonstrate it empirically, we consider a controlled example with a known signal and compare the results against the ground truth. The first experiment uses Gaussian representations. For random variable $Z \sim \mathcal{N}(0, I_4)$ and $\xi \sim \mathcal{N}(0, 1)$, we construct the target as:

$$Y = \sqrt{0.2}\, Z_1 + \sqrt{0.3}\, Z_1 Z_2 + \sqrt{0.5}\, \frac{\mathrm{He}_3(Z_3)}{\sqrt{6}} + \sigma\xi, \qquad \sigma \in \{0, 1\}$$

Here $\mathrm{He}_3(z) = z^3 - 3z$ and ξ is observation noise independent of Z. As the linear, quadratic, and cubic terms are orthogonal under the Gaussian law, the actual spectrum is $(s_1, s_2, s_3) = (0.2, 0.3, 0.5)/(1 + \sigma^2)$. Adding unit noise halves the predictable fraction without changing the regression function. Figure 2a shows that nested polynomial probes recover the expected degree contributions, while the unnecessary fourth degree adds essentially no held-out recovery. Cubic test R^2 is 1.000 without noise and 0.489 ± 0.015 with noise, close to the respective population ceilings of one and one-half. Fitting details and shuffled-label controls are given in Appendix C.1.

Another straightforward yet important observation is that target accessibility depends on the representation coordinates, as shown in figure 2b. For $Z \sim \mathcal{N}(0, I_2)$ and $Y = Z_1 Z_2$, affine test R^2 is -0.001 ± 0.001, while quadratic recovery is 1.000. After the transformation $T(Z) = (Z_1 Z_2, Z_2)$, affine recovery becomes 1.000. Since T is invertible whenever $Z_2 \neq 0$, it preserves information almost surely while changing the degree needed to read out the same target. This distinction between preserved information and predictor-dependent accessibility is consistent with the notion of usable information (Xu et al., 2020). Appendix C.2 gives the coordinate-check protocol and additional known-degree and sample-size controls.

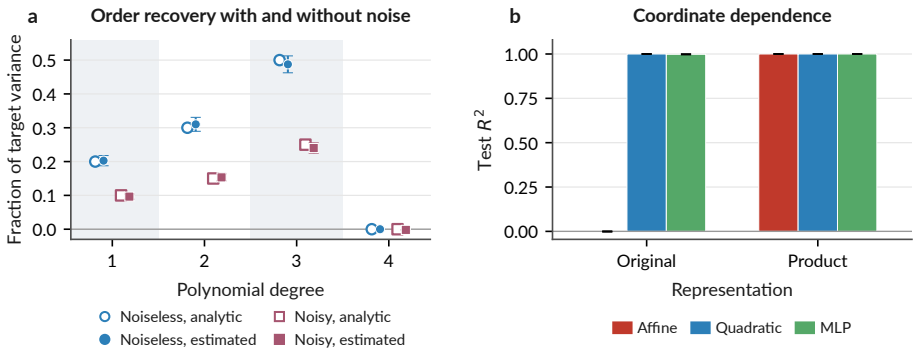


Figure 2: **(a)** Analytic degree energies (hollow markers) and estimated nested polynomial-probe increments (filled markers) without noise ($\sigma = 0$) and with noise ($\sigma = 1$), using ten seeds with 12,000 fitting and 6,000 test samples each. The population spectra sum to one and one-half, respectively. **(b)** Recovery of $Y = Z_1 Z_2$ from original coordinates and invertible product representation $T(Z) = (Z_1 Z_2, Z_2)$, using five seeds with 8,000 fitting and 4,000 test samples each. Both panels show means and one sample SD. Negative estimates are retained without renormalization.

5 Off-Gaussian Bound under VAE Representations

Variational autoencoders (VAEs) provide a tractable setting for checking the bound in Proposition 2 on learned representations. Although VAEs regularize encoder posteriors toward a standard Gaussian prior (Kingma & Welling, 2014), the aggregate latent distribution can remain non-Gaussian (Bauer & Mnih, 2019). For a fixed set of images, the aggregate posterior is a Gaussian mixture with an explicit regression function $f(z) = \mathbb{E}[Y \mid Z = z]$. This lets us evaluate the approximation residual τ_k numerically (Appendix D.3).

We train a two-dimensional β-VAE (Higgins et al., 2017) to reconstruct MNIST images (Deng, 2012), withholding digit labels until probing. The representation is a sample from the learned encoder, and we vary the regularizer weight β. Training and probe-fitting details are given in Appendices D.1 and D.2. Figure 3a shows that digit recovery decreases as β increases. At $\beta = 2$, the held-out affine, quadratic, and cubic probes achieve mean R^2 scores of 0.103, 0.176, and 0.205, respectively. The degree-two Gram deviation on the full-test mixture is 0.238 ± 0.045 across seeds.

For the bound checks, we use a separate mixture of encoder posteriors from 1,024 fixed test images. We evaluate the Gaussian-reference energies, approximation residuals, and optimal population probe increments on this same distribution (Appendix D.4). Figure 3b shows no bound violations beyond numerical tolerance across all 500 admissible checks with $\rho_k < 1$. Omitting the terms involving τ_k

gives 205 violations, showing why the residual matters in these checks. Convergence criteria and admissible settings are given in Appendix D.5.

The full bound is often conservative. At $\beta = 2$, the mean quadratic discrepancy is 0.0091, compared with a mean bound of 0.2564. Increasing β to 4 reduces the mean bound to 0.0413, while quadratic test R^2 falls to 0.075. Thus, a tighter bound need not imply better target recovery. These numerical checks concern the specified image mixture, not the unknown population distribution of handwriting.

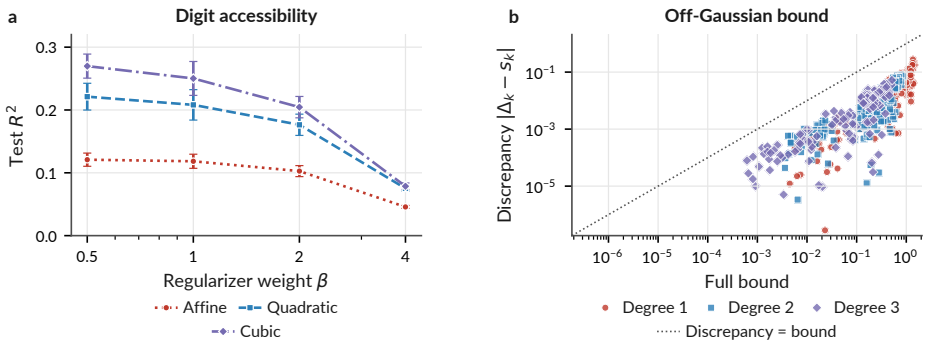


Figure 3: **(a)** Mean one-vs-rest test R^2 across ten digit classes, evaluated using posterior samples from 10,000 test images; error bars show ± 1 sample standard deviation across five training seeds. **(b)** Population-increment discrepancy versus the theoretical upper bound across 500 converged class/degree/model evaluations on a 1,024-image mixture. Markers denote polynomial degrees $k \in \{1, 2, 3\}$ (circles, squares, and diamonds, respectively); dotted line indicates $(y = x)$.

6 Decoding Board State in OthelloGPT

OthelloGPT is trained to predict moves from game sequences without being given board states or explicit game rules (Li et al., 2023). Earlier probing studies found that MLP probes recover board states more successfully than linear probes under black/white labels (Li et al., 2023). Linear recovery improves when the labels instead distinguish the pieces belonging to the player whose turn comes next from the opponent's pieces (Nanda et al., 2023b; Hazineh et al., 2023). We compare affine, quadratic, and MLP probes under both target definitions to test how much of this recovery gap can be explained by degree-two readouts.

We extract activations after each of the eight transformer blocks and evaluate them using affine, quadratic, and MLP probes. We define the target for each of the 64 board squares immediately following a move. Under the absolute encoding, we assign $+1$ to black, -1 to white, and 0 to empty. Under the player-relative encoding, $+1$ denotes pieces belonging to the player making the next legal move, -1 denotes the opponent's pieces, and empty squares remain 0. We compute the test R^2 independently for each square and report the average across the board. This measures regression performance on signed color labels, not three-class classification accuracy.

To reduce the cost of fitting quadratic probes, we also select a 17-dimensional subspace separately for each square and target definition (Appendix E.2). We report results for both the full 512-dimensional activation space and the selected subspaces. Data splits and probe-fitting settings are given in Appendices E.1 and E.3.

The empirical results are shown in Figure 4. Under the black/white target encoding within the full activation space (Figure 4a), affine recovery remains near zero at every layer. Conversely, the quadratic probe recovers substantial target variance across intermediate layers, peaking at $R^2 \approx 0.66$ at layer 4, before recovery falls near zero in the final layer. In contrast, the MLP maintains high recovery across all layers. With player-relative target labels (Figure 4c), the affine, quadratic, and MLP probes achieve similar recovery in the full space, although scores decline toward the final layer.

Further, applying dimensionality reduction (Figure 4b, d) largely preserves the relative performance hierarchy of the polynomial probes at intermediate depths, but recovery drops in the later layers.

Reduced-space MLP recovery also drops sharply for black/white targets. This comparison reflects both projection and a change in MLP architecture, since the reduced-space networks are fitted separately for each square (Appendix E.3). Finally, we check the Gram condition in Proposition 2 using selected, whitened subspaces at layers 3 and 7 (Appendix E.4). When positions from all moves are pooled, the empirical deviation ranges from 12.4 to 79.8, which exceeds the threshold $\rho_2 < 1$ stated in Proposition 2. We therefore interpret the probe scores as measures of predictive accessibility, not as calibrated Hermite-degree energies. A separate check using a known quadratic target is reported in Appendix E.5.

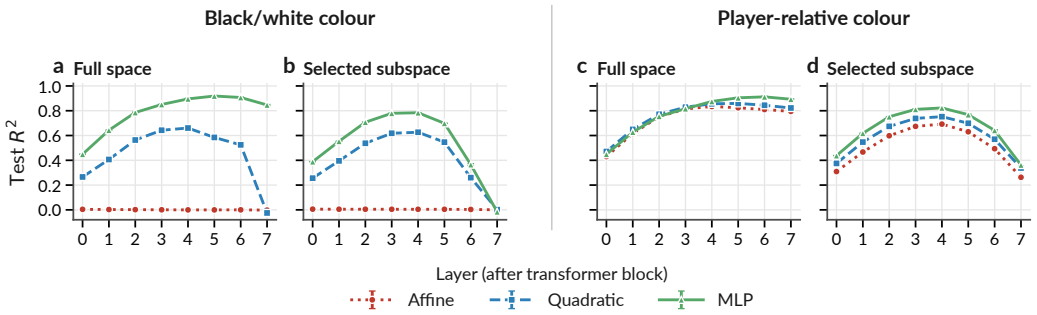


Figure 4: Panels compare each target before and after supervised reduction. Points show mean test R^2 across three refitted game splits; bars show sample standard deviations.

7 Order Spectrum in JEPA

Joint-Embedding Predictive Architectures (JEPAs) (LeCun, 2022) have emerged as a dominant paradigm in self-supervised learning, particularly for modeling sequence dynamics (Sobal et al., 2022; Assran et al., 2023). In this section, we examine their latent spaces through the lens of the order spectrum. In particular, we analyze the impact of SIGReg (Balestriero & LeCun, 2025). This regularizer encourages one-dimensional projections of the latent space to match a standard Gaussian. We test whether SIGReg is sufficient to yield a clean, theoretically bounded interpretation of Hermite energies. Specifically, we train a JEPA with GELU encoder and predictor networks using next-step latent prediction and probe its activations for energy. For a pendulum state, defined by (θ, p), the energy is given as $H(\theta, p) = p^2/2 + 1 - \cos\theta$. Appendix F.1 gives training details, and Appendix F.3 describes probe fitting and evaluation.

To determine whether the learned representations support a quantitative decomposition by Hermite degree, we first evaluate the spectral assumption underpinning Proposition 2. We find that despite SIGReg's regularization of one-dimensional marginals, the full multivariate Hermite-feature Gram matrix deviates significantly. Specifically, the empirical Gram deviation $\hat{\rho}_2$ ranges from 6.230 to 8.698, exceeding the threshold in the sufficient condition $\rho_2 < 1$ (Appendix F.4). The mean deviation decreases as the regularizer strength increases. In these experiments, SIGReg does not reduce the empirical Gram deviation below one, unlike the admissible VAE settings in Section 5. This does not establish a general limitation of projection-based regularization.

Figure 5 illustrates the fitted probes' predictive capacity. These empirical checks do not establish the sufficient condition in Proposition 2, so the bound does not justify interpreting the empirical R^2 increments as calibrated Hermite-degree energies. The fitted probes nevertheless show a clear gap between affine and quadratic recovery under SIGReg. Across all tested regularization weights, mean affine test R^2 ranges between 0.025 and 0.037. In contrast, the quadratic probe recovers the underlying physical energy, achieving a mean test R^2 between 0.986 and 0.991, closely tracking the mean MLP scores (0.996–0.997). Furthermore, we contrast SIGReg against a control baseline that penalizes only the first two moments (Zbontar et al., 2021) (Appendix F.2). Under this control, the affine probe recovers over 88% of the energy variance on average across model seeds. This affine accessibility emerges despite the baseline exhibiting an even higher empirical Gram deviation ($\hat{\rho}_2 = 10.737$–11.218) than the SIGReg models.

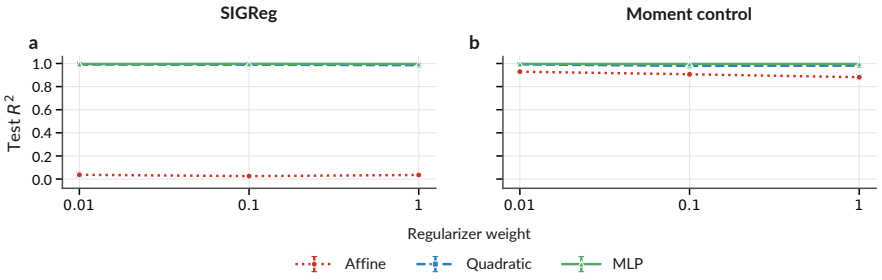


Figure 5: Data points indicate mean test R^2 across five training seeds, with error bars representing ± 1 standard deviation. Under SIGReg, fitted affine probes recover little energy variance, while quadratic probes recover nearly all of it. The moment control permits high affine recovery and does not show the same separation.

8 Quadratic Accessibility During Grokking

Finally, we apply our polynomial probing hierarchy to analyze how target accessibility evolves during training. We focus on the phenomenon of grokking, where generalization is substantially delayed relative to training accuracy (Power et al., 2022). This transition has been linked to changes in the learned representation (Liu et al., 2022; Kumar et al., 2024; Mohamadi et al., 2024). In the context of modular addition, models often generalize by learning Fourier features that compute sums via trigonometric identities, such as $\cos(x+y) = \cos x \cos y - \sin x \sin y$ (Nanda et al., 2023a; Mallinar et al., 2025). The multiplicative nature of this known solution naturally motivates evaluating quadratic accessibility as generalization emerges.

We train a one-layer Transformer on modular addition $(a+b) \bmod 97$. Across training checkpoints, we extract the final-token residual stream activations immediately preceding the Transformer MLP. The probing targets are the sine and cosine components of the dominant frequencies identified from the fully trained model's logits, which remain fixed across all evaluations. To isolate the role of regularization, we compare models trained with and without weight decay (Varma et al., 2023). Full training and probing protocols are detailed in Appendices G.1 and G.2.

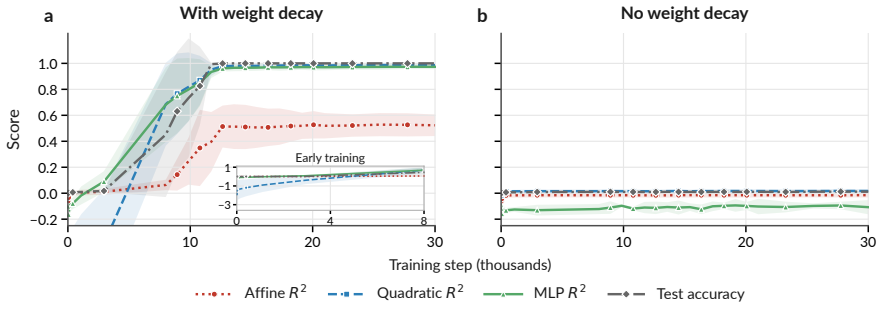


Figure 6: Probes are fitted to the residual stream immediately preceding the Transformer MLP and evaluated on pairs excluded from probe fitting. With weight decay, quadratic and MLP R^2 rise alongside model test accuracy, while affine R^2 remains lower. Without weight decay, probe scores and model test accuracy remain low. Curves denote means over three training seeds; shaded regions indicate ± 1 SD. The inset captures early negative R^2 values truncated from the main axes.

Figure 6 illustrates the probe R^2 alongside held-out model accuracy. Quadratic recovery rises as the model begins to generalize and exceeds affine recovery later in training. At the conclusion of training with weight decay, the affine R^2 ranges from 0.427 to 0.572, whereas the quadratic probe achieves 0.974–0.996, closely matching the MLP (0.952–0.988). Strikingly, across all weight-decay seeds, quadratic R^2 and model test accuracy cross the 0.5 threshold within two recorded checkpoints

of each other. Conversely, absent weight decay, the models purely memorize the training set: final test accuracy is 0.006–0.016, and the final quadratic R^2 is 0.005–0.034.

This final quadratic advantage holds across all selected frequency components and remains consistent even if target frequencies are extracted earlier in training (Appendix G.4). Interestingly, probing activations after the Transformer MLP shows that affine recovery nearly matches quadratic recovery in the weight-decay runs (Appendix G.3), aligning with the MLP's capacity to compute the required multiplicative interactions. While these results are highly consistent with the multiplicative structure of the hypothesized Fourier circuit, they do not strictly prove that the model natively implements this exact quadratic readout, nor do they validate a formal Hermite-energy interpretation of the representations (Appendix G.5).

9 Related Work

Probes study learned representations by predicting properties from frozen activations. Linear probes reveal how predictive information becomes linearly accessible across network depth (Alain & Bengio, 2017). Their interpretation also depends on the readout and evaluation procedure: control tasks assess memorization (Hewitt & Liang, 2019), while minimum-description-length probing accounts for the cost of learning a predictor (Voita & Titov, 2020). These perspectives motivate examining accessibility across a controlled hierarchy of readouts. Different probing objectives emphasize different aspects of accessibility. Information-theoretic probing motivates expressive predictors for estimating mutual information (Pimentel et al., 2020), whereas usable information formalizes what computationally constrained predictor families can exploit (Xu et al., 2020). Zhang et al. (2026) use polynomial expansions along data-interpolation paths to measure network complexity and study generalization, including grokking. Structured nonlinear probes extend this analysis beyond linear recovery. The structural probe of Hewitt & Manning (2019) predicts syntactic distances and depths using quadratic functions of representations. Oldfield et al. (2026) develop progressively trained polynomial activation probes whose degree-wise truncations support adaptive safety monitoring. Degree-wise variance decompositions are classical in polynomial-chaos analysis (Xiu & Karniadakis, 2002; Sudret, 2008), and nested predictor families are related to conditional probing (Hewitt et al., 2021). Our contribution is to connect polynomial degree to nested probe R^2 increments and characterize when these increments admit a Gaussian-reference spectral interpretation. Hermite-degree decompositions also underpin recent identifiability results for LeJEPA under Gaussian latent dynamics (Klindt et al., 2026). A complementary foundation comes from Lee & Kondor (2026), who characterize nonzero finite-dimensional, affine-invariant linear spaces of continuous scalar readouts as bounded-degree polynomial spaces. This symmetry-based perspective complements the variance interpretation of polynomial-probe increments developed in this work.

10 Conclusion

In this work, we introduced a principled framework to decompose the predictable variance of targets given latent representations across orthogonal Hermite polynomial degrees. By formalizing the **order spectrum**, we demonstrated that under a Gaussian representation assumption, the population R^2 of an optimal affine probe exactly captures the degree-one energy, while performance increments from higher-order nested probes systematically recover successive spectral energies. We also derived a rigorous upper bound on the population discrepancy between nested probe increments and Gaussian-reference degree energies in off-Gaussian regimes. Our results support polynomial probing with both a quantitative theoretical grounding and explicit validity boundaries. While the polynomial order of a successful probe establishes target accessibility within a given coordinate basis, formal spectral calibration fundamentally depends on representation geometry and target approximation residuals. Consequently, the order-spectrum theory provides a quantitative interpretation when the required representation geometry and approximation conditions hold; outside this regime, the associated polynomial probe hierarchy remains a descriptive measure of degree-limited predictive accessibility rather than a calibrated spectral decomposition. Future work should extend this framework to include finite-sample estimation guarantees and calibration bounds under broader representation families. Ultimately, coupling these spectral measurements with causal interventions will clarify whether the functional complexity required to read out a latent target accurately reflects the network's internal mechanistic computations.

AI USE STATEMENT

In this work, we used generative AI tools to develop experimental code, find relevant literature, draft the initial manuscript structure, polish the prose, and assist with proof development and verification. We did not use generative AI tools for research ideation, final proof checks, reference checks, or internal reviews. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text states the assumptions underlying our theoretical results, with proofs provided in the appendix. The experimental appendices describe the datasets, model configurations, data splits, probe-fitting procedures, and evaluation protocols. They also provide additional results and controls supporting the main experiments. Code for reproducing the experiments will be publicly released.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 66–75. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/bauer19a.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://doi.org/10.1162/coli_a_00422.

Lisa Bonheme and Marek Grzes. Be more active! understanding the differences between mean and sampled representations of variational autoencoders. *Journal of Machine Learning Research*, 24 (324):1–30, 2023. URL http://jmlr.org/papers/v24/21-1145.html.

Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Dean S. Hazineh, Zechen Zhang, and Jeffery Chiu. Linear latent world models in simple transformers: A case study on Othello-GPT, 2023. URL https://arxiv.org/abs/2310.07582.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis,

Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. Conditional probing: measuring usable information beyond a baseline. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.122. URL https://aclanthology.org/2021.emnlp-main.122/.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model?, 2026. URL https://arxiv.org/abs/2605.26379.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vt5mnLVIVo.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Su Hyeong Lee and Risi Kondor. Deep minds and shallow probes, 2026. URL https://arxiv.org/abs/2605.11448.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 34651–34663. Curran Associates, Inc., 2022. doi: 10.52202/068431-2511. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/dfc310e81992d2e4cedc09ac47eff13e-Paper-Conference.pdf.

Neil Rohit Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42834–42856. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/mallinar25a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/meng2023iclr-massediting/.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? A theoretical analysis on grokking modular addition. 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. 35934–35967. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/mohamadi24a.html.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/forum?id=9XFSbDPmdW.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

James Oldfield, Philip Torr, Ioannis Patras, Adel Bibi, and Fazl Barez. Beyond linear probes: Dynamic safety monitoring for language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AGWa8whf92.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420/.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Vlad Sobal, Jyothir S V, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features, 2022. URL https://arxiv.org/abs/2211.10831.

Bruno Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2007.04.002. URL https://www.sciencedirect.com/science/article/pii/S0951832007001329. Bayesian Networks in Dependability.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency, 2023. URL https://arxiv.org/abs/2309.02390.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Dongbin Xiu and George Em Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002. doi: 10.1137/S1064827501387826. URL https://doi.org/10.1137/S1064827501387826.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Tianren Zhang, Xiangxin Li, Minghao Xiao, Guanyu Chen, and Feng Chen. Quantifying and optimizing simplicity via polynomial representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=F0F03HOYb8.

A PROOF OF PROPOSITION 1

Proof. Assume $Z \sim \gamma = \mathcal{N}(0, I_d)$, as in Proposition 1. The population coefficient of determination R^2 for the optimal affine estimator $g(Z) = a^\top Z + b$ minimizing the mean squared error in predicting the true target Y is defined as:

$$R^2 = 1 - \frac{\min_{a \in \mathbb{R}^d, b \in \mathbb{R}} \mathbb{E}\left[(Y - (a^\top Z + b))^2\right]}{\mathrm{Var}(Y)}.$$

By adding and subtracting the conditional expectation $f(Z) = \mathbb{E}[Y \mid Z]$, we can decompose the mean squared error for any predictor $g(Z)$:

$$\mathbb{E}\left[(Y - g(Z))^2\right] = \mathbb{E}\left[(Y - f(Z) + f(Z) - g(Z))^2\right].$$

Because $f(Z)$ is the conditional expectation of Y given Z, the residual $Y - f(Z)$ is orthogonal to every square-integrable function of Z. Therefore, the cross-term vanishes, yielding:

$$\mathbb{E}\left[(Y - g(Z))^2\right] = \mathbb{E}\left[(Y - f(Z))^2\right] + \mathbb{E}\left[(f(Z) - g(Z))^2\right].$$

Since the irreducible error $\mathbb{E}[(Y - f(Z))^2]$ is independent of the choice of g, finding the optimal affine parameters (a^*, b^*) is equivalent to minimizing the squared $L^2(\gamma)$ distance to $f(Z)$, namely $\|f - g\|^2_{L^2(\gamma)}$.

Any affine function $g(z)$ resides in the direct sum of the first two Hermite subspaces $H_0 \oplus H_1$, where $H_0 = \mathrm{span}\{\mathrm{He}_0\}$ and $H_1 = \mathrm{span}\{\mathrm{He}_{e_i}\}_{i=1}^d$. Thus, the optimal affine predictor g^* is exactly the orthogonal projection of f onto $H_0 \oplus H_1$:

$$g^* = \mathrm{Proj}_{H_0 \oplus H_1} f = \mathrm{Proj}_{H_0} f + \mathrm{Proj}_{H_1} f.$$

By assumption, the target Y is centered ($\mathbb{E}[Y] = 0$). By the tower property, $f(Z)$ is also centered ($\mathbb{E}[f(Z)] = \mathbb{E}[\mathbb{E}[Y \mid Z]] = \mathbb{E}[Y] = 0$). This implies $\mathrm{Proj}_{H_0} f = 0$, and consequently, $g^* = \mathrm{Proj}_{H_1} f$.

Substituting g^* back into our MSE decomposition gives the minimal error for the optimal affine predictor:

$$\mathbb{E}\left[(Y - g^*(Z))^2\right] = \mathbb{E}\left[(Y - f(Z))^2\right] + \|f - \mathrm{Proj}_{H_1} f\|^2_{L^2(\gamma)}.$$

Because $f \in L^2(\gamma)$ admits the orthogonal decomposition $f = \sum_{k=1}^\infty \mathrm{Proj}_{H_k} f$ (with the zero-degree term being zero), the norm of the remaining residual expands as:

$$\|f - \mathrm{Proj}_{H_1} f\|^2_{L^2(\gamma)} = \sum_{k=2}^\infty \|\mathrm{Proj}_{H_k} f\|^2_{L^2(\gamma)}.$$

Furthermore, by the law of total variance and Parseval's identity applied to f, the total variance of Y can be decomposed as:

$$\mathrm{Var}(Y) = \mathbb{E}\left[(Y - f(Z))^2\right] + \mathrm{Var}(f(Z)) = \mathbb{E}\left[(Y - f(Z))^2\right] + \sum_{k=1}^\infty \|\mathrm{Proj}_{H_k} f\|^2_{L^2(\gamma)}.$$

Finally, substituting these components into our definition of R^2, we obtain:

$$\begin{aligned} R^2 &= 1 - \frac{\mathbb{E}\left[(Y - f(Z))^2\right] + \sum_{k=2}^\infty \|\mathrm{Proj}_{H_k} f\|^2_{L^2(\gamma)}}{\mathrm{Var}(Y)} \\ &= \frac{\mathrm{Var}(Y) - \left(\mathbb{E}\left[(Y - f(Z))^2\right] + \sum_{k=2}^\infty \|\mathrm{Proj}_{H_k} f\|^2_{L^2(\gamma)}\right)}{\mathrm{Var}(Y)} \\ &= \frac{\|\mathrm{Proj}_{H_1} f\|^2_{L^2(\gamma)}}{\mathrm{Var}(Y)}. \end{aligned} \tag{6}$$

According to Definition 1, this ratio is exactly $s_1(Y; Z)$, concluding the proof. □

B Proof of Proposition 2

Following Proposition 1, a conditional extension holds for non-Gaussian representations. Let $Z \in \mathbb{R}^d$ with $d \geq 1$, $\mu = \mathrm{Law}(Z)$, $\gamma = \mathcal{N}(0, I_d)$, and $0 < \sigma^2 = \mathrm{Var}(Y) < \infty$, with $\mathbb{E}Y = 0$. Fix an integer $k \geq 1$, assume $\mathbb{E}_\mu \|Z\|_2^{2k} < \infty$, and specify a measurable function $f \in L^2(\mu) \cap L^2(\gamma)$ on $\mathbb{R}^d$ satisfying $f(Z) = \mathbb{E}[Y \mid Z]$ almost surely. Let $\Phi_{\leq k}$ contain all $p_k = \binom{d+k}{k}$ Gaussian-orthonormal Hermite features, including the intercept, and suppose $G_k = \mathbb{E}_\mu[\Phi_{\leq k}\Phi_{\leq k}^\mathsf{T}]$ satisfies $\|G_k - I\|_{\mathrm{op}} \leq \rho < 1$. Writing $P^\gamma_{\leq k}$ for Gaussian orthogonal projection, $f_k = \mathrm{Proj}_{H_k} f$, and $\tau_k = \|f - P^\gamma_{\leq k} f\|_{L^2(\mu)}/\sigma$, the population nested increment $\Delta_k = R^2_{\leq k} - R^2_{\leq k-1}$ satisfies

$$|\Delta_k - s_k| \leq \rho s_k + 2\tau_k\sqrt{(1+\rho)s_k} + \tau_k^2, \qquad s_k = \frac{\|f_k\|^2_{L^2(\gamma)}}{\sigma^2}. \tag{7}$$

In particular, if f is a polynomial of degree at most k, then $|\Delta_k - s_k| \leq \rho s_k$ at that tested degree.

Proof. All vector norms below are Euclidean, matrix operator norms are induced by them, and tensor Frobenius norms sum over all ordered indices. Let $\Phi = \Phi_{\leq k}$ and $G = G_k$. The normalized probabilists' Hermite features are

$$h_\alpha(z) = \frac{\prod_{i=1}^d \mathrm{He}_{\alpha_i}(z_i)}{\sqrt{\alpha!}}, \qquad \alpha! = \prod_{i=1}^d \alpha_i!, \qquad |\alpha| \leq k.$$

Their Gaussian orthonormality gives

$$a = \mathbb{E}_\gamma[f\Phi], \qquad P^\gamma_{\leq k} f = a^\mathsf{T}\Phi, \qquad h = f - a^\mathsf{T}\Phi.$$

The moment assumption makes every feature square-integrable under μ, so $h \in L^2(\mu)$ and all expectations used below exist. Partition $\Phi = (X^\mathsf{T}, W^\mathsf{T})^\mathsf{T}$ into degrees at most $k-1$ and exactly k, respectively, and partition $a = (a_-, a_+)$ accordingly. These blocks have dimensions $p_{k-1} = \binom{d+k-1}{k-1}$ and $m_k = \binom{d+k-1}{k}$, with $p_{k-1} + m_k = p_k$. In particular, $\sigma^2 s_k = \|a_+\|_2^2$.

Write

$$G = \begin{pmatrix} A & B \\ B^\mathsf{T} & C \end{pmatrix}.$$

The Gram assumption implies $(1-\rho)I \preceq G \preceq (1+\rho)I$. Consequently, G and A are positive definite. Define

$$S = C - B^\mathsf{T}A^{-1}B, \qquad \Psi = W - B^\mathsf{T}A^{-1}X.$$

Completing the block quadratic form gives, for every $y \in \mathbb{R}^{m_k}$,

$$\begin{pmatrix} x \\ y \end{pmatrix}^\mathsf{T} G \begin{pmatrix} x \\ y \end{pmatrix} = (x + A^{-1}By)^\mathsf{T} A (x + A^{-1}By) + y^\mathsf{T} S y,$$

$$y^\mathsf{T} S y = \min_x \begin{pmatrix} x \\ y \end{pmatrix}^\mathsf{T} G \begin{pmatrix} x \\ y \end{pmatrix}.$$

The lower Gram bound applies to every x, while evaluating the minimum at $x = 0$ gives the upper bound. Thus

$$(1-\rho)I \preceq S \preceq (1+\rho)I. \tag{8}$$

In particular, S is invertible, both $\|A^{-1}\|_{\mathrm{op}}$ and $\|S^{-1}\|_{\mathrm{op}}$ are at most $(1-\rho)^{-1}$, and their spectral condition numbers are at most $(1+\rho)/(1-\rho)$. Direct expansion yields

$$\mathbb{E}_\mu[X\Psi^\mathsf{T}] = 0, \qquad \mathbb{E}_\mu[\Psi\Psi^\mathsf{T}] = S, \qquad \mathbb{E}_\mu[W\Psi^\mathsf{T}] = S.$$

For a feature vector F with positive-definite Gram matrix M and $c = \mathbb{E}[YF]$, completing the least-squares risk gives

$$\mathbb{E}[(Y - \beta^\mathsf{T}F)^2] = \sigma^2 - c^\mathsf{T}M^{-1}c + (\beta - M^{-1}c)^\mathsf{T} M (\beta - M^{-1}c).$$

Its minimum is $\sigma^2 - c^\mathsf{T}M^{-1}c$ and is attained at $\beta = M^{-1}c$. The map from (X, W) to (X, Ψ) is invertible and preserves the lower-degree span. The latter feature vector has a block- diagonal Gram

matrix $\mathrm{diag}(A, S)$. Applying the displayed risk identity to the full and lower-degree feature vectors therefore proves the exact nested formula

$$\sigma^2 \Delta_k = r^\mathsf{T} S^{-1} r, \qquad r = \mathbb{E}[Y\Psi]. \tag{9}$$

This is the additional explained variance after residualizing the new degree block against all lower-degree features.

The conditional-mean assumption and the decomposition of f give

$$r = \mathbb{E}_\mu[f\Psi] = Sa_+ + e, \qquad e = \mathbb{E}_\mu[h\Psi].$$

To bound the contribution of e, put $L = e^\mathsf{T} S^{-1} e \geq 0$. If $L > 0$, take $b = S^{-1}e$. Cauchy–Schwarz gives

$$L^2 = (\mathbb{E}_\mu[h\, b^\mathsf{T}\Psi])^2 \leq \mathbb{E}_\mu[h^2]\, \mathbb{E}_\mu[(b^\mathsf{T}\Psi)^2] = \|h\|^2_{L^2(\mu)} L.$$

Dividing by L proves the following inequality; it also holds when $L = 0$:

$$\|S^{-1/2} e\|_2^2 = L \leq \|h\|^2_{L^2(\mu)} = \sigma^2 \tau_k^2. \tag{10}$$

Expanding equation 9, and then using equation 8 and equation 10, yields

$$\begin{aligned} \left|\sigma^2 \Delta_k - \|a_+\|_2^2\right| &\leq \left|a_+^\mathsf{T}(S - I)a_+\right| + 2|a_+^\mathsf{T} e| + e^\mathsf{T} S^{-1} e \\ &\leq \rho \|a_+\|_2^2 + 2\|S^{1/2} a_+\|_2 \|S^{-1/2} e\|_2 + \sigma^2 \tau_k^2 \\ &\leq \rho \|a_+\|_2^2 + 2\sqrt{1+\rho}\, \|a_+\|_2 \sigma \tau_k + \sigma^2 \tau_k^2. \end{aligned}$$

Division by σ^2 proves equation 4. If the specified function f is a polynomial of total degree at most k, its Gaussian projection is f itself, so $h = 0$ and the stated relative bound follows at degree k.

It is important to note that the condition $\rho < 1$ provides a convenient sufficient criterion for positive definiteness and quantitative control of Gram conditioning. The discrepancy bound also holds whenever $G_k \succ 0$ and $\|G_k - I\|_{\mathrm{op}} \leq \rho$, including $\rho \geq 1$. Indeed, positive definiteness gives $A \succ 0$ and $S \succ 0$, while $S \preceq C \preceq (1+\rho)I$. For $\rho \geq 1$, these inequalities imply $\|S - I\|_{\mathrm{op}} \leq \rho$, so the remaining discrepancy argument applies unchanged.

We focus on $\rho < 1$ because, when $\tau_k = 0$ and $s_k > 0$, it guarantees

$$\frac{|\Delta_k - s_k|}{s_k} \leq \rho < 1.$$

For $\rho \geq 1$, the bound remains valid under positive definiteness but no longer guarantees relative discrepancy below 100%. □

C Synthetic Validation Details

C.1 Mixed-order targets with irreducible noise

Figure 2a uses seeds 0–9 and disjoint 12,000/6,000 fitting/test samples in four dimensions. Within each seed, independent random-number streams generate representations, observation noise, and a permutation of the fitting labels. Both noise conditions share the same representations and noise draw. We sample $Z \sim \mathcal{N}(0, I_4)$. The regression function is $f(Z) = \sqrt{0.2} Z_1 + \sqrt{0.3} Z_1 Z_2 + \sqrt{0.5/6}\mathrm{He}_3(Z_3)$. Since $\mathbb{E}[\mathrm{He}_3(Z_3)^2] = 3! = 6$, its orthogonal components have variances 0.2, 0.3, and 0.5. We generate targets as $Y = f(Z) + \sigma\xi$, where $\xi \sim \mathcal{N}(0,1)$ is independent of Z, and use $\sigma \in \{0, 1\}$. Thus $\mathrm{Var}(Y) = 1 + \sigma^2$ and the total predictable fraction is $1/(1+\sigma^2)$.

We use float64 ordinary least squares with an intercept and all multivariate Hermite terms up to the tested degree. The cumulative designs have 5, 15, 35, and 70 columns at degrees 1–4. No regularization, dimensionality reduction, or hyperparameter selection is used. Unnormalized probabilists' Hermite features span the same families as normalized features, so this does not change unregularized predictions in exact arithmetic. Test R^2 uses the test target mean in its denominator. The degree-one estimate subtracts the held-out score of the fitted constant predictor; higher estimates subtract consecutive cumulative scores. Finite-sample estimates can therefore be negative, and their sum can slightly exceed one. Values are not clipped or renormalized.

Table 1: Known and estimated degree energies. Means ± sample standard deviations over ten seeds.

	Noiseless ($\sigma = 0$)		Noisy ($\sigma = 1$)	
Degree	Analytic	Estimated	Analytic	Estimated
1	0.200	0.203 ± 0.015	0.100	0.096 ± 0.008
2	0.300	0.310 ± 0.021	0.150	0.153 ± 0.010
3	0.500	0.487 ± 0.025	0.250	0.240 ± 0.015
4	0.000	0.000 ± 0.000	0.000	-0.002 ± 0.001

The shuffled-label control permutes fitting targets and refits each degree, evaluating against unchanged test targets. Across both noise conditions and all degrees, mean shuffled cumulative R^2 ranges from -0.0076 to 0.0014. The known conditional mean scores 0.4908 ± 0.0150 on the noisy test sets; its finite-test score need not equal the population ceiling 0.5. Cubic recovery is 1.000 at every seed without noise. With noise, cubic and quartic fitted scores are 0.4892 ± 0.0148 and 0.4872 ± 0.0151. The fourth-degree increment is -0.0020 ± 0.0009, a finite-sample effect despite nested training function classes.

C.2 Pure-order and coordinate checks

The following experiments use seeds 0–4 and disjoint training and test sets. The coordinate experiments use $d = 2$ and 8,000/4,000 train/test samples. The known-degree and mixed-degree experiments use $d = 4$ with 10,000/5,000 and 12,000/6,000 samples, respectively. Inputs are independent standard Gaussian variables with no preprocessing. The continuous transformation uses $a \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$.

Affine probes use an intercept and all linear terms. Quadratic probes use an intercept, all linear terms, and all distinct products $Z_i Z_j$ for $i \le j$; both are fit by ordinary least squares. The MLP has two width-64 GELU layers and is trained by full-batch mean-squared error for 400 epochs using Adam with learning rate 10^{-3} and no weight decay. Inputs are not normalized, and we report the final iterate without early stopping or model selection. All table entries are held-out R^2, reported as mean ± sample standard deviation.

Table 2: Held-out R^2 on targets of known Hermite degree over five seeds.

Target	Degree	Affine	Quadratic	MLP
Z_1	1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
$0.7Z_1 - 0.4Z_2 + 0.2Z_3$	1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
$Z_1 Z_2$	2	0.000 ± 0.001	1.000 ± 0.000	0.998 ± 0.001
$\mathrm{He}_2(Z_1)$	2	-0.001 ± 0.001	1.000 ± 0.000	0.998 ± 0.001
$\mathrm{He}_2(Z_1) + \mathrm{He}_2(Z_2)$	2	-0.001 ± 0.002	1.000 ± 0.000	0.997 ± 0.001
$\mathrm{He}_3(Z_1)$	3	-0.001 ± 0.003	-0.009 ± 0.010	0.986 ± 0.007

For each target and probe family, we independently permute the training targets and evaluate the fitted probe against the unchanged test targets. At the largest sample size, the mean shuffled-target scores range from -0.038 to 0.009. The sample-size analysis fits the first $N \in \{100, 300, 1000, 3000, 8000\}$ training examples and evaluates every fit on the full test set. For the MLP on $\mathrm{He}_3(Z_1)$, held-out R^2 increases from 0.684 ± 0.142 at $N = 100$ to 0.986 ± 0.008 at $N = 8000$; its shuffled-target score remains non-positive.

For the mixed-degree target $Y = 0.6\mathrm{He}_1(Z_1) + 0.4\mathrm{He}_2(Z_1)$, Hermite orthogonality gives

$$s_1 = \frac{0.6^2}{0.6^2 + 2(0.4^2)} = 0.529, \qquad s_2 = \frac{2(0.4^2)}{0.6^2 + 2(0.4^2)} = 0.471.$$

The estimator fits cumulative ordinary least-squares models through degree 4 on the 12,000 training samples and computes marginal held-out R^2 gains on the 6,000 test samples.

The degree-through-four runs also test the pure cubic target $\mathrm{He}_3(Z_1)$. The cubic probe has $R^2 = 1.000$ at every seed, compared with -0.0008 ± 0.0011 for affine and -0.0048 ± 0.0064

for quadratic probes on that run's 12,000/6,000 split. These are separate fits from the 10,000/5,000 split in Table 2. The Gaussian law is specified by construction; spectrum recovery is checked against analytic energies.

Figure 2b reports the product reparameterization from Section 4, using the five-seed 8,000/4,000 protocol above. The inverse is $T^{-1}(u_1, u_2) = (u_1/u_2, u_2)$ for $u_2 \neq 0$. The MLP on the original coordinates reaches $R^2 = 0.998 \pm 0.001$. A continuous coordinate check in Figure 7a uses $Y = Z_1$ and $U_a = (\sinh(aZ_1)/a, Z_2)$, with $U_0 = Z$. The optimal population affine score is $a^2/\sinh(a^2)$, with limit one at $a = 0$, despite the transformation remaining invertible. At $a = 3$, the fitted MLP becomes unreliable, illustrating that a trained nonlinear probe need not attain optimal recovery. These transformed representations are non-Gaussian; the examples establish coordinate-dependent accessibility, not calibrated Hermite fractions after transformation.

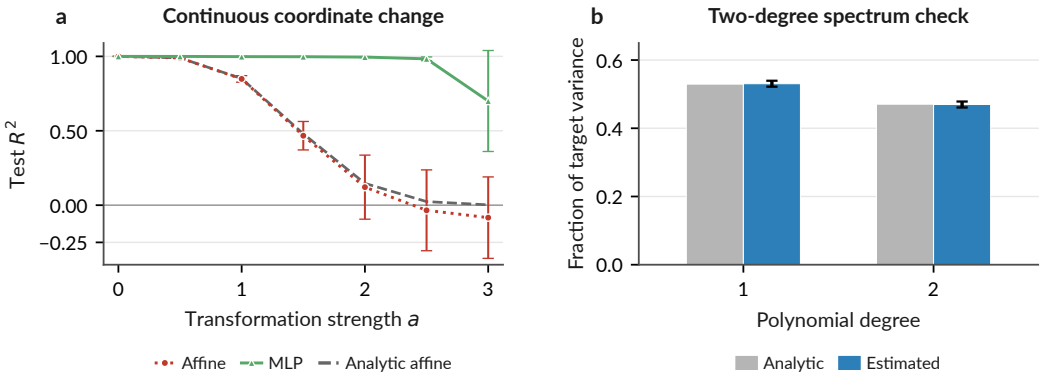


Figure 7: **(a)** Affine and MLP recovery of Z_1 under the continuous sinh transformation, alongside the analytic affine score. **(b)** Known and estimated degree-1/2 energies for $0.6\mathrm{He}_1(Z_1) + 0.4\mathrm{He}_2(Z_1)$. Error bars show sample standard deviations over five seeds. Panel a omits quadratic probes because joint sign symmetry makes the added quadratic terms orthogonal to the odd target and linear features, so optimal population quadratic and affine recovery coincide; finite-sample fits need not coincide.

C.3 Off-Gaussian stability sweep

For Figure 1, we use 61 evenly spaced skew values $\epsilon \in [-1.2, 1.2]$ and 39 tail values $\delta \in [0.55, 1.50]$, for 2,379 settings. Population expectations are evaluated by 160-node Gauss–Hermite quadrature, including the mean and variance used to standardize each transformation. We check numerical convergence by repeating the calculations at quadrature orders 100 and 240. Specifically, for $X \sim \mathcal{N}(0, 1)$ we set

$$Z_{\epsilon,\delta} = \mathrm{standardize}[\sinh(\delta(\mathrm{arcsinh}(X) - \epsilon))]\,.$$

Writing $m_j = \mathbb{E}[Z^j]$, population standardization gives $m_1 = 0$ and $m_2 = 1$. For the Gaussian-orthonormal feature vector $\Phi_{\leq 2} = (1, Z, (Z^2 - 1)/\sqrt{2})^\mathsf{T}$,

$$G_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & m_3/\sqrt{2} \\ 0 & m_3/\sqrt{2} & (m_4 - 1)/2 \end{pmatrix}, \quad \rho_2 = \frac{\sqrt{b^2 + 4a^2} + |b|}{2}, \quad a = \frac{m_3}{\sqrt{2}}, \quad b = \frac{m_4 - 3}{2}.$$

For $Y = 0.6Z + 0.4(Z^2 - 1)$,

$$\sigma^2 = 0.36 + 0.48m_3 + 0.16(m_4 - 1), \qquad s_2 = \frac{0.32}{\sigma^2},$$

and the exact nested increments are

$$\Delta_1 = \frac{(0.6 + 0.4m_3)^2}{\sigma^2}, \qquad \Delta_2 = 1 - \Delta_1.$$

We consequently compare $E_2 = |\Delta_2 - s_2|$ directly with $B_2 = \rho_2 s_2$. As a finite-sample check, we additionally draw 200,000 samples at 12 settings and fit the nested probes by least squares; the maximum absolute deviations from the population values are 0.0067 for Δ_1 and Δ_2, and 0.1071 for ρ_2.

C.4 Checking the nonzero approximation residual

To exercise the residual term in Proposition 2, we reuse the same 2,379 standardized distributions and set $q_c(z) = 0.6z + 0.4(z^2 - 1) + c\mathrm{He}_3(z)/\sqrt{6}$, $Y = q_c(Z) - \mathbb{E}_\mu q_c(Z)$ for $c \in \{0, 0.15, 0.30\}$. Write $v_c = \mathrm{Var}_\mu(q_c(Z))$. At degree two,

$$s_2 = \frac{0.32}{v_c}, \qquad \tau_2^2 = \frac{c^2 \mathbb{E}_\mu[\mathrm{He}_3(Z)^2]}{6v_c}.$$

Population centering changes only the constant Hermite coefficient. We compute Δ_2 from the two nested population least-squares projections, using the full uncentered Gram of Gaussian-normalized features including the intercept, rather than a centered correlation matrix.

Table 3: Numerical bound checks with and without the approximation-residual term. Each row uses the same 1,407 distributions satisfying $\rho_2 < 1$. Violations exceed an absolute tolerance of 10^{-10}.

c	Full-bound violations	Violations omitting the residual term
0	0	0
0.15	0	555
0.30	0	1,014

The full bound holds numerically at all tested admissible settings; the reduced expression $\rho_2 s_2$ has no general guarantee when $\tau_2 > 0$. Recomputing the entire grid at 100 and 240 quadrature nodes changes the 160-node full-bound values by at most 5.3×10^{-8} and 1.8×10^{-10}, respectively. Off Gaussian, s_2 is a Gaussian-reference energy divided by the actual target variance and can exceed one; these reference quantities are not a probability distribution over degrees.

D VAE Protocol and Off-Gaussian Validation

This appendix describes VAE training, held-out probe evaluation, and the numerical checks of the off-Gaussian bound. The probe scores measure recovery from sampled representations. The bound checks use a specified image-posterior mixture for which the regression function can be evaluated directly.

D.1 Data and VAE training

We average-pool MNIST images from 28×28 to 14×14 and scale pixel intensities to $[0, 1]$. A fixed permutation of the official training set, using seed 2027, provides 20,000 training images and 2,000 disjoint validation images. We hold out the official 10,000-image test set. All runs use the same image split.

The encoder is a 196–128–64 network with ReLU hidden activations and separate two-dimensional mean and log-variance heads. Log variances are clamped to $[-10, 6]$. The decoder is a 2–64–128–196 network with ReLU hidden activations and reconstruction logits as its output.

The training objective is summed-pixel binary cross entropy plus β times the analytic KL divergence to $\mathcal{N}(0, I_2)$. We use $\beta \in \{0.5, 1, 2, 4\}$ and initialization seeds 0–4, giving 20 models. Each model is trained for 20 epochs with Adam, a learning rate of 10^{-3}, and batch size 256. We select the checkpoint with the lowest validation objective.

Validation uses a fixed posterior-noise draw, with seed 8888, across epochs. This does not change the random-number stream used for training. Digit labels are not used for representation training or checkpoint selection. The five-seed study follows a two-seed pilot, so the overall study should not be treated as having an untouched test set.

D.2 Probe fitting and evaluation

We fit probes using the first 5,000 representation-training images and evaluate them on all 10,000 official test images. For each image, we draw four samples from its encoder posterior using sampling seed 989. We apply no additional noise, whitening, or coordinate selection. All distributional diagnostics concern sampled representations, whose dependence structure can differ from that of encoder means (Bonheme & Grzes, 2023).

We fit the complete polynomial families through degrees one, two, and three by ordinary least squares. Including the intercept, these families contain 3, 6, and 10 features. Each digit defines a binary one-vs-rest target; digit values are not treated as ordinal labels. We report the mean test R^2 across the ten targets. When reporting classification accuracy, we select the class with the largest regression output.

For a shuffled-label control, we permute image labels and fit cubic probes. All four posterior samples from an image retain the same shuffled label. Mean test R^2 ranges from -0.0127 to -0.0034 across the 20 models. Repeated samples from one image are not treated as independent images. Error bars show sample standard deviations across model-training seeds.

D.3 Image-mixture distribution and regression targets

To evaluate the bound, we use a fixed random subset of $M = 1024$ test images, selected with permutation seed 343 and shared by all models. Each model defines a mixture of the encoder posteriors for these images. This gives a known distribution and an explicit regression function, allowing us to evaluate the approximation residual in the bound.

Let image i have label y_i and encoder density

$$q_i(z) = \mathcal{N}(z; m_i, \operatorname{diag} v_i).$$

We sample an image index I uniformly from $\{1, \dots, M\}$ and then sample $Z \mid I = i \sim q_i$. The resulting density and class proportions are

$$q(z) = \frac{1}{M} \sum_i q_i(z), \qquad p_c = \frac{1}{M} \sum_i \mathbf{1}\{y_i = c\}.$$

For digit c, the centered target and its conditional mean are

$$Y_c = \mathbf{1}\{y_I = c\} - p_c, \qquad f_c(z) = \mathbb{E}[Y_c \mid Z = z] = \frac{\sum_{i:y_i=c} q_i(z)}{\sum_i q_i(z)} - p_c.$$

The target variance is $\sigma_c^2 = p_c(1 - p_c)$.

Positive encoder variances give every component full support. Thus f_c is defined over all of $\mathbb{R}^2$ and is bounded, so it belongs to both $L^2(q)$ and $L^2(\gamma)$ for $\gamma = \mathcal{N}(0, I_2)$.

The bound checks concern this finite mixture of learned posteriors, not the unknown population distribution of handwriting. It is also distinct from the mixture over all 10,000 test images used for the diagnostics in Figure 8.

D.4 Population quantities and bound evaluation

Let $\Phi_{\le k}$ contain the complete Gaussian-orthonormal Hermite features through degree k, including the constant. Under the image-mixture distribution, we compute

$$G_k = \mathbb{E}_q[\Phi_{\le k} \Phi_{\le k}^\top], \qquad c_{k,c} = \mathbb{E}[\Phi_{\le k}(Z) Y_c].$$

Four-node Gauss–Hermite quadrature per coordinate within each Gaussian component evaluates these moments exactly up to floating-point arithmetic. This integrates the Gram polynomials, which have degree at most six for the tested probe families.

The optimal population scores and their increments are

$$R^2_{\le k,c} = \frac{c_{k,c}^\top G_k^{-1} c_{k,c}}{\sigma_c^2}, \qquad \Delta_{k,c} = R^2_{\le k,c} - R^2_{\le k-1,c}.$$

The constant-only baseline has $R^2_{\leq 0,c} = 0$ because the targets are centered.

We separately compute the Gaussian-reference coefficients $\mathbb{E}_\gamma[\Phi_{\leq k} f_c]$. These determine the degree energy and approximation residual,

$$s_{k,c} = \frac{\|\mathrm{Proj}_{H_k} f_c\|^2_{L^2(\gamma)}}{\sigma_c^2}, \qquad \tau_{k,c} = \frac{\|f_c - P^\gamma_{\leq k} f_c\|_{L^2(q)}}{\sigma_c}.$$

The residual measures how well the truncated Gaussian expansion approximates the conditional mean. It is not the prediction error against noisy class labels.

For each setting with $\rho_k = \|G_k - I\|_{\mathrm{op}} < 1$, we compare the discrepancy $|\Delta_{k,c} - s_{k,c}|$ with the bound in Proposition 2,

$$B_{k,c} = \rho_k s_{k,c} + 2\tau_{k,c}\sqrt{(1+\rho_k)s_{k,c}} + \tau^2_{k,c}.$$

All quantities use the same 1,024-image mixture and its specified regression function. We do not substitute the full-test Gram from Figure 8 into these checks.

D.5 Numerical convergence and admissible settings

We cross-check the initial Gaussian-reference Gauss–Hermite quadrature with Gauss–Legendre integration on $[-10, 10]^2$, weighted by the standard Gaussian density. For mixture expectations, the integration rectangle extends nine component standard deviations beyond every component mean in each coordinate.

We compare reference/mixture quadrature orders of 240/300 and 400/450, increasing to 600/650 and 800/850 when needed. These convergence checks are performed regardless of whether $\rho_k < 1$. All 20 models meet the following criteria:

- Changes below 10^{-4} in the reported energies, residuals, discrepancies, bounds, and Bayes-recovery metrics.
- Mixture-mass error below 10^{-5}.
- Gram and cross-moment integration errors below 10^{-4} and 10^{-5}, respectively, compared with exact moments.

These checks assess numerical convergence; they are not rigorous interval certificates. We count a bound violation only when the discrepancy exceeds the bound by more than $\max(10^{-5}, 5\epsilon)$, where ϵ is the largest measured metric change between the final two integration resolutions.

We evaluate all ten digit targets at degrees one, two, and three. Settings with $\rho_k \geq 1$ are retained as diagnostics but are not included among the admissible bound checks. Table 4 reports every admissible, converged setting.

All 20 models satisfy the Gram condition at degree one. At $\beta = 0.5$, only seed 2 satisfies it at degree two, with $\rho_2 = 0.9025$ on the 1,024-image mixture. The same model has $\rho_2 = 1.0304$ on the separate full-test mixture, illustrating why the distributions must not be interchanged. At $\beta = 1$, all five seeds are admissible at degree two, but seed 1 is excluded at degree three because $\rho_3 = 1.1343$. All five seeds at $\beta = 2, 4$ are admissible at all three degrees.

The experiment artifacts record dataset and source hashes, split indices, checkpoint histories, encoder parameters, per-seed and per-digit diagnostics, integration resolutions, and which settings meet the Gram condition. Unit tests check Gaussian orthogonality, mixture posteriors, exact moments, a collapsed encoder, population projection ordering, and image-level shuffling.

A stronger KL penalty reduces both Gram deviation and retained digit information. A small absolute bound therefore does not by itself imply accurate relative degree fractions. The bound checks concern optimal population projections on the specified mixture, whereas Figure 3a reports finite-sample fitted probes.

Table 4: Full-bound checks on the 1,024-image posterior mixture. Means are over all ten digit targets and the eligible seeds. The checks are correlated. All 500 checks satisfy the full bound within numerical tolerance.

β	Degree	Eligible seeds	Mean discrepancy	Mean bound
0.5	1	5	0.046674	0.777634
0.5	2	1	0.022782	0.547319
1	1	5	0.031494	0.618634
1	2	5	0.020367	0.430082
1	3	4	0.016076	0.299932
2	1	5	0.022770	0.427840
2	2	5	0.009105	0.256410
2	3	5	0.009370	0.147091
4	1	5	0.005475	0.116554
4	2	5	0.001746	0.041261
4	3	5	0.000651	0.013074

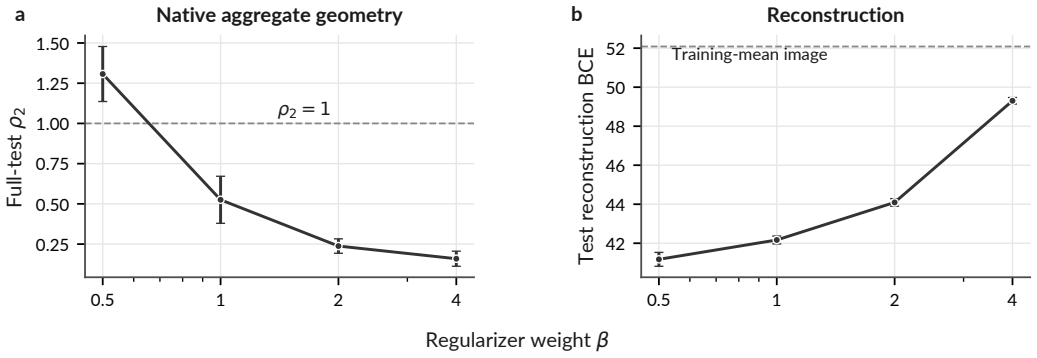


Figure 8: Degree-two Gram deviation for the mixture over all 10,000 test-image posteriors (left) and test reconstruction binary cross entropy (BCE) from sampled encodings (right). Points show means and bars show one sample standard deviation across five training seeds. The left dashed line marks $\rho_2 = 1$, not a test of exact Gaussianity. The right dashed line is the constant training-mean-image baseline (BCE 52.09). This full-test distribution is distinct from the 1,024-image mixture used for the bound checks.

E OTHELLO EXPERIMENTAL DETAILS

E.1 DATA AND EVALUATION

We retain the first 59 moves from each of 2,000 games without adding a beginning-of-sequence token. We split the games into four disjoint sets, using 350 for coordinate selection, 700 for probe fitting, 350 for hyperparameter selection, and 600 for testing. Means and scales are computed from the fitting set only. Targets and per-square scoring follow Section 6.

Error bars show standard deviations over three refitted splits of the same checkpoint. Test sets can overlap between splits. Test games are excluded from probe fitting but may have appeared during OthelloGPT pretraining. Exact dataset, checkpoint, activation-hook, and run identifiers are supplied with the code.

E.2 TARGET-AWARE DIMENSIONALITY REDUCTION

A full quadratic feature expansion grows as $O(d^2)$. To reduce the cost of explicit polynomial fitting, we also evaluate probes in target-specific subspaces. On the coordinate-selection set A, we form the

target-weighted second moment from centered activations z_i and centered targets y_i,

$$\widehat{M}_A = \frac{1}{|A|} \sum_{i \in A} y_i z_i z_i^\top.$$

We retain its 16 leading eigenvectors, ranked by absolute eigenvalue. We then append the linear covariance direction

$$\widehat{b}_A = \frac{1}{|A|} \sum_{i \in A} y_i z_i$$

and orthonormalize the combined vectors. This gives a basis $V \in \mathbb{R}^{512 \times 17}$ and reduced coordinates $r = V^\top z$.

We select a separate basis for each board square and target definition. These bases do not form a single shared 17-dimensional board representation. We freeze each basis before probe fitting. Coordinate selection, probe fitting, hyperparameter selection, and testing use the separate game sets described above. Test labels are not used to select the basis or fit the probes.

E.3 Probe fitting and hyperparameters

Full-space probes use all 512 activation coordinates. Reduced-space probes use the 17 coordinates selected for their target. Hyperparameters and stopping epochs are chosen by mean validation R^2. Probes are then evaluated once on the test set without refitting on validation data. The quadratic search also includes an affine-only candidate. Table 5 lists the fitting settings.

Table 5: Hyperparameters and optimization settings for the OthelloGPT probes.

Probe	**Setting**
Affine	Ridge regression with an unpenalized intercept. Penalties $\{0, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$.
Quadratic	All degree-at-most-two terms with an unpenalized affine part. Penalties $\{0.001, 0.01, 0.1, 1, 10, 100\}$. Full-space fits use the equivalent quadratic kernel. Reduced-space fits use explicit monomials with the same normalization. The affine-only candidate uses ordinary least squares rather than the full affine ridge search.
MLP	Two hidden layers of 512 GELU units, trained with Adam and batch size 1,024. Learning rates $\{10^{-3}, 3 \times 10^{-4}\}$. Training lasts at most 300 epochs, with validation every five epochs. Early stopping follows 20 epochs without improvement and cannot occur before epoch 30.

The full-space MLP predicts all 64 squares jointly. The reduced-space MLP uses a separate scalar-output network for each square because the selected coordinates differ between targets. The comparison therefore reflects both projection and a change in fitting architecture.

Nine of 48 selected full-space MLP fits reach their best validation score at the epoch limit. The MLP scores provide a nonlinear baseline rather than an estimate of optimal nonlinear recovery. Pilot diagnostics, including preliminary test scores, informed the search ranges. These comparisons are therefore exploratory rather than independent confirmation.

E.4 Off-Gaussian diagnostics on learned activations

We use partition seed 0, layers 3 and 7, both target definitions, and squares 0, 27, 36, and 63. For each combination, we select either 4 or 16 moment directions and append the linearly independent covariance direction, giving dimensions 5 and 17. The resulting 32 settings use the same game partitions and basis-selection procedure as the decoding study. Each test calculation pools 35,400 positions from 600 games. Basis selection uses target labels, but the Gram calculation uses only the frozen coordinates.

We fit the mean and full covariance on the probe-fitting games and whiten the projected coordinates. For the resulting coordinates z, we use

$$\Phi_{\le 2}(z) = \Big(1,\, (z_i)_i,\, ((z_i^2-1)/\sqrt{2})_i,\, (z_i z_j)_{i<j}\Big)^{\mathsf{T}}$$

and compute the uncentered empirical feature Gram and its deviation,

$$\widehat{G}_2 = \frac{1}{N}\sum_{i=1}^{N} \Phi_{\le 2}(z_i)\Phi_{\le 2}(z_i)^{\mathsf{T}}, \qquad \hat{\rho}_2 = \|\widehat{G}_2 - I\|_{\mathrm{op}}.$$

The intercept is included. This is not a centered correlation matrix. Whitening changes the Gaussian reference and is used only for this diagnostic, not to reinterpret the regularized fits in Figure 4.

Across pooled positions, $\hat{\rho}_2$ ranges from 12.4 to 79.8. Independent Gaussian samples of the same size give 0.061–0.192. This comparison is a numerical check, not a null test for correlated game data.

To investigate the large deviations, we freeze the leading eigenvector of the fitting-set Gram and evaluate squared feature projections on test positions. At layer 7, the first post-move position contributes 98.4–99.4% of this directional second moment across all 16 settings. No single game contributes more than 0.45%. These shares concern one fixed direction and do not decompose $\hat{\rho}_2$. Layer 3 shows a more varied pattern across moves.

We also examine three fixed stages, covering moves 1–20, 21–40, and 41 onward. We fit a separate whitening map on the fitting games within each stage while keeping the selected subspaces unchanged. All 96 combinations are retained, with no rank failures. Table 6 reports the results.

Table 6: Degree-two Gram deviations after fitting-set stage-specific whitening. Each row includes 16 settings.

Layer	Moves	$\hat{\rho}_2$ range	Settings with $\hat{\rho}_2 < 1$
3	1–20	7.436–23.268	0/16
3	21–40	0.448–21.653	2/16
3	41 onward	0.445–13.562	2/16
7	1–20	9.752–26.441	0/16
7	21–40	0.493–3.599	2/16
7	41 onward	0.594–9.704	2/16

The eight settings below one remain below one in both fixed 300-game test halves. These empirical checks do not establish a population upper bound. The stage analysis is exploratory and uses the same games as the pilot.

E.5 Controlled-target bound check

To check the bound using a target defined over the full coordinate space, we set

$$q(z) = 0.6 z_1 + 0.4\,\frac{\|z\|^2 - d}{\sqrt{2d}}.$$

Its unnormalized Gaussian-reference degree-two energy is 0.16, and $\tau_2 = 0$. On each stage's empirical test distribution, we compute $s_2 = 0.16/\operatorname{Var}(q)$ and the exact affine least-squares projection. Since q is quadratic, the degree-two increment equals the affine residual variance divided by $\operatorname{Var}(q)$.

In the eight settings with $\hat{\rho}_2 < 1$, the absolute discrepancy ranges from 0.00044 to 0.04785, below the bounds $\hat{\rho}_2 s_2$ of 0.13586–0.22255. The other 88 settings also show no violations for this target, but do not satisfy the stated sufficient condition. Fitting the affine predictor on the separate fitting games increases its normalized test residual by 0.00099–0.00334 in the eight admissible settings. We do not substitute this fitted gap for the exact projection increment.

This calculation checks the bound on the empirical activation distribution, not an independent population distribution. For the actual board-state target, we have not specified the conditional mean

away from the activation support or measured its Gaussian-reference approximation residual. The controlled-target check therefore does not establish a Hermite-energy interpretation of the board-state probe scores.

F JEPA Pendulum Experimental Details

F.1 Data and representation learning

We simulate the nonlinear pendulum using velocity Verlet with step size 0.05. Initial states are rejection-sampled with energies in $[0.2, 3.0]$, covering both oscillation and full rotation. Encoder training uses 2,000 trajectories of length 128. We reserve 400 separate trajectories for probe development and generate another 400 with a fixed independent seed for the reported confirmation scores. These evaluations use the same energy range as training, rather than a held-out energy interval.

States are standardized using encoder-training statistics. The encoder and predictor each have three hidden layers of width 128, with LayerNorm and GELU activations. The encoder produces 32-dimensional representations. The predictor uses a time-step embedding and predicts a residual added to the current latent state. We train for 200 epochs using next-step latent prediction with either SIGReg or the moment control, across five model seeds. Training uses AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, and gradient clipping at norm 1. The learning rate increases linearly over the first 5% of updates, followed by cosine decay.

F.2 Regularization and moment control

SIGReg compares empirical characteristic functions along random projections with those of a standard Gaussian. Each evaluation uses 1,024 random projections and 17 quadrature points over $[0, 3]$. The moment control instead penalizes nonzero means and covariance deviations from the identity,

$$\frac{\|\mathbb{E}[z]\|_2^2}{d} + \frac{\|\mathrm{Cov}(z) - I\|_F^2}{d}.$$

The moment penalty is computed across the batch separately at each position in the two-state training window, then averaged. It constrains only the first two moments. We test regularizer weights 0.01, 0.1, and 1 for both objectives, but these weights are not calibrated across objectives.

A zero-regularization run collapses to variance near 10^{-9}. Probes fitted after rescaling show that information can remain numerically recoverable. We exclude this run from the main comparison because polynomial fitting becomes sensitive to numerical precision and conditioning.

F.3 Probe fitting and evaluation

The 400 development trajectories are divided into 160 for probe fitting, 40 for hyperparameter selection, and 200 for the earlier test evaluation. No trajectory appears in more than one split. The reported confirmation scores use the separate set of 400 trajectories described in Appendix F.1.

Probe inputs and polynomial feature columns are standardized using fitting trajectories only. Polynomial targets are centered; MLP targets are centered and scaled using fitting-set statistics. We fit affine and full quadratic probes by ordinary least squares, with an unpenalized intercept and a standard singular-value cutoff.

As numerical checks, we also select ridge penalties from $\{0, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$ using the validation split and repeat polynomial evaluation with encoder activations computed in float64. Neither check changes the observed gap between affine and quadratic recovery under SIGReg.

The MLP probe has two GELU hidden layers. We test widths $\{64, 128\}$ and learning rates $\{10^{-3}, 3 \times 10^{-4}\}$, using two initialization seeds per configuration. Training uses Adam for at most 2,000 full-batch updates. Validation is checked every 25 updates, with stopping after 250 updates without improvement and no earlier than update 300. The checkpoint with the lowest validation loss is retained. We select the configuration by its mean best validation loss across the two seeds and evaluate both selected fits without further selection on test performance. Figure 5 reports initialization seed 0 for each trained representation; both initialization results are retained in the saved

evaluation records. Of the 60 selected MLP fits across the two objectives, three regularizer weights, five model seeds, and two probe initializations, 45 attain their lowest recorded validation loss at the 2,000-update limit.

Scores pool positions within the evaluation trajectories. To estimate uncertainty from the evaluation data, we resample whole trajectories and use the same resampled trajectories for all probes. Separately, the error bars in Figure 5 show sample standard deviations across the five independently trained representations.

F.4 Full multivariate He-Gram calculation

For each of the five checkpoints at each regularizer weight, we encode all $400 \times 128 = 51{,}200$ confirmation states. The calculation uses the raw 32-dimensional encoder outputs, without additional latent centering, scaling, or whitening. For each output z, we construct the complete degree-at-most-two Gaussian-orthonormal Hermite feature vector

$$\Phi_{\le 2}(z) = \left(1,\ (z_i)_{i=1}^{32},\ \left(\frac{z_i^2 - 1}{\sqrt{2}}\right)_{i=1}^{32},\ (z_i z_j)_{1 \le i < j \le 32}\right)^{\mathsf{T}}.$$

This gives $1 + 32 + 32 + 496 = 561$ features. With $N = 51{,}200$, the empirical Gram matrix and its deviation are

$$\widehat{G}_2 = \frac{1}{N}\sum_{n=1}^{N} \Phi_{\le 2}(z_n)\Phi_{\le 2}(z_n)^{\mathsf{T}}, \qquad \hat{\rho}_2 = \|\widehat{G}_2 - I_{561}\|_{\mathrm{op}} = \max_j |\lambda_j(\widehat{G}_2) - 1|.$$

The eigenvalues are computed from the full symmetric matrix. The intercept is included, and the feature columns are not centered or standardized. This is an uncentered second-moment matrix, not a correlation matrix or a random-projection estimate.

Across the 15 SIGReg checkpoints, $\hat{\rho}_2$ ranges from 6.230 to 8.698. Across the 15 moment-control checkpoints, it ranges from 10.737 to 11.218. None of these empirical matrices satisfies $\hat{\rho}_2 < 1$. These finite-sample calculations do not certify the population Gram deviation or establish the sufficient condition in Proposition 2. The probe scores are therefore reported as predictive accessibility, not calibrated Hermite-degree energies.

F.5 Projected He-Gram diagnostic

On the confirmation trajectories, we evaluate 48 fixed random directions. For each direction, we standardize the projected representation using its confirmation-set mean and standard deviation, then compute the largest absolute off-diagonal correlation among its first three Hermite polynomials. We average these values over directions to obtain the projected He-Gram statistic. Bootstrap resampling uses whole trajectories rather than individual positions, with the directions and standardization held fixed.

Mean diagnostic values are 0.16–0.18 under SIGReg and 0.37–0.39 under the moment control. Lower values indicate weaker correlations between the tested Hermite degrees.

This statistic measures departures from Hermite orthogonality along the tested directions. It is not the Gram deviation of the complete multivariate quadratic feature set and does not establish the conditions of Proposition 2. It therefore cannot, by itself, justify interpreting probe increments as calibrated Hermite-degree energies.

G Grokking Experiment and Controls

G.1 Training, targets, and evaluation

For each seed, we randomly assign $3{,}763$ of the $97^2 = 9{,}409$ input pairs to model training and reserve the other $5{,}646$ for model test accuracy. The model has one attention layer with four heads, width 128, a width-512 GELU MLP, and a linear unembedding. It is trained for 30,000 full-batch AdamW steps with learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.98)$, and weight decay 1.0 or 0. Each condition uses seeds 0–2.

The model-training pairs are split again into 2,634 probe-fitting and 1,129 probe-test pairs. Probe-test pairs are therefore unseen during probe fitting but were used to train the transformer. Model generalization is measured on the separate model-test set.

We record the final-token residual after attention and immediately after adding the MLP output. At the final checkpoint, we average the logit vectors by the correct answer residue over all 9,409 input pairs. We center this profile across residues, take its discrete Fourier transform, and sum squared magnitudes across output classes. Among the four highest-power nonzero frequencies, we retain those whose power is at least 10% of the largest power. These frequencies are held fixed across checkpoints.

For each selected frequency k, the two scalar targets are

$$\cos\left(\frac{2\pi k(a+b)}{97}\right) \qquad \text{and} \qquad \sin\left(\frac{2\pi k(a+b)}{97}\right).$$

Reported scores average their componentwise held-out R^2. For the weight-decay runs, the final frequency sets are $\{2, 9, 40\}$, $\{18, 21, 46, 47\}$, and $\{3, 42\}$ for seeds 0–2. Without weight decay, they are $\{20, 40, 41, 42\}$, $\{14, 22, 33, 48\}$, and $\{23, 29, 42, 43\}$, respectively.

Frequency selection uses both model-training and model-test pairs, although probe coefficients are fitted only on the probe-fitting set. The analysis therefore concerns model-selected targets rather than independently specified targets.

G.2 Probe fitting

All probes use the complete 128-dimensional activation. Inputs are standardized using the corresponding fitting data only. During penalty selection, standardization uses only the internal fitting subset.

The affine probe is ridge regression with penalty 10^{-3} and an unpenalized intercept, fitted by a float64 singular value decomposition. The degree-at-most-two probe is exact polynomial-kernel ridge regression with

$$K(z, z') = \left(1 + \frac{z^\top z'}{128}\right)^2,$$

which spans the full affine and quadratic feature family.

At the final checkpoint, we randomly split the 2,634 probe-fitting pairs into 2,107 fitting and 527 validation pairs, using the model seed plus 101. For each activation location, we select the quadratic penalty from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ by mean validation R^2 across target components. We then fit on all 2,634 probe-fitting pairs, keeping the selected penalty fixed across checkpoints. The 1,129 probe-test pairs are not used for penalty selection. Quadratic fitting and validation use float64 arithmetic.

The MLP probe has one width-256 ReLU hidden layer and jointly predicts all selected sine and cosine targets. It is initialized afresh for each checkpoint and activation location, then trained for 120 full-batch Adam updates with learning rate 10^{-3}, mean squared error, and no weight decay or early stopping. We evaluate the final iterate. MLP fitting uses float32 arithmetic.

A shuffled-target quadratic probe is fitted using the same procedure and selected penalty. Its final R^2 is negative in every run. We retain negative R^2 values rather than clipping them.

G.3 Recovery before and after the transformer MLP

The main figure reports recovery from the residual before the transformer MLP. Figure 9 also shows recovery after adding its output.

In the weight-decay runs, final post-MLP affine R^2 reaches 0.959–0.988, close to quadratic recovery of 0.974–0.998. Without weight decay, post-MLP affine scores are 0.067–0.100. The gap between affine and quadratic recovery is therefore smaller after the MLP in the weight-decay runs. This residual directly feeds the model's linear output head.

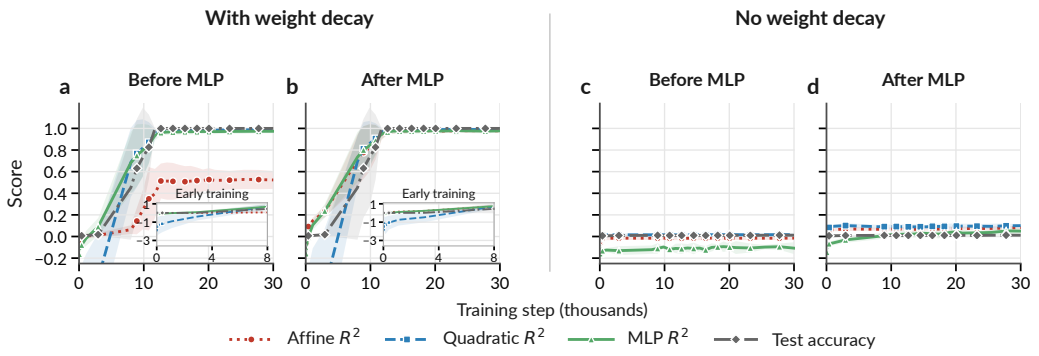


Figure 9: Probe scores are evaluated on pairs excluded from probe fitting. The before-MLP panels repeat the main comparison for reference. After the MLP, affine recovery nearly matches quadratic recovery in the weight-decay runs. Curves and bands show means and one standard deviation over three seeds.

G.4 Frequency-selection and fitting controls

Selecting frequencies from the final model could favor the reported trajectory. We therefore repeat the analysis using frequencies selected at the first checkpoint whose model-test accuracy is at least 0.30. At this checkpoint, quadratic penalties are also selected separately for each activation location using the same grid and internal split sizes as above, but with split seed equal to the model seed plus 701. The selected frequencies and penalties are then held fixed. Their selection uses no final-checkpoint information, although the checkpoint choice depends on model-test accuracy.

Of the early-selected frequencies, 2/3, 3/4, and 2/3 also occur in the final frequency sets across seeds. For these early-selected targets, the final pre-MLP affine/quadratic scores are $0.372/0.633$, $0.425/0.966$, and $0.484/0.995$. Post-MLP affine scores are 0.645, 0.853, and 0.986. Quadratic recovery therefore remains higher than affine recovery before the MLP in all three seeds. Figure 10 shows the complete trajectories.

For the early-selected targets, varying the quadratic ridge penalty from 10^{-3} through 1 gives final pre-MLP R^2 ranges of 0.626–0.645, 0.907–0.966, and 0.992–0.995 across seeds.

For the primary final-frequency analysis, every sine or cosine component has pre-MLP quadratic R^2 of at least 0.911 in the weight-decay runs. The corresponding affine scores range from -0.024 to 0.753. Post-MLP affine scores range from 0.862 to 0.995. Thus, the mean quadratic advantage before the MLP is not driven by only a subset of the selected targets.

G.5 Distributional diagnostics and interpretation

The projected He-Gram diagnostic uses 32 random unit directions and the probe-fitting activations. Each projected activation is standardized before computing the largest absolute off-diagonal correlation among its first three Hermite polynomials. The statistic is the mean of these values over directions.

In the weight-decay runs, the projected He-Gram diagnostic at the final checkpoint ranges from 0.203 to 0.276 before the MLP and from 0.296 to 0.469 after it. These projected diagnostics do not establish the full Gram condition or control the target-dependent approximation residual required by Proposition 2.

We therefore interpret the results as comparisons of held-out predictive performance, not as calibrated Hermite-degree fractions. A successful quadratic probe demonstrates recovery within degree two; a low fitted score does not establish that every readout in that family must fail.

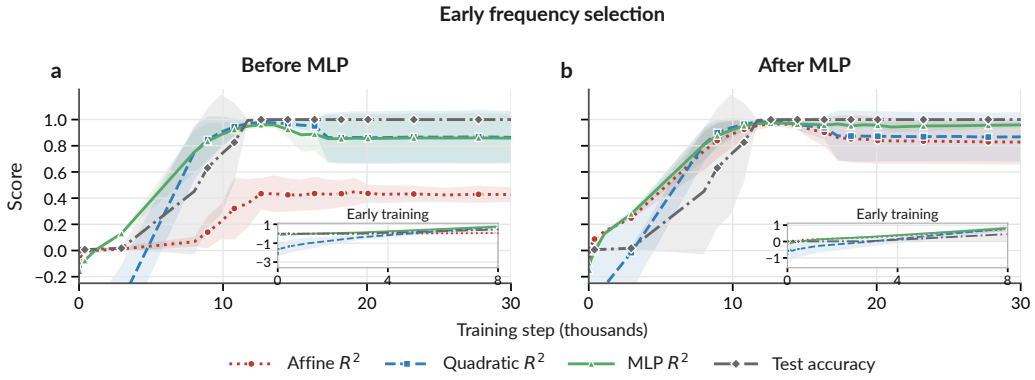


Figure 10: Frequencies are chosen at the first checkpoint with model-test accuracy at least 0.30 and then held fixed. Probe scores are evaluated on pairs excluded from probe fitting. Curves and bands show the mean and one standard deviation over three seeds.