# EQUIVARIANT DIFFUSION CAN HAVE A RELATIVELY LARGER GENERALIZATION REGIME THAN NON-EQUIVARIANT ONE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Equivariant diffusion models use known symmetries to guide denoising with promising performance in practice, but how symmetry theoretically shapes the course of learning remains unclear. Understanding this effect is important for explaining when training favors useful structure over less relevant variation. This paper identifies selective delay as a mechanism by which symmetry separates these stages of learning. Averaging over symmetry-related inputs can slow the learning of less relevant variation more strongly than the learning of the target's main structure. This increases the relative separation between the two stages, even when both become slower. We establish this mechanism in a tractable denoising model, characterize when it produces a wider generalization regime under discrete and continuous symmetries, and verify its predictions numerically. These results connect symmetry to learning time and provide a principle for choosing symmetries that favor task-relevant structure early in training.

## 1 INTRODUCTION

Diffusion models generate data by learning to reverse a noise-corruption process (Ho et al., 2020; Song et al., 2021). Their ability to generalize depends on the structure that architecture and training encourage them to learn (Kadkhodaie et al., 2024; Vastola, 2025). Symmetry provides a natural way to encode such structure. In molecular generation, for example, equivariant denoisers respect how a molecule transforms under rotations and reflections (Hoogeboom et al., 2022). Understanding the role of symmetry is therefore central to explaining how these models learn from structured data.

Learning evolves over training time. Recent work shows that diffusion training can learn the underlying distribution well before memorizing individual samples (Bonnaire et al., 2025), and optimization choices can suppress memorization (Wu et al., 2025). This motivates a temporal view of generalization: useful structure is learned during a regime that precedes overfitting samples. However, equivariant diffusion has not studied in this aspect. Can symmetry enlarge this *generalization regime*? To address the learning mechanism behind this question, we study how symmetry changes the relative timing of learning the target's main structure and less relevant variation.

Our central insight is that symmetry can create this separation through *selective delay*. Averaging over symmetry-related inputs removes variation along those transformations. When that variation contributes little to the denoising task, averaging can slow its learning more strongly than the learning of the target's main structure. The two learning stages then become more separated in relative time, even if both become slower. This links the choice of symmetry to the order of learning: a suitable symmetry emphasizes the task's dominant structure by delaying the learning of less relevant variation. The resulting criterion gives a way to assess symmetry choices through their effect on training dynamics.

**Contributions.** We make this mechanism precise in a frozen-feature denoising model, comparing training with and without symmetry averaging on the same task. Our main contributions are threefold:

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.

1. We give an exact construction for the finite-cyclic group in which symmetry separates learning stages that coincide in the original model. The result shows how the choice of symmetry determines the relative delay and the duration of the resulting generalization regime.
2. We establish conditions under which this separation increases for general finite symmetry groups, and extend the result to continuous symmetries on a torus. This turns the mechanism from an explicit example into a criterion applicable to a broader class of symmetry choices.
3. We validate the predicted delays numerically and compare examples satisfying the criterion with matched controls. The experiments connect the theoretical conditions to the observed changes in learning times.

**Organization.** Section 2 reviews related work, and Section 3 introduces the background. Section 4 defines the matched models and target. Section 5 develops the dynamics and regime-widening results. Sections 6 and 7 present numerical validation and conclusions. The appendices provide proofs, notation tables, and experimental details.

## 2 Related Work

This section covers diffusion learning dynamics, symmetry in diffusion models, invariant features, and spectral accounts of learning order and speed.

**Diffusion learning dynamics.** Score matching and denoising score matching (DSM) provide the regression objectives underlying diffusion models (Hyvärinen, 2005; Vincent, 2011; Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021). Subsequent analyses study optimization, feature learning, and generalization of neural score estimators (Han et al., 2024; 2025), while random-feature and kernel theories characterize frozen representations through high-dimensional and Gaussian-equivalence methods (Rahimi & Recht, 2007; Goldt et al., 2022; Mei et al., 2021). Sampling guarantees connect score error to generated distributions (Chen et al., 2023; Li et al., 2024); geometry-adaptive representations and diffusion trajectories connect training to generalization and memorization (Kadkhodaie et al., 2024; Vastola, 2025; Biroli et al., 2024; Bonnaire et al., 2025; Wu et al., 2025). This literature establishes the diffusion objective and its statistical consequences, but does not isolate how orbit averaging changes relative fitting times for different target components.

**Symmetry in diffusion models and invariant features.** Equivariant diffusion models impose transformation laws on scores, feature fields, or transition rules, with applications to molecular and geometric generation (Hoogeboom et al., 2022; Satorras et al., 2021; Finzi et al., 2021). Representation-theoretic frameworks cover compact groups and general group actions (Kondor & Trivedi, 2018; Cohen et al., 2019; Finzi et al., 2021); recent diffusion-specific work studies equivariant score guarantees, target or stochastic symmetrization, quotient spaces, and canonicalization (Chen et al., 2024; Lu et al., 2025; Tong et al., 2025; Zhang et al., 2025; Xu et al., 2026; Zhou et al., 2026). Invariant features and data augmentation analyze related effects on risk, regularization, sample complexity, and kernels (Mroueh et al., 2015; Dao et al., 2019; Chen et al., 2020; Mei et al., 2021; Tahmasebi & Jegelka, 2023), including all-layer optimization (Nordenfors et al., 2025). The orbit average here modifies a shared frozen feature map while matching the target and trainable readout across models, so it differs from full score equivariance and from augmentation analyses centered on risk or sample efficiency.

**Spectral bias and learning order.** Spectral-bias analyses relate kernel or feature-covariance eigenvalues to learning order and time scale (Bordelon et al., 2020; Wang & Pehlevan, 2025; George et al., 2026); in kernel and wide-network regimes, target projections determine mode-wise learning curves (Canatar et al., 2021). This motivates the *target contribution*: the squared target–feature coupling carried by a Gram-invariant block. The matched Base-RFM and Avg-RFM analysis uses this quantity to identify subspaces carrying most or little of the squared initial update norm, and shows how averaging can increase the ratio of their threshold-crossing times by delaying the latter more strongly.

# 3 PRELIMINARIES

**Notations**. Boldface letters denote vectors and matrices. We use $\|\cdot\|$ and $\|\cdot\|_{\mathrm{F}}$ to denote the Euclidean norm of vectors and the Frobenius norm of matrices, respectively. For symmetric matrices, $\mathbf{A} \succeq 0$ means that $\mathbf{A}$ is positive semidefinite. We summarize key notations in Appendix A.

## 3.1 DIFFUSION DENOISING

Let $\mathbf{X} \sim \mu_0$ be a clean sample in $\mathbb{R}^d$. We use the Ornstein–Uhlenbeck forward diffusion $\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t$, with $\mathbf{X}_0 = \mathbf{X}$, where $\mathbf{B}_t$ is standard Brownian motion independent of $\mathbf{X}$. This is a constant-rate instance of variance-preserving diffusion (Song et al., 2021). At each fixed time $t > 0$, its transition law can be sampled as

$$\mathbf{X}_t = e^{-t}\mathbf{X} + \sqrt{\Delta_t}\boldsymbol{\xi}, \tag{1}$$

where $\Delta_t = 1 - e^{-2t}$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is independent of $\mathbf{X}$, and $\mathbf{I}_d$ is the identity matrix. Let $p_t$ denote the marginal density of $\mathbf{X}_t$. Denoising score matching (DSM) (Vincent, 2011; Song et al., 2021) estimates the score $\nabla \log p_t$ by fitting a predictor $\widehat{\mathbf{s}} : \mathbb{R}^d \to \mathbb{R}^d$ with the loss

$$\mathcal{L}_t^{\mathrm{DSM}}(\widehat{\mathbf{s}}) = \mathbb{E}_{\mathbf{X},\boldsymbol{\xi}}\left[\left\|\widehat{\mathbf{s}}(\mathbf{X}_t) + \boldsymbol{\xi}/\sqrt{\Delta_t}\right\|^2\right]. \tag{2}$$

Conditioned on $\mathbf{X}$, the noisy sample has distribution $\mathbf{X}_t \mid \mathbf{X} \sim \mathcal{N}(e^{-t}\mathbf{X}, \Delta_t\mathbf{I}_d)$; write its conditional density as $p_t(\mathbf{z} \mid \mathbf{X})$. Hence, up to a term independent of $\mathbf{z}$, $\log p_t(\mathbf{z} \mid \mathbf{X}) = -\left\|\mathbf{z} - e^{-t}\mathbf{X}\right\|^2/(2\Delta_t)$, and differentiating with respect to $\mathbf{z}$ gives $\nabla_{\mathbf{z}} \log p_t(\mathbf{z} \mid \mathbf{X}) = -(\mathbf{z} - e^{-t}\mathbf{X})/\Delta_t$. At $\mathbf{z} = \mathbf{X}_t$, this equals the target $\mathbf{Y}_t := -\boldsymbol{\xi}/\sqrt{\Delta_t}$; the negative sign points back toward the conditional Gaussian mean. Thus the loss regresses $\mathbf{Y}_t$ on $\mathbf{X}_t$. Its population minimizer over all square-integrable predictors is the conditional mean $\mathbb{E}[\mathbf{Y}_t \mid \mathbf{X}_t = \mathbf{z}] = \nabla \log p_t(\mathbf{z})$ (Vincent, 2011; George et al., 2026).

## 3.2 GROUP ACTIONS AND ORBIT AVERAGING

Group averaging gives the same features to inputs related by symmetry. This section defines that average and the part of the denoising vector that is unchanged by the group.

**Input transformations.** Orthogonal transformations preserve Euclidean distances and isotropic Gaussian noise, making them compatible with the corruption model in Eq. (1). Let $G$ be a finite subgroup of the orthogonal matrix group $O(d) = \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_d\}$. Finiteness permits exact averaging by summing over all group elements. The inclusion $G \hookrightarrow O(d)$ is its defining orthogonal representation: each $g \in G$ is already a $d \times d$ matrix. The same matrix acts on inputs by $\mathbf{z} \mapsto g\mathbf{z}$ and on denoising vectors by $\mathbf{Y}_t \mapsto g\mathbf{Y}_t$. The *orbit* $\{g\mathbf{z} : g \in G\}$ collects all inputs related to $\mathbf{z}$ by these transformations.

**Feature averaging.** To give symmetry-related inputs identical features, each feature is averaged uniformly over the group (Kondor & Trivedi, 2018). Let $f(\mathbf{z}) \in \mathbb{R}^p$ denote the features of an input $\mathbf{z} \in \mathbb{R}^d$. Their average is

$$(\mathrm{Avg}_G f)(\mathbf{z}) = \frac{1}{|G|}\sum_{g\in G} f(g\mathbf{z}). \tag{3}$$

For any $h \in G$, replacing $\mathbf{z}$ by $h\mathbf{z}$ only permutes the summands. Hence $(\mathrm{Avg}_G f)(h\mathbf{z}) = (\mathrm{Avg}_G f)(\mathbf{z})$: the averaged features are $G$-invariant, or constant on each input orbit.

**Outputs unchanged by symmetry.** The quantity predicted in Section 3.1 is the denoising vector $\mathbf{Y}_t = -\boldsymbol{\xi}/\sqrt{\Delta_t}$. To identify which part can be learned from invariant features, consider the output vectors unchanged by every group element. They form the *fixed-point subspace* $\mathrm{Fix}(G) = \{\mathbf{v} \in \mathbb{R}^d : g\mathbf{v} = \mathbf{v} \text{ for all } g \in G\}$. Averaging the transformed output vectors defines the *Reynolds operator* (Derksen & Kemper, 2015),

$$\mathbf{P}_G = \frac{1}{|G|}\sum_{g\in G} g \in \mathbb{R}^{d\times d}. \tag{4}$$

Applied to $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{P}_G$ retains its component in $\mathrm{Fix}(G)$, which is unchanged by every group element. Appendix B.4 ("Group averaging and the fixed subspace") establishes this property. Section 4.2 uses it to choose the regression target.

The extension to compact groups replaces finite averages by Haar integrals (Appendix E).

# 4 Problem Setup

We consider two random-feature models (RFMs) at a fixed diffusion time to isolate the effect of orbit averaging on parameter learning. Base-RFM uses the original features, whereas Avg-RFM averages them over input orbits; both use the same regression target (Figure 1).

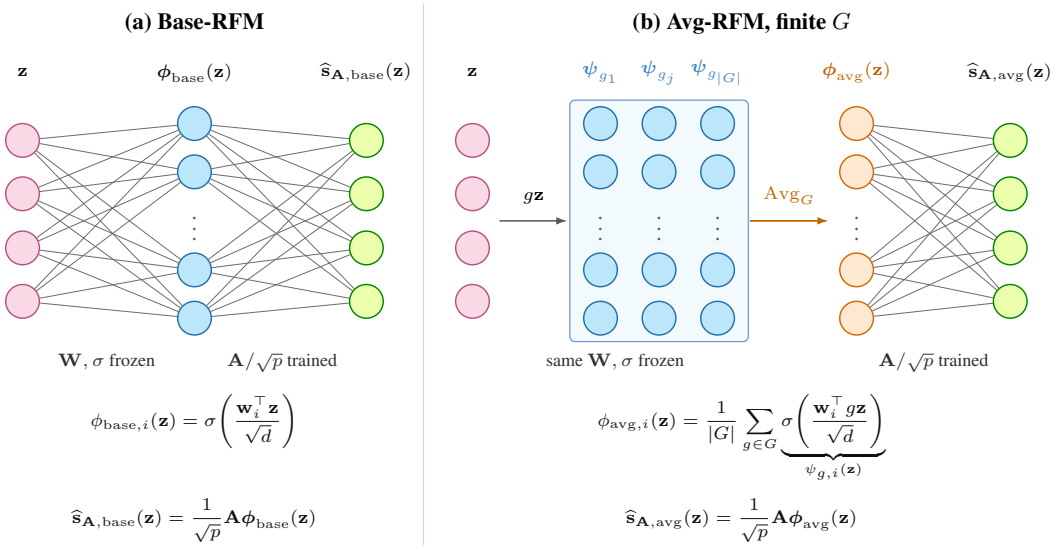


Figure 1: **Base-RFM and Avg-RFM.** Both models use the same frozen weights and activation; only the output matrix $\mathbf{A}$ is trained. In (b), the blue columns represent $\boldsymbol{\psi}_g(\mathbf{z}) = \boldsymbol{\phi}_{\mathrm{base}}(g\mathbf{z})$, whose average forms $\boldsymbol{\phi}_{\mathrm{avg}}(\mathbf{z})$. Vertical dots omit feature coordinates. Section 4.1 gives the formal definitions.

## 4.1 Two matched models

Let $G \subseteq O(d)$ be a finite group. Base-RFM and Avg-RFM use the same (frozen) weight matrix $\mathbf{W} \in \mathbb{R}^{p\times d}$, with rows $\mathbf{w}_i^\top$, and the same coordinate-wise activation $\sigma : \mathbb{R} \to \mathbb{R}$. For a noisy input $\mathbf{z} \in \mathbb{R}^d$, their feature maps are

$$\boldsymbol{\phi}_{\mathrm{base}}(\mathbf{z}) = \sigma(\mathbf{W}\mathbf{z}/\sqrt{d}), \qquad \boldsymbol{\phi}_{\mathrm{avg}}(\mathbf{z}) = (\mathrm{Avg}_G\, \boldsymbol{\phi}_{\mathrm{base}})(\mathbf{z}) = \frac{1}{|G|}\sum_{g\in G} \boldsymbol{\phi}_{\mathrm{base}}(g\mathbf{z}). \tag{5}$$

Avg-RFM thus applies group pooling to the Base-RFM features (Anselmi et al., 2013; Mroueh et al., 2015; Mei et al., 2021). In Figure 1, $\boldsymbol{\psi}_g(\mathbf{z}) = \boldsymbol{\phi}_{\mathrm{base}}(g\mathbf{z})$ denotes the features evaluated at the transformed input $g\mathbf{z}$.

For either model $M \in \{\mathrm{base}, \mathrm{avg}\}$, a trainable linear output layer gives the prediction

$$\widehat{\mathbf{s}}_{\mathbf{A},M}(\mathbf{z}) = p^{-1/2}\mathbf{A}\boldsymbol{\phi}_M(\mathbf{z}), \tag{6}$$

where $\mathbf{A} \in \mathbb{R}^{d\times p}$ is trained by gradient descent (or flow) from $\mathbf{A} = \mathbf{0}$. To isolate the effect of feature averaging on their learning dynamics, we next specify a common regression target and loss.

## 4.2 Target and loss

At a fixed diffusion time $t > 0$, let $\mathbf{Y}_t := -\boldsymbol{\xi}/\sqrt{\Delta_t}$ be the DSM target for the noisy input $\mathbf{X}_t$. The operator $\mathbf{P}_G$ from Section 3.2 yields the projected target $\mathbf{Y}_{t,G} := \mathbf{P}_G\mathbf{Y}_t$, the component of $\mathbf{Y}_t$ in $\mathrm{Fix}(G)$.

In this paper, $\mathbf{Y}_{t,G}$ is chosen because Avg-RFM gives the same prediction to all inputs on an orbit. Applying $g \in G$ to the clean input and its noise can change the DSM target from $\mathbf{Y}_t$ to $g\mathbf{Y}_t$, while Avg-RFM's prediction stays the same. Under the assumption that the clean-data distribution is $G$-invariant, averaging these targets yields $\mathbf{P}_G\mathbf{Y}_t$, removing variation that the invariant features cannot predict (Proposition B.13 in Appendix B.4). Base-RFM uses the same target to isolate the effect of feature averaging on a common regression task. The assumption $\dim \mathrm{Fix}(G) > 0$ excludes an identically zero target.

By the DSM identity,

$$\mathbb{E}[\mathbf{Y}_{t,G} \mid \mathbf{X}_t = \mathbf{z}] = \mathbf{P}_G \nabla \log p_t(\mathbf{z}). \tag{7}$$

Thus the regression aims to estimate the score component in $\mathrm{Fix}(G)$, which is constant on input orbits under the symmetry assumption. See Appendix B.4 ("Conditional mean and input invariance") for the derivation and invariance proof.

For training, fix clean samples $\mathbf{x}^1, \ldots, \mathbf{x}^n$ and the frozen weights $\mathbf{W}$. The symmetry assumption concerns the data distribution, so the sample set need not be $G$-invariant. For each sample $\nu$, draw independent noise $\boldsymbol{\xi}^\nu \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and set $\mathbf{x}_t^\nu = e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\boldsymbol{\xi}^\nu$ and $\mathbf{y}_{t,G}^\nu = -\mathbf{P}_G\boldsymbol{\xi}^\nu/\sqrt{\Delta_t}$. Both models use the noise-averaged squared loss

$$\mathcal{L}_{t,G,M}^{(n)}(\mathbf{A}) = \frac{\Delta_t d}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}^\nu}\left[\left\|\widehat{\mathbf{s}}_{\mathbf{A},M}(\mathbf{x}_t^\nu) - \mathbf{y}_{t,G}^\nu\right\|^2\right]. \tag{8}$$

# 5 Main Results

This section presents the parameter dynamics and defines the generalization regime. It then gives a cyclic-group example of regime widening, followed by sufficient conditions for general finite groups. The final part extends these conditions to continuous averaging on the torus.

## 5.1 Dynamics and the generalization regime

This subsection derives the parameter dynamics and identifies the feature directions along which the target drives learning. It then asks when directions accounting for most of this effect are learned before those accounting for little, and defines the resulting generalization regime.

### 5.1.1 Parameter dynamics

At fixed diffusion time $t > 0$, gradient descent on Eq. (8) updates the output weights of either model $M \in \{\mathrm{base}, \mathrm{avg}\}$ as

$$\mathbf{A}_M^{k+1} = \mathbf{A}_M^k - h\nabla_{\mathbf{A}}\mathcal{L}_{t,G,M}^{(n)}(\mathbf{A}_M^k), \qquad \mathbf{A}_M^0 = \mathbf{0}. \tag{9}$$

Here $k$ indexes the updates and $h > 0$ is the step size. The quantity $\tau_k = kh$ is the discrete update time at step $k$, whereas $\tau$ is the continuous time variable used by the gradient flow. As $h \to 0$, the discrete times $\tau_k$ become dense in this continuous time. Substituting Eq. (6) into the loss (8) gives the gradient flow

$$\dot{\mathbf{A}}_M(\tau) = -\nabla_{\mathbf{A}}\mathcal{L}_{t,G,M}^{(n)}(\mathbf{A}_M(\tau)) = -c_t\big(\mathbf{A}_M(\tau)\mathbf{U}_M - \sqrt{p}\,\mathbf{R}_{t,M}\big), \tag{10}$$

where $c_t = 2\Delta_t d/p$. Writing $\boldsymbol{\phi}_M^\nu = \boldsymbol{\phi}_M(\mathbf{x}_t^\nu)$ and $\mathbf{y}^\nu = \mathbf{y}_{t,G}^\nu$ for simplicity, and the two matrices are

$$\mathbf{U}_M = \mathbf{U}_M^{(n)} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}^\nu}[\boldsymbol{\phi}_M^\nu(\boldsymbol{\phi}_M^\nu)^\top], \qquad \mathbf{R}_{t,M} = \mathbf{R}_{t,M}^{(n)} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}^\nu}[\mathbf{y}^\nu(\boldsymbol{\phi}_M^\nu)^\top]. \tag{11}$$

These expectations are over diffusion noise, with the clean samples and weights fixed. The derivation and exact solution appear in Proposition D.5 (Appendix D.2).

Here $\mathbf{U}_M \succeq 0$ is the feature Gram matrix. The $i$th column of $\mathbf{R}_{t,M}$ averages the target vector weighted by feature coordinate $i$. Their roles in learning are described next.

### 5.1.2 LEARNING DIRECTIONS

Intuitively, the gradient flow can be viewed as a collection of feature directions being controlled by the target. At initialization, $\mathbf{R}_{t,M}$ tells us where this push is strong or weak, while $\mathbf{U}_M$ controls how quickly the corresponding parameter errors relax. We therefore first identify the directions receiving most of the initial push, and then compare their relative-error crossing times in the generalization regime below.

At zero initialization, $\mathbf{A}_M\mathbf{U}_M = \mathbf{0}$, so $\dot{\mathbf{A}}_M(0) = c_t\sqrt{p}\,\mathbf{R}_{t,M}$. The $j$th column of $\mathbf{R}_{t,M}$ is therefore the initial velocity of the output weights attached to feature $j$, up to the common factor $c_t\sqrt{p}$. A vector $\mathbf{v} \in \mathbb{R}^p$ combines feature directions, so its initial velocity is $c_t\sqrt{p}\,\mathbf{R}_{t,M}\mathbf{v}$. To record the squared velocity of every such combination with one matrix, define

$$\mathbf{T}_{t,M} = \mathbf{R}_{t,M}^\top\mathbf{R}_{t,M} \in \mathbb{R}^{p\times p}, \qquad \left\|\dot{\mathbf{A}}_M(0)\mathbf{v}\right\|^2 = c_t^2 p\,\mathbf{v}^\top\mathbf{T}_{t,M}\mathbf{v}. \tag{12}$$

$\mathbf{T}_{t,M}$ is the Gram matrix of the column velocities: its diagonal entries give individual feature strengths, while its off-diagonal entries say whether two features push in similar or opposite directions. The common factor $c_t^2 p$ is the same for every direction and does not affect these comparisons.

Learning can involve combinations of features rather than individual coordinates. Let $\mathbf{P} \in \mathbb{R}^{p\times p}$ be the orthogonal projector onto a subspace of feature directions. It selects a bundle of directions: $\mathbf{A}_M\mathbf{P}$ is the part of the output layer attached to that bundle, and $\mathbf{R}_{t,M}\mathbf{P}$ is the target push entering it. For $\mathbf{R}_{t,M} \neq 0$, the fraction of the total initial push in this bundle is

$$\ell_{\mathbf{T}_{t,M}}(\mathbf{P}) = \frac{\|\mathbf{R}_{t,M}\mathbf{P}\|_{\mathrm{F}}^2}{\|\mathbf{R}_{t,M}\|_{\mathrm{F}}^2} = \frac{\mathrm{Tr}(\mathbf{P}\mathbf{T}_{t,M})}{\mathrm{Tr}(\mathbf{T}_{t,M})}. \tag{13}$$

The common factor $c_t^2 p$ cancels, so this ratio asks how the total initial push is distributed rather than introducing a new learning rate. Because $\mathbf{P}$ is an orthogonal projector, the ratio lies in $[0, 1]$. For example, if $\mathbf{R}_{t,M} = (3, -1)$ and $\mathbf{P}_1 = \mathrm{diag}(1, 0)$, $\mathbf{P}_2 = \mathrm{diag}(0, 1)$, then $\ell_{\mathbf{T}_{t,M}}(\mathbf{P}_1) = 9/(9+1) = 0.9$ and $\ell_{\mathbf{T}_{t,M}}(\mathbf{P}_2) = 0.1$.

We use two orthogonal feature subspaces to make this comparison concrete. Let a subspace $S$ carry most of the initial push and another subspace $N$ carry little, with projectors $\mathbf{P}_{S,M}$ and $\mathbf{P}_{N,M}$, respectively. They are subspaces of the feature-coordinate space $\mathbb{R}^p$, whereas $G$ and $\mathbf{P}_G$ act on the input and output space $\mathbb{R}^d$. To compare their dynamics, choose them so that $\mathbf{P}_{S,M}\mathbf{P}_{N,M} = 0$, $\mathbf{P}_{B,M}\mathbf{U}_M = \mathbf{U}_M\mathbf{P}_{B,M}$, and $\mathbf{R}_{t,M}\mathbf{P}_{B,M} \neq 0$ for $B \in \{S, N\}$. The first condition keeps the two bundles separate, the second makes each bundle evolve by a closed equation, and the third ensures that both bundles actually move. (Appendix D.2 derives these conditions and the corresponding blockwise dynamics.)

To turn "most" and "little" into checkable conditions, we use tolerances $0 < \alpha_S, \alpha_N < 1/2$:

$$\ell_{\mathbf{T}_{t,M}}(\mathbf{P}_{S,M}) \geq 1 - \alpha_S, \qquad \ell_{\mathbf{T}_{t,M}}(\mathbf{P}_{N,M}) \leq \alpha_N. \tag{14}$$

Here $\alpha_S$ is the allowed fraction of the push outside $S$, while $\alpha_N$ is the allowed fraction inside $N$. For example, $\alpha_S = \alpha_N = 0.1$ means that at least $90\%$ of the initial push lies in $S$ and at most $10\%$ lies in $N$.

These fractions describe where learning starts, not how fast it proceeds. The Gram matrix $\mathbf{U}_M$ controls the error-decay rates, so the block with the larger initial push need not reach a common relative-accuracy threshold first. The next subsection uses that common ruler to define the two crossing times and the interval called the generalization regime.

### 5.1.3 GENERALIZATION REGIME

With the two directions selected, we use one relative-error ruler to compare their progress. The initial errors of the two subspaces can differ, so each error is normalized by its own starting size before the common ruler is applied. The generalization regime is the interval in which the $S$ block has crossed this ruler while the $N$ block has not. For model $M \in \{\mathrm{base}, \mathrm{avg}\}$ and subspace $B \in \{S, N\}$, let $\mathbf{A}_{\star,M} = \lim_{\tau\to\infty}\mathbf{A}_M(\tau)$ be the flow limit and define the squared relative error

$$E_{B,M}(\tau) = \frac{\|(\mathbf{A}_M(\tau) - \mathbf{A}_{\star,M})\mathbf{P}_{B,M}\|_{\mathrm{F}}^2}{\|(\mathbf{A}_M(0) - \mathbf{A}_{\star,M})\mathbf{P}_{B,M}\|_{\mathrm{F}}^2}. \tag{15}$$

Here $\mathbf{P}_{B,M}$ projects onto subspace $B$ for model $M$. The nonzero-update condition above ensures a positive denominator. For a common threshold $\rho \in (0, 1)$, define the two times associated with the starts of generalization and memorization at which the normalized error reaches $\rho^2$:

$$\begin{aligned} \tau_{\mathrm{gen},M}(\rho) &= \inf\{\tau \geq 0 : E_{S,M}(\tau) \leq \rho^2\}, \\ \tau_{\mathrm{mem},M}(\rho) &= \inf\{\tau \geq 0 : E_{N,M}(\tau) \leq \rho^2\}. \end{aligned} \tag{16}$$

Under the preceding conditions, these times are positive and finite (Appendix D.2).

Their time gap and ratio are

$$A_M(\rho) = \tau_{\mathrm{mem},M}(\rho) - \tau_{\mathrm{gen},M}(\rho), \qquad \bar{W}_M(\rho) = \frac{\tau_{\mathrm{mem},M}(\rho)}{\tau_{\mathrm{gen},M}(\rho)}. \tag{17}$$

When $A_M > 0$, the *generalization regime* $[\tau_{\mathrm{gen},M}, \tau_{\mathrm{mem},M})$ covers a time period $A_M$ during which $S$ has reached the threshold but $N$ has not. The ratio $\bar{W}_M$ measures their relative separation between $\tau_{\mathrm{mem},M}$ and $\tau_{\mathrm{gen},M}$.

For corresponding subspaces, compare averaging with the baseline through the log delays

$$\begin{aligned} \Delta_S(\rho) &= \log \frac{\tau_{\mathrm{gen,avg}}(\rho)}{\tau_{\mathrm{gen,base}}(\rho)}, \qquad \Delta_N(\rho) = \log \frac{\tau_{\mathrm{mem,avg}}(\rho)}{\tau_{\mathrm{mem,base}}(\rho)}, \\ \Delta_{\mathrm{win}}(\rho) &= \Delta_N(\rho) - \Delta_S(\rho) = \log \frac{\bar{W}_{\mathrm{avg}}(\rho)}{\bar{W}_{\mathrm{base}}(\rho)}. \end{aligned} \tag{18}$$

Positive $\Delta_S$ or $\Delta_N$ indicates slower learning in that subspace; $\Delta_{\mathrm{win}} > 0$ means a larger time ratio. The cyclic construction below achieves this by delaying $N$ by a larger factor than $S$.

### 5.2 Finite-cyclic Group Case

To obtain this larger delay for $N$, the construction below places more of its feature weights in directions changed by the group.

Let $G = C_m$ be a cyclic group of order $m \geq 2$, generated by repeatedly applying one transformation. Split the features into nonempty subspace $S$ and $N$. Each feature weight has norm $\sqrt{d}$ and combines a component fixed by $G$ with one changed by $G$. The fixed directions are orthogonal across features; the changed components lie in mutually orthogonal subspaces on which $G$ applies the same sign flip or rotation. Within block $B \in \{S, N\}$, let $\theta_B$ be the fraction of the squared weight norm in the changed component, outside $\mathrm{Fix}(G)$.

**Assumptions.** We use gradient flow on the expected loss for the shared projected target, averaging over clean data $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and independent diffusion noise. Assume $\sigma$ is bounded, with $\mathbb{E}[\sigma(Z)] = 0$ and $\mathbb{E}[Z\sigma(Z)] \neq 0$ for $Z \sim \mathcal{N}(0, 1)$, giving centered features and nonzero updates toward the target. Require $0 \leq \theta_S < \theta_N \leq \theta^* < 1$, where $\theta^*$ is the largest fraction keeping a weight's inner products with all its group transforms nonnegative.

To quantify averaging's effect, define the base feature variance and the fraction retained for a feature with squared-weight fraction $\theta$ outside $\mathrm{Fix}(G)$:

$$\lambda_{\mathrm{base}} = \mathbb{E}[\sigma(Z)^2], \qquad q_G(\theta) = \frac{\mathbb{E}[\phi_{\mathrm{avg},i}(\mathbf{X}_t)^2]}{\lambda_{\mathrm{base}}}. \tag{19}$$

Here $\phi_{\mathrm{avg},i}$ is the $i$th averaged feature in Eq. (5); smaller $q_G(\theta)$ means more feature variance is removed.

The following theorem shows how these retained fractions separate the learning times of $S$ and $N$, opening a positive learning window.

**Theorem 5.1** (Cyclic generalization regime, informal, proofs in Appendix C.5)**.** *Under these assumptions, with $S, N$ satisfying the target-contribution conditions in Eq.* (14)*, both base blocks reach any common threshold $\rho \in (0, 1)$ at the same time $\tau_0(\rho) = -\log \rho/(c_t \lambda_{\mathrm{base}})$, with the common flow*

*factor $c_t$ from Eq.* (10)*. Then, the retained fractions satisfy* $0 < q_G(\theta_N) < q_G(\theta_S) \le 1$*, giving*

$$\bar{W}_{\mathrm{base}}(\rho) = 1, \qquad \boxed{\frac{\bar{W}_{\mathrm{avg}}(\rho)}{\bar{W}_{\mathrm{base}}(\rho)} = \frac{q_G(\theta_S)}{q_G(\theta_N)} > 1}, \tag{20}$$

$$A_{\mathrm{base}}(\rho) = 0, \qquad \boxed{A_{\mathrm{avg}}(\rho) = \tau_0(\rho)\left[\frac{1}{q_G(\theta_N)} - \frac{1}{q_G(\theta_S)}\right] > 0}. \tag{21}$$

*Remark* 5.2 (Mechanism of regime widening). This theorem shows that, the equivariant model Avg-RFM widens the generalization regime compared to the non-equivariant one base-RFM. Specifically, orbit averaging widens the generalization regime from an empty interval to a positive-duration window: $S$ reaches the threshold while $N$ has not. In this construction, each block's learning rate is proportional to its feature variance. Since $N$ places more weight in directions changed by the group, averaging removes more of its variance and delays it by a larger factor. The time ratio consequently rises from 1 to $q_G(\theta_S)/q_G(\theta_N) > 1$. If $\theta_S = 0$, averaging leaves the $S$ features and their learning time unchanged.

### 5.3 Finite Group Case

For general finite groups, spectral bounds can establish the same selective delay even when learning rates vary within each block.

**Assumptions.** Fix the data and weights in Eq. (11). For each $M \in \{\mathrm{avg}, \mathrm{base}\}$, let $\mathbf{P}_{S,M}, \mathbf{P}_{N,M}$ project onto disjoint spectral blocks of $\mathbf{U}_M$, so they evolve independently. For $B \in \{S, N\}$, require $\mathbf{R}_{t,M}\mathbf{P}_{B,M} \neq 0$, excluding blocks with no learning signal, and eigenvalues in $[\underline{\lambda}_{B,M}, \overline{\lambda}_{B,M}] \subset (0, \infty)$, ensuring finite learning times. Both models satisfy Eq. (14) with $0 \le \alpha_S, \alpha_N < 1/2$, so $S$ carries most of the squared initial update norm and $N$ little. To match blocks despite rotation, require nonzero projectors $\mathbf{Q}_S, \mathbf{Q}_N$ onto orthogonal reference subspaces and thresholds $\pi_S, \pi_N \in (1/2, 1]$ such that, for every $B, M$,

$$\min\left\{\frac{\mathrm{Tr}(\mathbf{P}_{B,M}\mathbf{Q}_B)}{\mathrm{rank}(\mathbf{P}_{B,M})}, \frac{\mathrm{Tr}(\mathbf{P}_{B,M}\mathbf{Q}_B)}{\mathrm{rank}(\mathbf{Q}_B)}\right\} \ge \pi_B.$$

The two ratios keep each block concentrated in, and covering, the same reference subspace across models.

Smaller Gram eigenvalues mean slower error decay. To assess how much averaging preserves $S$ and slows $N$, define

$$s_0 = \max\left\{0, \log\frac{\overline{\lambda}_{S,\mathrm{base}}}{\underline{\lambda}_{S,\mathrm{avg}}}, \log\frac{\overline{\lambda}_{S,\mathrm{avg}}}{\underline{\lambda}_{S,\mathrm{base}}}\right\}, \qquad d_0 = \log\frac{\underline{\lambda}_{N,\mathrm{base}}}{\overline{\lambda}_{N,\mathrm{avg}}}. \tag{22}$$

Here $s_0$ bounds the log change in $S$'s rates in either direction; a small value means near preservation. A positive $d_0$ means even $N$'s fastest averaged rate is below its slowest base rate. Both use only Gram spectra. Inverse-eigenvalue bounds on learning times turn these spectral comparisons into the following result.

**Theorem 5.3** (Finite-group regime widening, informal, proof in Appendix D.4)**.** *Under these assumptions, if $d_0 > s_0$, then for every common threshold $\rho \in (0, 1)$, $|\Delta_S(\rho)| \le s_0$ and $\Delta_N(\rho) \ge d_0$, we have*

$$\boxed{\frac{\bar{W}_{\mathrm{avg}}(\rho)}{\bar{W}_{\mathrm{base}}(\rho)} \ge e^{d_0 - s_0} > 1, \qquad \Delta_{\mathrm{win}}(\rho) \ge d_0 - s_0 > 0}.$$

*Remark* 5.4. Similarly, this theorem shows that, under suitable (mild) conditions, the equivariant model widens the generalization regime compared to the non-equivariant one. Specifically, the cyclic construction in Section 5.2 gives $d_0 - s_0 = \log[q_G(\theta_S)/q_G(\theta_N)] > 0$. The condition is also intuitive: when symmetry largely preserves the features supporting $S$ while suppressing more orbit-dependent variation in $N$, it is plausible for $N$'s guaranteed delay to exceed $S$'s largest possible delay, provided the spectral blocks are sufficiently narrow. The theorem then guarantees a larger learning-time ratio at every common threshold, with a nonempty averaged regime when $\bar{W}_{\mathrm{base}} \ge 1$.

### 5.4 Infinite Continuous Group on the Torus

Here we consider the compact continuous group on the torus. Haar-RFM replaces the finite sum in Eq. (3) by a Haar integral (see its explicit definitions and error bounds in Appendix E). Here $K = \mathbb{T}^r$ has a continuously differentiable orthogonal action, and a grid with $m$ equally spaced points per circle approximates the integral. The transfer must control both Gram and target errors, while all models use the common target projected onto $\mathrm{Fix}(K)$. We can have the following theorem.

**Corollary 5.5** (Regime widening under continuous averaging, informal, proof in Appendix E)**.** *Assume bounded Lipschitz $\sigma$ and finite second moments of noisy inputs and targets. For fixed data and weights, grid-to-Haar errors in Gram and target moments are $O(m^{-1})$. If these bounds hold in conditions of Theorem 5.3 for Haar-RFM and Base-RFM with a positive widening margin, then*

$$\bar{W}_{\mathrm{avg}}(\rho) > \bar{W}_{\mathrm{base}}(\rho) \qquad \textit{for every } \rho \in (0,1).$$

This theorem shows that, under suitable (mild) conditions, the equivariant model can similarly widen the generalization regime compared to a non-equivariant one.

## 6 Experiments

We test the predicted selective delays. See Appendix F for experimental details and further checks.

### 6.1 Cyclic-Group Results

Population evaluations for $m \in \{2, 3, 4, 5, 8, 16\}$ (Figure 2) illustrate Theorem 5.1: averaging opens a window $A_{\mathrm{avg}} > 0$ from $A_{\mathrm{base}} = 0$. For $C_2$, $S$'s crossing time rises by 6.2–10.4% while $N$'s doubles, giving $\bar{W}_{\mathrm{avg}}/\bar{W}_{\mathrm{base}} = 1.81$–$1.88$ at every common threshold. The gain nearly saturates at $m = 8$.

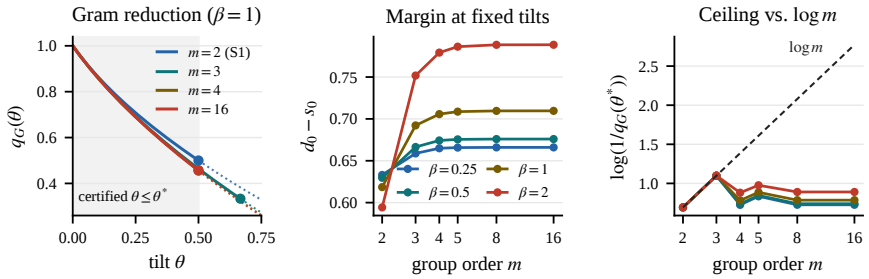


Figure 2: **Cyclic mechanism check.** Retained variance $q_G(\theta)$, $\Delta_{\mathrm{win}}$ at $(\theta_S, \theta_N) = (0.05, 0.5)$, and the maximum $\Delta_N$ compared with $\log m$.

### 6.2 Finite-Group Results

All runs on structured $C_4, C_8, C_{16}$ examples satisfy Theorem 5.3 and exhibit its predicted widening. At $\rho = 0.1$, $\Delta_S = 0$ and mean $\Delta_{\mathrm{win}} = \Delta_N$ ranges from 1.626 to 1.635: $N$ is delayed while $S$ is preserved, increasing $\bar{W}_{\mathrm{avg}}/\bar{W}_{\mathrm{base}}$ above one. All 90 matched controls give $\Delta_{\mathrm{win}} = 0$, hence no ratio gain. (Appendix F.3).

## 7 Conclusion

We identify selective delay as a mechanism by which symmetry can widen the generalization regime in diffusion denoising. In matched frozen-feature models with a shared projected target, orbit averaging delays low-contribution directions more than high-contribution ones, increasing their learning-time ratio even when both stages slow. Exact cyclic constructions and sufficient conditions for finite groups and compact tori establish when this occurs, supported by numerical experiments. This criterion suggests evaluating symmetries by relative learning times. Extending the analysis to learned equivariant features, full-score training across diffusion times, and links to generalization and memorization remains open.

## REFERENCES


Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations in hierarchical architectures. *arXiv preprint arXiv:1311.4158*, 2013. URL https://arxiv.org/abs/1311.4158.

Zhidong Bai and Yongquan Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

Zhidong Bai, Jack W. Silverstein, and Yongquan Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26(2):166–168, 1988.

Bachir Bekka, Pierre de la Harpe, and Alain Valette. *Kazhdan's Property (T)*, volume 11 of *New Mathematical Monographs*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511542749.

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15:9957, 2024.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *NeurIPS*, 2025. URL https://neurips.cc/virtual/2025/poster/119372. Oral.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *ICML*, pp. 1024–1034, 2020.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *ICLR*, 2023. URL https://iclr.cc/virtual/2023/oral/12680. Oral.

Ziyu Chen, Markos A. Katsoulakis, and Benjamin J. Zhang. Equivariant score-based generative models provably learn distributions with symmetries efficiently. *arXiv preprint arXiv:2410.01244*, 2024. URL https://arxiv.org/abs/2410.01244.

Taco S. Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant CNNs on homogeneous spaces. In *NeurIPS*, pp. 9142–9153, 2019.

Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *ICML*, pp. 1528–1537, 2019.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Harm Derksen and Gregor Kemper. Invariant theory of finite groups. In *Computational Invariant Theory*, Encyclopaedia of Mathematical Sciences, pp. 71–152. Springer, 2015. doi: 10.1007/978-3-662-48422-7_3.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *ICML*, pp. 3318–3328, 2021.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves. In *AISTATS*, volume 300 of *PMLR*, pp. 1405–1413, 2026. URL https://proceedings.mlr.press/v300/george26a.html.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *MSML*, pp. 426–471, 2022.

Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. In *ICLR*, 2025. URL https://openreview.net/forum?id=JjdU6ysnCr.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *ICLR*, 2024. URL https://openreview.net/forum?id=h8GeqOxtd4.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *ICML*, pp. 8867–8887, 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Svante Janson. *Gaussian Hilbert Spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997. doi: 10.1017/CBO9780511526169.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *ICLR*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cbaf319a4712385b5ba8a414808b5713-Abstract-Conference.html.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *ICML*, pp. 2747–2755, 2018.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *ICLR*, 2024. URL https://iclr.cc/virtual/2024/poster/19463.

Haoye Lu, Spencer Szabados, and Yaoliang Yu. Diffusion models under group transformations. In *AISTATS*, volume 258 of *PMLR*, pp. 3538–3546, 2025. URL https://proceedings.mlr.press/v258/lu25a.html.

Vladimir A. Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *COLT*, volume 134 of *PMLR*, pp. 3351–3418, 2021. URL https://proceedings.mlr.press/v134/mei21a.html.

Youssef Mroueh, Stephen Voinea, and Tomaso A. Poggio. Learning with group invariant features: A kernel perspective. In *NeurIPS*, volume 28, 2015. URL https://arxiv.org/abs/1506.02544.

Oskar Nordenfors, Fredrik Ohlsson, and Axel Flinth. Optimization dynamics of equivariant and augmented neural networks. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=PTTa3U29NR.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NeurIPS*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *ICML*, volume 139 of *PMLR*, pp. 9323–9332, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pp. 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *NeurIPS*, pp. 55616–55646, 2023.

Vinh Tong, Yun Ye, Trung-Dung Hoang, Anji Liu, Guy Van den Broeck, and Mathias Niepert. Symmetry-preserving diffusion models via target symmetrization. In *ICLR 2025 DeLTa Workshop*, 2025. URL https://arxiv.org/abs/2502.09890. Poster.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

John J. Vastola. Generalization through variance: how noise shapes inductive biases in diffusion models. In *ICLR*, 2025. URL https://iclr.cc/virtual/2025/poster/30806.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *NeurIPS*, volume 38, 2025. URL https://neurips.cc/virtual/2025/poster/117950. Spotlight.

Yu-Han Wu, Pierre Marion, Gérard Biau, and Claire Boyer. Taking a big step: Large learning rates in denoising score matching prevent memorization. In *COLT*, pp. 5718–5756, 2025. URL https://proceedings.mlr.press/v291/wu25a.html.

Yixian Xu, Yusong Wang, Shengjie Luo, Kaiyuan Gao, Tianyu He, Di He, and Chang Liu. Quotient-space diffusion models. In *ICLR*, 2026. URL https://arxiv.org/abs/2604.21809.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Leo Zhang, Kianoosh Ashouritaklimi, Yee Whye Teh, and Rob Cornish. SymDiff: Equivariant diffusion via stochastic symmetrisation. In *ICLR*, 2025. URL https://openreview.net/forum?id=i1NNCrRxdM.

Cai Zhou, Zijie Chen, Zian Li, Jike Wang, Kaiyi Jiang, Pan Li, Rose Yu, Muhan Zhang, Stephen Bates, and Tommi Jaakkola. Rethinking diffusion models with symmetries through canonicalization with applications to molecular graph generation. In *ICML*, volume 306 of *PMLR*, 2026. URL https://arxiv.org/abs/2602.15022.

## APPENDIX OVERVIEW

Proof roadmap . . . . . 14

**A Notation . . . . . 16**

**B Exact Decomposition and Assumptions . . . . . 16**
B.1 Weight-aware Mehler representation . . . . . 17
B.2 Concentration conditional on the weights . . . . . 18
B.3 Scalar calibration . . . . . 20
B.4 Target scope, orbit variance, and reflection-example details . . . . . 21

**C Finite-Group Proofs and Approximation Bounds . . . . . 29**
C.1 Exact low-order Hermite terms . . . . . 29
C.2 Conditional approximation error . . . . . 30
C.3 Approximation checks . . . . . 30
C.4 Spectral and target-contribution bounds . . . . . 32
C.5 Cyclic groups: block structure and the exact regime . . . . . 33

**D Frozen-Feature Dynamics and Time-Gap Bounds . . . . . 38**
D.1 The Gaussian comparison model . . . . . 38
D.2 Denoising loss and parameter dynamics . . . . . 40
D.3 Exact spectral-subspace dynamics . . . . . 42
D.4 Proof of the finite-group separation result . . . . . 43

**E Compact-Torus Approximation . . . . . 46**
E.1 Proof for the Compact Torus . . . . . 47

**F Experiments and Results . . . . . 48**
F.1 Reported Quantities . . . . . 49
F.2 Numerical Procedure . . . . . 49
F.3 Experimental Results . . . . . 50
F.4 QM9 Learned-Feature Results . . . . . 54
F.5 Experiment Summary . . . . . 55
F.6 Statistical Summary . . . . . 55

## PROOF ROADMAP

The roadmap uses two levels. Figure A.1 records the proof blocks and their interfaces; Figure A.2 expands the individual lemmas and theorems. An arrow means that the statement at its tail is used in the proof at its head. A dashed arrow marks an instantiation or a reuse of a general result rather than a prerequisite for the statement at the head. The cyclic branch is therefore parallel to the general finite-group certificate.

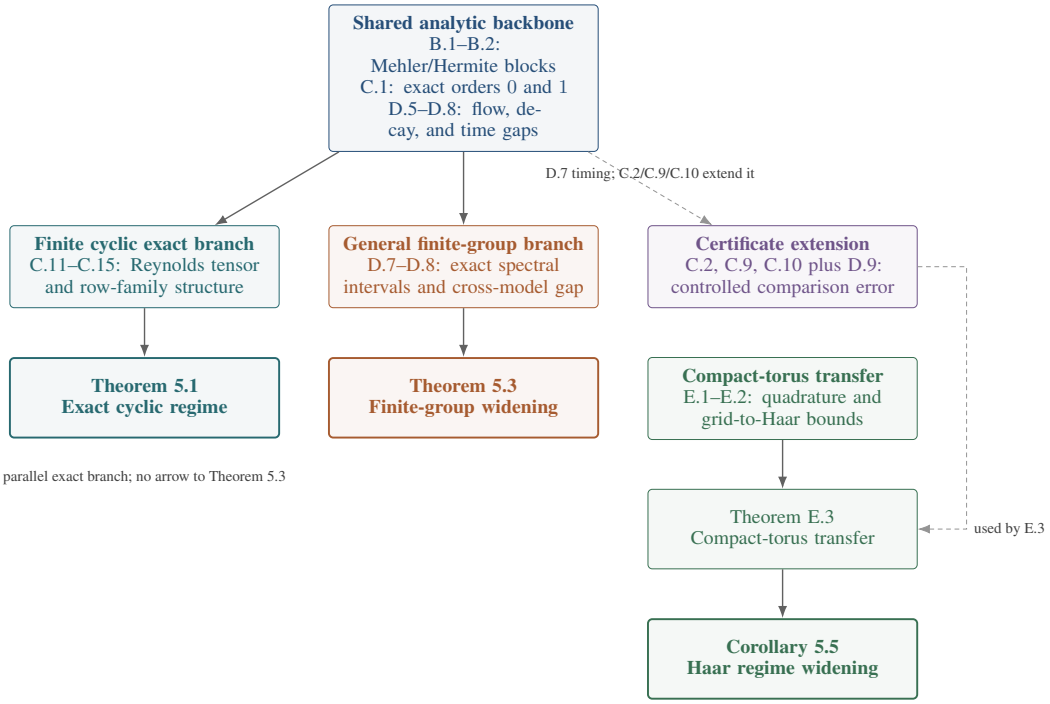


Figure A.1: **Structural proof roadmap.** Colors identify the proof block: blue for shared analytic ingredients, teal for the exact cyclic construction, orange for the general finite-group theorem, purple for the comparison-matrix certificate extension, and green for the compact-torus transfer. Bold boxes are the headline results stated in the main text; bold lines inside ordinary boxes are only block titles. The light fills only group related statements; no result is encoded by shading. The standard Mehler/Hermite and orbit-averaging constructions are attributed in the main text to Janson (1997); Kondor & Trivedi (2018); Derksen & Kemper (2015); the theorem-specific combinations and transfers in the colored blocks are proved in this paper.

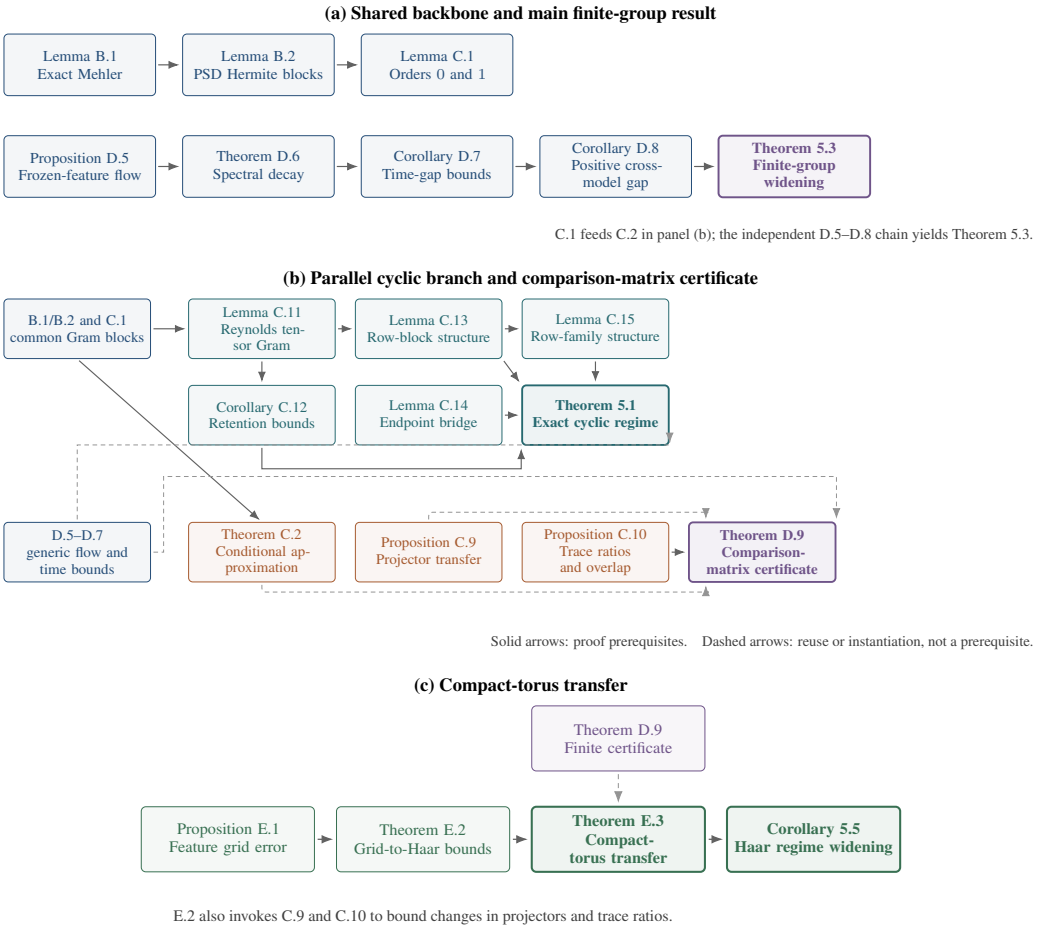


Figure A.2: **Expanded proof roadmap.** Panel (a) separates the common analytic chain from the exact finite-group theorem: D.8 feeds Theorem 5.3. Panel (b) shows the independent cyclic proof and the distinct D.9 comparison-matrix certificate. Panel (c) shows that E.3 reuses D.9 to transfer the certificate from a tensor grid to Haar averaging. The colors and boldface have the same meaning as in Figure A.1; the light fills are grouping cues only, and the labels are tied to the actual appendix references rather than hard-coded numbers.

# A NOTATION

Table A.1 summarizes the key notations in this paper.

Table A.1: **Key Notations.**

| Symbol | Meaning |
|---|---|
| $M;\ B \in \{S, N\}$ | Model (base or avg); block (high- or low-contribution). |
| $t, \tau$ | Diffusion time; parameter along the gradient flow. |
| $\boldsymbol{\phi}_{\text{base}}, \boldsymbol{\phi}_{\text{avg}}$ | Base and orbit-averaged features. |
| $\mathbf{P}_G, \mathbf{P}_K$ | Output projectors onto the finite or compact group's fixed space. |
| $\mathbf{Y}_{t,G}, \mathbf{Y}_{t,K}$ | DSM targets projected onto the corresponding fixed space. |
| $\mathbf{U}_M$ | Feature Gram of model $M$. |
| $\mathbf{R}_{t,M}$ | Cross-moment matrix of targets and features; Eq. (11). |
| $\mathbf{T}_{t,M}$ | Matrix $\mathbf{R}_{t,M}^\top \mathbf{R}_{t,M}$ measuring the squared target term along feature directions. |
| Superscripts $(n), \text{pop}$ | Finite-design and population moments, with weights fixed. |
| $\mathbf{A}_M, \mathbf{A}_{\star,M}$ | Current and minimum-norm stationary parameter matrices. |
| $\mathbf{D}_M$ | Parameter error $\mathbf{A}_M - \mathbf{A}_{\star,M}$. |
| $\mathbf{P}_{B,M}, \mathbf{Q}_B$ | Gram-invariant block projector and shared reference projector. |
| $\ell_{\mathbf{T}}(\mathbf{P})$ | Fraction of $\|\mathbf{R}\|_{\text{F}}^2$ in subspace $\mathbf{P}$; Eq. (13). |
| $\alpha_S, \alpha_N$ | Maximum fractions outside $S$ and inside $N$ in Eq. (14). |
| $\pi_B$ | Required overlap with reference $\mathbf{Q}_B$ in both directions. |
| $\rho, E_{B,M}(\tau)$ | Error threshold; squared block error relative to initialization. |
| $\tau_{B,M}(\rho), \bar{\tau}_{B,M}(\rho)$ | First time $E_{B,M} \le \rho^2$, and scaled time $c_t \tau_{B,M}$. |
| $\tau_{\text{gen},M}, \tau_{\text{mem},M}$ | Learning times of $S$ and $N$, respectively, at the common parameter-error threshold. |
| $A_M(\rho)$ | Time gap $\tau_{\text{mem},M} - \tau_{\text{gen},M}$; window duration when positive. |
| $\bar{W}_M(\rho)$ | Low-to-high learning-time ratio $\tau_{\text{mem},M}/\tau_{\text{gen},M}$. |
| $\Delta_B, \Delta_{\text{win}}$ | Block log delay (avg/base) and regime log increase $\Delta_N - \Delta_S$. |
| $\underline{\lambda}_{B,M}, \overline{\lambda}_{B,M}$ | Lower and upper eigenvalue bounds on block $B$ of model $M$. |
| $s_0, d_0, M_{\text{gap}}$ | Upper bound on $\|\Delta_S\|$, lower bound on $\Delta_N$, and margin $d_0 - s_0$. |
| $\mathbf{\Pi}_B, r_B$ | Coordinate-block projector and block size in the cyclic construction. |
| $\theta_B$ | Squared weight fraction outside the fixed space in block $B$. |
| $q_G(\theta)$ | Fraction of feature variance retained after orbit averaging. |
| $\lambda_{\text{base}}, \kappa_\sigma$ | Base feature variance and its first-order Hermite fraction. |
| $\mathbf{U}_{\text{cond}}$ | Conditional Gram with exact Hermite orders 0 and 1. |
| $\mathbf{U}_{\text{cal}}$ | Scalar-calibrated Gram approximation. |
| $\mathbf{U}_{\text{G}}$ | Gaussian comparison Gram. |
| $\mathbf{U}_M^\star, \varepsilon_M$ | Comparison Gram and bound on $\|\mathbf{U}_M - \mathbf{U}_M^\star\|_{\text{op}}$. |
| $\mathbf{U}_K, \mathbf{U}^{(m)}$ | Haar and grid Grams; $m$ is the number of points per circle. |
| $\varepsilon_{K,m}^U, \varepsilon_{K,m}^R, \varepsilon_{K,m}^T$ | Grid-to-Haar error bounds for the Gram, the matrix $\mathbf{R}$, and the target-contribution matrix. |
| $\widetilde{E}_{\mathbf{Q},M}, \widetilde{\tau}_{\mathbf{Q},M}$ | Uniform-direction spectral decay and its scaled crossing time. |
| $\widetilde{\Delta}_B^{\text{ref}}, \widetilde{\Delta}_B^{\text{exact}}$ | Spectral log delays on shared references or exact Gram blocks. |

# B EXACT DECOMPOSITION AND ASSUMPTIONS

This appendix distinguishes the four matrices used in the analysis. For Avg-RFM, the exact matrix governing the frozen parameter dynamics is $\mathbf{U}_{\text{avg}}$ (for Base-RFM, replace it throughout by $\mathbf{U}_{\text{base}}$). The conditional approximation $\mathbf{U}_{\text{cond}}$ replaces only the centered sample fluctuations in diffusion-noise Hermite orders $k \ge 2$. The calibrated matrix $\mathbf{U}_{\text{cal}}$ replaces the exact first-order Hermite term by $\mathbf{B}_{\text{cal}}$ and the remaining first- and higher-order residual by $\mathbf{D}_{\text{cal}}$. The Gaussian matrix $\mathbf{U}_{\text{G}}$ remains descriptive unless it has a separate explicit operator-norm error bound. Matching spectral subspaces, checking the block trace ratios, and checking overlap in both directions are later steps with their own conditions.

A naive Hermite-tail argument would incorrectly infer from

$$\sup_{q\in[q_-,q_+],\, s\in\mathbb{R}} \sum_{k>K_{\mathrm{H}}} \frac{\mu_{k,q}(s)^2}{k!} \xrightarrow[K_{\mathrm{H}}\to\infty]{} 0$$

that the sum of the degreewise suprema converges. We instead split each Hermite tail into diagonal and off-diagonal matrix entries. The uniform total tail controls the diagonal, while orbit coherence supplies a geometric factor for the off-diagonal entries. The resulting bound is measurable with respect to the weights and preserves conditional independence across clean-sample blocks.

### B.1 Weight-aware Mehler representation

For the finite orthogonal action, the Reynolds projector is

$$\mathbf{P}_G = \frac{1}{|G|}\sum_{g\in G} g. \tag{B.1}$$

For row $i$, sample $\nu$, and group element $g$, define

$$\begin{aligned} q_i &= \frac{\|\mathbf{w}_i\|^2}{d}, \\ S_{i,g}^{\nu} &= \frac{e^{-t}\mathbf{w}_i^\top g\mathbf{x}^{\nu}}{\sqrt{d}}, \\ \mu_{k,q}(s) &= \mathbb{E}\Big[\sigma\Big(s+\sqrt{\Delta_t q}\, Z\Big)\,\mathrm{He}_k(Z)\Big], \quad Z\sim\mathcal{N}(0,1). \end{aligned} \tag{B.2}$$

Here $\mathrm{He}_k$ is the probabilists' Hermite polynomial of degree $k$ and $\gamma=\mathcal{N}(0,1)$. Since $\{\mathrm{He}_k/\sqrt{k!} : k\ge 0\}$ is an orthonormal basis of $L^2(\gamma)$, orthogonality gives

$$\sigma\Big(s+\sqrt{\Delta_t q}\, z\Big) = \sum_{k\ge 0}\frac{\mu_{k,q}(s)}{k!}\mathrm{He}_k(z) \quad \text{in } L^2(\gamma).$$

Thus $\mu_{k,q}(s)/k!$ multiplies $\mathrm{He}_k$ in this expansion. These definitions are conditional on the fixed clean design and weights. Assume $q_i>0$ for the rows used below. A zero-weight row yields the constant feature $\sigma(0)$ and is excluded from this representation or handled separately in the $k=0$ block. Let

$$Z_{i,g} = \frac{\mathbf{w}_i^\top g\boldsymbol{\xi}}{\sqrt{dq_i}}, \qquad \mathbf{v}_{i,g} = \frac{g^\top \mathbf{w}_i}{\|\mathbf{w}_i\|}.$$

Then each $Z_{i,g}$ is standard normal and

$$\mathbb{E}[Z_{i,g}Z_{j,g'} \mid \mathbf{W}] = (\mathbf{C}_{g,g'})_{ij} = \langle \mathbf{v}_{i,g}, \mathbf{v}_{j,g'}\rangle .$$

For each sample $\nu$ and Hermite order $k\ge 0$, define the $k$th noise-averaged Gram block by

$$(\mathbf{U}_k^{\nu})_{ij} = \frac{1}{|G|^2}\sum_{g,g'\in G}\frac{(\mathbf{C}_{g,g'})_{ij}^k}{k!}\,\mu_{k,q_i}(S_{i,g}^{\nu})\mu_{k,q_j}(S_{j,g'}^{\nu}).$$

With this definition, the exact noise-averaged sample Gram is $\mathbf{U}^{\nu}=\sum_{k\ge 0}\mathbf{U}_k^{\nu}$ whenever the series is integrable.

**Lemma B.1** (Exact group-averaged Mehler decomposition)**.** *Assume $\sigma(s+\sqrt{\Delta_t q}\, Z)\in L^2(\gamma)$ for every relevant $(s,q)$. For fixed data, weights, and group elements,*

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}}\left[\sigma\Big(S_{i,g}^{\nu}+\sqrt{\Delta_t q_i}Z_{i,g}\Big)\,\sigma\Big(S_{j,g'}^{\nu}+\sqrt{\Delta_t q_j}Z_{j,g'}\Big)\right] \\ &\qquad = \sum_{k\ge 0}\frac{(\mathbf{C}_{g,g'})_{ij}^k}{k!}\mu_{k,q_i}(S_{i,g}^{\nu})\mu_{k,q_j}(S_{j,g'}^{\nu}). \end{aligned} \tag{B.3}$$

*The series is absolutely convergent, and averaging over $(g,g')$ and samples gives the positive-semidefinite Hermite expansion of Lemma B.2.*

*Proof.* Set $f_i(z) = \sigma(s_i + \sqrt{\Delta_t q_i}\, z)$ and $f_j(z) = \sigma(s_j + \sqrt{\Delta_t q_j}\, z)$, where $s_i = S^\nu_{i,g}$ and $s_j = S^\nu_{j,g'}$. First, truncate their Hermite expansions at degree $K$. For jointly standard Gaussian variables with correlation $c$,

$$\mathbb{E}[\mathrm{He}_k(Z_1)\mathrm{He}_\ell(Z_2)] = \delta_{k\ell} k! c^k.$$

Applying this identity to the two finite truncations and then letting $K \to \infty$ proves Eq. (B.3): the truncations converge in $L^2$ under the respective Gaussian marginals, so their product converges in expectation by Cauchy–Schwarz. Next, Cauchy–Schwarz and Parseval give absolute convergence of the limiting series:

$$\sum_{k\geq 0} \frac{|c|^k |\mu_{k,q_i}(s_i)\mu_{k,q_j}(s_j)|}{k!} \leq \left\|\sigma(s_i + \sqrt{\Delta_t q_i} Z)\right\|_{L^2} \left\|\sigma(s_j + \sqrt{\Delta_t q_j} Z)\right\|_{L^2}.$$

Finally, in the lemma $c = (\mathbf{C}_{g,g'})_{ij}$, so the displayed identity is exactly Eq. (B.3). □

**Lemma B.2** (Positive-semidefinite Hermite-order blocks)**.** *For every $k \geq 0$ and sample $\nu$, $\mathbf{U}^\nu_k \succeq 0$. More precisely, in the $k$th symmetric tensor power of $\mathbb{R}^d$, define*

$$\boldsymbol{\zeta}^\nu_{i,k} = \frac{1}{|G|} \sum_{g\in G} \frac{\mu_{k,q_i}(S^\nu_{i,g})}{\sqrt{k!}} \mathbf{v}^{\otimes k}_{i,g}.$$

*Then $(\mathbf{U}^\nu_k)_{ij} = \langle \boldsymbol{\zeta}^\nu_{i,k}, \boldsymbol{\zeta}^\nu_{j,k} \rangle$.*

*Proof.* Use $\left\langle \mathbf{v}^{\otimes k}_{i,g}, \mathbf{v}^{\otimes k}_{j,g'} \right\rangle = \langle \mathbf{v}_{i,g}, \mathbf{v}_{j,g'} \rangle^k = (\mathbf{C}_{g,g'})^k_{ij}$ and expand the tensor inner product. □

*Remark* B.3 (Convergence and conditional expectation). For bounded $\sigma$, the partial sums $\sum_{k=0}^K \mathbf{U}^\nu_k$ increase in the positive-semidefinite order and have trace at most $p\|\sigma\|^2_\infty$. They therefore converge in trace norm, and hence in operator norm, to the noise-averaged sample Gram. Applying monotone convergence to quadratic forms gives $\mathbb{E}[\sum_{k\geq 2} \mathbf{U}^1_k \mid \mathbf{W}] = \sum_{k\geq 2} \mathbb{E}[\mathbf{U}^1_k \mid \mathbf{W}]$.

### B.2 Concentration conditional on the weights

**Assumption B.4** (Conditional sampling and weight regularity)**.** *The clean samples $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are i.i.d. and independent of $\mathbf{W}$. For deterministic constants $0 < q_- < q_+ < \infty$ and $0 \leq \mu_d < 1$, let $\mathcal{E}_{W,d}$ be the event*

$$q_- \leq \min_i q_i \leq \max_i q_i \leq q_+, \qquad \chi_G(\mathbf{W}) \leq \mu_d,$$

*where*

$$\chi_G(\mathbf{W}) = \max_{i\neq j,\, h\in G} \frac{|\mathbf{w}_i^\top h \mathbf{w}_j|}{\|\mathbf{w}_i\|\, \|\mathbf{w}_j\|}.$$

*We write $\eta_d = \mathbb{P}(\mathcal{E}^c_{W,d})$.*

**Assumption B.5** (Bounded saturating activation)**.** *The activation $\sigma$ is bounded and continuous and has finite limits at $+\infty$ and $-\infty$. Write $B_\sigma = \sup_x |\sigma(x)|$.*

**Proposition B.6** (Uniform Hermite tail)**.** *Under Assumption B.5, for every fixed $0 < q_- < q_+ < \infty$,*

$$\mathfrak{r}_{K_\mathrm{H}} = \sup_{q\in[q_-,q_+],\, s\in\mathbb{R}} \sum_{k>K_\mathrm{H}} \frac{\mu_{k,q}(s)^2}{k!} \xrightarrow[K_\mathrm{H}\to\infty]{} 0. \tag{B.4}$$

*Proof.* Set $f_{s,q}(z) = \sigma(s + \sqrt{\Delta_t q}\, z)$. Parseval gives

$$\sum_{k>K_\mathrm{H}} \frac{\mu_{k,q}(s)^2}{k!} = \|f_{s,q} - \Pi_{K_\mathrm{H}} f_{s,q}\|^2_{L^2(\gamma)},$$

where $\Pi_{K_\mathrm{H}}$ is Hermite projection through degree $K_\mathrm{H}$. The map $(s,q) \mapsto f_{s,q}$ is continuous into $L^2(\gamma)$. The finite limits of $\sigma$ extend this map continuously to the compactification $[-\infty,+\infty] \times [q_-, q_+]$, with constant limiting functions at the two endpoints. Its image is therefore compact in $L^2(\gamma)$. Strong convergence $\Pi_{K_\mathrm{H}} f \to f$ is uniform on compact subsets of a Hilbert space, proving (B.4). □

**Proposition B.7** (Deterministic bounds for finite Hermite orders and the tail)**.** *Work on $\mathcal{E}_{W,d}$. For every clean sample and every $K_{\mathrm{H}} \geq 2$, define*

$$\mathbf{U}^{\nu}_{2:K_{\mathrm{H}}} = \sum_{k=2}^{K_{\mathrm{H}}} \mathbf{U}^{\nu}_{k}, \qquad \mathbf{U}^{\nu}_{>K_{\mathrm{H}}} = \sum_{k>K_{\mathrm{H}}} \mathbf{U}^{\nu}_{k}.$$

*Then*

$$\left\|\mathbf{U}^{\nu}_{2:K_{\mathrm{H}}}\right\|_{\mathrm{op}} \leq B_{\sigma}^{2}\left(1 + (p-1)\frac{\mu_d^2}{1-\mu_d}\right) = L_d, \tag{B.5}$$

$$\left\|\mathbf{U}^{\nu}_{>K_{\mathrm{H}}}\right\|_{\mathrm{op}} \leq \mathfrak{r}_{K_{\mathrm{H}}} + B_{\sigma}^{2}(p-1)\frac{\mu_d^{K_{\mathrm{H}}+1}}{1-\mu_d} = T_{K_{\mathrm{H}},d}. \tag{B.6}$$

*The same tail bound holds for $\left\|\mathbb{E}[\mathbf{U}^{1}_{>K_{\mathrm{H}}} \mid \mathbf{W}]\right\|_{\mathrm{op}}$.*

*Proof.* For a diagonal entry, $|(\mathbf{C}_{g,g'})_{ii}| \leq 1$. Cauchy–Schwarz over Hermite degree gives

$$\sum_{k=2}^{K_{\mathrm{H}}} \frac{|\mu_{k,q_i}(s)\mu_{k,q_i}(s')|}{k!} \leq \left(\sum_{k\geq 2}\frac{\mu_{k,q_i}(s)^2}{k!}\right)^{1/2}\left(\sum_{k\geq 2}\frac{\mu_{k,q_i}(s')^2}{k!}\right)^{1/2} \leq B_{\sigma}^{2}.$$

The same argument restricted to $k > K_{\mathrm{H}}$ gives the diagonal tail bound $\mathfrak{r}_{K_{\mathrm{H}}}$.

For $i \neq j$, $|(\mathbf{C}_{g,g'})_{ij}| \leq \mu_d$. Parseval implies $|\mu_{k,q}(s)|/\sqrt{k!} \leq B_\sigma$, hence

$$\sum_{k=2}^{K_{\mathrm{H}}} \frac{|(\mathbf{C}_{g,g'})_{ij}|^k |\mu_{k,q_i}(s)\mu_{k,q_j}(s')|}{k!} \leq B_{\sigma}^{2}\frac{\mu_d^2}{1-\mu_d},$$

and the corresponding tail is at most $B_{\sigma}^{2}\mu_d^{K_{\mathrm{H}}+1}/(1-\mu_d)$. Averaging over group pairs preserves these entrywise bounds. The operator norm of a symmetric matrix is bounded by its maximum absolute row sum, which yields (B.5)–(B.6). Finally, Jensen gives $\left\|\mathbb{E}[\mathbf{U}^{1}_{>K_{\mathrm{H}}} \mid \mathbf{W}]\right\|_{\mathrm{op}} \leq \mathbb{E}[\left\|\mathbf{U}^{1}_{>K_{\mathrm{H}}}\right\|_{\mathrm{op}} \mid \mathbf{W}] \leq T_{K_{\mathrm{H}},d}$. □

**Proposition B.8** (Gaussian-row orbit coherence)**.** *Let the rows of $\mathbf{W}$ be i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and suppose $p|G|$ grows at most exponentially in $o(d)$. Then the normalized squared row norms $q_i = \|\mathbf{w}_i\|^2/d$ lie in a fixed interval $[q_-, q_+]$ with probability tending to one and*

$$\chi_G(\mathbf{W}) = O_{\mathbb{P}}\left(\sqrt{\frac{\log(p|G|)}{d}}\right). \tag{B.7}$$

*Proof.* Chi-square concentration controls all row norms. The normalized rows are independent and uniform on the sphere. For fixed $i \neq j$ and $h \in G$, the inner product $\mathbf{w}_i^{\top} h \mathbf{w}_j/(\|\mathbf{w}_i\|\,\|\mathbf{w}_j\|)$ has a sub-Gaussian tail at scale $d^{-1/2}$ because $h$ is orthogonal. A union bound over $p(p-1)|G|$ triples proves (B.7). □

**Corollary B.9** (Conditions for vanishing conditional-approximation error)**.** *Assume $\sigma$ is* tanh *or logistic sigmoid, the rows of $\mathbf{W}$ are i.i.d. standard Gaussian, $p \leq Cd$, $|G| \leq d^C$, and $n/(\log d)^3 \to \infty$. Then there exists $K_{\mathrm{H},d} \to \infty$ for which the conditional-approximation error bound $\varepsilon_{\mathrm{br}}$ in* (C.5) *tends to zero in probability, and hence $\|\mathbf{U}_{\mathrm{avg}} - \mathbf{U}_{\mathrm{cond}}\|_{\mathrm{op}} = o_{\mathbb{P}}(1)$.*

*Proof.* Take $\delta_d = d^{-2}$ in Theorem C.2. Choose a deterministic sequence $C_d \to \infty$ slowly enough that the event

$$\chi_G(\mathbf{W}) \leq \widetilde{\mu}_d \stackrel{\mathrm{def}}{=} C_d\sqrt{\frac{\log(p|G|)}{d}}$$

has probability tending to one; also intersect with the row-norm event. Then

$$L_d = O\big(1 + p\widetilde{\mu}_d^2\big) = O(C_d^2 \log d).$$

If $C_d$ grows sufficiently slowly, the two Bernstein terms, now with $\log(2p/\delta_d) = O(\log d)$, vanish under $n/(\log d)^3 \to \infty$. Proposition B.6 gives $\mathfrak{r}_{K_{\mathrm{H}}} \to 0$. Because $\widetilde{\mu}_d \to 0$, a diagonal choice of deterministic $K_{\mathrm{H},d} \to \infty$ can satisfy simultaneously $\mathfrak{r}_{K_{\mathrm{H},d}} \to 0$ and $p\widetilde{\mu}_d^{K_{\mathrm{H},d}+1} \to 0$. Thus $T_{K_{\mathrm{H},d},d} \to 0$ on an event whose probability tends to one. Since $\eta_d + \delta_d \to 0$, the resulting operator error is $o_{\mathbb{P}}(1)$. □

**Assumption B.10** (Gaussian data used only for calibration and the benchmark)**.** *When the scalar calibration or Gaussian benchmark is invoked, the clean samples are i.i.d.* $\mathbf{x}^\nu \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_d)$*, independently of* $\mathbf{W}$*, and* $\mathbf{\Sigma}_d g = g\mathbf{\Sigma}_d$ *for every* $g \in G$*. Let* $d_{\mathrm{proj},d} = \mathrm{rank}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)$*. We assume* $d_{\mathrm{proj},d} \to \infty$*,* $p/d_{\mathrm{proj},d}$ *and* $n/d_{\mathrm{proj},d}$ *remain in a compact subset of* $(0,\infty)$*,* $\sup_d \|\mathbf{\Sigma}_d\|_{\mathrm{op}} < \infty$*, and the average nonzero eigenvalue of the projected covariance is bounded away from zero.*

The commutation condition makes this Gaussian model $G$-invariant but is not required by Theorem C.2.

**Assumption B.11** (Condition for using the calibrated approximation)**.** *Whenever an asymptotic theorem uses* $\mathbf{U}_{\mathrm{cal}}$ *rather than merely measuring it, the combined calibration error in Proposition C.7 satisfies* $\varepsilon_{\mathrm{cal},d} = o_{\mathbb{P}}(1)$*. The stronger componentwise conditions* $\|\mathbf{G}_1 - \mathbf{B}_{\mathrm{cal}}\|_{\mathrm{op}} = o_{\mathbb{P}}(1)$ *and* $\|\mathbf{M}_{\geq 2} - \mathbf{D}_{\mathrm{cal}}\|_{\mathrm{op}} = o_{\mathbb{P}}(1)$ *are one sufficient condition, but they are not required by the descriptive scalar comparison.*

**Assumption B.12** (Separated and target-identified subspaces)**.** *Whenever a comparison cluster is connected to an exact-Gram subspace, its gap from the rest of the comparison spectrum exceeds twice the stated operator-error bound. Whenever a cluster is called high-contribution or low-contribution, it satisfies the target-contribution condition in Section 5.1.2; a common-reference claim also satisfies the overlap condition in Section 5.3. Whenever a tight-cluster equality is used, the relative cluster spread tends to zero.*

### B.3 Scalar calibration

Assume the Gaussian setting above, unit normalized row norms $q_i = \|\mathbf{w}_i\|^2/d = 1$ for the scalar comparison, and $\mathrm{Tr}(\mathbf{\Sigma}_d) > 0$. Define

$$\gamma_G = \frac{\mathrm{Tr}(\mathbf{P}_G)}{d}, \qquad \kappa_{\mathrm{sig}} = \frac{\mathrm{Tr}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)}{\mathrm{Tr}(\mathbf{\Sigma}_d)}, \qquad \sigma_{\mathbf{x}}^2 = \frac{\mathrm{Tr}(\mathbf{\Sigma}_d)}{d}. \tag{B.8}$$

Whenever $\gamma_G > 0$, they satisfy

$$\frac{\mathrm{Tr}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)}{\mathrm{Tr}(\mathbf{P}_G)} = \frac{\kappa_{\mathrm{sig}}}{\gamma_G}\sigma_{\mathbf{x}}^2. \tag{B.9}$$

Three variance scales must be kept separate:

$$\begin{aligned} \sigma_{\mathrm{pre},t}^2 &= e^{-2t}\frac{\mathrm{Tr}(\mathbf{\Sigma}_d)}{d}, \\ \sigma_{\mathrm{proj},t}^2 &= e^{-2t}\frac{\mathrm{Tr}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)}{d}, \\ \sigma_{\mathrm{fix},t}^2 &= e^{-2t}\frac{\mathrm{Tr}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)}{\mathrm{Tr}(\mathbf{P}_G)} = \frac{\sigma_{\mathrm{proj},t}^2}{\gamma_G} \qquad (\gamma_G > 0). \end{aligned} \tag{B.10}$$

For a fixed row $\mathbf{w}$ satisfying $\|\mathbf{w}\|^2 = d$, the variance of the clean preactivation $e^{-t}\mathbf{w}^\top\mathbf{X}/\sqrt{d}$ equals $e^{-2t}\mathbf{w}^\top\mathbf{\Sigma}_d\mathbf{w}/d$. For an isotropically distributed normalized row $\mathbf{w}/\sqrt{d}$, the average of this variance is $\sigma_{\mathrm{pre},t}^2$, while the corresponding orbit-averaged linear term has average variance scale $\sigma_{\mathrm{proj},t}^2$. The fixed-coordinate density $\sigma_{\mathrm{fix},t}^2$ is a per-coordinate quantity and should not be substituted for the marginal preactivation variance $\sigma_{\mathrm{pre},t}^2$.

Write $\bar{\mu}_k(s) = \mu_{k,1}(s)$ and let $S \sim \mathcal{N}(0, \sigma_{\mathrm{pre},t}^2)$ with $\sigma_{\mathrm{pre},t}^2 > 0$. Define

$$\begin{aligned} m_t &= \mathbb{E}\bar{\mu}_0(S), & a_t &= \frac{\mathbb{E}[S\bar{\mu}_0(S)]}{\sigma_{\mathrm{pre},t}^2}, \\ r_t(S) &= \bar{\mu}_0(S) - m_t - a_t S, & v_t^2 &= \mathbb{E}r_t(S)^2, \\ b_t &= \mathbb{E}\bar{\mu}_1(S), & s_t^2 &= \mathrm{Var}(\bar{\mu}_1(S)) + \sum_{k\geq 2}\frac{\mathbb{E}\bar{\mu}_k(S)^2}{k!}. \end{aligned} \tag{B.11}$$

The last line is nonnegative and gives the scalar identity

$$\sum_{k\geq 0}\frac{\mathbb{E}\bar{\mu}_k(S)^2}{k!} = m_t^2 + a_t^2\sigma_{\mathrm{pre},t}^2 + v_t^2 + b_t^2 + s_t^2. \tag{B.12}$$

If $v_t^2 > 0$, define

$$\Psi_t(c) = \frac{\mathbb{E}[r_t(S_1)r_t(S_2)]}{v_t^2} \tag{B.13}$$

where $S_1, S_2$ are jointly Gaussian with marginal variance $\sigma_{\mathrm{pre},t}^2$ and correlation $c$. Then $\Psi_t(0) = 0$ and $\Psi_t(1) = 1$. Because $r_t$ is orthogonal to both constants and the linear Gaussian coordinate, its normalized Hermite expansion starts at degree two. Mehler's identity (Janson, 1997), equivalently the expansion in Lemma B.1, therefore gives the uniform bound

$$|\Psi_t(c)| \leq |c|^2, \qquad |c| \leq 1. \tag{B.14}$$

The scalar approximation used in the experiments is

$$\mathbf{B}_{\mathrm{cal}} = b_t^2\frac{\mathbf{W}\mathbf{P}_G\mathbf{W}^\top}{d}, \qquad \mathbf{D}_{\mathrm{cal}} = \gamma_G s_t^2\mathbf{I}_p. \tag{B.15}$$

It is a fitted structural approximation, not an exact consequence of the Gaussian setting: the clean nonlinear residual, first-order fluctuations, and higher Hermite orders need not be captured by the same scalar factor. Any theorem using this matrix must include the measured $\varepsilon_{\mathrm{cal},d}$ in its total error bound.

### B.4 Target scope, orbit variance, and reflection-example details

This subsection establishes the target restriction and the orbit-variance identity, then relates feature averaging to data augmentation and controls the finite-sample correction. It also records the aligned reflection as a spectral proxy and gives the activation and finite-group geometry checks used by the validation protocol. The cyclic population theorem is proved in Appendix C.5. Throughout, $t > 0$ and $\mathbf{W}$ are fixed. We use the feature notation of Section 4.1,

$$\boldsymbol{\phi} = \boldsymbol{\phi}_{\mathrm{avg}}, \qquad \boldsymbol{\psi}_g(\mathbf{z}) = \boldsymbol{\phi}_{\mathrm{base}}(g\mathbf{z}),$$

and all expectations below are over the variables explicitly displayed. When a population clean law is used, it is denoted by $\mu_0$; the finite design itself remains fixed unless an i.i.d. sampling statement is made.

**Group averaging and the fixed subspace.** Let $G \subseteq O(d)$ be finite and $\mathbf{P}_G = |G|^{-1}\sum_{g\in G} g$. For each $h \in G$, left multiplication permutes the group elements, so $h\mathbf{P}_G = \mathbf{P}_G$. Hence $\mathbf{P}_G\mathbf{v} \in \mathrm{Fix}(G)$ for every $\mathbf{v} \in \mathbb{R}^d$. Conversely, if $\mathbf{u} \in \mathrm{Fix}(G)$, then every summand fixes $\mathbf{u}$, giving $\mathbf{P}_G\mathbf{u} = \mathbf{u}$. Thus $\mathrm{Im}(\mathbf{P}_G) = \mathrm{Fix}(G)$ and $\mathbf{P}_G^2 = \mathbf{P}_G$. Orthogonality of the group action and closure under inverses also give

$$\mathbf{P}_G^\top = \frac{1}{|G|}\sum_{g\in G} g^\top = \frac{1}{|G|}\sum_{g\in G} g^{-1} = \mathbf{P}_G.$$

For any $\mathbf{v} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathrm{Fix}(G)$, it follows that

$$\langle \mathbf{v} - \mathbf{P}_G\mathbf{v}, \mathbf{u}\rangle = \langle \mathbf{v}, \mathbf{u} - \mathbf{P}_G\mathbf{u}\rangle = 0.$$

Thus $\mathbf{P}_G$ is the orthogonal projector onto $\mathrm{Fix}(G)$: it retains the component in that subspace and removes the perpendicular component. In particular, write the DSM target as $\mathbf{Y}_t = \mathbf{Y}_{t,G} + \mathbf{Y}_{t,\perp}$, where

$$\mathbf{Y}_{t,G} = \mathbf{P}_G\mathbf{Y}_t \in \mathrm{Fix}(G),$$
$$\mathbf{Y}_{t,\perp} := (\mathbf{I}_d - \mathbf{P}_G)\mathbf{Y}_t \in \mathrm{Fix}(G)^\perp.$$

**Target restriction and loss decomposition.** For the diffusion model in Section 3.1, a $G$-invariant clean-data law and independent isotropic Gaussian noise give $(\mathbf{X}, \boldsymbol{\xi}) \overset{d}{=} (g\mathbf{X}, g\boldsymbol{\xi})$. Applying the definitions of $\mathbf{X}_t$ and $\mathbf{Y}_t$ to these pairs gives the joint invariance assumed in the following proposition.

**Proposition B.13** (Fixed-subspace target)**.** *Let $G \subseteq O(d)$ be finite and use the averaged features $\boldsymbol{\phi} = \boldsymbol{\phi}_{\mathrm{avg}}$ in Eq. (5). Suppose $(\mathbf{X}_t, \mathbf{Y}_t)$ and $(g\mathbf{X}_t, g\mathbf{Y}_t)$ have the same distribution for every $g \in G$, and $\mathbf{X}_t$, $\mathbf{Y}_t$, and $\boldsymbol{\phi}(\mathbf{X}_t)$ have finite second moments. The cross-moment matrix $\mathbf{R}_t = \mathbb{E}[\mathbf{Y}_t \boldsymbol{\phi}(\mathbf{X}_t)^\top]$ satisfies $\mathbf{P}_G \mathbf{R}_t = \mathbf{R}_t$, and the residual satisfies*

$$\mathbb{E}\big[\mathbf{Y}_{t,\perp} \boldsymbol{\phi}(\mathbf{X}_t)^\top\big] = (\mathbf{I}_d - \mathbf{P}_G)\mathbf{R}_t = \mathbf{0}. \tag{B.16}$$

*Among linear predictors using $\boldsymbol{\phi}$, there is a population squared-loss minimizer with output in $\mathrm{Fix}(G)$, and the zero predictor is optimal for $\mathbf{Y}_{t,\perp}$. If $\mathrm{Fix}(G) = \{0\}$, zero is optimal for the full target.*

*Proof.* Feature invariance follows because right multiplication by any $h \in G$ permutes the finite sum in Eq. (5). Joint $G$-invariance in law and feature invariance give

$$\begin{aligned}\mathbf{R}_t &= \mathbb{E}[\mathbf{Y}_t \boldsymbol{\phi}(\mathbf{X}_t)^\top] \\ &= \mathbb{E}[g\mathbf{Y}_t \boldsymbol{\phi}(g\mathbf{X}_t)^\top] \\ &= g\mathbf{R}_t.\end{aligned}$$

Averaging over $g$ yields $\mathbf{P}_G \mathbf{R}_t = \mathbf{R}_t$, which implies Eq. (B.16).

Let $\mathbf{P} = \mathbf{P}_G$ and $\widehat{\mathbf{Y}} = p^{-1/2}\mathbf{A}\boldsymbol{\phi}(\mathbf{X}_t)$. Decompose the squared error into the orthogonal output spaces $\mathrm{Im}(\mathbf{P})$ and $\mathrm{Im}(\mathbf{I} - \mathbf{P})$:

$$\begin{aligned}\|\widehat{\mathbf{Y}} - \mathbf{Y}_t\|^2 = \|\mathbf{P}\widehat{\mathbf{Y}} - \mathbf{P}\mathbf{Y}_t\|^2 \\ + \|(\mathbf{I} - \mathbf{P})\widehat{\mathbf{Y}} - (\mathbf{I} - \mathbf{P})\mathbf{Y}_t\|^2.\end{aligned}$$

The cross term in the second square has expectation

$$\begin{aligned}\mathbb{E}\left\langle (\mathbf{I} - \mathbf{P})\widehat{\mathbf{Y}}, (\mathbf{I} - \mathbf{P})\mathbf{Y}_t \right\rangle &= p^{-1/2}\,\mathrm{Tr}\big((\mathbf{I} - \mathbf{P})\mathbf{A}\mathbb{E}[\boldsymbol{\phi}(\mathbf{X}_t)\mathbf{Y}_t^\top](\mathbf{I} - \mathbf{P})\big) \\ &= 0,\end{aligned}$$

because $(\mathbf{I} - \mathbf{P})\mathbf{R}_t = 0$. For $\mathcal{L}(\mathbf{A}) = \mathbb{E}[\|p^{-1/2}\mathbf{A}\boldsymbol{\phi}(\mathbf{X}_t) - \mathbf{Y}_t\|^2]$, expanding the remaining terms gives the full risk identity

$$\begin{aligned}\mathcal{L}(\mathbf{A}) = \mathbb{E}\Big[\|p^{-1/2}\mathbf{P}_G\mathbf{A}\boldsymbol{\phi}(\mathbf{X}_t) - \mathbf{P}_G\mathbf{Y}_t\|^2\Big] \\ + \mathbb{E}\Big[\|p^{-1/2}(\mathbf{I}_d - \mathbf{P}_G)\mathbf{A}\boldsymbol{\phi}(\mathbf{X}_t)\|^2\Big] \\ + \mathbb{E}\big[\|(\mathbf{I}_d - \mathbf{P}_G)\mathbf{Y}_t\|^2\big].\end{aligned} \tag{B.17}$$

Left-multiplying the parameter matrix by $\mathbf{P}$ removes the nonnegative second term without changing the first. For the residual target alone, Eq. (B.16) similarly gives

$$\mathbb{E}\|\widehat{\mathbf{Y}} - \mathbf{Y}_{t,\perp}\|^2 = \mathbb{E}\|\widehat{\mathbf{Y}}\|^2 + \mathbb{E}\|\mathbf{Y}_{t,\perp}\|^2,$$

so predicting zero minimizes its population squared loss. If $\mathbf{P} = 0$, then $\mathbf{R}_t = 0$, so the zero predictor is optimal. □

**Conditional mean and input invariance.** Since $\mathbf{P}_G$ is deterministic, the DSM identity in Section 3.1 gives

$$\mathbb{E}[\mathbf{Y}_{t,G} \mid \mathbf{X}_t = \mathbf{z}] = \mathbf{P}_G\mathbb{E}[\mathbf{Y}_t \mid \mathbf{X}_t = \mathbf{z}] = \mathbf{P}_G\nabla \log p_t(\mathbf{z}),$$

which proves Eq. (7). The $G$-invariant clean-data law and isotropic noise also make $p_t$ $G$-invariant. The following transformation law relates this symmetry to that of the score.

**Lemma B.14** (Score transformation law)**.** *If a positive differentiable density satisfies $p_t(g\mathbf{x}) = p_t(\mathbf{x})$ for an orthogonal $g$, then*

$$\nabla \log p_t(g\mathbf{x}) = g\nabla \log p_t(\mathbf{x}).$$

*Proof.* Differentiation with respect to $\mathbf{x}$ gives $g^\top \nabla \log p_t(g\mathbf{x}) = \nabla \log p_t(\mathbf{x})$. Multiplication by $g$ proves the claim. □

Combining Lemma B.14 with $\mathbf{P}_G g = \mathbf{P}_G$ gives

$$\mathbf{P}_G \nabla \log p_t(g\mathbf{z}) = \mathbf{P}_G g \nabla \log p_t(\mathbf{z}) = \mathbf{P}_G \nabla \log p_t(\mathbf{z}).$$

Thus the conditional mean of the projected target is constant on input orbits and takes values in $\mathrm{Fix}(G)$. This establishes its compatibility with the symmetry of the averaged features in Section 4.2.

**Orbit moments and population comparison.** For the fixed clean design of Section 5.1.1, let $g$ be uniform on $G$ and independent of the diffusion noise. Assume all displayed second moments are finite. Define the orbit second moment and orbit-variance matrix by

$$\begin{aligned} \mathbf{U}_{\mathrm{orb}}^{(n)} &= \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}^\nu, g}[\boldsymbol{\psi}_g(\mathbf{x}_t^\nu)\boldsymbol{\psi}_g(\mathbf{x}_t^\nu)^\top], \\ \mathbf{V}_G^{(n)} &= \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}^\nu, g}[(\boldsymbol{\psi}_g(\mathbf{x}_t^\nu) - \boldsymbol{\phi}(\mathbf{x}_t^\nu))(\boldsymbol{\psi}_g(\mathbf{x}_t^\nu) - \boldsymbol{\phi}(\mathbf{x}_t^\nu))^\top]. \end{aligned} \tag{B.18}$$

Both matrices act on $\mathbb{R}^p$. The Grams $\mathbf{U}_{\mathrm{base}}^{(n)}$ and $\mathbf{U}_{\mathrm{avg}}^{(n)}$ are defined in Eq. (11). To make the population notation explicit, draw $\mathbf{X} \sim \mu_0$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and $g$ uniformly from $G$, independently, set $\mathbf{X}_t = e^{-t}\mathbf{X} + \sqrt{\Delta_t}\boldsymbol{\xi}$, and define

$$\begin{aligned} \mathbf{U}_M^{\mathrm{pop}} &= \mathbb{E}_{\mathbf{X},\boldsymbol{\xi}}[\boldsymbol{\phi}_M(\mathbf{X}_t)\boldsymbol{\phi}_M(\mathbf{X}_t)^\top], && M \in \{\mathrm{base}, \mathrm{avg}\}, \\ \mathbf{U}_{\mathrm{orb}}^{\mathrm{pop}} &= \mathbb{E}_{\mathbf{X},\boldsymbol{\xi},g}[\boldsymbol{\psi}_g(\mathbf{X}_t)\boldsymbol{\psi}_g(\mathbf{X}_t)^\top], \\ \mathbf{V}_G^{\mathrm{pop}} &= \mathbb{E}_{\mathbf{X},\boldsymbol{\xi},g}[(\boldsymbol{\psi}_g(\mathbf{X}_t) - \boldsymbol{\phi}(\mathbf{X}_t))(\boldsymbol{\psi}_g(\mathbf{X}_t) - \boldsymbol{\phi}(\mathbf{X}_t))^\top]. \end{aligned} \tag{B.19}$$

Thus $\mathbf{U}_{\mathrm{avg}}^{\mathrm{pop}}$ and $\mathbf{U}_{\mathrm{base}}^{\mathrm{pop}}$ are the two cases of the first line, while $\mathbf{U}_{\mathrm{orb}}^{\mathrm{pop}}$ and $\mathbf{V}_G^{\mathrm{pop}}$ are the second and third lines. All four matrices act on $\mathbb{R}^p$ and keep $\mathbf{W}$ fixed. For $M \in \{\mathrm{base}, \mathrm{avg}\}$, write $\boldsymbol{\phi}_M^\nu = \boldsymbol{\phi}_M(\mathbf{x}_t^\nu)$. The target moments are $\mathbf{R}_{t,M}^{(n)} = n^{-1}\sum_\nu \mathbb{E}[\mathbf{y}_{t,G}^\nu(\boldsymbol{\phi}_M^\nu)^\top]$ and $\mathbf{T}_{t,M}^{(n)} = (\mathbf{R}_{t,M}^{(n)})^\top \mathbf{R}_{t,M}^{(n)}$, with the population versions obtained by replacing the design average by $\mathbf{X} \sim \mu_0$. These are the same moments as in Eqs. (11) and (C.14); the superscript pop only distinguishes them from finite-design moments, and no Gaussian clean-data assumption is needed here.

**Proposition B.15** (Orbit variance and population target contribution)**.** *For every fixed design and weight matrix with the preceding finite moments,*

$$\mathbf{U}_{\mathrm{orb}}^{(n)} = \mathbf{U}_{\mathrm{avg}}^{(n)} + \mathbf{V}_G^{(n)}, \qquad \mathbf{V}_G^{(n)} \succeq 0. \tag{B.20}$$

*If the clean population is $G$-invariant and the diffusion noise is isotropic, then, assuming the population second moments are finite,*

$$\mathbf{U}_{\mathrm{avg}}^{\mathrm{pop}} = \mathbf{U}_{\mathrm{base}}^{\mathrm{pop}} - \mathbf{V}_G^{\mathrm{pop}} \preceq \mathbf{U}_{\mathrm{base}}^{\mathrm{pop}}. \tag{B.21}$$

*For the common target $\mathbf{Y}_{t,G} = \mathbf{P}_G \mathbf{Y}_t$, these population assumptions also give*

$$\mathbf{R}_{t,\mathrm{avg}}^{\mathrm{pop}} = \mathbf{R}_{t,\mathrm{base}}^{\mathrm{pop}}, \qquad \mathbf{T}_{t,\mathrm{avg}}^{\mathrm{pop}} = \mathbf{T}_{t,\mathrm{base}}^{\mathrm{pop}}. \tag{B.22}$$

*Proof of Proposition B.15.* For a fixed noisy input, suppress its argument in $\boldsymbol{\phi}$ and $\boldsymbol{\psi}_g$. Since $\boldsymbol{\phi} = \mathbb{E}_g \boldsymbol{\psi}_g$, the covariance decomposition around this group mean is

$$\mathbb{E}_g[\boldsymbol{\psi}_g \boldsymbol{\psi}_g^\top] = \boldsymbol{\phi}\boldsymbol{\phi}^\top + \mathbb{E}_g[(\boldsymbol{\psi}_g - \boldsymbol{\phi})(\boldsymbol{\psi}_g - \boldsymbol{\phi})^\top].$$

Averaging over diffusion noise and samples proves Eq. (B.20).

At population level, $G$-invariance of the clean law and isotropy of the diffusion noise imply $g\mathbf{X}_t \overset{d}{=} \mathbf{X}_t$ for every $g$. Consequently, for each fixed $g$,

$$\mathbb{E}_{\mathbf{X}_t}[\boldsymbol{\psi}_g(\mathbf{X}_t)\boldsymbol{\psi}_g(\mathbf{X}_t)^\top] = \mathbb{E}_{\mathbf{X}_t}[\boldsymbol{\psi}_e(\mathbf{X}_t)\boldsymbol{\psi}_e(\mathbf{X}_t)^\top].$$

Averaging this equality over $g$ gives $\mathbf{U}_{\mathrm{orb}}^{\mathrm{pop}} = \mathbf{U}_{\mathrm{base}}^{\mathrm{pop}}$; combining it with Eq. (B.20) proves Eq. (B.21). The same reindexing is exact for an equally weighted orbit-complete empirical design.

For this common target, $\mathbf{P}_G g = \mathbf{P}_G$ and joint invariance of $(\mathbf{X}_t, \mathbf{Y}_t)$ give, for every $g \in G$,

$$\begin{aligned}\mathbb{E}[\mathbf{Y}_{t,G}\boldsymbol{\phi}_{\mathrm{base}}(g\mathbf{X}_t)^\top] &= \mathbb{E}[\mathbf{P}_G g \mathbf{Y}_t \boldsymbol{\phi}_{\mathrm{base}}(g\mathbf{X}_t)^\top] \\ &= \mathbb{E}[\mathbf{P}_G \mathbf{Y}_t \boldsymbol{\phi}_{\mathrm{base}}(\mathbf{X}_t)^\top] \\ &= \mathbf{R}_{t,\mathrm{base}}^{\mathrm{pop}}.\end{aligned}$$

Averaging over $g$ proves the first equality in Eq. (B.22); the second follows from $\mathbf{T}_{t,M}^{\mathrm{pop}} = (\mathbf{R}_{t,M}^{\mathrm{pop}})^\top \mathbf{R}_{t,M}^{\mathrm{pop}}$. All population statements condition on the same fixed weights.

□

For an arbitrary fixed design, Eq. (B.20) compares $\mathbf{U}_{\mathrm{avg}}^{(n)}$ with $\mathbf{U}_{\mathrm{orb}}^{(n)}$; $\mathbf{U}_{\mathrm{orb}}^{(n)} = \mathbf{U}_{\mathrm{base}}^{(n)}$ and equality of the target moments need not hold. Even at population level, the equal target matrices in Eq. (B.22) can give different initial update fractions on model-specific spectral projectors. The Gram contraction alone therefore does not determine the relative block delays. Section 5.2 supplies common active blocks and a strict comparison of their variance retention.

The orbit identity also isolates the fixed-feature content of data augmentation. We state the general definition first, so that the subsequent quadratic identity is not confused with an all-layer equivalence.

**Definition B.16** (Equivariance and Haar data augmentation)**.** For the orthogonal action used here, a vector predictor $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is $G$-equivariant if $f_\theta(g\mathbf{z}) = g f_\theta(\mathbf{z})$ for every $g \in G$. A scalar function $h : \mathbb{R}^d \to \mathbb{R}$ is $G$-invariant if $h(g\mathbf{z}) = h(\mathbf{z})$. For an input–target pair $(\mathbf{Z}, \mathbf{Y})$, loss $\ell$, and normalized Haar measure $\mu_G$, data augmentation replaces $\mathcal{R}(\theta) = \mathbb{E}[\ell(f_\theta(\mathbf{Z}), \mathbf{Y})]$ by

$$\mathcal{R}_{\mathrm{aug}}(\theta) = \int_G \mathbb{E}[\ell(f_\theta(g\mathbf{Z}), g\mathbf{Y})] \, \mathrm{d}\mu_G(g). \tag{B.23}$$

For finite $G$, the integral is the uniform average over group elements.

For Base-RFM with this target, the augmented regression risk is

$$\mathcal{L}_{t,G,\mathrm{aug}}^{(n)}(\mathbf{A}) = \Delta_t d \, \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}^\nu, g}\left[ \left\| p^{-1/2} \mathbf{A} \boldsymbol{\psi}_g(\mathbf{x}_t^\nu) - g\mathbf{y}_{t,G}^\nu \right\|^2 \right]. \tag{B.24}$$

Since $g\mathbf{y}_{t,G}^\nu = \mathbf{y}_{t,G}^\nu$ and $\mathbb{E}_g \boldsymbol{\psi}_g = \boldsymbol{\phi}$,

$$\mathcal{L}_{t,G,\mathrm{aug}}^{(n)}(\mathbf{A}) = \mathcal{L}_{t,G,\mathrm{avg}}^{(n)}(\mathbf{A}) + \frac{\Delta_t d}{p} \operatorname{Tr}\Big( \mathbf{A} \mathbf{V}_G^{(n)} \mathbf{A}^\top \Big). \tag{B.25}$$

Thus the Hessians with respect to $\mathbf{A}$ are governed by $\mathbf{U}_{\mathrm{orb}}^{(n)}$ and $\mathbf{U}_{\mathrm{avg}}^{(n)}$, whose positive semidefinite difference is $\mathbf{V}_G^{(n)}$. This is an identity for the frozen-feature regression considered here; it does not assert an equivalence for a network in which earlier layers are trained.

**Finite-sample orbit correction.** Now let the clean samples be i.i.d. from a $G$-invariant law, independently of the fixed weights, retain the isotropic Gaussian corruption, and assume $B_\sigma = \sup_z |\sigma(z)| < \infty$. Define the correction

$$\mathbf{E}_{\mathrm{orb},n} = \mathbf{U}_{\mathrm{orb}}^{(n)} - \mathbf{U}_{\mathrm{base}}^{(n)}.$$

For $\delta \in (0,1)$, Eq. (B.20) and a concentration bound give, with probability at least $1-\delta$ over the clean samples,

$$\begin{aligned}\mathbf{U}_{\mathrm{avg}}^{(n)} &= \mathbf{U}_{\mathrm{base}}^{(n)} - \mathbf{V}_G^{(n)} + \mathbf{E}_{\mathrm{orb},n}, \\ \|\mathbf{E}_{\mathrm{orb},n}\|_{\mathrm{op}} &\le \varepsilon_{\mathrm{orb}}(n,\delta),\end{aligned} \tag{B.26}$$

where

$$\varepsilon_{\mathrm{orb}}(n,\delta) = 2pB_\sigma^2 \sqrt{\frac{2\log(2p/\delta)}{n}} + \frac{4pB_\sigma^2 \log(2p/\delta)}{3n}. \tag{B.27}$$

To obtain the bound, write $\boldsymbol{\psi}_g^\nu = \boldsymbol{\psi}_g(\mathbf{x}_t^\nu)$ and set

$$\mathbf{Z}_\nu = \mathbb{E}_{\boldsymbol{\xi}^\nu, g}[\boldsymbol{\psi}_g^\nu(\boldsymbol{\psi}_g^\nu)^\top] - \mathbb{E}_{\boldsymbol{\xi}^\nu}[\boldsymbol{\psi}_e^\nu(\boldsymbol{\psi}_e^\nu)^\top].$$

Then $\mathbf{E}_{\mathrm{orb},n} = n^{-1}\sum_\nu \mathbf{Z}_\nu$. Population invariance gives $\mathbb{E}\mathbf{Z}_\nu = 0$, boundedness gives $\|\mathbf{Z}_\nu\|_{\mathrm{op}} \le 2pB_\sigma^2$, and the matrix-variance parameter is at most $4np^2B_\sigma^4$. Self-adjoint matrix Bernstein (Tropp, 2012; 2015) and division by $n$ give the stated radius.

**Reflection examples.** The two-element construction is the $(C_2, \mathrm{sign})$ instance of Theorem 5.1: $C_2 = \{e, s\}$ is the cyclic group of order two, and the sign representation is $s \mapsto -1$. The general cyclic construction and its proof are in Appendix C.5; the present calculation is a separate aligned negative control.

This control can contain a block with zero target contribution, for which the relative parameter error is undefined. To describe Gram-induced decay even on such a block, let $\mathbf{Q}$ be a nonzero orthogonal projector on feature space, with rank $r$. Define the uniform-direction spectral diagnostic and its crossing time in scaled time $\bar\tau = c_t\tau$ by

$$\begin{aligned} \widetilde{E}_{\mathbf{Q},M}(\bar\tau) &= \frac{1}{r}\operatorname{Tr}\big(\mathbf{Q}e^{-2\mathbf{U}_M\bar\tau}\big), \\ \widetilde{\tau}_{\mathbf{Q},M}(\rho) &= \inf\{\bar\tau \ge 0 : \widetilde{E}_{\mathbf{Q},M}(\bar\tau) \le \rho^2\}, \qquad 0 < \rho < 1. \end{aligned} \tag{B.28}$$

The trace averages squared decay over uniformly distributed unit directions in $\mathrm{Ran}(\mathbf{Q})$; it does not weight directions by the initial parameter error or target contribution. The definition also permits projectors that do not commute with $\mathbf{U}_M$. The crossing time is $+\infty$ if the threshold is never reached. For common reference projectors $\mathbf{Q}_S, \mathbf{Q}_N$ and positive finite crossing times, write $\widetilde{\Delta}_B^{\mathrm{ref}} = \log(\widetilde{\tau}_{\mathbf{Q}_B,\mathrm{avg}}/\widetilde{\tau}_{\mathbf{Q}_B,\mathrm{base}})$ and $\widetilde{\Delta}_{\mathrm{win}}^{\mathrm{ref}} = \widetilde{\Delta}_N^{\mathrm{ref}} - \widetilde{\Delta}_S^{\mathrm{ref}}$ at a common $\rho$. These are spectral diagnostics, not parameter-convergence claims.

For this control, let $\mathcal{S} \subseteq \mathbb{R}^d$ be the fixed output subspace, let $\mathcal{N} = \mathcal{S}^\perp$, and let $\mathbf{P}_\mathcal{S}$ and $\mathbf{P}_\mathcal{N}$ be their orthogonal projectors. Set

$$G = \{\mathbf{I}_d, \mathbf{J}\}, \qquad \mathbf{J} = \mathbf{P}_\mathcal{S} - \mathbf{P}_\mathcal{N}, \qquad \mathrm{Fix}(G) = \mathcal{S}.$$

Here $\mathbf{P}_\mathcal{S} + \mathbf{P}_\mathcal{N} = \mathbf{I}_d$, so $\mathbf{J}$ is an orthogonal involution: it fixes $\mathcal{S}$ and changes the sign on $\mathcal{N}$. Choose orthonormal feature directions $\{\mathbf{u}_i : i \in \mathcal{I}_S\}$ in $\mathcal{S}$ and $\{\mathbf{v}_j : j \in \mathcal{I}_N\}$ in $\mathcal{N}$; take the corresponding rows of $\mathbf{W}$ to be $\mathbf{w}_i = \sqrt{d}\,\mathbf{u}_i$ for $i \in \mathcal{I}_S$ and $\mathbf{w}_j = \sqrt{d}\,\mathbf{v}_j$ for $j \in \mathcal{I}_N$. The corresponding feature-coordinate projectors are $\mathbf{\Pi}_S, \mathbf{\Pi}_N \in \mathbb{R}^{p\times p}$, where $\mathcal{I}_S \dot\cup \mathcal{I}_N = \{1, \dots, p\}$ and $\mathbf{\Pi}_S + \mathbf{\Pi}_N = \mathbf{I}_p$. Thus $p = |\mathcal{I}_S| + |\mathcal{I}_N|$. The rows of $\mathbf{W}$ are supported on these two frames, so the $S$ and $N$ labels refer to feature coordinates, whereas $\mathcal{S}$ and $\mathcal{N}$ refer to output/input subspaces. Use the projected Gaussian target $\mathbf{Y}_{t,S} = -\mathbf{P}_\mathcal{S}\boldsymbol{\xi}/\sqrt{\Delta_t}$. For $Z \sim \mathcal{N}(0,1)$ define

$$\begin{aligned} \sigma_{\mathrm{e}}(z) &= \frac{\sigma(z)+\sigma(-z)}{2}, \quad \sigma_{\mathrm{o}}(z) = \frac{\sigma(z)-\sigma(-z)}{2}, \\ v_{\mathrm{e}} &= \mathbb{E}[\sigma_{\mathrm{e}}(Z)^2], \qquad v_{\mathrm{o}} = \mathbb{E}[\sigma_{\mathrm{o}}(Z)^2], \\ a_1 &= \mathbb{E}[Z\sigma(Z)]. \end{aligned}$$

In this paragraph assume $\mathbb{E}[\sigma(Z)] = 0$ and $v_{\mathrm{e}}, v_{\mathrm{o}} > 0$.

**Proposition B.17** (Aligned reflection gives a spectral proxy only)**.** *At population level, with the Gaussian clean law $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and rows supported on the two orthonormal frames above, the aligned construction has*

$$\begin{aligned} \mathbf{U}_{\mathrm{avg}}^{\mathrm{pop}} &= (v_{\mathrm{e}} + v_{\mathrm{o}})\mathbf{\Pi}_S + v_{\mathrm{e}}\mathbf{\Pi}_N, \\ \mathbf{U}_{\mathrm{base}}^{\mathrm{pop}} &= (v_{\mathrm{e}} + v_{\mathrm{o}})\mathbf{I}_p. \end{aligned}$$

*The target-contribution matrix for $\mathbf{Y}_{t,S}$ is supported on $\mathbf{\Pi}_S$, so the $N$ block has zero target contribution and its parameter-convergence time is undefined. The uniform-direction trace proxy in Eq.* (B.28) *remains well defined for both coordinate projectors, which is why this example is only a spectral proxy.*

Write $\lambda_S = v_{\mathrm{e}} + v_{\mathrm{o}}$ and $\lambda_N = v_{\mathrm{e}}$ for the averaged population block rates. For the fixed coordinate blocks, use $\widetilde{\tau}_{B,M}^{\mathrm{block}} = \widetilde{\tau}_{\mathbf{\Pi}_B,M}$, $B \in \{S, N\}$, from Eq. (B.28), with log delays $\widetilde{\Delta}_B^{\mathrm{block}} = \log(\widetilde{\tau}_{B,\mathrm{avg}}^{\mathrm{block}}/\widetilde{\tau}_{B,\mathrm{base}}^{\mathrm{block}})$ and $\widetilde{\Delta}_{\mathrm{win}}^{\mathrm{block}} = \widetilde{\Delta}_N^{\mathrm{block}} - \widetilde{\Delta}_S^{\mathrm{block}}$, all at the same $\rho$.

**Proposition B.18** (Finite-design control for the aligned reflection proxy)**.** *Assume additionally that $\sigma$ is bounded, with $B_\sigma = \sup_z |\sigma(z)| < \infty$, and that the two population variances above satisfy $v_{\mathrm{e}}, v_{\mathrm{o}} > 0$. For every fixed clean design, the diffusion-noise-averaged Base-RFM and Avg-RFM cross-moments coincide, and their $N$-feature columns vanish for the target $\mathbf{Y}_{t,S}$. Thus $\mathbf{T}^{(n)}_{t,\mathrm{avg}} = \mathbf{T}^{(n)}_{t,\mathrm{base}} =: \mathbf{T}^{(n)}_t$. Whenever this matrix is nonzero, the $S$ and $N$ coordinate blocks account for all and none of the squared initial update norm, respectively.*

*The population gap between the two Avg-RFM blocks is $\gamma_{SN} := \lambda_S - \lambda_N = v_{\mathrm{o}}$. Fix $\delta, \rho \in (0,1)$, let $\eta = \eta_n(\delta)$ be the radius in Eq.* (B.32)*, set $\zeta_\eta = (2\eta/\gamma_{SN})^2$, and let $s_{\eta,\rho}, d_{\eta,\rho}$ be as in Eq.* (B.33)*. For an i.i.d. clean-design draw, the joint Gram-deviation event has probability at least $1-\delta$. On this event, if*

$$\eta < \min\{\gamma_{SN}/2, \lambda_N\}, \qquad \zeta_\eta < \min\{\rho^2, 1-\rho^2\}, \qquad \mathbf{T}^{(n)}_t \neq 0,$$

*then the two Avg-RFM empirical clusters are isolated and account for at least $1-\zeta_\eta$ and at most $\zeta_\eta$ of the squared initial update norm, respectively, and*

$$\begin{aligned} \left|\widetilde{\Delta}^{\mathrm{block}}_S(\rho)\right| &\le s_{\eta,\rho}, \\ \widetilde{\Delta}^{\mathrm{block}}_N(\rho) &\ge d_{\eta,\rho}, \\ \widetilde{\Delta}^{\mathrm{block}}_{\mathrm{win}}(\rho) &\ge d_{\eta,\rho} - s_{\eta,\rho}. \end{aligned}$$

*Moreover, $s_{\eta,\rho} \to 0$ and $d_{\eta,\rho} \to \log(\lambda_S/\lambda_N) > 0$ as $\eta \to 0$. Thus the bound shows a strict finite-design gap whenever $d_{\eta,\rho} > s_{\eta,\rho}$.*

*Proof of Proposition B.17 and Proposition B.18.* Write the standard Gaussian noised coordinates as $Z_i = \mathbf{u}_i^\top \mathbf{X}_t$ and $H_j = \mathbf{v}_j^\top \mathbf{X}_t$. The orthonormality of the two frames and $\mathbf{X}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ make all $Z_i, H_j$ independent standard Gaussians. Signal features are fixed by $G$ and equal $\sigma(Z_i)$. On $\mathcal{N}$, the reflection sends $H_j$ to $-H_j$, so the averaged $N$-block feature is

$$\frac{\sigma(H_j) + \sigma(-H_j)}{2} = \sigma_{\mathrm{e}}(H_j).$$

The symmetry of the standard Gaussian gives

$$\mathbb{E}[\sigma_{\mathrm{e}}(Z)] = \tfrac{1}{2}\{\mathbb{E}[\sigma(Z)] + \mathbb{E}[\sigma(-Z)]\} = \mathbb{E}[\sigma(Z)] = 0.$$

Thus the averaged $N$ features are centered. Independence of distinct frame coordinates then makes all population off-diagonal Gram entries zero. The even and odd parts are orthogonal under the symmetric Gaussian law, so $\mathbb{E}[\sigma(Z)^2] = v_{\mathrm{e}} + v_{\mathrm{o}} =: \lambda_S$. The N-block diagonal variance after averaging is $\lambda_N := v_{\mathrm{e}}$. Without averaging, every coordinate uses $\sigma$ and has variance $\lambda_S$. Thus

$$\begin{aligned} \mathbf{U}^{\mathrm{pop}}_{\mathrm{avg}} &= \lambda_S \mathbf{\Pi}_S + \lambda_N \mathbf{\Pi}_N, \\ \mathbf{U}^{\mathrm{pop}}_{\mathrm{base}} &= \lambda_S \mathbf{I}_p. \end{aligned} \tag{B.29}$$

Conditional Gaussian regression gives $\mathbb{E}[\boldsymbol{\xi} \mid \mathbf{X}_t] = \sqrt{\Delta_t}\, \mathbf{X}_t$. Therefore the $i$th signal column of the cross-moment is

$$\begin{aligned} \mathbb{E}\left[-\frac{\mathbf{P}_{\mathcal{S}}\boldsymbol{\xi}}{\sqrt{\Delta_t}}\sigma(Z_i)\right] &= -\mathbb{E}[\mathbf{P}_{\mathcal{S}}\mathbf{X}_t\sigma(Z_i)] \\ &= -a_1 \mathbf{u}_i, \end{aligned}$$

whereas every $N$-block column is zero: $\mathbf{P}_{\mathcal{S}}\mathbf{X}_t$ is independent of $H_j$ and $\mathbb{E}[\sigma_{\mathrm{e}}(H_j)] = 0$. Orthonormality gives

$$\mathbf{T}^{\mathrm{pop}}_{t,\mathrm{avg}} = \mathbf{T}^{\mathrm{pop}}_{t,\mathrm{base}} = a_1^2 \mathbf{\Pi}_S. \tag{B.30}$$

Both population Grams commute with the fixed feature blocks. Therefore their block spectral-decay times are $\log(1/\rho)/\lambda_B$ for Avg-RFM and $\log(1/\rho)/\lambda_S$ for Base-RFM. Taking the two model ratios gives the population log-delays stated in the theorem.

Signal features are identical in the two models, and every N-block column of either cross-moment vanishes. Thus, after integrating out the diffusion noise, the same support statement holds for every

clean design:

$$\begin{aligned} \mathbf{R}_{t,\mathrm{avg}}^{(n)} &= \mathbf{R}_{t,\mathrm{base}}^{(n)} =: \mathbf{R}_t^{(n)}, \\ \mathbf{R}_t^{(n)}\mathbf{\Pi}_N &= 0, \\ \mathbf{T}_{t,\mathrm{avg}}^{(n)} &= \mathbf{T}_{t,\mathrm{base}}^{(n)} =: \mathbf{T}_t^{(n)} = \mathbf{\Pi}_S\mathbf{T}_t^{(n)}\mathbf{\Pi}_S. \end{aligned} \tag{B.31}$$

An N-block feature depends only on the N-block clean coordinate and the N-block noise coordinates, which are independent of $\mathbf{P}_S\boldsymbol{\xi}$; its cross-moment is therefore zero. For a signal feature, the S-frame noise coordinates are independent except for the matching coordinate, so all output coordinates other than the matching $\mathbf{u}_i$ coordinate have zero cross-moment. Thus every signal column is a scalar multiple of $\mathbf{u}_i$.

*Finite-sample bounds.* Let $B_\sigma = \sup_z |\sigma(z)|$, fix $\delta \in (0,1)$, and define

$$\eta = \eta_n(\delta) = B_\sigma\sqrt{\frac{2p\lambda_S\log(4p/\delta)}{n}} + \frac{2pB_\sigma^2\log(4p/\delta)}{3n}. \tag{B.32}$$

For a common threshold $\rho \in (0,1)$, set

$$\begin{aligned} \zeta_\eta &= \left(\frac{2\eta}{\gamma_{SN}}\right)^2, \qquad h_0 = \log\frac{1}{\rho^2}, \\ h_- &= \log\frac{1}{\rho^2+\zeta_\eta}, \qquad h_+ = \log\frac{1}{\rho^2-\zeta_\eta}, \\ s_{\eta,\rho} &= \log\frac{\lambda_S+\eta}{\lambda_S-\eta} + \max\left\{\log\frac{h_+}{h_0}, \log\frac{h_0}{h_-}\right\}, \\ d_{\eta,\rho} &= \log\frac{\lambda_S-\eta}{\lambda_N+\eta} + \log\frac{h_-}{h_0}. \end{aligned} \tag{B.33}$$

For concentration, define the conditional noise-averaged contribution of clean sample $\nu$ by

$$\mathbf{M}_\nu^M = \mathbb{E}_{\boldsymbol{\xi}^\nu}[\boldsymbol{\phi}_M(\mathbf{x}_t^\nu)\boldsymbol{\phi}_M(\mathbf{x}_t^\nu)^\top], \qquad M \in \{\mathrm{avg}, \mathrm{base}\},$$

so that $\mathbf{U}_M^{(n)} = n^{-1}\sum_\nu \mathbf{M}_\nu^M$. Each $\mathbf{M}_\nu^M$ is positive semidefinite with $\|\mathbf{M}_\nu^M\|_{\mathrm{op}} \le L \overset{\mathrm{def}}{=} pB_\sigma^2$, and both population Grams have operator norm $\lambda_S$. If $\mathbf{Y}_\nu^M = \mathbf{M}_\nu^M - \mathbb{E}\mathbf{M}_\nu^M$, then $-L\mathbf{I} \preceq \mathbf{Y}_\nu^M \preceq L\mathbf{I}$ and

$$\mathbb{E}[(\mathbf{Y}_\nu^M)^2] = \mathbb{E}[(\mathbf{M}_\nu^M)^2] - (\mathbb{E}\mathbf{M}_\nu^M)^2 \preceq \mathbb{E}[(\mathbf{M}_\nu^M)^2] \preceq L\,\mathbb{E}\mathbf{M}_\nu^M.$$

The variance parameter is therefore at most $npB_\sigma^2\lambda_S$. Matrix Bernstein applied to the two models, with a union bound, gives Eq. (B.34).

$$\begin{aligned} \left\|\mathbf{U}_{\mathrm{avg}} - \mathbf{U}_{\mathrm{avg}}^{\mathrm{pop}}\right\|_{\mathrm{op}} &\le \eta_n(\delta), \\ \|\mathbf{U}_{\mathrm{base}} - \mathbf{U}_{\mathrm{base}}^{\mathrm{pop}}\|_{\mathrm{op}} &\le \eta_n(\delta). \end{aligned} \tag{B.34}$$

When $\eta < \gamma_{SN}/2$, Weyl's inequality (Bhatia, 1997) separates the two averaged clusters, and Davis–Kahan (Davis & Kahan, 1970; Bhatia, 1997; Yu et al., 2015) gives

$$\left\|\widehat{\mathbf{\Pi}}_S - \mathbf{\Pi}_S\right\|_{\mathrm{op}} \le \frac{2\eta}{\gamma_{SN}}. \tag{B.35}$$

For equal-rank orthogonal projectors,

$$\left\|\mathbf{\Pi}_S(\mathbf{I} - \widehat{\mathbf{\Pi}}_S)\mathbf{\Pi}_S\right\|_{\mathrm{op}} = \left\|\widehat{\mathbf{\Pi}}_S - \mathbf{\Pi}_S\right\|_{\mathrm{op}}^2.$$

Since every nonzero empirical target-contribution matrix is positive semidefinite and supported on $\mathbf{\Pi}_S$, this proves

$$\begin{aligned} \ell_{\mathbf{T}_t^{(n)}}(\widehat{\mathbf{\Pi}}_S) &\ge 1 - \left(\frac{2\eta}{\gamma_{SN}}\right)^2, \\ \ell_{\mathbf{T}_t^{(n)}}(\widehat{\mathbf{\Pi}}_N) &\le \left(\frac{2\eta}{\gamma_{SN}}\right)^2, \\ \mathrm{prec}(\widehat{\mathbf{\Pi}}_S \mid \mathbf{\Pi}_S) = \mathrm{rec}(\widehat{\mathbf{\Pi}}_S \mid \mathbf{\Pi}_S) &\ge 1 - \left(\frac{2\eta}{\gamma_{SN}}\right)^2. \end{aligned} \tag{B.36}$$

The averaged S- and N-block eigenvalues lie in $[\lambda_S - \eta, \lambda_S + \eta]$ and $[\lambda_N - \eta, \lambda_N + \eta]$. Let $\widehat{\mathbf{\Pi}}_S$ be the empirical spectral projector onto the cluster near $\lambda_S$, and set $\widehat{\mathbf{\Pi}}_N = \mathbf{I}_p - \widehat{\mathbf{\Pi}}_S$. Eq. (B.35) and the complementary projector identity give, for both blocks,

$$\left\|\widehat{\mathbf{\Pi}}_B - \mathbf{\Pi}_B\right\|_{\mathrm{op}} \le \delta_\eta := \frac{2\eta}{\gamma_{SN}}.$$

The fixed coordinate block $\mathbf{\Pi}_B$ need not commute with the empirical averaged Gram. Applying Theorem D.10 with $(\mathbf{P}_B, \mathbf{Q}_B) = (\widehat{\mathbf{\Pi}}_B, \mathbf{\Pi}_B)$ therefore gives, whenever $\eta < \min\{\gamma_{SN}/2, \lambda_N\}$ and $\zeta_\eta = \delta_\eta^2 < \min\{\rho^2, 1-\rho^2\}$,

$$\frac{h_-}{2(\lambda_B + \eta)} \le \widetilde{\tau}^{\mathrm{block}}_{B,\mathrm{avg}}(\rho) \le \frac{h_+}{2(\lambda_B - \eta)}, \qquad B \in \{S, N\}, \tag{B.37}$$

where $h_-, h_+$ are defined in Eq. (B.33). For Base-RFM, the global operator bound implies $(\lambda_S - \eta)\mathbf{I} \preceq \mathbf{U}_{\mathrm{base}} \preceq (\lambda_S + \eta)\mathbf{I}$. Functional calculus therefore gives

$$\frac{h_0}{2(\lambda_S + \eta)} \le \widetilde{\tau}^{\mathrm{block}}_{B,\mathrm{base}}(\rho) \le \frac{h_0}{2(\lambda_S - \eta)}, \qquad B \in \{S, N\}. \tag{B.38}$$

Combining the upper bound on the Avg-RFM signal time with the lower bound on the Base-RFM signal time gives

$$\widetilde{\Delta}^{\mathrm{block}}_S(\rho) \le \log\frac{\lambda_S + \eta}{\lambda_S - \eta} + \log\frac{h_+}{h_0},$$

Combining the upper Base-RFM signal bound with the lower Avg-RFM signal bound gives the same inequality for $-\widetilde{\Delta}^{\mathrm{block}}_S$, with $\log(h_0/h_-)$ in the second term. Combining the lower Avg-RFM N-block bound with the upper Base-RFM N-block bound gives

$$\widetilde{\Delta}^{\mathrm{block}}_N(\rho) \ge \log\frac{\lambda_S - \eta}{\lambda_N + \eta} + \log\frac{h_-}{h_0}.$$

These expressions are $s_{\eta,\rho}$ and $d_{\eta,\rho}$ in Eq. (B.33). Subtracting the signal bound from the N-block bound proves the finite-design regime statement. □

**Lemma B.19** (Activation used in the reflection calculation)**.** *Let $Z \sim \mathcal{N}(0,1)$, set $c_2 = \mathbb{E}[\tanh^2 Z]$, and define $\sigma_\beta(z) = \tanh z + \beta(\tanh^2 z - c_2)$. For every $\beta \neq 0$, this activation (used in the analytic-check protocol of Appendix F.3) is bounded and centered under the standard Gaussian law, and it satisfies $v_{\mathrm{e}}, v_{\mathrm{o}} > 0$ and $a_1 := \mathbb{E}[Z\sigma_\beta(Z)] \neq 0$, as required by the reflection calculations above.*

*Proof.* Boundedness is immediate. For $Z \sim \mathcal{N}(0,1)$, oddness gives $\mathbb{E}[\tanh Z] = 0$, while the choice $c_2 = \mathbb{E}[\tanh^2 Z]$ centers the even term. Parity gives

$$\sigma_{\mathrm{o}}(z) = \tanh z, \qquad \sigma_{\mathrm{e}}(z) = \beta(\tanh^2 z - c_2),$$

and hence

$$v_{\mathrm{o}} = \mathbb{E}[\tanh^2 Z] > 0, \qquad v_{\mathrm{e}} = \beta^2 \operatorname{Var}(\tanh^2 Z) > 0$$

for every $\beta \neq 0$, including the positive values in the experiment. The even term has no first-order Hermite component, while Gaussian integration by parts gives $a_1 = \mathbb{E}[Z \tanh Z] = \mathbb{E}[\operatorname{sech}^2 Z] > 0$. □

**Proposition B.20** (Projected signal-density identity)**.** *Let $d_{\mathrm{proj},d} = \operatorname{rank}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)$ and use the quantities in Eq.* (B.8)*. If $\operatorname{Tr}(\mathbf{P}_G) > 0$, then*

$$\frac{\operatorname{Tr}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)}{\operatorname{Tr}(\mathbf{P}_G)} = \frac{\kappa_{\mathrm{sig}}}{\gamma_G}\sigma_{\mathbf{x}}^2.$$

*If the covariance is degenerate on part of $\operatorname{Fix}(G)$, the quantity averaged over the active projected subspace is instead $\operatorname{Tr}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)/d_{\mathrm{proj},d}$; the two denominators coincide only when the covariance is nondegenerate on all of $\operatorname{Fix}(G)$.*

*Proof.* Because $\mathbf{P}_G$ is an orthogonal projector, $\operatorname{Tr}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)$ is the trace of the covariance restricted to $\operatorname{Fix}(G)$. Substituting $\operatorname{Tr}(\mathbf{P}_G) = d\gamma_G$, $\operatorname{Tr}(\mathbf{\Sigma}_d) = d\sigma_{\mathbf{x}}^2$, and the definition of $\kappa_{\mathrm{sig}}$ in Eq. (B.8) gives the first identity. Restricting the trace to its nonzero eigenspace gives the second statement. □

**Proposition B.21** (Aligned block-regular orbit correlations)**.** *Let $G = C_m$ act through $q$ copies of the regular representation, so $d = qm$, $\operatorname{rank}(\mathbf{P}_G) = q$, and $\gamma_G = q/d = 1/m$. For*

$$\mathbf{\Sigma}_d = \lambda_{\text{inv}}\mathbf{P}_G + \lambda_\perp(\mathbf{I} - \mathbf{P}_G),$$

*assume $\lambda_{\text{inv}}, \lambda_\perp \geq 0$, $\operatorname{Tr}(\mathbf{\Sigma}_d) > 0$, and $v_t^2 > 0$. Define $\rho_{g,g'} = \operatorname{Tr}(\mathbf{\Sigma}_d g^{-1}g')/\operatorname{Tr}(\mathbf{\Sigma}_d)$ and $\kappa_{\text{noise},t} = m^{-2}\sum_{g,g'\in G}\Psi_t(\rho_{g,g'})$. Then $\rho_{e,e} = 1$ and, for every $h \neq e$,*

$$\rho_{e,h} = \frac{\lambda_{\text{inv}} - \lambda_\perp}{\lambda_{\text{inv}} + (m-1)\lambda_\perp}.$$

*Consequently*

$$\kappa_{\text{noise},t} = \frac{1}{m}\left[1 + (m-1)\Psi_t\left(\frac{\lambda_{\text{inv}} - \lambda_\perp}{\lambda_{\text{inv}} + (m-1)\lambda_\perp}\right)\right].$$

*Proof.* For nonidentity $h$, the regular representation has $\operatorname{Tr}(h) = 0$, while $\mathbf{P}_G h = \mathbf{P}_G$ and $\operatorname{Tr}(\mathbf{P}_G) = q$. Hence $\operatorname{Tr}(\mathbf{\Sigma}_d h) = q(\lambda_{\text{inv}} - \lambda_\perp)$ and $\operatorname{Tr}(\mathbf{\Sigma}_d) = q[\lambda_{\text{inv}} + (m-1)\lambda_\perp]$, which gives the displayed value of $\rho_{e,h}$. In the double sum, each relative element $h = g^{-1}g'$ occurs exactly $m$ times. One identity element and $m-1$ nonidentity elements therefore give the stated formula for $\kappa_{\text{noise},t}$. □

**Corollary B.22** (Residual factor in the isotropic case)**.** *If $\lambda_{\text{inv}} = \lambda_\perp$ and $v_t^2 > 0$, then $\kappa_{\text{noise},t} = 1/m = \gamma_G$ exactly. More generally, suppose $m \to \infty$, $\lambda_\perp > 0$, and $\sup_m \lambda_{\text{inv}}/\lambda_\perp < \infty$. Then $\kappa_{\text{noise},t} - \gamma_G \to 0$.*

*Proof.* In the isotropic case the off-identity correlation is zero. Under the bounded variance ratio, that correlation is $O(m^{-1})$. Hence Eq. (B.14) makes its contribution to $\kappa_{\text{noise},t} - 1/m$ of order $m^{-2}$. □

# C Finite-Group Proofs and Approximation Bounds

## C.1 Exact low-order Hermite terms

Let

$$\mathbf{G}_k = \frac{1}{n}\sum_{\nu=1}^{n}\mathbf{U}_k^\nu, \qquad (\mathbf{H}_0)_{i\nu} = \frac{1}{|G|}\sum_{g\in G}\mu_{0,q_i}(S_{i,g}^\nu),$$

and define the conditional population contribution of all higher Hermite orders and the conditional approximation by

$$\mathbf{M}_{\geq 2}(\mathbf{W}) = \mathbb{E}\left[\sum_{k\geq 2}\mathbf{U}_k^1 \,\middle|\, \mathbf{W}\right], \qquad \mathbf{U}_{\text{cond}} = \mathbf{G}_0 + \mathbf{G}_1 + \mathbf{M}_{\geq 2}(\mathbf{W}). \tag{C.1}$$

The expectation in $\mathbf{M}_{\geq 2}(\mathbf{W})$ is over a fresh clean sample. The diffusion noise has already been integrated out in each Hermite-order block.

**Lemma C.1** (Exact Hermite orders 0 and 1)**.** *Let $\mathbf{Q} = \operatorname{diag}(q_1, \ldots, q_p)$. Then*

$$\mathbf{G}_0 = \frac{1}{n}\mathbf{H}_0\mathbf{H}_0^\top,$$

*and*

$$\mathbf{G}_1 = \frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{|G|^2}\sum_{g,g'\in G}\mathbf{D}_{1,g}^\nu\mathbf{C}_{g,g'}\mathbf{D}_{1,g'}^\nu, \qquad \mathbf{D}_{1,g}^\nu = \operatorname{diag}\big(\mu_{1,q_i}(S_{i,g}^\nu)\big). \tag{C.2}$$

*Moreover,*

$$\frac{1}{|G|^2}\sum_{g,g'}\mathbf{C}_{g,g'} = \mathbf{Q}^{-1/2}\frac{\mathbf{W}\mathbf{P}_G\mathbf{W}^\top}{d}\mathbf{Q}^{-1/2}. \tag{C.3}$$

*Proof.* The first two identities are the $k = 0$ and $k = 1$ cases of Lemma B.1. For Eq. (C.3), average $g(g')^{-1}$ over $G^2$ and use that every relative group element occurs equally often. □

For arbitrary rowwise constants $m_i, a_i$, let $\mathbf{m} = (m_1, \dots, m_p)^\top$ and let $\mathbf{1}_n$ be the all-ones vector. Define $\widetilde{r}_{0,i}(s) = \mu_{0,q_i}(s) - m_i - a_i s$ and write $\mathbf{X}_{\mathrm{des}} = [\mathbf{x}^1, \dots, \mathbf{x}^n] \in \mathbb{R}^{d\times n}$. Then

$$\mathbf{H}_0 = \mathbf{m}\mathbf{1}_n^\top + e^{-t}\mathbf{A}_{\mathrm{lin}}\frac{\mathbf{W}\mathbf{P}_G\mathbf{X}_{\mathrm{des}}}{\sqrt{d}} + \mathbf{R}^{(0)}, \tag{C.4}$$

where $\mathbf{A}_{\mathrm{lin}} = \mathrm{diag}(a_i)$ and $(\mathbf{R}^{(0)})_{i\nu} = |G|^{-1}\sum_g \widetilde{r}_{0,i}(S^\nu_{i,g})$. Thus the projected clean linear design enters through the zeroth-order Hermite block. All mean–linear, mean–residual, and linear–residual cross terms remain in $\mathbf{H}_0\mathbf{H}_0^\top/n$.

### C.2 Conditional approximation error

**Theorem C.2** (Finite-group conditional approximation)**.** *Under Assumptions B.4–B.5, for any $K_{\mathrm{H}} \geq 2$, with probability at least $1 - \eta_d - \delta$,*

$$\|\mathbf{U}_{\mathrm{avg}} - \mathbf{U}_{\mathrm{cond}}\|_{\mathrm{op}} \leq \varepsilon_{\mathrm{br}}(n, K_{\mathrm{H}}, \delta),$$

*with the explicit bound $\varepsilon_{\mathrm{br}}(n, K_{\mathrm{H}}, \delta)$ given in* (C.5) *below. The conditional approximation $\mathbf{U}_{\mathrm{cond}}$ replaces only centered sample fluctuations in the higher diffusion-noise Hermite terms. It does not average over the fixed clean-design or first-noise-response channels; Corollary B.9 gives a checkable Gaussian-row regime in which $\varepsilon_{\mathrm{br}} \to 0$ without assuming Gaussian clean data.*

*Proof of Theorem C.2.* Fix $K_{\mathrm{H}} \geq 2$ and condition on a weight realization in $\mathcal{E}_{W,d}$. The matrices $\mathbf{U}^\nu_{2:K_{\mathrm{H}}} = \sum_{k=2}^{K_{\mathrm{H}}}\mathbf{U}^\nu_k$ are independent across $\nu$. Let

$$\mathbf{Y}_{K_{\mathrm{H}},\nu} = \mathbf{U}^\nu_{2:K_{\mathrm{H}}} - \mathbb{E}[\mathbf{U}^\nu_{2:K_{\mathrm{H}}} \mid \mathbf{W}].$$

Proposition B.7 gives $\|\mathbf{Y}_{K_{\mathrm{H}},\nu}\|_{\mathrm{op}} \leq 2L_d$ and

$$\big\|\mathbb{E}[\mathbf{Y}^2_{K_{\mathrm{H}},\nu} \mid \mathbf{W}]\big\|_{\mathrm{op}} \leq \mathbb{E}[\|\mathbf{Y}_{K_{\mathrm{H}},\nu}\|^2_{\mathrm{op}} \mid \mathbf{W}] \leq 4L_d^2.$$

The self-adjoint matrix Bernstein inequality (Tropp, 2012; 2015) therefore yields, conditionally with probability at least $1 - \delta$,

$$\left\|\frac{1}{n}\sum_{\nu=1}^n \mathbf{Y}_{K_{\mathrm{H}},\nu}\right\|_{\mathrm{op}} \leq 2L_d\sqrt{\frac{2\log(2p/\delta)}{n}} + \frac{4L_d\log(2p/\delta)}{3n}.$$

The finite-sample and conditional-mean tails satisfy

$$\left\|\frac{1}{n}\sum_{\nu=1}^n \mathbf{U}^\nu_{>K_{\mathrm{H}}}\right\|_{\mathrm{op}} + \big\|\mathbb{E}[\mathbf{U}^1_{>K_{\mathrm{H}}} \mid \mathbf{W}]\big\|_{\mathrm{op}} \leq 2T_{K_{\mathrm{H}},d}.$$

Define the conditional-approximation error bound

$$\varepsilon_{\mathrm{br}}(n, K_{\mathrm{H}}, \delta) \stackrel{\mathrm{def}}{=} 2L_d\sqrt{\frac{2\log(2p/\delta)}{n}} + \frac{4L_d\log(2p/\delta)}{3n} + 2T_{K_{\mathrm{H}},d}. \tag{C.5}$$

Combining these bounds and adding the failure probability $\eta_d$ proves Theorem C.2. Corollary B.9 separately gives a concrete asymptotic regime in which this bound vanishes. □

### C.3 Approximation checks

**Proposition C.3** (First-order Hermite approximation error)**.** *Let $\mathbf{b} = \mathrm{diag}(b_1, \dots, b_p)$ and*

$$\mathbf{B}_{\mathrm{row}} = \mathbf{b}\left(\mathbf{Q}^{-1/2}\frac{\mathbf{W}\mathbf{P}_G\mathbf{W}^\top}{d}\mathbf{Q}^{-1/2}\right)\mathbf{b}.$$

*Suppose*

$$\delta_{1,d} = \max_{i,\nu,g} |\mu_{1,q_i}(S^\nu_{i,g}) - b_i|,$$
$$b_* = \max_i |b_i|,$$
$$c_{1,d} = \max_{g,g'} \|\mathbf{C}_{g,g'}\|_{\mathrm{op}}.$$

*The resulting operator error satisfies*

$$\|\mathbf{G}_1 - \mathbf{B}_{\mathrm{row}}\|_{\mathrm{op}} \le c_{1,d}(2b_*\delta_{1,d} + \delta^2_{1,d}). \tag{C.6}$$

*Proof.* For each $\nu, g$, write

$$\mathbf{D}^\nu_{1,g} = \mathbf{b} + \mathbf{E}^\nu_g, \qquad \|\mathbf{E}^\nu_g\|_{\mathrm{op}} \le \delta_{1,d}.$$

Using Eqs. (C.2) and (C.3), and expanding around $\mathbf{b}$, gives

$$\mathbf{G}_1 - \mathbf{B}_{\mathrm{row}} = \frac{1}{n|G|^2} \sum_{\nu=1}^{n} \sum_{g,g' \in G} \Big( \mathbf{E}^\nu_g \mathbf{C}_{g,g'} \mathbf{b} + \mathbf{b} \mathbf{C}_{g,g'} \mathbf{E}^\nu_{g'} + \mathbf{E}^\nu_g \mathbf{C}_{g,g'} \mathbf{E}^\nu_{g'} \Big).$$

By the triangle inequality and submultiplicativity of the operator norm,

$$\big\| \mathbf{E}^\nu_g \mathbf{C}_{g,g'} \mathbf{b} + \mathbf{b} \mathbf{C}_{g,g'} \mathbf{E}^\nu_{g'} + \mathbf{E}^\nu_g \mathbf{C}_{g,g'} \mathbf{E}^\nu_{g'} \big\|_{\mathrm{op}} \le c_{1,d}\big(2b_*\delta_{1,d} + \delta^2_{1,d}\big),$$

because $\|\mathbf{b}\|_{\mathrm{op}} = b_*$ and $\|\mathbf{C}_{g,g'}\|_{\mathrm{op}} \le c_{1,d}$. Averaging over $\nu, g, g'$ does not increase this uniform bound, which proves Eq. (C.6). □

*Remark* C.4. At fixed diffusion time, $S^\nu_{i,g}$ has order-one variation, so a constant first-order Hermite approximation need not be asymptotically exact. Eq. (C.6) is a sufficient condition, not a consequence of group algebra.

**Corollary C.5** (A small-preactivation sufficient condition)**.** *If $(q, s) \mapsto \mu_{1,q}(s)$ is Lipschitz with constant $L_{1,d}$ on the relevant domain and the row-normalized approximation uses $b_i = \mu_{1,q_i}(0)$, then*

$$\delta_{1,d} \le L_{1,d} \max_{i,\nu,g} |S^\nu_{i,g}|.$$

*For the alternative local scalar choice*

$$b_{\mathrm{loc}} = \mu_{1,1}(0),$$

*the bound gains the term $L_{1,d} \max_i |q_i - 1|$. The reported scalar approximation instead uses $b_t = \mathbb{E}\bar{\mu}_1(S)$ from Eq.* (B.11)*. Its error is measured and is not claimed to vanish in the fixed-time aligned experiments.*

**Proposition C.6** (Condition for a diagonal residual approximation)**.** *Let $\mathbf{R}_{\mathrm{cal}} = (\mathbf{G}_1 - \mathbf{B}_{\mathrm{cal}}) + \mathbf{M}_{\ge 2}(\mathbf{W})$ and let $\mathbf{D} = \mathrm{diag}(d_1, \dots, d_p)$ be any proposed diagonal shift. Set $\mathbf{U}(\mathbf{D}) = \mathbf{G}_0 + \mathbf{B}_{\mathrm{cal}} + \mathbf{D}$, and define*

$$\varepsilon(\mathbf{D}) = \|\mathbf{U}_{\mathrm{cond}} - \mathbf{U}(\mathbf{D})\|_{\mathrm{op}} = \|\mathbf{R}_{\mathrm{cal}} - \mathbf{D}\|_{\mathrm{op}}.$$

*Then*

$$\varepsilon(\mathbf{D}) \le \|\mathbf{R}_{\mathrm{cal}} - \mathrm{diag}(\mathbf{R}_{\mathrm{cal}})\|_{\mathrm{op}} + \max_i |(\mathbf{R}_{\mathrm{cal}})_{ii} - d_i|. \tag{C.7}$$

*Thus a diagonal shift is accurate when the combined residual is asymptotically diagonal and its diagonal entries are uniformly approximated.*

*Proof.* Since

$$\mathbf{R}_{\mathrm{cal}} - \mathbf{D} = \big(\mathbf{R}_{\mathrm{cal}} - \mathrm{diag}(\mathbf{R}_{\mathrm{cal}})\big) + \big(\mathrm{diag}(\mathbf{R}_{\mathrm{cal}}) - \mathbf{D}\big),$$

the triangle inequality gives

$$\varepsilon(\mathbf{D}) \le \|\mathbf{R}_{\mathrm{cal}} - \mathrm{diag}(\mathbf{R}_{\mathrm{cal}})\|_{\mathrm{op}} + \|\mathrm{diag}(\mathbf{R}_{\mathrm{cal}}) - \mathbf{D}\|_{\mathrm{op}}.$$

The second matrix is diagonal, so its operator norm equals $\max_i |(\mathbf{R}_{\mathrm{cal}})_{ii} - d_i|$, which proves Eq. (C.7). □

**Proposition C.7** (Total approximation error)**.** *Write $\varepsilon_{\mathrm{cal},d} = \|\mathbf{U}_{\mathrm{cond}} - \mathbf{U}_{\mathrm{cal}}\|_{\mathrm{op}}$ for the error of the scalar approximation $\mathbf{U}_{\mathrm{cal}} = \mathbf{G}_0 + \mathbf{B}_{\mathrm{cal}} + \mathbf{D}_{\mathrm{cal}}$. The operator-norm distance from $\mathbf{U}_{\mathrm{cal}}$ to the exact Gram is at most the conditional-approximation error of Theorem C.2 plus $\varepsilon_{\mathrm{cal},d}$. In particular, a small conditional-approximation error does not imply an accurate scalar approximation.*

*Proof of Proposition C.7.* On the event from Theorem C.2,

$$\begin{aligned}\|\mathbf{U}_{\mathrm{avg}} - \mathbf{U}_{\mathrm{cal}}\|_{\mathrm{op}} &\le \|\mathbf{U}_{\mathrm{avg}} - \mathbf{U}_{\mathrm{cond}}\|_{\mathrm{op}} + \|\mathbf{U}_{\mathrm{cond}} - \mathbf{U}_{\mathrm{cal}}\|_{\mathrm{op}} \\ &\le \varepsilon_{\mathrm{br}}(n, K_{\mathrm{H}}, \delta) + \varepsilon_{\mathrm{cal},d}.\end{aligned}$$

The second term is exactly the calibration error defined in the proposition. Thus the claimed bound holds with the same probability as the conditional-approximation bound. □

**Corollary C.8** (Full approximation)**.** *Suppose the conditional-approximation radius in Theorem C.2 satisfies $\varepsilon_{\mathrm{br}} = o_{\mathbb{P}}(1)$, as in Corollary B.9, and suppose Assumption B.11 holds. Then*

$$\|\mathbf{U}_{\mathrm{avg}} - \mathbf{U}_{\mathrm{cal}}\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

### C.4 Spectral and target-contribution bounds

For nonzero projectors $\mathbf{P}$ and $\mathbf{Q}$, define the two overlap fractions

$$\mathrm{prec}(\mathbf{P} \mid \mathbf{Q}) = \frac{\mathrm{Tr}(\mathbf{PQ})}{\mathrm{rank}(\mathbf{P})}, \qquad \mathrm{rec}(\mathbf{P} \mid \mathbf{Q}) = \frac{\mathrm{Tr}(\mathbf{PQ})}{\mathrm{rank}(\mathbf{Q})}. \tag{C.8}$$

Thus precision is the fraction of $\mathbf{P}$ captured by $\mathbf{Q}$, whereas recall is the fraction of $\mathbf{Q}$ captured by $\mathbf{P}$.

**Proposition C.9** (Spectral projector error bound)**.** *Let $\mathbf{U}$ and $\mathbf{U}_\star$ be symmetric and set*

$$\varepsilon_\star = \|\mathbf{U} - \mathbf{U}_\star\|_{\mathrm{op}}.$$

*Then $\max_j |\lambda_j(\mathbf{U}) - \lambda_j(\mathbf{U}_\star)| \le \varepsilon_\star$. If a cluster of $\mathbf{U}_\star$ has gap $\Delta_\star$ from the rest of its spectrum and $\varepsilon_\star < \Delta_\star/2$, the corresponding exact cluster is isolated. If $\mathbf{P}$ and $\mathbf{P}_\star$ are the two cluster projectors, then*

$$\|\mathbf{P} - \mathbf{P}_\star\|_{\mathrm{op}} \le \frac{2\varepsilon_\star}{\Delta_\star}. \tag{C.9}$$

*For every matrix $\mathbf{A}$,*

$$|\|\mathbf{AP}\|_{\mathrm{F}} - \|\mathbf{AP}_\star\|_{\mathrm{F}}| \le \|\mathbf{A}\|_{\mathrm{F}} \frac{2\varepsilon_\star}{\Delta_\star}. \tag{C.10}$$

*Proof.* Weyl's eigenvalue perturbation inequality (Bhatia, 1997) gives the eigenvalue bound. Under the stated gap condition, the Davis–Kahan sin-$\Theta$ theorem (Davis & Kahan, 1970; Yu et al., 2015) gives the projector bound in (C.9). The reverse triangle inequality and $\|\mathbf{A}(\mathbf{P} - \mathbf{P}_\star)\|_{\mathrm{F}} \le \|\mathbf{A}\|_{\mathrm{F}} \|\mathbf{P} - \mathbf{P}_\star\|_{\mathrm{op}}$ then prove the energy bound in (C.10). □

**Proposition C.10** (Target-contribution and subspace-overlap bounds)**.** *Let $\mathbf{T} \succeq 0$ be nonzero and define $\ell_{\mathbf{T}}(\mathbf{P}) = \mathrm{Tr}(\mathbf{PT})/\mathrm{Tr}(\mathbf{T})$. For any orthogonal projectors $\mathbf{P}, \mathbf{Q}$,*

$$|\ell_{\mathbf{T}}(\mathbf{P}) - \ell_{\mathbf{T}}(\mathbf{Q})| \le \|\mathbf{P} - \mathbf{Q}\|_{\mathrm{op}}. \tag{C.11}$$

*For any reference projector $\mathbf{R}$,*

$$|\,\mathrm{Tr}[(\mathbf{P} - \mathbf{Q})\mathbf{R}]| \le \mathrm{rank}(\mathbf{R})\, \|\mathbf{P} - \mathbf{Q}\|_{\mathrm{op}}. \tag{C.12}$$

*If $\mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{Q}) = r > 0$ and $\mathrm{rank}(\mathbf{R}) > 0$, then*

$$\begin{aligned}|\,\mathrm{prec}(\mathbf{P} \mid \mathbf{R}) - \mathrm{prec}(\mathbf{Q} \mid \mathbf{R})| &\le \frac{\mathrm{rank}(\mathbf{R})}{r}\, \|\mathbf{P} - \mathbf{Q}\|_{\mathrm{op}}, \\ |\,\mathrm{rec}(\mathbf{P} \mid \mathbf{R}) - \mathrm{rec}(\mathbf{Q} \mid \mathbf{R})| &\le \|\mathbf{P} - \mathbf{Q}\|_{\mathrm{op}}.\end{aligned} \tag{C.13}$$

*When all three projectors have rank $r$, both differences are at most the projector error.*

*Proof.* For a self-adjoint $\mathbf{A}$ and $\mathbf{T} \succeq 0$, $|\,\mathrm{Tr}(\mathbf{AT})| \le \|\mathbf{A}\|_{\mathrm{op}} \mathrm{Tr}(\mathbf{T})$. Taking $\mathbf{A} = \mathbf{P} - \mathbf{Q}$ and dividing by the positive $\mathrm{Tr}(\mathbf{T})$ proves (C.11). Applying the same inequality with $\mathbf{T} = \mathbf{R}$, and using $\mathrm{Tr}(\mathbf{R}) = \mathrm{rank}(\mathbf{R})$, proves (C.12). Finally, dividing this trace bound by $r$ and by $\mathrm{rank}(\mathbf{R})$, respectively, proves the two inequalities in (C.13). □

### C.5 CYCLIC GROUPS: BLOCK STRUCTURE AND THE EXACT REGIME

**Construction and population moments.** Let $G = C_m$ act orthogonally on $\mathbb{R}^d$. Choose $p$ orthonormal vectors $\mathbf{u}_i \in \mathrm{Fix}(G)$ and $p$ mutually orthogonal copies $\mathcal{M}_i$ of one nontrivial real irreducible representation of dimension $l \in \{1, 2\}$. Choose unit $\mathbf{v}_i \in \mathcal{M}_i$ with the same orbit profile $c(g) = \langle \mathbf{v}_i, g\mathbf{v}_i \rangle$ for every $i$. Thus the representation must contain these fixed and nontrivial components, requiring $d \geq p(1 + l)$. Partition the feature indices into nonempty $\mathcal{I}_S, \mathcal{I}_N$, with sizes $r_S, r_N$ and coordinate projectors $\mathbf{\Pi}_S, \mathbf{\Pi}_N$, using the mixed fixed and moved components specified in Section 5.2. The mixed-row correlation is $1 - \theta(1 - c(g))$. Because $|G|^{-1} \sum_g c(g) = 0$ and $c(e) = 1$, $c_- := \min_g c(g) < 0$, so the nonnegative-correlation range ends at

$$\theta^* = \frac{1}{1 - c_-} = \frac{1}{1 + \max\{0, -c_-\}} < 1.$$

For the sign representation of $C_2$, $c_- = -1$ and $\theta^* = 1/2$.

For the explicit examples in Section 5.2, set $l = 1$ and $\mathbf{B}_2 = (-1)$ when $m = 2$; for $m \geq 3$, set $l = 2$ and

$$\mathbf{B}_m = \begin{pmatrix} \cos(2\pi/m) & -\sin(2\pi/m) \\ \sin(2\pi/m) & \cos(2\pi/m) \end{pmatrix}.$$

In orthonormal coordinates, take the generator

$$g_0 = \mathrm{diag}\left( \mathbf{I}_p, \underbrace{\mathbf{B}_m, \ldots, \mathbf{B}_m}_{p \text{ copies}}, \mathbf{I}_{d-p(1+l)} \right),$$
$$G = \{g_0^j : 0 \leq j < m\},$$

omitting the final block when $d = p(1 + l)$. The first $p$ coordinate vectors give the $\mathbf{u}_i$; take $\mathbf{v}_i$ as the first coordinate vector in the $i$th $\mathbf{B}_m$ block. Then $g_0$ has order $m$, distinct row orbits are orthogonal, and $c(g_0^j) = \cos(2\pi j/m)$. The proof below applies more generally to any common nontrivial real irreducible component satisfying the stated orthogonality and profile conditions.

The population matrices used in Section 5.2 are

$$\begin{aligned} \mathbf{U}_M &= \mathbb{E}[\boldsymbol{\phi}_M(\mathbf{X}_t)\boldsymbol{\phi}_M(\mathbf{X}_t)^\top], \\ \mathbf{R}_{t,M} &= \mathbb{E}[\mathbf{Y}_{t,G}\boldsymbol{\phi}_M(\mathbf{X}_t)^\top], \qquad \mathbf{T}_{t,M} = \mathbf{R}_{t,M}^\top \mathbf{R}_{t,M}, \qquad M \in \{\mathrm{base}, \mathrm{avg}\}. \end{aligned} \tag{C.14}$$

Expectations average over independent $\mathbf{X}, \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, with the constructed weights held fixed and $\mathbf{X}_t = e^{-t}\mathbf{X} + \sqrt{\Delta_t}\boldsymbol{\xi}$ at fixed $t > 0$. In particular, $\mathbf{X}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathbb{E}[\mathbf{Y}_{t,G} \mid \mathbf{X}_t] = -\mathbf{P}_G\mathbf{X}_t$. The associated learning times and their ratios also omit the population superscript. Proposition B.15 shows that averaging preserves the common projected-target cross-moment while removing within-orbit feature variance. Orthogonality of distinct row orbits makes both Grams diagonal, so retained variance equals the averaged-to-base eigenvalue ratio.

Assume $\sigma$ is bounded, $\mathbb{E}[\sigma(Z)] = 0$, and $a_1 = \mathbb{E}[Z\sigma(Z)] \neq 0$ for $Z \sim \mathcal{N}(0, 1)$. Write $\lambda_{\mathrm{base}} = \mathbb{E}[\sigma(Z)^2]$ and $\kappa_\sigma = a_1^2/\lambda_{\mathrm{base}} > 0$. For a row with non-fixed fraction $\theta$, define $q_G(\theta) = \mathbb{E}[\phi_{\mathrm{avg},i}(\mathbf{X}_t)^2]/\lambda_{\mathrm{base}}$. To compute this fraction, let $\mathrm{He}_k$ denote the probabilists' Hermite polynomial of degree $k$ and set $\mu_k = \mathbb{E}[\sigma(Z)\mathrm{He}_k(Z)]$, $m_k = \mu_k^2/k!$. Centering gives $\mu_0 = 0$, while $\lambda_{\mathrm{base}} = \sum_{k \geq 1} m_k$ and $m_1 = a_1^2$. At Gaussian correlation $c$, the feature cross-moment is $\sum_{k \geq 1} m_k c^k$, giving the retention formula

$$\begin{aligned} q_G(\theta) &= \frac{1}{\lambda_{\mathrm{base}}} \sum_{k \geq 1} m_k \frac{1}{|G|} \sum_{g \in G} \big[1 - \theta(1 - c(g))\big]^k, \\ \underline{q}_G &= q_G(\theta^*). \end{aligned} \tag{C.15}$$

**Theorem** (Formal version of Theorem 5.1)**.** *Use the zero-initialized flow in Eq.* (10) *with the preceding population matrices and projected target* $\mathbf{Y}_{t,G}$*. Under the stated construction and Gaussian assumptions, suppose* $0 \leq \theta_S < \theta_N \leq \theta^*$ *and, for* $0 < \alpha_S, \alpha_N < 1/2$*,*

$$\frac{r_N(1 - \theta_N)}{r_S(1 - \theta_S) + r_N(1 - \theta_N)} \leq \min\{\alpha_S, \alpha_N\}. \tag{C.16}$$

*The population matrices satisfy*

$$\begin{aligned}\mathbf{U}_{\text{base}} &= \lambda_{\text{base}}\mathbf{I}_p,\\ \mathbf{U}_{\text{avg}} &= \lambda_{\text{base}}\big[q_G(\theta_S)\mathbf{\Pi}_S + q_G(\theta_N)\mathbf{\Pi}_N\big],\\ \mathbf{T}_{t,\text{base}} = \mathbf{T}_{t,\text{avg}} &= a_1^2\big[(1-\theta_S)\mathbf{\Pi}_S + (1-\theta_N)\mathbf{\Pi}_N\big].\end{aligned} \tag{C.17}$$

*The common blocks $\mathbf{P}_{B,\text{base}} = \mathbf{P}_{B,\text{avg}} = \mathbf{\Pi}_B$ are active and invariant under both Grams, and satisfy Eq.* (14)*. Moreover, $0 < q_G(\theta_N) < q_G(\theta_S) \le 1$ and*

$$q_G(\theta_S) - q_G(\theta_N) \ge \kappa_\sigma(\theta_N - \theta_S) > 0. \tag{C.18}$$

*For every $\rho \in (0,1)$, in the original optimization time $\tau$,*

$$\tau_{B,\text{base}}(\rho) = \tau_0(\rho) = \frac{-\log\rho}{c_t\lambda_{\text{base}}}, \qquad \tau_{B,\text{avg}}(\rho) = \frac{\tau_0(\rho)}{q_G(\theta_B)}, \qquad B \in \{S,N\}. \tag{C.19}$$

*Consequently, $\Delta_B = -\log q_G(\theta_B)$ and*

$$\Delta_{\text{win}}(\rho) = \log\frac{q_G(\theta_S)}{q_G(\theta_N)} \ge \log\left(1 + \frac{\kappa_\sigma(\theta_N-\theta_S)}{q_G(\theta_N)}\right) > 0. \tag{C.20}$$

*The ratio and duration are given by Eqs.* (20)*–*(21)*. In particular, the base window is empty, whereas the averaged window is $[\tau_0/q_G(\theta_S), \tau_0/q_G(\theta_N))$, with duration gain*

$$A_{\text{avg}}(\rho) - A_{\text{base}}(\rho) \ge \frac{\tau_0(\rho)\kappa_\sigma(\theta_N-\theta_S)}{q_G(\theta_S)q_G(\theta_N)} > 0. \tag{C.21}$$

*Finally, $q_G(\theta_N) \ge \underline{q}_G \ge 1/|G|$, so $\Delta_N \le \log(1/\underline{q}_G) \le \log|G|$.*

This subsection proves Theorem 5.1. Every statement is exact and at the population level, under the Gaussian law fixed in Section 5.2: $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ independent of $\mathbf{X}$, and $\mathbf{X}_t = e^{-t}\mathbf{X} + \sqrt{\Delta_t}\boldsymbol{\xi}$ with $t > 0$. Nothing here is a finite-design result, and nothing is claimed for a non-cyclic group. Rows have $\|\mathbf{w}_i\|^2 = d$, so the row scale of Eq. (B.2) is $q_i = 1$. The retention ratio $q_G(\theta)$ of Eq. (C.15) is a different quantity. It always carries the group as a subscript. Write $\hat{\mathbf{w}}_i = \mathbf{w}_i/\sqrt{d}$ for the unit row, so that $\hat{\mathbf{w}}_i = \mathbf{v}_{i,e}$ in the notation of Appendix B.1, and set

$$\begin{aligned}c_{ij}(g) &= \langle \hat{\mathbf{w}}_i, g\hat{\mathbf{w}}_j\rangle,\\ \mathbf{P}_G^{(k)} &= \frac{1}{|G|}\sum_{g\in G} g^{\otimes k} \quad \text{on } (\mathbb{R}^d)^{\otimes k},\\ \mathbf{P}_G^{(1)} &= \mathbf{P}_G.\end{aligned} \tag{C.22}$$

Each $g^{\otimes k}$ is orthogonal and $g \mapsto g^{\otimes k}$ is a representation, so $\mathbf{P}_G^{(k)}$ is a symmetric idempotent.

**Lemma C.11** (Reynolds tensor Gram identity)**.** *Let $\sigma$ be bounded and centered with Hermite masses $m_k = \mu_k^2/k!$. Then, at population level,*

$$\begin{aligned}(\mathbf{U}_{\text{base}})_{ij} &= \sum_{k\ge1} m_k \langle \hat{\mathbf{w}}_i, \hat{\mathbf{w}}_j\rangle^k,\\ (\mathbf{U}_{\text{avg}})_{ij} &= \sum_{k\ge1} m_k \left\langle \hat{\mathbf{w}}_i^{\otimes k}, \mathbf{P}_G^{(k)}\hat{\mathbf{w}}_j^{\otimes k}\right\rangle\\ &= \frac{1}{|G|}\sum_{g\in G}\sum_{k\ge1} m_k c_{ij}(g)^k.\end{aligned} \tag{C.23}$$

*In particular every Hermite-order block is positive semidefinite, with $\boldsymbol{\zeta}_{i,k} = \sqrt{m_k}\,\mathbf{P}_G^{(k)}\hat{\mathbf{w}}_i^{\otimes k}$.*

*Proof.* Since $g$ is orthogonal, $\langle \hat{\mathbf{w}}_i, g\mathbf{z}\rangle = \langle g^\top\hat{\mathbf{w}}_i, \mathbf{z}\rangle$ and $g^\top\hat{\mathbf{w}}_i$ is a unit vector, so every summand of the averaged feature is $\sigma$ at a standard normal. Mehler's identity in the form $\mathbb{E}[\sigma(\langle\mathbf{a},\mathbf{X}\rangle)\sigma(\langle\mathbf{b},\mathbf{X}\rangle)] =$

$\sum_{k\geq 1} m_k \langle \mathbf{a}, \mathbf{b}\rangle^k$ for unit $\mathbf{a}, \mathbf{b}$ (Lemma B.1 at $S = 0$, $q = 1$, with absolute convergence from Lemma B.2) gives

$$(\mathbf{U}_{\text{avg}})_{ij} = \frac{1}{|G|^2} \sum_{g,h\in G} \sum_{k\geq 1} m_k \left\langle g^\top \hat{\mathbf{w}}_i, h^\top \hat{\mathbf{w}}_j \right\rangle^k.$$

Now $\left\langle g^\top \hat{\mathbf{w}}_i, h^\top \hat{\mathbf{w}}_j \right\rangle = c_{ij}(gh^{-1})$, and for each fixed $g$ the map $h \mapsto gh^{-1}$ is a bijection of $G$, so the double average collapses to $|G|^{-1} \sum_g$. Finally $\sum_g c_{ij}(g)^k = \sum_g \left\langle \hat{\mathbf{w}}_i^{\otimes k}, g^{\otimes k} \hat{\mathbf{w}}_j^{\otimes k} \right\rangle = |G| \left\langle \hat{\mathbf{w}}_i^{\otimes k}, \mathbf{P}_G^{(k)} \hat{\mathbf{w}}_j^{\otimes k} \right\rangle$, and idempotence plus symmetry of $\mathbf{P}_G^{(k)}$ turns each order into a Gram of the $\boldsymbol{\zeta}_{i,k}$. The base case is $G = \{e\}$. □

This is the Gaussian population version of Lemma B.2, not the result of setting $S_{i,g} = 0$ in it. There the expansion comes after conditioning on a clean design, and the factor $\mu_{k,q_i}(S_{i,g}^\nu)$ depends on $g$. Orbit averaging acts as $\mathbf{P}_G^{(k)}$ only when that factor does not depend on $g$. It does not here, because the centered population model leaves no fixed clean preactivation, but this is an assumption on the model rather than a consequence of the group action.

**Corollary C.12** (Retention, trivial energy, and floors)**.** *Write* $\mathsf{r}_k(i) = \left\| \mathbf{P}_G^{(k)} \hat{\mathbf{w}}_i^{\otimes k} \right\|^2$ *and* $\mathsf{e}_0(i) = \|\mathbf{P}_G \hat{\mathbf{w}}_i\|^2$. *Then* $\mathsf{r}_k(i) \in [0, 1]$,

$$\frac{(\mathbf{U}_{\text{avg}})_{ii}}{(\mathbf{U}_{\text{base}})_{ii}} = \frac{1}{\lambda_{\text{base}}} \sum_{k\geq 1} m_k \mathsf{r}_k(i),$$
$$\mathsf{r}_1(i) = \mathsf{e}_0(i),$$
$$\mathsf{r}_k(i) \geq \mathsf{e}_0(i)^k,$$

*and* $\mathsf{r}_k(i) \geq 1/|G|$ *for every even* $k$.

*Proof.* The range and the ratio are immediate from Lemma C.11, and $\mathsf{r}_1(i) = \langle \hat{\mathbf{w}}_i, \mathbf{P}_G \hat{\mathbf{w}}_i \rangle = \|\mathbf{P}_G \hat{\mathbf{w}}_i\|^2$. For the power bound, assume $\mathsf{e}_0(i) > 0$ (otherwise the claim is trivial) and put $\boldsymbol{\psi} = (\mathbf{P}_G \hat{\mathbf{w}}_i)^{\otimes k}$, which is $G$-invariant, so $\mathbf{P}_G^{(k)} \boldsymbol{\psi} = \boldsymbol{\psi}$ and $\left\langle \mathbf{P}_G^{(k)} \hat{\mathbf{w}}_i^{\otimes k}, \boldsymbol{\psi} \right\rangle = \left\langle \hat{\mathbf{w}}_i^{\otimes k}, \boldsymbol{\psi} \right\rangle = \mathsf{e}_0(i)^k = \|\boldsymbol{\psi}\|^2$. Cauchy–Schwarz then gives $\mathsf{r}_k(i) \geq \mathsf{e}_0(i)^{2k} / \mathsf{e}_0(i)^k$. For even $k$ every term $c_{ii}(g)^k$ in Eq. (C.23) is nonnegative and the identity contributes 1. □

**Lemma C.13** (Block structure is an assumption on the rows)**.** *Let* $\mathcal{O}_i = \operatorname{span}\{g\hat{\mathbf{w}}_i : g \in G\}$. *If* $\mathcal{O}_i \perp \mathcal{O}_j$ *for all* $i \neq j$, *then* $c_{ij}(g) = 0$ *for all* $g$ *and all* $i \neq j$, *so* $\mathbf{U}_{\text{base}}$ *and* $\mathbf{U}_{\text{avg}}$ *are diagonal in the feature index. The target-contribution matrix is diagonal as well only under the projected Gaussian target of Section 5.2, or another hypothesis making the projected row directions orthogonal. Orbit orthogonality alone does not imply it. Mutual orthogonality requires* $\sum_i \dim \mathcal{O}_i \leq d$. *For a generic row in a representation containing the regular representation* $\dim \mathcal{O}_i = |G|$, *which forces* $p \leq d/|G|$.

*Proof.* Orthogonality of $\mathcal{O}_i$ and $\mathcal{O}_j$ gives $\langle \hat{\mathbf{w}}_i, g\hat{\mathbf{w}}_j \rangle = 0$ because $g\hat{\mathbf{w}}_j \in \mathcal{O}_j$, and Lemma C.11 then kills every off-diagonal entry. The dimension count is the direct sum of the $\mathcal{O}_i$. □

Write $\mathcal{V}_0 = \operatorname{Fix}(G)$. A row that mixes $\mathcal{V}_0$ with a non-trivial component lies in no single irreducible component, and two such rows that share an invariant direction have $c_{ij}(g) \neq 0$. Block structure in the feature index is therefore an assumption on the rows, never a consequence of splitting $\mathbb{R}^d$.

**Lemma C.14** (Endpoint bridge)**.** *Fix* $M$, *suppose* $(\mathbf{U}_M)_{ii} = \lambda_{B,M}$ *for every* $i \in \mathcal{I}_B$, *and set* $\mathbf{U}_M^\star = \lambda_{S,M} \boldsymbol{\Pi}_S + \lambda_{N,M} \boldsymbol{\Pi}_N$ *and*

$$\varepsilon_M = \|\mathbf{U}_M - \mathbf{U}_M^\star\|_{\text{op}} \leq \max_i \sum_{j\neq i} |(\mathbf{U}_M)_{ij}|.$$

*If* $|\lambda_{S,M} - \lambda_{N,M}| > 2\varepsilon_M$, *then Weyl's inequality separates the two clusters, exactly* $\operatorname{rank} \boldsymbol{\Pi}_B$ *eigenvalues belong to block* $B$, *and* $\underline{\lambda}_{B,M} \geq \lambda_{B,M} - \varepsilon_M$, $\overline{\lambda}_{B,M} \leq \lambda_{B,M} + \varepsilon_M$. *If in addition every*

$\lambda_{B,M} - \varepsilon_M$ *is strictly positive and both clusters are nonempty, substitution into Eq.* (22) *gives*

$$\begin{aligned} d_0 &\geq \log \frac{\lambda_{N,\mathrm{base}} - \varepsilon_{\mathrm{base}}}{\lambda_{N,\mathrm{avg}} + \varepsilon_{\mathrm{avg}}}, \\ s_0 &\leq \max\left\{0, \log \frac{\lambda_{S,\mathrm{base}} + \varepsilon_{\mathrm{base}}}{\lambda_{S,\mathrm{avg}} - \varepsilon_{\mathrm{avg}}}, \log \frac{\lambda_{S,\mathrm{avg}} + \varepsilon_{\mathrm{avg}}}{\lambda_{S,\mathrm{base}} - \varepsilon_{\mathrm{base}}}\right\}. \end{aligned} \tag{C.24}$$

*When* $\varepsilon_M = 0$ *and* $q_G \leq 1$*, these collapse to* $-\log q_G$*.*

*Proof.* $\mathbf{U}_M - \mathbf{U}_M^\star$ has zero diagonal, so Gershgorin bounds its operator norm by the largest absolute row sum. Weyl's inequality moves each eigenvalue of $\mathbf{U}_M^\star$ by at most $\varepsilon_M$. Since the two eigenvalues of $\mathbf{U}_M^\star$ are $\lambda_{S,M}$ and $\lambda_{N,M}$ with multiplicities $\operatorname{rank}\mathbf{\Pi}_S$ and $\operatorname{rank}\mathbf{\Pi}_N$, the assumed gap keeps the two perturbed clusters disjoint and preserves their counts. The endpoint substitutions are monotone in each argument. For the last claim, $\lambda_{B,\mathrm{base}} = \lambda_{\mathrm{base}}$ and $\lambda_{B,\mathrm{avg}} = \lambda_{\mathrm{base}} q_G(\theta_B)$, while $q_G \leq 1$ because $|1 - \theta(1 - c(g))| \leq 1$ forces $\mathsf{r}_k \leq 1$ in Eq. (C.15). Hence the second argument of the maximum defining $s_0$ is nonpositive. □

Lemma C.14 is the only entry point for non-orthogonal rows and for $p \gg d$, where $\varepsilon_M > 0$ by construction. In that regime Theorem 5.3 is stated for the spectral projectors $\mathbf{P}_{B,M}$ of $\mathbf{U}_M$, not for the design projectors $\mathbf{\Pi}_B$, so its conditions must be re-verified after transporting them: Proposition C.9 gives $\|\mathbf{P}_{B,M} - \mathbf{\Pi}_B\|_{\mathrm{op}}$ in terms of $\varepsilon_M$ and the separated-cluster gap, and $\left|\ell_{\mathbf{T}_{t,M}}(\mathbf{P}_{B,M}) - \ell_{\mathbf{T}_{t,M}}(\mathbf{\Pi}_B)\right| \leq \|\mathbf{P}_{B,M} - \mathbf{\Pi}_B\|_{\mathrm{op}}$. In the family below $\varepsilon_M = 0$, so $\mathbf{P}_{B,M} = \mathbf{\Pi}_B$ for the averaged Gram and this step is vacuous.

**Lemma C.15** (Structure of the row family)**.** *For the rows in this mixed fixed/moved construction with the common profile* $c$ *defined above,*

$$\begin{aligned} &\mathsf{e}_0(i) = 1 - \theta_i, && c_{ii}(g) = 1 - \theta_i\big(1 - c(g)\big), \\ &c_{ij}(g) = 0 \quad (i \neq j), && \frac{1}{|G|}\sum_{g \in G} c(g) = 0. \end{aligned} \tag{C.25}$$

*Consequently* $\mathsf{r}_k(i) = \mathsf{r}_k(\theta_i)$ *for the function* $\mathsf{r}_k(\theta) = |G|^{-1}\sum_g [1 - \theta(1 - c(g))]^k$ *appearing in Eq.* (C.15)*, and* $\dim \mathcal{O}_i \leq 1 + l$*, with equality when* $\theta_i > 0$*. The family therefore needs* $p(1 + l) \leq d$*. Because the profile has mean zero and* $c(e) = 1$*, necessarily* $c_- < 0$*, so* $\theta^* < 1$ *and every admissible block has* $\theta_B < 1$*.*

*Proof.* $g\mathbf{u}_j = \mathbf{u}_j$ because $\mathbf{u}_j \in \mathrm{Fix}(G)$, and $g\mathbf{v}_j \in \mathcal{M}_j$. Since $\mathcal{V}_0 \perp \mathcal{M}_i$ for every $i$, all four cross terms of $\langle \hat{\mathbf{w}}_i, g\hat{\mathbf{w}}_j \rangle$ vanish for $i \neq j$: $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$, $\langle \mathbf{u}_i, g\mathbf{v}_j \rangle = 0$, $\langle \mathbf{v}_i, \mathbf{u}_j \rangle = 0$, and $\langle \mathbf{v}_i, g\mathbf{v}_j \rangle = 0$ because $\mathcal{M}_i \perp \mathcal{M}_j$. For $i = j$ the two cross terms still vanish and the remainder is $(1 - \theta_i) + \theta_i c(g)$. Taking $g$ averages, $\mathsf{e}_0(i) = \|\mathbf{P}_G \hat{\mathbf{w}}_i\|^2 = 1 - \theta_i$ and $|G|^{-1}\sum_g c(g) = \langle \mathbf{v}_i, \mathbf{P}_G \mathbf{v}_i \rangle = 0$ because $\mathbf{v}_i \perp \mathrm{Fix}(G)$. Finally $\mathcal{O}_i \subseteq \mathbb{R}\mathbf{u}_i \oplus \mathcal{M}_i$, with equality when $\theta_i > 0$. □

*Proof of Theorem 5.1. Grams.* Lemmas C.11, C.13 and C.15 give $(\mathbf{U}_{\mathrm{base}})_{ij} = \delta_{ij}\lambda_{\mathrm{base}}$ and $(\mathbf{U}_{\mathrm{avg}})_{ij} = \delta_{ij}\lambda_{\mathrm{base}} q_G(\theta_i)$, which is Eq. (C.17).

*Target and activity.* Stein's identity (Stein, 1981) gives $\mathbb{E}[\mathbf{X}\sigma(\langle \mathbf{a}, \mathbf{X} \rangle)] = a_1 \mathbf{a}$ for unit $\mathbf{a}$, so the base cross-moment matrix has $i$th column $-a_1 \mathbf{P}_G \hat{\mathbf{w}}_i = -a_1\sqrt{1 - \theta_i}\,\mathbf{u}_i$. For the averaged model $|G|^{-1}\sum_g g^\top = \mathbf{P}_G$ and $\mathbf{P}_G^2 = \mathbf{P}_G$, so $\mathbf{R}_{t,\mathrm{avg}} = \mathbf{R}_{t,\mathrm{base}}$. For either model, $\mathbf{T}_{t,M} = \mathbf{R}_{t,M}^\top \mathbf{R}_{t,M}$ is diagonal because the $\mathbf{u}_i$ are orthonormal. This is the target-matrix identity in Eq. (C.17), and the block traces give the initial update fractions

$$\begin{aligned} \ell_{\mathbf{T}_{t,M}}(\mathbf{\Pi}_N) &= \frac{r_N(1 - \theta_N)}{r_S(1 - \theta_S) + r_N(1 - \theta_N)}, \\ \ell_{\mathbf{T}_{t,M}}(\mathbf{\Pi}_S) &= 1 - \ell_{\mathbf{T}_{t,M}}(\mathbf{\Pi}_N). \end{aligned}$$

The condition in Eq. (C.16) therefore gives both $\ell_{\mathbf{T}_{t,M}}(\mathbf{\Pi}_S) \geq 1 - \alpha_S$ and $\ell_{\mathbf{T}_{t,M}}(\mathbf{\Pi}_N) \leq \alpha_N$. In particular, the latter is equivalent to

$$\theta_N \geq 1 - \frac{\alpha_N}{1 - \alpha_N}\frac{r_S}{r_N}(1 - \theta_S), \qquad 0 < \alpha_N < 1/2.$$

This verifies the block conditions in Eq. (14); the timing margin is proved below. Also, $\mathrm{Tr}(\mathbf{\Pi}_B \mathbf{T}_{t,M}) = r_B a_1^2(1-\theta_B) > 0$, since $\theta_B \le \theta^* < 1$ by Lemma C.15, and both Grams are positive definite because $\lambda_{\mathrm{base}} \ge a_1^2 > 0$ and $q_G(\theta) \ge \kappa_\sigma(1-\theta) > 0$ by Corollary C.12 at $k=1$. Hence $\mathbf{A}_{\star,M}\mathbf{\Pi}_B = \sqrt{p}\,\mathbf{R}_{t,M}\mathbf{\Pi}_B\mathbf{U}_M^{-1} \neq 0$ and both blocks are active in both models.

*The decrease.* Put $t_g(\theta) = 1-\theta(1-c(g))$. Since $c(g) \le 1$ and $c_- < 0$, the condition $0 \le \theta \le \theta^* = 1/(1-c_-)$ gives $0 \le t_g(\theta) \le 1$ for every $g$. For $0 \le \theta_S < \theta_N \le \theta^*$ and every $g$, $t_g(\theta_S) - t_g(\theta_N) = (\theta_N - \theta_S)(1-c(g)) \ge 0$ with both values in $[0,1]$, so $t_g(\theta_S)^k \ge t_g(\theta_N)^k$ termwise and $\mathsf{r}_k(\theta_S) \ge \mathsf{r}_k(\theta_N)$ for every $k$. At $k=1$, Eq. (C.25) gives $\mathsf{r}_1(\theta_S) - \mathsf{r}_1(\theta_N) = \theta_N - \theta_S$ exactly. The series $\sum_k m_k \mathsf{r}_k$ converges absolutely because $\mathsf{r}_k \le 1$ and $\sum_k m_k = \lambda_{\mathrm{base}} < \infty$, so the difference may be taken termwise and

$$\begin{aligned} q_G(\theta_S) - q_G(\theta_N) &= \frac{1}{\lambda_{\mathrm{base}}} \sum_{k\ge 1} m_k \big[\mathsf{r}_k(\theta_S) - \mathsf{r}_k(\theta_N)\big] \\ &\ge \frac{m_1}{\lambda_{\mathrm{base}}}(\theta_N - \theta_S), \end{aligned}$$

which is Eq. (C.18) with $m_1 = a_1^2$. No differentiation is used, so no interchange of a derivative with an infinite sum needs justification.

*Endpoints.* The two averaged eigenvalues $\lambda_{\mathrm{base}} q_G(\theta_S) > \lambda_{\mathrm{base}} q_G(\theta_N) > 0$ are distinct, so $\mathbf{\Pi}_S$ and $\mathbf{\Pi}_N$ are the exact spectral projectors of $\mathbf{U}_{\mathrm{avg}}$ and Lemma C.14 applies to it with $\varepsilon_{\mathrm{avg}} = 0$. The base Gram is $\lambda_{\mathrm{base}}\mathbf{I}_p$: it commutes with every coordinate projector and acts by the same eigenvalue on each, so $\mathbf{\Pi}_S$ and $\mathbf{\Pi}_N$ are common *reducing* subspaces rather than canonical spectral projectors of the base Gram, and the strict cluster-gap premise of Lemma C.14 is not available there. The reducing-block convention applies here: the parameter flow leaves each such subspace invariant, so $E_{B,M}(\tau) = \exp(-2\lambda_{B,M}\bar{\tau})$ is a direct calculation on the block, with $\lambda_{B,\mathrm{base}} = \lambda_{\mathrm{base}}$ and $\lambda_{B,\mathrm{avg}} = \lambda_{\mathrm{base}} q_G(\theta_B)$. Solving $E_{B,M}(\tau) = \rho^2$ gives $\tau_{B,M} = -\log\rho/(c_t\lambda_{B,M})$, which proves Eq. (C.19) and hence Eq. (20). Since $q_G \le 1$, the clipped definition of $s_0$ in Eq. (22) collapses to $-\log q_G(\theta_S)$, and substituting the two block eigenvalues into the definition of $\Delta_B$ gives Eq. (C.20), with

$$\begin{aligned} \Delta_{\mathrm{win}} &= \log \frac{q_G(\theta_S)}{q_G(\theta_N)} \\ &= \log\left(1 + \frac{q_G(\theta_S) - q_G(\theta_N)}{q_G(\theta_N)}\right) \\ &\ge \log\left(1 + \frac{\kappa_\sigma(\theta_N - \theta_S)}{q_G(\theta_N)}\right) > 0. \end{aligned}$$

The inequality $d_0 > s_0$ is a conclusion here, not a hypothesis: Eq. (C.16) only labels the blocks, and activity is checked separately. This direct calculation uses the invariant-block convention of Appendix D.2. It does not invoke Theorem 5.3, whose spectral-projector premise is not satisfied by a split of the base Gram's repeated eigenspace.

*Window duration.* Equation (C.19) gives $\tau_{\mathrm{gen,base}} = \tau_{\mathrm{mem,base}} = \tau_0$, so $A_{\mathrm{base}} = 0$. For Avg-RFM, the strict inequality $q_G(\theta_S) > q_G(\theta_N) > 0$ gives $\tau_{\mathrm{gen,avg}} < \tau_{\mathrm{mem,avg}}$. The block errors are strictly decreasing, so the generalization regime is exactly $[\tau_0/q_G(\theta_S), \tau_0/q_G(\theta_N))$. Subtracting its endpoints proves Eq. (21), and rewriting their difference as

$$A_{\mathrm{avg}} - A_{\mathrm{base}} = \tau_0 \frac{q_G(\theta_S) - q_G(\theta_N)}{q_G(\theta_S) q_G(\theta_N)}$$

and applying Eq. (C.18) proves Eq. (C.21). Only $\tau_0$ depends on $\rho$, establishing the claimed scaling of duration with $-\log\rho$ and the threshold independence of the ratio gain. If $\theta_S = 0$, its features are invariant and $q_G(0) = 1$, so $\tau_{\mathrm{gen,avg}} = \tau_{\mathrm{gen,base}}$.

*Upper bound.* Monotonicity gives $q_G(\theta) \ge q_G(\theta^*) = \underline{q}_G$ on $[0,\theta^*]$. At $\theta^*$ every $t_g \ge 0$, so every term of $|G|\ \mathsf{r}_k(\theta^*)$ is nonnegative while the identity element contributes 1. Summing against $m_k$ gives $\underline{q}_G \ge \lambda_{\mathrm{base}}/(|G|\,\lambda_{\mathrm{base}}) = 1/|G|$ and hence $d_0 = -\log q_G(\theta_N) \le \log(1/\underline{q}_G) \le \log|G|$. □

*Remark* C.16 (When the ceiling equals $\log|G|$, and a uniform bound). $\underline{q}_G = 1/|G|$ holds exactly when every non-identity element satisfies $t_g(\theta^*) = 0$, that is $c(g) = c_-$ for all $g \neq e$. For the $C_2$

sign action this is automatic, and for the primitive rotation profile $c(g_0^j) = \cos(2\pi j/m)$ it holds only at $m = 3$. So the clean upper bound $\log|G|$ occurs only at $m \in \{2, 3\}$, where it is independent of the activation. For $m \geq 4$ the floor is strictly larger and the upper bound is activation dependent and strictly below $\log m$. A uniform bound is still available: for these profiles $\theta^* = 1/2$ when $m$ is even and $\theta^* = 1/(1+\cos(\pi/m)) \leq 2/3$ when $m$ is odd, so $\underline{q}_G \geq \kappa_\sigma(1-\theta^*) \geq \kappa_\sigma/3$ and

$$\sup_{m\geq 2} \log(1/\underline{q}_{C_m}) \leq \log(3/\kappa_\sigma)$$

for each fixed activation and primitive profile. This is a statement about the primitive cyclic family only. The numerical agreement of $m = 8$ and $m = 16$ in Table F.3 is a saturation diagnostic, not a proof of convergence, and no supremum over arbitrary finite groups or profiles is claimed.

*Remark* C.17 (The two-element case, and a warning for larger groups). Taking $l = 1$ with the sign character gives $c(\mathbf{J}) = -1$, $\theta^* = 1/2$, $\mathsf{r}_k = [1 + (1-2\theta)^k]/2$, $\underline{q}_G = 1/2$, and $d_0 \leq \log 2$: every formula in Theorem 5.1 reduces to the tilted-reflection formula, so the two-element case is the $(C_2, \text{sign})$ instance and not a separate result. The same computation is a warning for larger groups. If $n_-$ elements of any $G$ act by $-1$ on a one-dimensional component, then $\mathsf{r}_k = (|G| - n_-)/|G| + (n_-/|G|)(1-2\theta)^k$, which for a cyclic group of even order equals $[1 + (1-2\theta)^k]/2$ for every $m$. A row whose residual $\max_k \left|\mathsf{r}_k - [1 + (1-2\theta)^k]/2\right|$ is at numerical zero therefore carries no information beyond the two-element case and must not be reported as general finite-group evidence. The flag is a statement about the row self-profile, not about cross-row or non-cyclic structure. The archived grid raises it on exactly the 48 cells with $m = 2$.

**Scope of Theorem 5.1.** Five boundaries are worth naming. (i) Everything above is population exact. The orbit-concentration radius of Eq. (B.27) would need $n$ of order $10^8$–$10^{10}$ at these dimensions and gaps, so no finite-design version is claimed. (ii) The base Gram is isotropic, so $\bar{W}_{\text{base}} = 1$. The prescribed blocks coincide across models, so their common-reference overlaps equal one without any subspace transport. This follows because both Grams are diagonal in the prescribed coordinate basis. (iii) $p \gg d$ needs non-orthogonal rows, $\varepsilon_M > 0$, and a non-diagonal target-contribution matrix, and is not attempted. (iv) Only the real-irreducible components of a finite cyclic group are covered. the executed two-dimensional rotation is real-irreducible but of complex type in the Schur classification, and for a non-cyclic group the real dimension $l$, the profile $c$, the multiplicity, and the existence of a common profile must all be re-established first. (v) The construction assumes the representation contains at least $p$ copies of both the trivial representation and the chosen nontrivial component, which has to be checked for each concrete representation.

# D Frozen-Feature Dynamics and Time-Gap Bounds

This appendix separates the Gaussian comparison model from the exact parameter path. The comparison model is a generalized sample covariance used only to compare with the exact Gram, following Gaussian-equivalence and random-feature DSM learning-curve analyses (Goldt et al., 2022; George et al., 2026). The dynamics results themselves require no Gaussian assumption.

## D.1 The Gaussian comparison model

Under Assumption B.10 and $v_t^2 > 0$, define

$$\kappa_{\text{noise},t} = \frac{1}{|G|^2} \sum_{g,g' \in G} \Psi_t(\rho_{g,g'}), \qquad \rho_{g,g'} = \frac{\text{Tr}(\mathbf{\Sigma}_d g^{-1} g')}{\text{Tr}(\mathbf{\Sigma}_d)}. \tag{D.1}$$

The Gaussian residual covariance is positive semidefinite, so this orbit average is the normalized variance of an averaged residual and satisfies $\kappa_{\text{noise},t} \geq 0$; hence the square root in (D.3) is well-defined. For the block-regular cyclic covariance,

$$\rho_{e,h} = \frac{\lambda_{\text{inv}} - \lambda_\perp}{\lambda_{\text{inv}} + (m-1)\lambda_\perp}, \qquad h \neq e. \tag{D.2}$$

For this subsection, write $\mathbf{X}_{\text{des}} = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{d\times n}$. The scalar Gaussian comparison model is

$$\mathbf{L} = a_t e^{-t} \frac{\mathbf{W}\mathbf{P}_G \mathbf{X}_{\text{des}}}{\sqrt{d}}, \qquad \mathbf{R}_G = \sqrt{\kappa_{\text{noise},t}}\, v_t \mathbf{\Omega}, \qquad \mathbf{F}_G = \mathbf{L} + \mathbf{R}_G, \tag{D.3}$$

where $\mathbf{\Omega}$ has i.i.d. standard Gaussian entries and is independent of $\mathbf{X}_{\mathrm{des}}$, with

$$\mathbf{H}_G = m_t \mathbf{1}_p \mathbf{1}_n^\top + \mathbf{F}_G, \qquad \mathbf{U}_{\mathrm{G}} = \frac{1}{n}\mathbf{H}_G\mathbf{H}_G^\top + \mathbf{B}_{\mathrm{cal}} + \mathbf{D}_{\mathrm{cal}}. \tag{D.4}$$

**Theorem D.1** (Exact representation of the Gaussian comparison model)**.** *Conditional on* $\mathbf{W}$*, there exists* $\mathbf{Z} \in \mathbb{R}^{p\times n}$ *with i.i.d. standard Gaussian entries such that*

$$\mathbf{F}_G \overset{d}{=} \mathbf{\Sigma}_F(\mathbf{W})^{1/2}\mathbf{Z}, \qquad \mathbf{\Sigma}_F(\mathbf{W}) = \frac{a_t^2 e^{-2t}}{d}\mathbf{W}\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G\mathbf{W}^\top + \kappa_{\mathrm{noise},t}v_t^2\mathbf{I}_p. \tag{D.5}$$

*Let* $\mathbf{P}_{\mathbf{1}_p^\perp} = \mathbf{I}_p - p^{-1}\mathbf{1}_p\mathbf{1}_p^\top$*. Then*

$$\mathbf{P}_{\mathbf{1}_p^\perp}\mathbf{U}_{\mathrm{G}}\mathbf{P}_{\mathbf{1}_p^\perp} \overset{d}{=} \mathbf{P}_{\mathbf{1}_p^\perp}\left[\frac{1}{n}\mathbf{\Sigma}_F^{1/2}\mathbf{Z}\mathbf{Z}^\top\mathbf{\Sigma}_F^{1/2} + \mathbf{B}_{\mathrm{cal}} + \mathbf{D}_{\mathrm{cal}}\right]\mathbf{P}_{\mathbf{1}_p^\perp}. \tag{D.6}$$

*The projected signal and residual terms do not form two additive independent Wishart matrices; their cross terms are part of the single generalized sample covariance.*

*Proof.* Conditional on $\mathbf{W}$, each column of $\mathbf{L}$ and $\mathbf{R}_G$ is centered Gaussian with covariances

$$\mathrm{Cov}(\mathbf{L}_\nu \mid \mathbf{W}) = a_t^2 e^{-2t}\mathbf{W}\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G\mathbf{W}^\top/d,$$
$$\mathrm{Cov}(\mathbf{R}_{G,\nu} \mid \mathbf{W}) = \kappa_{\mathrm{noise},t}v_t^2\mathbf{I}_p.$$

Their sum has covariance $\mathbf{\Sigma}_F(\mathbf{W})$, independently across samples.

Projection by $\mathbf{P}_{\mathbf{1}_p^\perp}$ removes the rank-one mean of the comparison model and its cross terms. It does not remove the signal–residual cross terms in $\mathbf{F}_G\mathbf{F}_G^\top/n$. □

**Proposition D.2** (Residual-only Marchenko–Pastur reference)**.** *Assume* $p, n \to \infty$*,* $p/n \to \gamma \in (0,\infty)$*,* $\kappa_{\mathrm{noise},t}v_t^2 \to \sigma_R^2 \geq 0$*, and* $\gamma_G s_t^2 \to d_R \geq 0$*. For i.i.d.* $\mathbf{\Omega} \in \mathbb{R}^{p\times n}$*, the lower edge of the nonzero Wishart branch of*

$$\frac{\kappa_{\mathrm{noise},t}v_t^2}{n}\mathbf{\Omega}\mathbf{\Omega}^\top + \gamma_G s_t^2\mathbf{I}_p$$

*converges almost surely to* $d_R + \sigma_R^2(1-\sqrt{\gamma})^2$*. The finite-dimensional scalar reference used in the numerical comparisons is*

$$\lambda_{\mathrm{res},-}^{G,0} = \gamma_G s_t^2 + \kappa_{\mathrm{noise},t}v_t^2\left(1-\sqrt{p/n}\right)^2. \tag{D.7}$$

*If* $p > n$*, there is also an atom of multiplicity* $p-n$ *at the diagonal shift* $\gamma_G s_t^2$*.*

*Proof.* Apply the Marchenko–Pastur/Bai–Yin edge law to $n^{-1}\mathbf{\Omega}\mathbf{\Omega}^\top$ and then scale and shift (Marchenko & Pastur, 1967; Bai et al., 1988; Bai & Yin, 1993; Vershynin, 2018). Convergence of the scalar parameters gives the displayed limit; retaining their finite-dimensional values gives the reference value in (D.7). □

**Proposition D.3** (Trace scale of the projected linear term)**.** *For i.i.d. standard Gaussian weights,*

$$\mathbb{E}_{\mathbf{W}}\left[\frac{1}{p}\,\mathrm{Tr}\left(\frac{a_t^2e^{-2t}}{d}\mathbf{W}\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G\mathbf{W}^\top\right)\right] = a_t^2e^{-2t}\kappa_{\mathrm{sig}}\sigma_{\mathbf{x}}^2. \tag{D.8}$$

*Neither this identity nor* (D.7) *determines a separated eigenvalue or the lower edge of the full* $\mathbf{U}_{\mathrm{G}}$*.*

*Proof.* Use $\mathbb{E}[\mathbf{W}^\top\mathbf{W}] = p\mathbf{I}_d$, cyclicity of trace, and (B.8). □

**Corollary D.4** (Projected dimensions and limits of the comparison model)**.** *The active projected aspect ratios are* $p/d_{\mathrm{proj}}$ *and* $n/d_{\mathrm{proj}}$*, where* $d_{\mathrm{proj}} = \mathrm{rank}(\mathbf{P}_G\mathbf{\Sigma}_d\mathbf{P}_G)$*, and the projected variance density is* (B.9)*. Equations* (D.7) *and* (D.8) *describe the two components separately. The lower edge of the full comparison matrix requires analysis of the generalized covariance in* (D.5) *together with* $\mathbf{B}_{\mathrm{cal}}$ *and* $\mathbf{D}_{\mathrm{cal}}$*.*

### D.2 Denoising loss and parameter dynamics

**Proposition D.5** (Frozen-feature parameter dynamics)**.** *Fix the clean samples and feature weights, and assume that the features have finite second moments under the noise distribution for each sample. Let $\mathbf{y}_t^\nu(\boldsymbol{\xi})$ be any square-integrable vector target; in the main text this is the projected DSM target $\mathbf{y}_{t,G}^\nu$. Consider*

$$\mathcal{L}_t(\mathbf{A}) = \Delta_t d \, \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \left[ \left\| p^{-1/2} \mathbf{A} \boldsymbol{\phi}(\mathbf{x}_t^\nu) - \mathbf{y}_t^\nu(\boldsymbol{\xi}) \right\|^2 \right]. \tag{D.9}$$

*Let $\mathbf{U} = n^{-1} \sum_\nu \mathbb{E}[\boldsymbol{\phi}(\mathbf{x}_t^\nu)\boldsymbol{\phi}(\mathbf{x}_t^\nu)^\top]$ and $\mathbf{R}_t = n^{-1} \sum_\nu \mathbb{E}[\mathbf{y}_t^\nu \boldsymbol{\phi}(\mathbf{x}_t^\nu)^\top]$. The loss minimizer with the smallest Frobenius norm is*

$$\mathbf{A}_\star = \sqrt{p}\, \mathbf{R}_t \mathbf{U}^\dagger, \tag{D.10}$$

*where $\mathbf{U}^\dagger$ denotes the pseudoinverse. For any initialization, the deviation $\mathbf{D}(\tau) = \mathbf{A}(\tau) - \mathbf{A}_\star$ satisfies, with $c_t = 2\Delta_t d/p$,*

$$\dot{\mathbf{D}}(\tau) = -c_t \mathbf{D}(\tau)\mathbf{U}, \qquad \mathbf{D}(\tau) = \mathbf{D}(0) e^{-c_t \mathbf{U} \tau}.$$

*Under zero initialization, $\mathbf{D}(0) = -\mathbf{A}_\star$, so*

$$\mathbf{A}(\tau) = \mathbf{A}_\star (\mathbf{I}_p - e^{-c_t \mathbf{U} \tau}), \qquad \mathbf{D}(\tau) = -\mathbf{A}_\star e^{-c_t \mathbf{U} \tau}. \tag{D.11}$$

*In particular, $\mathbf{A}(\tau) \to \mathbf{A}_\star$ even if $\mathbf{U}$ is singular. Using the base, group-averaged, tensor-grid, or Haar-averaged features for $\boldsymbol{\phi}$ gives the corresponding model's dynamics.*

*Proof.* Differentiating (D.9) gives

$$\nabla_{\mathbf{A}} \mathcal{L}_t(\mathbf{A}) = \frac{2\Delta_t d}{p} \mathbf{A}\mathbf{U} - \frac{2\Delta_t d}{\sqrt{p}} \mathbf{R}_t.$$

This gradient is affine and globally Lipschitz. With optimization times $\tau_k = kh$ and $t, n, p, d$ fixed, the linearly interpolated gradient-descent iterates therefore converge uniformly on bounded time intervals to gradient flow as $h \to 0$. The loss is convex, so its minimizers solve $\mathbf{A}\mathbf{U} = \sqrt{p}\,\mathbf{R}_t$. If $\mathbf{v} \in \ker(\mathbf{U})$, then $0 = \mathbf{v}^\top \mathbf{U} \mathbf{v} = n^{-1} \sum_\nu \mathbb{E}[(\mathbf{v}^\top \boldsymbol{\phi}(\mathbf{x}_t^\nu))^2]$, so $\mathbf{v}^\top \boldsymbol{\phi}(\mathbf{x}_t^\nu) = 0$ almost surely and hence $\mathbf{R}_t \mathbf{v} = 0$. Therefore $\mathbf{R}_t = \mathbf{R}_t \mathbf{U}^\dagger \mathbf{U}$, which proves stationarity. The displayed solution lies in the row space of $\mathbf{U}$ and is therefore the minimum-Frobenius-norm stationary solution. Subtracting its gradient equation gives the displayed parameter path; zero initialization yields (D.11). For an eigenvector $\mathbf{v}$ with eigenvalue $\lambda > 0$, this path gives $\mathbf{D}(\tau)\mathbf{v} = e^{-c_t \lambda \tau} \mathbf{D}(0)\mathbf{v}$. On $\ker(\mathbf{U})$, both $\mathbf{A}(\tau)$ and $\mathbf{A}_\star$ vanish under zero initialization, proving convergence also in the singular case. For any orthogonal projector $\mathbf{P}$ commuting with $\mathbf{U}$, $\mathbf{R}_t \mathbf{P} \neq 0$ is then equivalent to $\|\mathbf{D}(0)\mathbf{P}\|_{\mathrm{F}} > 0$. For such an active block, the squared relative error is a positive weighted sum of decaying exponentials and decreases strictly from one to zero, so every threshold $\rho \in (0, 1)$ has a positive, finite crossing time. □

**Meaning of the two target quantities.** The definitions in Section 5.1.2 follow from the dynamics. For either model, suppress $M$ and let $\mathbf{T}_t = \mathbf{R}_t^\top \mathbf{R}_t$. The normalized fraction $\ell_{\mathbf{T}_t}$ is defined for $\mathbf{T}_t \neq 0$, and all projectors in its definition are orthogonal. For any unit vector $\mathbf{v}$, the initial change in the corresponding output weights $\mathbf{A}\mathbf{v}$ satisfies

$$\dot{\mathbf{A}}(0)\mathbf{v} = c_t \sqrt{p}\, \mathbf{R}_t \mathbf{v}, \qquad \left\| \dot{\mathbf{A}}(0)\mathbf{v} \right\|^2 = c_t^2 p \, \|\mathbf{R}_t \mathbf{v}\|^2 = c_t^2 p \mathbf{v}^\top \mathbf{T}_t \mathbf{v}.$$

Thus $\mathbf{T}_t$ measures the squared initial update in any feature direction, after removing the common factor $c_t^2 p$. In particular, for the coordinate vector $\mathbf{e}_i$,

$$[\mathbf{T}_t]_{ii} = \sum_{a=1}^{d} [\mathbf{R}_t]_{ai}^2, \qquad \left\| \dot{\mathbf{A}}(0)\mathbf{e}_i \right\|^2 = c_t^2 p [\mathbf{T}_t]_{ii}.$$

For an orthogonal projector $\mathbf{P}$ with orthonormal basis matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$, write $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$. Then

$$\mathrm{Tr}(\mathbf{P}\mathbf{T}_t) = \mathrm{Tr}(\mathbf{V}^\top \mathbf{T}_t \mathbf{V}) = \sum_{j=1}^{r} \|\mathbf{R}_t \mathbf{v}_j\|^2 = \|\mathbf{R}_t \mathbf{P}\|_{\mathrm{F}}^2 .$$

Taking $\mathbf{P} = \mathbf{I}_p$ gives the denominator in Eq. (13). Orthogonality gives

$$\|\mathbf{R}_t\|_{\mathrm{F}}^2 = \|\mathbf{R}_t\mathbf{P}\|_{\mathrm{F}}^2 + \|\mathbf{R}_t(\mathbf{I}_p - \mathbf{P})\|_{\mathrm{F}}^2,$$

so the ratio lies in $[0, 1]$. This also explains why $\alpha_S$ bounds the fraction outside $S$ and $\alpha_N$ the fraction inside $N$. At zero initialization, $\dot{\mathbf{A}}(0) = c_t\sqrt{p}\,\mathbf{R}_t$, hence

$$\ell_{\mathbf{T}_t}(\mathbf{P}) = \frac{\left\|\dot{\mathbf{A}}(0)\mathbf{P}\right\|_{\mathrm{F}}^2}{\left\|\dot{\mathbf{A}}(0)\right\|_{\mathrm{F}}^2} = \frac{\mathrm{Tr}(\mathbf{P}\mathbf{T}_t)}{\mathrm{Tr}(\mathbf{T}_t)}.$$

Thus the same ratio measures the fraction of the squared initial update norm in the subspace. It labels the parameter directions by their target term. The actual convergence times also depend on $\mathbf{U}$.

**Independent evolution and nonzero learning.** The learning-time comparison requires the weights associated with $S$ and $N$ to evolve independently and to change during training. The projectors in Section 5.1.2 act on feature coordinates from the right of $\mathbf{A}$, while $\mathbf{P}_G$ acts on output coordinates. Choose nonzero orthogonal projectors $\mathbf{P}_S, \mathbf{P}_N$ satisfying

$$\mathbf{P}_S\mathbf{P}_N = 0, \qquad \mathbf{P}_B\mathbf{U} = \mathbf{U}\mathbf{P}_B \quad (B \in \{S, N\}).$$

They need not exhaust feature space. Commutation ensures that each subspace is invariant under $\mathbf{U}$ and gives the closed equation

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathbf{A}\mathbf{P}_B) = -c_t\big[(\mathbf{A}\mathbf{P}_B)\mathbf{U} - \sqrt{p}\,\mathbf{R}_t\mathbf{P}_B\big].$$

Thus $\mathbf{A}\mathbf{P}_B$ evolves independently of the remaining parameters. This condition also permits a subspace of a repeated eigenspace, as used by the cyclic construction.

The subspace $B$ is *active* when $\mathbf{R}_t\mathbf{P}_B \neq 0$, equivalently $\mathrm{Tr}(\mathbf{P}_B\mathbf{T}_t) > 0$. If this condition fails, the closed equation and zero initialization give $\mathbf{A}(\tau)\mathbf{P}_B = 0$ for all $\tau$. The low-contribution inequality allows this case, so activity is an additional requirement. The preceding proof shows that activity is equivalent to $\|\mathbf{D}(0)\mathbf{P}_B\|_{\mathrm{F}} > 0$ and ensures a positive, finite crossing time at every $\rho \in (0, 1)$.

In the example with two features from Section 5.1.2, the coordinate projectors $\mathrm{diag}(1, 0)$ and $\mathrm{diag}(0, 1)$ satisfy the condition for independent evolution when $\mathbf{U}$ is diagonal. Both are active because $\mathbf{R}_t = (3, -1)$ has two nonzero entries. This additional condition makes the example suitable for comparing learning times as well as contribution fractions.

**Learning times and window comparisons.** For Section 5.1.3, restore the model index $M$ in Eq. (D.10) to obtain $\mathbf{A}_{\star,M} = \sqrt{p}\,\mathbf{R}_{t,M}\mathbf{U}_M^{\dagger}$. With $\mathbf{D}_M(\tau) = \mathbf{A}_M(\tau) - \mathbf{A}_{\star,M}$, the squared relative error used in Eq. (16) is

$$E_{B,M}(\tau) = \frac{\|\mathbf{D}_M(\tau)\mathbf{P}_{B,M}\|_{\mathrm{F}}^2}{\|\mathbf{D}_M(0)\mathbf{P}_{B,M}\|_{\mathrm{F}}^2}, \qquad B \in \{S, N\}. \tag{D.12}$$

The subspaces satisfy the conditions stated above, with zero initialization giving $\mathbf{D}_M(0) = -\mathbf{A}_{\star,M}$. For statements applying to either subspace, the generic notation $\tau_{B,M}(\rho) = \inf\{\tau \geq 0 : E_{B,M}(\tau) \leq \rho^2\}$ denotes the corresponding learning time, and $\bar{\tau}_{B,M} = c_t\tau_{B,M}$ its normalized version. The relative error is a weighted average of $e^{-2c_t\lambda\tau}$ over the subspace's positive Gram eigenvalues, with weights proportional to the squared initial errors in their eigendirections. Theorem D.6 gives the spectral formula and learning-time bounds. Thus both the decay rates and the initial-error weights affect the learning times. The names $\tau_{\mathrm{gen},M}$ and $\tau_{\mathrm{mem},M}$ refer to $S$ and $N$ in this parameter-convergence comparison; these definitions do not assert a test-error threshold or a sample-copying event.

Since each block's error decreases strictly, the set of times at which $S$ has reached the threshold and $N$ has not is exactly $[\tau_{\mathrm{gen},M}, \tau_{\mathrm{mem},M})$ when $\tau_{\mathrm{gen},M} < \tau_{\mathrm{mem},M}$, and is empty otherwise. The time gap in Eq. (17) satisfies

$$A_M(\rho) = \tau_{\mathrm{gen},M}(\rho)\big(\bar{W}_M(\rho) - 1\big).$$

Thus the window is nonempty exactly when $A_M > 0$, equivalently $\bar{W}_M > 1$, and its duration is $\max\{0, A_M\}$. Multiplying both learning times by the same factor multiplies $A_M$ by that factor and

leaves $\bar{W}_M$ unchanged. Likewise, $\bar{\tau} = c_t\tau$ changes the time unit but not the ratio or the cross-model log delays. Consequently, $\Delta_{\mathrm{win}} > 0$ establishes an increase in the time ratio; alone it does not establish an increase in the absolute window duration or a nonempty averaged-model window. The latter follows when also $\bar{W}_{\mathrm{base}} \geq 1$.

The cross-model comparison uses corresponding subspaces: Section 5.2 constructs common subspaces, while Section 5.3 matches model-specific subspaces through shared reference subspaces. These matching conditions give the two models' learning times a common interpretation.

### D.3 Exact spectral-subspace dynamics

**Theorem D.6** (Spectral formula and subspace decay)**.** *Let $\mathbf{U} \succeq 0$ be any model Gram with $\mathbf{D}(\tau) = \mathbf{D}(0)e^{-c_t\mathbf{U}\tau}$. Let $\{(\lambda_j, \mathbf{u}_j)\}$ be its eigensystem and $\mathbf{P}_B$ its spectral projector for an active cluster $B$ with $\|\mathbf{D}(0)\mathbf{P}_B\|_{\mathrm{F}} > 0$. Let $\nu_B$ be the probability measure assigning weight proportional to $\|\mathbf{D}(0)\mathbf{u}_j\|^2$ to each eigenvalue in $B$. Then*

$$E_B(\tau) = \int_B e^{-2c_t\lambda\tau}\,\mathrm{d}\nu_B(\lambda). \tag{D.13}$$

*If $0 < \lambda_B^- \leq \lambda \leq \lambda_B^+$ on $B$, then*

$$\frac{\log(1/\rho)}{c_t\lambda_B^+} \leq \tau_B(\rho) \leq \frac{\log(1/\rho)}{c_t\lambda_B^-}. \tag{D.14}$$

*Separately, for any nonzero spectral projector $\mathbf{P}_B$, the spectral-decay crossing time in* (B.28)*, with $\mathbf{Q} = \mathbf{P}_B$, satisfies the same endpoint bounds after the change of variables $\bar{\tau} = c_t\tau$. This statement does not require activity.*

*Proof.* Diagonalize $\mathbf{U}$ in the displayed parameter path. Orthogonality gives (D.13). Bound the integrand between the exponentials at the two cluster endpoints and solve for the amplitude threshold $\rho$. For the spectral quantity, the same diagonalization gives a uniform weight on the eigenvalues in $\mathbf{P}_B$. Applying the same endpoint bounds proves the last statement without using $\mathbf{D}(0)$. □

**Corollary D.7** (Time-gap bounds under an explicit matrix error)**.** *The paper uses the following deterministic implication. Let $\mathbf{U}$ be the exact model Gram, let $\mathbf{U}_\star$ be symmetric, and set $\varepsilon_\star = \|\mathbf{U} - \mathbf{U}_\star\|_{\mathrm{op}}$. Suppose $\mathbf{U}_\star$ has comparison clusters $I_S = [a_S^\star, b_S^\star]$ and $I_N = [a_N^\star, b_N^\star]$. For each $B \in \{S, N\}$, the cluster is separated by more than $2\varepsilon_\star$ from the rest of $\mathrm{spec}(\mathbf{U}_\star)$, with $a_B^\star > \varepsilon_\star$, and the corresponding exact-Gram cluster is active. For any $\rho \in (0,1)$, define $\bar{W}(\rho) = \bar{\tau}_{\mathrm{mem}}(\rho)/\bar{\tau}_{\mathrm{gen}}(\rho) = \tau_{\mathrm{mem}}(\rho)/\tau_{\mathrm{gen}}(\rho)$. Then the corresponding exact-Gram clusters satisfy*

$$\frac{\log(1/\rho)}{c_t(b_B^\star + \varepsilon_\star)} \leq \tau_B(\rho) \leq \frac{\log(1/\rho)}{c_t(a_B^\star - \varepsilon_\star)}, \tag{D.15}$$

$$\frac{a_S^\star - \varepsilon_\star}{b_N^\star + \varepsilon_\star} \leq \bar{W}(\rho) \leq \frac{b_S^\star + \varepsilon_\star}{a_N^\star - \varepsilon_\star}. \tag{D.16}$$

*Proof.* Weyl's inequality places each corresponding exact cluster inside $[a_B^\star - \varepsilon_\star, b_B^\star + \varepsilon_\star]$, and the gap condition isolates it. Apply Theorem D.6; divide the lower bound for $\tau_{\mathrm{mem}}$ by the upper bound for $\tau_{\mathrm{gen}}$, and conversely. □

**Corollary D.8** (Positive cross-model regime gap under explicit error bounds)**.** *Suppose both the orbit-averaged and unaveraged models satisfy Corollary D.7. For each model, set $\underline{\lambda}_{B,M} = a_{B,M}^\star - \varepsilon_{\star,M}$ and $\overline{\lambda}_{B,M} = b_{B,M}^\star + \varepsilon_{\star,M}$, and define $s_0, d_0$ by* (22)*. Then, for every common threshold $\rho \in (0,1)$,*

$$|\Delta_S(\rho)| \leq s_0, \qquad \Delta_N(\rho) \geq d_0. \tag{D.17}$$

*If additionally*

$$d_0 > s_0, \tag{D.18}$$

*then $\Delta_{\mathrm{win}}(\rho) > 0$ for every common threshold $\rho \in (0,1)$. Equivalently, the within-model ratio satisfies $\bar{W}_{\mathrm{avg}}(\rho) > \bar{W}_{\mathrm{base}}(\rho)$, but the conclusion supported by the bounds is the directional pair $(\Delta_S, \Delta_N)$. The algebra permits model-specific spectral subspaces; interpreting them as the same directions additionally requires condition (iv) of Theorem D.9.*

*Proof.* Corollary D.7 gives the endpoint bounds used in (22). Hence $|\Delta_S| \le s_0$ and $\Delta_N \ge d_0$. If $d_0 > s_0$, then $\Delta_{\mathrm{win}} = \Delta_N - \Delta_S \ge d_0 - s_0 > 0$. □

### D.4 Proof of the finite-group separation result

**Theorem D.9** (Two-model comparison)**.** *Consider any pair of continuous-time frozen-feature linear models with a common feature dimension and normalized time, indexed for convenience by $M \in \{\mathrm{avg}, \mathrm{base}\}$. Let their exact Grams be $\mathbf{U}_M$ and let $\mathbf{T}_M \neq 0$ be positive-semidefinite target-contribution matrices used only to label the two spectral subspaces. Choose nonzero orthogonal common-reference projectors $\mathbf{Q}_S, \mathbf{Q}_N$, with $\mathbf{Q}_S\mathbf{Q}_N = 0$, overlap thresholds $\pi_S, \pi_N \in (1/2, 1]$, and a symmetric comparison matrix $\mathbf{U}^\star_M$ with error radius $\varepsilon_M$ for each model. For $B \in \{S, N\}$, let $I^\star_{B,M} = [a^\star_{B,M}, b^\star_{B,M}]$ be disjoint comparison clusters, with projectors $\mathbf{P}^\star_{B,M}$. Set $\underline{\lambda}_{B,M} = a^\star_{B,M} - \varepsilon_M$ and $\overline{\lambda}_{B,M} = b^\star_{B,M} + \varepsilon_M$. We impose:*

(i) operator control: $\|\mathbf{U}_M - \mathbf{U}^\star_M\|_{\mathrm{op}} \le \varepsilon_M$;

(ii) spectral isolation: *each cluster has gap $\gamma^\star_{B,M} > 2\varepsilon_M$ and $a^\star_{B,M} - \varepsilon_M > 0$; write $\mathbf{P}_{B,M}$ for the corresponding exact-Gram projector;*

(iii) target contribution: *for fixed $0 \le \alpha_S, \alpha_N < 1/2$, either the exact projectors satisfy*

$$\ell_{\mathbf{T}_M}(\mathbf{P}_{S,M}) \ge 1 - \alpha_S, \qquad \ell_{\mathbf{T}_M}(\mathbf{P}_{N,M}) \le \alpha_N,$$

*or, with $\delta^{\mathrm{proj}}_{B,M} = 2\varepsilon_M/\gamma^\star_{B,M}$, the comparison projectors satisfy*

$$0 \le \delta^{\mathrm{proj}}_{S,M} \le \alpha_S, \qquad 0 \le \delta^{\mathrm{proj}}_{N,M} \le \alpha_N,$$
$$\ell_{\mathbf{T}_M}(\mathbf{P}^\star_{S,M}) \ge 1 - \alpha_S + \delta^{\mathrm{proj}}_{S,M},$$
$$\ell_{\mathbf{T}_M}(\mathbf{P}^\star_{N,M}) \le \alpha_N - \delta^{\mathrm{proj}}_{N,M}.$$

(iv) common references: *for every $B, M$, $\mathrm{prec}(\mathbf{P}_{B,M} \mid \mathbf{Q}_B) \ge \pi_B$ and $\mathrm{rec}(\mathbf{P}_{B,M} \mid \mathbf{Q}_B) \ge \pi_B$;*

(v) positive margin: *$s_0, d_0$ are defined by* (22) *and satisfy $d_0 > s_0$.*

*If all four exact spectral subspaces are active, then for every common $\rho \in (0, 1)$,*

$$|\Delta_S(\rho)| \le s_0, \qquad \Delta_N(\rho) \ge d_0, \qquad \Delta_{\mathrm{win}}(\rho) \ge d_0 - s_0 > 0.$$

*Equivalently, $\bar{W}_{\mathrm{avg}}(\rho) > \bar{W}_{\mathrm{base}}(\rho)$. For initialization-independent decay on the exact spectral subspace, define*

$$\widetilde{\Delta}^{\mathrm{exact}}_B(\rho) = \log \frac{\widetilde{\tau}_{\mathbf{P}_{B,\mathrm{avg}},\mathrm{avg}}(\rho)}{\widetilde{\tau}_{\mathbf{P}_{B,\mathrm{base}},\mathrm{base}}(\rho)}, \qquad \widetilde{\Delta}^{\mathrm{exact}}_{\mathrm{win}}(\rho) = \widetilde{\Delta}^{\mathrm{exact}}_N(\rho) - \widetilde{\Delta}^{\mathrm{exact}}_S(\rho).$$

*When each spectral subspace has nonzero rank, the same bounds hold with $(\Delta_S, \Delta_N, \Delta_{\mathrm{win}})$ replaced by the corresponding tilded quantities.*

*Proof of Theorem D.9.* Assume $\|\mathbf{U}_M - \mathbf{U}^\star_M\|_{\mathrm{op}} \le \varepsilon_M$ for $M \in \{\mathrm{avg}, \mathrm{base}\}$. Fix a spectral cluster $B \in \{S, N\}$ and a model $M$. Proposition C.9 applied with $(\mathbf{U}, \mathbf{U}_\star, \varepsilon_\star, \Delta_\star) = (\mathbf{U}_M, \mathbf{U}^\star_M, \varepsilon_M, \gamma^\star_{B,M})$ shows that the corresponding exact-Gram cluster is isolated, has spectrum contained in

$$[a^\star_{B,M} - \varepsilon_M,\ b^\star_{B,M} + \varepsilon_M],$$

and satisfies

$$\left\|\mathbf{P}_{B,M} - \mathbf{P}^\star_{B,M}\right\|_{\mathrm{op}} \le \delta^{\mathrm{proj}}_{B,M} = \frac{2\varepsilon_M}{\gamma^\star_{B,M}}.$$

The endpoint assumptions give nonnegative bounded intervals and the positive lower endpoint $a^\star_{B,M} - \varepsilon_M > 0$, so the decay-time bounds below are finite.

If target contribution is checked directly on the exact projectors, condition (iii) already gives the required labels. Otherwise, Proposition C.10 and the comparison slack give

$$\ell_{\mathbf{T}_M}(\mathbf{P}_{S,M}) \ge \ell_{\mathbf{T}_M}(\mathbf{P}^\star_{S,M}) - \left\|\mathbf{P}_{S,M} - \mathbf{P}^\star_{S,M}\right\|_{\mathrm{op}} \ge 1 - \alpha_S,$$

and

$$\ell_{\mathbf{T}_M}(\mathbf{P}_{N,M}) \le \ell_{\mathbf{T}_M}(\mathbf{P}^\star_{N,M}) + \left\|\mathbf{P}_{N,M} - \mathbf{P}^\star_{N,M}\right\|_{\mathrm{op}} \le \alpha_N.$$

These inequalities identify the exact-Gram high-contribution and low-contribution spectral subspaces before any timing comparison is made. Condition (iv) of Theorem D.9 fixes their common cross-model meaning. If this overlap is checked on comparison projectors, (C.13) supplies the rank-aware slack; this is not an additional timing inequality.

Now apply Corollary D.7 to each corresponding exact spectral subspace. For parameter convergence, the required condition is precisely $\|\mathbf{D}_M(0)\mathbf{P}_{B,M}\|_{\mathrm{F}} > 0$, so Theorem D.6 applies. Thus for every common $\rho \in (0,1)$,

$$\frac{\log(1/\rho)}{b^\star_{B,M} + \varepsilon_M} \le \bar\tau_{B,M}(\rho) \le \frac{\log(1/\rho)}{a^\star_{B,M} - \varepsilon_M}.$$

For the exact-subspace spectral-decay quantity, the same display holds with $\bar\tau_{B,M}$ replaced by $\widetilde\tau_{\mathbf{P}_{B,M},M}$. The following ratio argument therefore applies separately to parameter convergence and initialization-independent spectral decay. Taking the two possible high-contribution ratios gives

$$\Delta_S(\rho) \le \log \frac{b^\star_{S,\mathrm{base}} + \varepsilon_{\mathrm{base}}}{a^\star_{S,\mathrm{avg}} - \varepsilon_{\mathrm{avg}}}, \qquad -\Delta_S(\rho) \le \log \frac{b^\star_{S,\mathrm{avg}} + \varepsilon_{\mathrm{avg}}}{a^\star_{S,\mathrm{base}} - \varepsilon_{\mathrm{base}}},$$

and therefore $|\Delta_S(\rho)| \le s_0$. The low-contribution bounds give

$$\Delta_N(\rho) = \log \frac{\bar\tau_{\mathrm{mem,avg}}(\rho)}{\bar\tau_{\mathrm{mem,base}}(\rho)} \ge \log \frac{a^\star_{N,\mathrm{base}} - \varepsilon_{\mathrm{base}}}{b^\star_{N,\mathrm{avg}} + \varepsilon_{\mathrm{avg}}} = d_0.$$

Consequently

$$\Delta_{\mathrm{win}}(\rho) = \Delta_N(\rho) - \Delta_S(\rho) \ge d_0 - s_0.$$

If $d_0 > s_0$, the log increase in the learning-time ratio is strictly positive at every common active threshold. Since

$$\log \frac{\bar W_{\mathrm{avg}}(\rho)}{\bar W_{\mathrm{base}}(\rho)} = \Delta_{\mathrm{win}}(\rho),$$

this is equivalent to $\bar W_{\mathrm{avg}}(\rho) > \bar W_{\mathrm{base}}(\rho)$.

Regardless of whether the operator-error bounds come from Theorem C.2, Proposition C.7, or a direct check, the same deterministic argument holds on any event where all required conditions hold; the parameter-convergence case additionally requires the four activity conditions for these spectral subspaces. □

**Theorem D.10** (Spectral decay on an approximate projector)**.** *Let $\mathbf{U} \succeq 0$ be a model Gram and $\mathbf{P}_B$ a rank-$r$ spectral projector of $\mathbf{U}$ with eigenvalues in $[\lambda_B^-, \lambda_B^+]$, and let $\mathbf{Q}_B$ be any rank-$r$ orthogonal projector. Set $\delta_B = \|\mathbf{P}_B - \mathbf{Q}_B\|_{\mathrm{op}}$ and*

$$\bar R_{\mathbf{Q}_B}(\bar\tau) = \frac{1}{r}\,\mathrm{Tr}\big(\mathbf{Q}_B e^{-2\mathbf{U}\bar\tau}\big)\,.$$

*Define its normalized crossing time by*

$$\bar\tau_{\mathbf{Q}_B}(\rho) = \inf\{\bar\tau \ge 0 : \bar R_{\mathbf{Q}_B}(\bar\tau) \le \rho^2\}, \qquad \rho \in (0,1).$$

*Then*

$$\Big(e^{-2\lambda_B^+\bar\tau} - \delta_B^2\Big)_+ \le \bar R_{\mathbf{Q}_B}(\bar\tau) \le e^{-2\lambda_B^-\bar\tau} + \delta_B^2. \tag{D.19}$$

*If $0 < \lambda_B^- \le \lambda_B^+$, the lower crossing-time bound below requires $\rho^2 + \delta_B^2 < 1$, and the upper bound requires $\rho^2 > \delta_B^2$:*

$$\frac{\log(1/[\rho^2 + \delta_B^2])}{2\lambda_B^+} \le \bar\tau_{\mathbf{Q}_B}(\rho) \le \frac{\log(1/[\rho^2 - \delta_B^2])}{2\lambda_B^-}. \tag{D.20}$$

*The lower and upper conditions are logically separate. In particular, bounding the upper crossing at a small threshold requires projector error on the $\rho^2$ scale.*

*Proof.* Let $\mathbf{E}_{\bar\tau} = e^{-2\mathbf{U}\bar\tau}$, so $0 \preceq \mathbf{E}_{\bar\tau} \preceq \mathbf{I}$. Because $\mathbf{P}_B$ is spectral, it commutes with $\mathbf{E}_{\bar\tau}$. Writing $\mathbf{P}_B^\perp = \mathbf{I} - \mathbf{P}_B$,

$$\begin{aligned}\operatorname{Tr}[(\mathbf{Q}_B - \mathbf{P}_B)\mathbf{E}_{\bar\tau}] = &- \operatorname{Tr}[\mathbf{P}_B(\mathbf{I} - \mathbf{Q}_B)\mathbf{P}_B\mathbf{E}_{\bar\tau}] \\ &+ \operatorname{Tr}[\mathbf{P}_B^\perp\mathbf{Q}_B\mathbf{P}_B^\perp\mathbf{E}_{\bar\tau}].\end{aligned}$$

The two displayed traces are nonnegative and each is at most $r\delta_B^2$: equal-rank projector geometry gives $\|\mathbf{P}_B(\mathbf{I} - \mathbf{Q}_B)\mathbf{P}_B\|_{\mathrm{op}} = \delta_B^2$, and $\operatorname{Tr}[\mathbf{P}_B^\perp\mathbf{Q}_B\mathbf{P}_B^\perp] = \operatorname{Tr}[\mathbf{P}_B(\mathbf{I} - \mathbf{Q}_B)\mathbf{P}_B]$. Hence

$$\left|\frac{1}{r}\operatorname{Tr}[(\mathbf{Q}_B - \mathbf{P}_B)\mathbf{E}_{\bar\tau}]\right| \le \delta_B^2.$$

The exact spectral-decay quantity on $\mathbf{P}_B$ lies between $e^{-2\lambda_B^+\bar\tau}$ and $e^{-2\lambda_B^-\bar\tau}$. This proves (D.19); solving the two threshold inequalities gives (D.20). □

Parameter convergence on an approximate projector depends on both the projector error and the scaled error factor

$$r_B = \|\mathbf{P}_B - \mathbf{Q}_B\|_{\mathrm{op}} \frac{\|\mathbf{D}(0)\|_{\mathrm{F}}}{\|\mathbf{D}(0)\mathbf{P}_B\|_{\mathrm{F}}}, \tag{D.21}$$

The following theorem quantifies this dependence. By contrast, the spectral-decay quantity incurs the quadratic additive error $\|\mathbf{P}_B - \mathbf{Q}_B\|_{\mathrm{op}}^2$.

For parameter-error energy, whenever $\|\mathbf{D}(0)\mathbf{Q}_B\|_{\mathrm{F}} > 0$, define

$$E_{\mathbf{Q}_B}(\tau) = \frac{\|\mathbf{D}(\tau)\mathbf{Q}_B\|_{\mathrm{F}}^2}{\|\mathbf{D}(0)\mathbf{Q}_B\|_{\mathrm{F}}^2}, \qquad \tau_{\mathbf{Q}_B}(\varepsilon) = \inf\{\tau \ge 0 : E_{\mathbf{Q}_B}(\tau) \le \varepsilon^2\}, \quad \varepsilon \in (0,1).$$

These are unscaled convergence times; their normalized version is $\bar\tau_{\mathbf{Q}_B} = c_t\tau_{\mathbf{Q}_B}$.

**Theorem D.11** (Parameter-error energy on an approximate projector)**.** *Let $\mathbf{U} \succeq 0$ be a model Gram with $\mathbf{D}(\tau) = \mathbf{D}(0)e^{-c_t\mathbf{U}\tau}$, and let $\mathbf{P}_B$ be a spectral projector of $\mathbf{U}$ with eigenvalues in $[\lambda_B^-, \lambda_B^+]$, assume $\|\mathbf{D}(0)\mathbf{P}_B\|_{\mathrm{F}} > 0$, and let $\mathbf{Q}_B$ be any orthogonal projector. Set $\delta_B = \|\mathbf{P}_B - \mathbf{Q}_B\|_{\mathrm{op}}$ and $r_B$ as in* (D.21)*. If $r_B < 1$, then*

$$\left(\frac{(e^{-c_t\lambda_B^+\tau} - r_B)_+}{1 + r_B}\right)^2 \le \frac{\|\mathbf{D}(\tau)\mathbf{Q}_B\|_{\mathrm{F}}^2}{\|\mathbf{D}(0)\mathbf{Q}_B\|_{\mathrm{F}}^2} \le \left(\frac{e^{-c_t\lambda_B^-\tau} + r_B}{1 - r_B}\right)^2. \tag{D.22}$$

*If additionally $0 < \lambda_B^- \le \lambda_B^+$, $\varepsilon(1+r_B)+r_B < 1$ and $\varepsilon(1-r_B)-r_B > 0$, the approximate-projector crossing time obeys*

$$\frac{\log(1/[\varepsilon(1 + r_B) + r_B])}{c_t\lambda_B^+} \le \tau_{\mathbf{Q}_B}(\varepsilon) \le \frac{\log(1/[\varepsilon(1 - r_B) - r_B])}{c_t\lambda_B^-}. \tag{D.23}$$

*Proof.* Let $a_\tau = \|\mathbf{D}(\tau)\mathbf{P}_B\|_{\mathrm{F}}$, $b_\tau = \|\mathbf{D}(\tau)\mathbf{Q}_B\|_{\mathrm{F}}$, and $F_0 = \|\mathbf{D}(0)\|_{\mathrm{F}}$. Since $e^{-c_t\mathbf{U}\tau}$ is a contraction,

$$|a_\tau - b_\tau| \le \|\mathbf{D}(\tau)(\mathbf{P}_B - \mathbf{Q}_B)\|_{\mathrm{F}} \le \delta_B F_0.$$

The same bound holds at $\tau = 0$. Writing $r_B = \delta_B F_0/a_0$ gives

$$\frac{(a_\tau/a_0 - r_B)_+}{1 + r_B} \le \frac{b_\tau}{b_0} \le \frac{a_\tau/a_0 + r_B}{1 - r_B}.$$

Theorem D.6 bounds $a_\tau/a_0$ between $e^{-c_t\lambda_B^+\tau}$ and $e^{-c_t\lambda_B^-\tau}$. Squaring proves (D.22); solving the two threshold inequalities gives (D.23). □

# E COMPACT-TORUS APPROXIMATION

Let $K$ be a compact group with a continuous orthogonal representation $\varrho_d : K \to O(d)$. Its normalized Haar measure $\mu_K$ is a probability measure invariant under left and right group multiplication (Bekka et al., 2008). For features $f(\mathbf{z}) \in \mathbb{R}^p$ satisfying $\int_K \|f(\varrho_d(k)\mathbf{z})\| \, \mathrm{d}\mu_K(k) < \infty$ for every $\mathbf{z}$, the compact analogues of the averages in Section 3.2 are

$$(\mathrm{Avg}_K f)(\mathbf{z}) = \int_K f(\varrho_d(k)\mathbf{z}) \, \mathrm{d}\mu_K(k),$$
$$\mathbf{P}_K = \int_K \varrho_d(k) \, \mathrm{d}\mu_K(k).$$

Right invariance of Haar measure makes $\mathrm{Avg}_K f$ constant on input orbits. The matrix $\mathbf{P}_K$ sends every vector into $\mathrm{Fix}(K) = \{\mathbf{v} : \varrho_d(k)\mathbf{v} = \mathbf{v} \text{ for all } k \in K\}$ and fixes every vector in that space. Invariance under inversion gives $\mathbf{P}_K^\top = \mathbf{P}_K$, so this is the orthogonal projector onto $\mathrm{Fix}(K)$.

The compact extension directly controls the nonlinear feature integral and the cross-moment matrix of targets and features. Agreement of the linear representation's fixed space alone does not guarantee nonlinear accuracy.

For the torus $K = \mathbb{T}^r = [0, 2\pi)^r$, normalized Haar measure is $\mathrm{d}\mu_K(\boldsymbol{\theta}) = (2\pi)^{-r}\mathrm{d}\boldsymbol{\theta}$. Assume the orthogonal representation $\varrho_d(\boldsymbol{\theta})$ is continuously differentiable. For the tensor-grid subgroup

$$\Gamma_m = \left\{ \frac{2\pi}{m}(j_1, \ldots, j_r) : 0 \le j_\ell < m \right\},$$

define

$$\phi_{K,i}(\mathbf{z}) = \int_K \sigma\left( \frac{\mathbf{w}_i^\top \varrho_d(\boldsymbol{\theta})\mathbf{z}}{\sqrt{d}} \right) \mathrm{d}\mu_K(\boldsymbol{\theta}),$$
$$\phi_i^{(m)}(\mathbf{z}) = \frac{1}{|\Gamma_m|} \sum_{\boldsymbol{\theta} \in \Gamma_m} \sigma\left( \frac{\mathbf{w}_i^\top \varrho_d(\boldsymbol{\theta})\mathbf{z}}{\sqrt{d}} \right).$$

Let

$$L_\varrho = \sum_{\ell=1}^{r} \sup_{\boldsymbol{\theta} \in K} \left\| \partial_{\theta_\ell} \varrho_d(\boldsymbol{\theta}) \right\|_{\mathrm{op}}, \qquad \varepsilon_{\phi,m} = \frac{\pi L_\sigma L_\varrho}{m\sqrt{d}} \|\mathbf{W}\|_{\mathrm{F}}.$$

**Proposition E.1** (Feature-level grid error)**.** *Suppose $|\sigma| \le B_\sigma$ and $\sigma$ is $L_\sigma$-Lipschitz. Then for every $\mathbf{z} \in \mathbb{R}^d$,*

$$\left\| \boldsymbol{\phi}^{(m)}(\mathbf{z}) - \boldsymbol{\phi}_K(\mathbf{z}) \right\| \le \varepsilon_{\phi,m} \|\mathbf{z}\|. \tag{E.1}$$

*Proof.* Partition the torus into the $m^r$ periodic cells centered at the grid points. Every coordinate of a point in a cell differs from its center by at most $\pi/m$. For row $i$, the derivative in coordinate $\ell$ has magnitude at most $L_\sigma \|\mathbf{w}_i\| \|\mathbf{z}\| \left\|\partial_{\theta_\ell} \varrho_d\right\|_{\mathrm{op}} / \sqrt{d}$. The fundamental theorem of calculus bounds the error within each cell by the sum of the coordinate displacements. Averaging over cells and summing the squared rowwise bounds proves the result. □

Let $\mathbf{U}_K$ and $\mathbf{U}^{(m)}$ be the Haar and grid versions of the empirical noise-averaged Gram. In the compact comparison, grid, Haar, and Base-RFM all use the common target $\mathbf{y}_{t,K}^\nu = \mathbf{P}_K \mathbf{y}_t^\nu$; the following bounds in fact hold for any square-integrable common target. Define

$$\mathbf{R}_K = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}[\mathbf{y}_{t,K}^\nu \boldsymbol{\phi}_K(\mathbf{x}_t^\nu)^\top], \qquad \mathbf{R}^{(m)} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}[\mathbf{y}_{t,K}^\nu \boldsymbol{\phi}^{(m)}(\mathbf{x}_t^\nu)^\top],$$
$$\mathbf{T}_K = (\mathbf{R}_K)^\top \mathbf{R}_K, \qquad \mathbf{T}^{(m)} = (\mathbf{R}^{(m)})^\top \mathbf{R}^{(m)}.$$

Set

$$\overline{R}_{t,n} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \|\mathbf{x}_t^\nu\|, \quad \overline{Y}_{t,n} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \left\|\mathbf{y}_{t,K}^\nu\right\|, \quad \overline{Q}_{t,n} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}[\left\|\mathbf{y}_{t,K}^\nu\right\| \|\mathbf{x}_t^\nu\|].$$

**Theorem E.2** (Direct Grid-to-Haar Bounds)**.** *Under Proposition E.1,*

$$\left\|\mathbf{U}^{(m)} - \mathbf{U}_K\right\|_{\mathrm{op}} \le \varepsilon^U_{K,m} \stackrel{\mathrm{def}}{=} 2B_\sigma\sqrt{p}\,\varepsilon_{\phi,m}\overline{R}_{t,n}, \tag{E.2}$$

$$\left\|\mathbf{R}^{(m)} - \mathbf{R}_K\right\|_{\mathrm{F}} \le \varepsilon^R_{K,m} \stackrel{\mathrm{def}}{=} \varepsilon_{\phi,m}\overline{Q}_{t,n}, \tag{E.3}$$

$$\left\|\mathbf{T}^{(m)} - \mathbf{T}_K\right\|_{*} \le \varepsilon^T_{K,m} \stackrel{\mathrm{def}}{=} 2B_\sigma\sqrt{p}\,\overline{Y}_{t,n}\varepsilon^R_{K,m}. \tag{E.4}$$

*Consequently any grid cluster indexed by $B$ whose gap from the rest of the grid spectrum exceeds $2\varepsilon^U_{K,m}$ has a corresponding Haar cluster satisfying the Weyl and Davis–Kahan conclusions of Proposition C.9. Let $\mathbf{P}^{(m)}_B$ be that grid projector and $\mathbf{P}_{B,K}$ the corresponding Haar projector. If $\mathbf{T}_K$ and $\mathbf{T}^{(m)}$ are nonzero, then*

$$\left|\ell_{\mathbf{T}_K}(\mathbf{P}_{B,K}) - \ell_{\mathbf{T}^{(m)}}(\mathbf{P}^{(m)}_B)\right| \le \left\|\mathbf{P}_{B,K} - \mathbf{P}^{(m)}_B\right\|_{\mathrm{op}} + \frac{2\varepsilon^T_{K,m}}{\min\{\mathrm{Tr}(\mathbf{T}_K), \mathrm{Tr}(\mathbf{T}^{(m)})\}}. \tag{E.5}$$

*Thus the spectrum, target contribution, and decay time each need their own Gram or target error bound. These results do not imply a closed-form Gaussian spectral edge or learning-time ratio for the compact model.*

*Proof.* For vectors $\mathbf{a}, \mathbf{b}$, $\left\|\mathbf{a}\mathbf{a}^\top - \mathbf{b}\mathbf{b}^\top\right\|_{\mathrm{op}} \le (\|\mathbf{a}\| + \|\mathbf{b}\|)\,\|\mathbf{a} - \mathbf{b}\|$. Bound both feature norms by $B_\sigma\sqrt{p}$, apply (E.1) to every noised sample, and average to obtain (E.2). The same feature bound gives

$$\left\|\mathbf{R}^{(m)} - \mathbf{R}_K\right\|_{\mathrm{F}} \le \frac{1}{n}\sum_\nu \mathbb{E}[\left\|\mathbf{y}^\nu_{t,K}\right\| \left\|\boldsymbol{\phi}^{(m)} - \boldsymbol{\phi}_K\right\|] \le \varepsilon_{\phi,m}\overline{Q}_{t,n}.$$

Moreover $\|\mathbf{R}_K\|_{\mathrm{F}}, \left\|\mathbf{R}^{(m)}\right\|_{\mathrm{F}} \le B_\sigma\sqrt{p}\,\overline{Y}_{t,n}$. Expanding $\mathbf{T}^{(m)} - \mathbf{T}_K$ and using $\left\|\mathbf{A}^\top\mathbf{B}\right\|_* \le \|\mathbf{A}\|_{\mathrm{F}}\,\|\mathbf{B}\|_{\mathrm{F}}$ proves (E.4).

For the target-contribution bound, first change the projector while holding $\mathbf{T}_K$ fixed, using Proposition C.10. For any projector $\mathbf{P}$ and nonzero positive semidefinite $\mathbf{S}, \mathbf{T}$,

$$\left|\frac{\mathrm{Tr}(\mathbf{P}\mathbf{S})}{\mathrm{Tr}(\mathbf{S})} - \frac{\mathrm{Tr}(\mathbf{P}\mathbf{T})}{\mathrm{Tr}(\mathbf{T})}\right| \le \frac{2\,\|\mathbf{S} - \mathbf{T}\|_*}{\min\{\mathrm{Tr}(\mathbf{S}), \mathrm{Tr}(\mathbf{T})\}}.$$

Apply this with $(\mathbf{S}, \mathbf{T}) = (\mathbf{T}_K, \mathbf{T}^{(m)})$. The spectral and timing conclusions follow from Proposition C.9 and Theorem D.6. □

*Proof of Corollary 5.5.* Proposition E.1 gives the $O(m^{-1})$ feature bound. Theorem E.2 converts it into the stated bounds for the Gram matrix, cross-moment matrix, target-contribution matrix, and projectors. Adding these errors to the tensor-grid conditions verifies the hypotheses of Theorem E.3, which gives the Haar conclusion. □

### E.1 Proof for the Compact Torus

**Theorem E.3** (Compact-Torus Conditions from a Tensor Grid)**.** *Let $K = \mathbb{T}^r$ act through the continuously differentiable orthogonal representation above. Haar-RFM, tensor-grid RFM, and Base-RFM use the common target $\mathbf{Y}_{t,K} = \mathbf{P}_K\mathbf{Y}_t$. Define*

$$\mathbf{R}_{\mathrm{base}} = \frac{1}{n}\sum_{\nu=1}^n \mathbb{E}_{\boldsymbol{\xi}}[\mathbf{y}^\nu_{t,K}\boldsymbol{\phi}_{\mathrm{base}}(\mathbf{x}^\nu_t)^\top], \qquad \mathbf{T}_{\mathrm{base}} = \mathbf{R}^\top_{\mathrm{base}}\mathbf{R}_{\mathrm{base}},$$

*and assume $\mathbf{T}_K, \mathbf{T}^{(m)}, \mathbf{T}_{\mathrm{base}} \neq 0$.*

*Use the tensor-grid Gram itself as the Haar comparison matrix:*

$$\mathbf{U}^\star_K = \mathbf{U}^{(m)}, \qquad \varepsilon_K = \varepsilon^U_{K,m}, \qquad \mathbf{P}^\star_{B,K} = \mathbf{P}^{(m)}_B.$$

*Suppose it has two disjoint clusters $I^\star_{B,K} = [a^\star_{B,K}, b^\star_{B,K}]$, $B \in \{S, N\}$, each with gap $\gamma^\star_{B,K} > 2\varepsilon_K$ and $a^\star_{B,K} > \varepsilon_K$. Let $\mathbf{P}_{B,K}$ be the corresponding Haar projector and set*

$$\delta^{\mathrm{target}}_{B,K} = \left\|\mathbf{P}_{B,K} - \mathbf{P}^{(m)}_B\right\|_{\mathrm{op}} + \frac{2\varepsilon^T_{K,m}}{\min\{\mathrm{Tr}(\mathbf{T}_K), \mathrm{Tr}(\mathbf{T}^{(m)})\}}.$$

*For fixed* $0 \le \alpha_S, \alpha_N < 1/2$*, assume*

$$\ell_{\mathbf{T}^{(m)}}(\mathbf{P}_S^{(m)}) \ge 1 - \alpha_S + \delta_{S,K}^{\text{target}}, \qquad \ell_{\mathbf{T}^{(m)}}(\mathbf{P}_N^{(m)}) \le \alpha_N - \delta_{N,K}^{\text{target}}.$$

*For Base-RFM, choose a comparison matrix* $\mathbf{U}_{\text{base}}^\star$ *and error* $\varepsilon_{\text{base}}$ *satisfying the operator and spectral conditions of Theorem D.9; write* $\mathbf{P}_{B,\text{base}}$ *for the corresponding exact projectors and require their exact target contributions to be at least* $1 - \alpha_S$ *and at most* $\alpha_N$*, respectively. Choose nonzero orthogonal common-reference projectors* $\mathbf{Q}_S, \mathbf{Q}_N$ *and thresholds* $\pi_S, \pi_N \in (1/2, 1]$ *such that the two-way overlap conditions of Theorem D.9 hold for* $\mathbf{P}_{B,K}$ *and* $\mathbf{P}_{B,\text{base}}$*. Assume all four exact spectral subspaces are active.*

*Let* $\bar{\tau}_{B,K}(\rho)$ *and* $\bar{\tau}_{B,\text{base}}(\rho)$ *be the normalized parameter-error crossing times on these exact projectors, and define*

$$\begin{aligned}
\Delta_{B,K}(\rho) &= \log \frac{\bar{\tau}_{B,K}(\rho)}{\bar{\tau}_{B,\text{base}}(\rho)}, \\
\Delta_{\text{win},K}(\rho) &= \Delta_{N,K}(\rho) - \Delta_{S,K}(\rho), \\
\bar{W}_H(\rho) &= \frac{\bar{\tau}_{\text{mem},H}(\rho)}{\bar{\tau}_{\text{gen},H}(\rho)}, \quad H \in \{K, \text{base}\}.
\end{aligned}$$

*Define*

$$\begin{aligned}
s_K &= \max\left\{0, \log \frac{b_{S,\text{base}}^\star + \varepsilon_{\text{base}}}{a_{S,K}^\star - \varepsilon_K}, \log \frac{b_{S,K}^\star + \varepsilon_K}{a_{S,\text{base}}^\star - \varepsilon_{\text{base}}}\right\}, \\
d_K &= \log \frac{a_{N,\text{base}}^\star - \varepsilon_{\text{base}}}{b_{N,K}^\star + \varepsilon_K}.
\end{aligned} \tag{E.6}$$

*If* $d_K > s_K$*, then for every common* $\rho \in (0, 1)$*,*

$$|\Delta_{S,K}(\rho)| \le s_K, \qquad \Delta_{N,K}(\rho) \ge d_K, \qquad \Delta_{\text{win},K}(\rho) \ge d_K - s_K > 0.$$

*Equivalently,* $\bar{W}_K(\rho) > \bar{W}_{\text{base}}(\rho)$*. The exact-subspace spectral-decay analogue replaces each* $\bar{\tau}_{B,H}$ *by* $\widetilde{\tau}_{\mathbf{P}_{B,H},H}$ *and requires nonzero subspace rank instead of activity.*

*Proof.* Theorem E.2 gives $\left\|\mathbf{U}_K - \mathbf{U}^{(m)}\right\|_{\text{op}} \le \varepsilon_K$, matches each grid cluster to the exact Haar projector $\mathbf{P}_{B,K}$, and gives the following Haar target contributions from the displayed grid bounds: $\ell_{\mathbf{T}_K}(\mathbf{P}_{S,K}) \ge 1 - \alpha_S$ and $\ell_{\mathbf{T}_K}(\mathbf{P}_{N,K}) \le \alpha_N$. Thus Haar-RFM and Base-RFM satisfy the same exact-projector target conditions and all other hypotheses of Theorem D.9, with its Avg-RFM label instantiated by Haar-RFM. That theorem yields the parameter-convergence and spectral-decay conclusions. □

*Remark* E.4 (Scope). The bounds concern the chosen torus representation and tensor-grid subgroup. Dimension-free rates, arbitrary i.i.d. Haar point clouds, and a closed-form compact Gaussian edge require separate analysis. Taking $m$ above a linear harmonic cutoff may identify the linear fixed space, while control of the nonlinear activation requires the full-Gram and target-contribution bounds in Eqs. (E.2) and (E.4).

# F Experiments and Results

This section defines the quantities reported by the experiments and explains how they are computed. Parameter-convergence times in (16) use active exact spectral subspaces, whereas a common reference subspace gives the spectral-decay quantity in (B.28). Independent Monte Carlo estimates assess numerical repeatability. The finite-cyclic and compact-grid studies report spectral and numerical trends; the structured finite-group study evaluates the theorem conditions; and the QM9 study compares learned-feature training behavior in a molecular setting.

### F.1 Reported Quantities

All time quantities use $\bar{\tau} = c_t \tau$, where $c_t = 2\Delta_t d/p$. The primary amplitude threshold is $\rho = 0.1$; the sensitivity analysis uses $\rho \in \{0.3, 0.1, 0.03, 0.01\}$. Crossing times are obtained from the eigensystem by bisection to relative tolerance $10^{-8}$.

For model-specific exact spectral projectors $\mathbf{P}_{B,M}$, define

$$\widetilde{\Delta}_B^{\mathrm{exact}}(\rho) = \log \frac{\widetilde{\tau}_{\mathbf{P}_{B,\mathrm{avg}},\mathrm{avg}}(\rho)}{\widetilde{\tau}_{\mathbf{P}_{B,\mathrm{base}},\mathrm{base}}(\rho)}, \qquad \widetilde{W}_M^{\mathrm{exact}}(\rho) = \frac{\widetilde{\tau}_{\mathbf{P}_{N,M},M}(\rho)}{\widetilde{\tau}_{\mathbf{P}_{S,M},M}(\rho)}. \tag{F.1}$$

The corresponding log increase in the learning-time ratio is $\widetilde{\Delta}_{\mathrm{win}}^{\mathrm{exact}}(\rho) = \widetilde{\Delta}_N^{\mathrm{exact}}(\rho) - \widetilde{\Delta}_S^{\mathrm{exact}}(\rho)$.

For active exact spectral subspaces, the parameter-convergence quantities are $\Delta_S, \Delta_N$, and $\Delta_{\mathrm{win}}$ in (18). For common references that are not exact spectral subspaces, they are the distinct $\widetilde{\Delta}_S^{\mathrm{ref}}, \widetilde{\Delta}_N^{\mathrm{ref}}$, and $\widetilde{\Delta}_{\mathrm{win}}^{\mathrm{ref}}$ of (B.28). These reference-based quantities describe spectral decay on the prespecified common subspaces.

Under the zero initialization used in the main text, the initial parameter error is $\mathbf{D}_M(0) = -\sqrt{p}\,\mathbf{R}_{t,M}^{(n)}\mathbf{U}_M^\dagger$, so the relevant feature-space weight is $\mathbf{U}_M^\dagger \mathbf{T}_{t,M}\mathbf{U}_M^\dagger$. Accordingly, we also report

$$\ell_{\star,t,M}(\mathbf{P}) = \frac{\mathrm{Tr}\Big(\mathbf{P}\mathbf{U}_M^\dagger \mathbf{T}_{t,M}\mathbf{U}_M^\dagger \mathbf{P}\Big)}{\mathrm{Tr}\Big(\mathbf{U}_M^\dagger \mathbf{T}_{t,M}\mathbf{U}_M^\dagger\Big)}, \qquad \mathbf{T}_{t,M} = (\mathbf{R}_{t,M}^{(n)})^\top \mathbf{R}_{t,M}^{(n)}.$$

This parameter-error weighting is reported separately from the initialization-independent spectral-decay measure and never selects a subspace after timing is inspected.

The finite-sample evaluation uses an error bound $\varepsilon_M \geq \|\mathbf{U}_M - \mathbf{U}_M^\star\|_{\mathrm{op}}$ with an explicit deterministic guarantee or confidence level. It then computes $s_0, d_0$ from (22). The theorem conditions hold when the spectral, target-contribution, target-error, shared-reference overlap, active-subspace, and $d_0 > s_0$ conditions in Theorem D.9 all hold. Any practical effect-size threshold is reported separately. The evaluation uses $\alpha_S = 0.1$, $\alpha_N = 0.05$, $\pi_S = 0.9$, and $\pi_N = 0.8$.

### F.2 Numerical Procedure

Within each statistical unit, Avg-RFM and Base-RFM use the same clean data and base weights. Their numerical integration and target estimates use recorded, model-specific seed offsets.

For each model, we construct two independent numerical Gram estimates $\widehat{\mathbf{U}}_M^{(1)}, \widehat{\mathbf{U}}_M^{(2)}$, average them, and compute their global discrepancy

$$E_{\mathrm{rep},M} = \frac{1}{2}\left\|\widehat{\mathbf{U}}_M^{(1)} - \widehat{\mathbf{U}}_M^{(2)}\right\|_{\mathrm{op}}. \tag{F.2}$$

For any reference $\mathbf{Q}$, we also compute the restricted discrepancy

$$E_{\mathrm{rep},M}(\mathbf{Q}) = \frac{1}{2}\left\|\mathbf{Q}(\widehat{\mathbf{U}}_M^{(1)} - \widehat{\mathbf{U}}_M^{(2)})\mathbf{Q}\right\|_{\mathrm{op}}.$$

These discrepancies quantify repeatability; rigorous operator-error bounds come from the explicit theorem quantity $\varepsilon_M$. The completed finite-cyclic grid used $N_\xi = 1024$, adaptively increasing to 16384, and target batches of 8192 with $N_{\xi,T} = 512$, adaptively increasing to 4096. The active tolerance is

$$\lambda_{\mathrm{act}} = \max\{10^{-10}, 10^{-6}\lambda_{\max}(\widehat{\mathbf{U}}_M), 2E_{\mathrm{rep},M}\}. \tag{F.3}$$

Eigenvalues above $\lambda_{\mathrm{act}}$ are classified as numerically active. Each common reference must have a finite crossing time. We assess exact clusters using four repeatability criteria: the separating gap exceeds $\max\{4E_{\mathrm{rep},M}, 10^{-4}\lambda_{\max}(\widehat{\mathbf{U}}_M)\}$, ranks agree between the two estimates, endpoint relative discrepancy is below $5\%$, and projector discrepancy is below $0.05$. These criteria assess numerical resolution. The theorem's operator-error condition uses the explicit bound $\varepsilon_M$.

For the finite-group studies, the target is the projected denoising vector $\mathbf{Y}_{t,G} = -\mathbf{P}_G\boldsymbol{\xi}/\sqrt{\Delta_t}$. The compact study instead uses $\mathbf{Y}_{t,K} = -\mathbf{P}_K\boldsymbol{\xi}/\sqrt{\Delta_t}$. The finite-cyclic study estimates two independent

cross-moments $\mathbf{R}_1, \mathbf{R}_2$ and averages the two $\mathbf{R}_j^\top \mathbf{R}_j$ target-contribution matrices. The target-based reference projector is constructed from these independent cross-moment estimates. For each pair, we average the cross-moment matrices, $\bar{\mathbf{R}} = (\mathbf{R}_1 + \mathbf{R}_2)/2$, and form $\bar{\mathbf{R}}^\top \bar{\mathbf{R}}$; the estimate-specific matrices are used to quantify repeatability. Let $\mathbf{T}_{\mathrm{avg}}$ and $\mathbf{T}_{\mathrm{base}}$ denote the resulting averaged target-contribution matrices. Define their normalized common target-contribution matrix

$$\mathbf{T}_{\mathrm{common}} = \frac{1}{2}\left(\frac{\mathbf{T}_{\mathrm{avg}}}{\operatorname{Tr}\mathbf{T}_{\mathrm{avg}}} + \frac{\mathbf{T}_{\mathrm{base}}}{\operatorname{Tr}\mathbf{T}_{\mathrm{base}}}\right).$$

For the finite-cyclic study, $\mathbf{Q}_S$ is the projector onto the leading rank-$q$ eigenspace of $\mathbf{T}_{\mathrm{common}}$. To construct $\mathbf{Q}_N$, draw an independent matrix $\mathbf{X}_N \in \mathbb{R}^{N_N \times d}$ with standard Gaussian rows, where $N_N = \max\{4q, 1024\}$, and set

$$\mathbf{X}_N^{\perp} = \mathbf{X}_N(\mathbf{I} - \mathbf{P}_G), \qquad \mathbf{Z}_N = (\mathbf{I}_p - \mathbf{Q}_S)\mathbf{W}(\mathbf{X}_N^{\perp})^\top/\sqrt{d}.$$

Then $\mathbf{Q}_N$ is the projector onto the leading rank-$q$ left-singular subspace of $\mathbf{Z}_N$. The resulting reference projectors are held fixed when evaluating spectral decay.

### F.3 Experimental Results

The studies address complementary questions. Study A evaluates the two-element population formula, and Study A′ extends the calculation to cyclic groups of order up to 16. Study B measures common-reference spectral decay under a natural finite-group action. Study C examines orbit variance and its scalar reduction. Study D compares compact tensor-grid quantities with a finite reference. Study E evaluates the theorem conditions in a structured finite-group construction.

**A: Tilted-Reflection Population.** Study A uses $\mathbb{R}^d = \mathcal{S} \oplus \mathcal{N}$, $G = \{\mathbf{I}, \mathbf{J}\}$, and $\mathbf{J} = \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{N}}$. Each feature uses the mixed fixed/moved weights from Section 5.2, with block dimensions satisfying $r_S/r_N = 5$, $(\theta_S, \theta_N) = (0.05, 0.5)$, and $\alpha_S = \alpha_N = 0.1$. The activation is $\sigma_\beta(z) = \tanh z + \beta(\tanh^2 z - c_2)$, where $c_2 = \mathbb{E}[\tanh^2 Z]$ for $Z \sim \mathcal{N}(0,1)$, and $\beta \in \{0.25, 0.5, 1, 2\}$.

The dimensional settings are $(r_N, r_S, d, p) = (8, 40, 160, 48), (16, 80, 320, 96), (32, 160, 640, 192)$, so both orthonormal frames fit. Population values are computed from the Hermite expansion and Gauss–Hermite quadrature. A Monte Carlo evaluation with 200,000 samples per grid point quantifies numerical agreement; the largest discrepancies are $2.94 \cdot 10^{-3}$ for the averaged Gram value and $3.61 \cdot 10^{-3}$ for the target-contribution value. This comparison evaluates the population calculation and its numerical implementation.

Table F.1: **Tilted-reflection population quantities.**

| $\beta$ | $q(\theta_S)$ | $q(\theta_N)$ | $s_0$ | $d_0$ | $d_0 - s_0$ |
|---|---|---|---|---|---|
| 0.25 | 0.9417 | 0.5000 | 0.0601 | 0.6931 | 0.6331 |
| 0.50 | 0.9385 | 0.5000 | 0.0635 | 0.6931 | 0.6297 |
| 1.00 | 0.9281 | 0.5000 | 0.0746 | 0.6931 | 0.6185 |
| 2.00 | 0.9059 | 0.5000 | 0.0988 | 0.6931 | 0.5943 |

Values are rounded to four decimals. Here $s_0 = -\log q(\theta_S)$ and $d_0 = -\log q(\theta_N)$; the general theorem clips these endpoint values at zero.

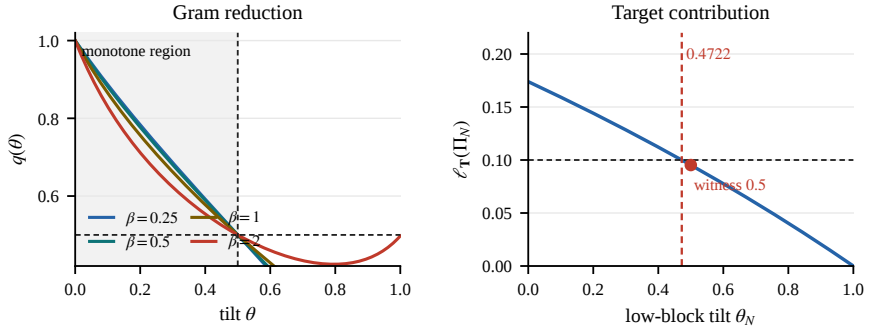

Figure F.1: **Tilted-reflection population calculation.** The plot shows the averaged-to-base Gram ratio and the tilt threshold ensuring the block conditions in Eq. (C.16). It compares the computed quantities with their population formulas.

The accompanying rerun script reproduces the reported output, and the manifest records the file hashes, making the population calculation reproducible.

**A′: Cyclic-Group Population.** Study A′ evaluates the same family on $G = C_m$ for $m \in \{2, 3, 4, 5, 8, 16\}$, with the sign character at $m = 2$ and one primitive two-dimensional rotation component for $m \geq 3$, so that $d = 2p$ and $d = 3p$ respectively. The blocks use $r_S/r_N = 8$, $(\theta_S, \theta_N) = (0.05, 0.5)$ or $\theta_N = \theta^*$, $\alpha_S = \alpha_N = 0.1$, and the same activation family $\sigma_\beta$ as Study A. Three feature dimensions $p \in \{9, 36, 144\}$, with $d \leq 432$, give 288 parameter combinations satisfying the construction constraints. Population quantities are evaluated by 200-point two-dimensional Gauss–Hermite quadrature, and a 64-order Hermite series provides an independent numerical comparison. A Monte Carlo evaluation with 200,000 samples per parameter combination has maximum discrepancies $1.113 \cdot 10^{-3}$ for the averaged Gram value and $8.419 \cdot 10^{-4}$ for the target-contribution value. These comparisons evaluate the population calculation and its numerical implementation.

Table F.2 reports the margin, and Table F.3 reports its sharp upper bound. The $m = 2$ column reproduces Table F.1 to four decimals, providing a consistency comparison for the general formula. In the selected parameter setting, the initial update fractions are $0.9383$ and $0.0617$. Across six groups, four activations, and four tilt pairs, the margin ranges from $0.3530$ to $1.0380$. At $\beta = 2$, the margin increases from $0.5943$ at $m = 2$ to $0.7885$ at $m = 16$. The $m = 16$ upper bound is $0.8905$, compared with $\log 16 = 2.7726$; the largest upper bound in the grid is $1.0986$ at $m = 3$. Thus, the improvement over the two-element case is not monotone in group order. The endpoint quantities depend only on $q_G(\theta_S)$, $q_G(\theta_N)$ and the rank ratio, and across the three scales they agree to $6.1 \cdot 10^{-15}$; the averaged Gram is diagonal to $4.3 \cdot 10^{-18}$ at $d = 432$. Two additional configurations illustrate the role of the construction assumptions. Equal invariant energy in both blocks gives zero margin. A row family that places the blocks in *different* irreducible components satisfies the target and activity conditions but has a negative margin at $\beta = 2$. This case motivates the shared correlation-profile assumption in Proposition 5.1.

Table F.2: **Cyclic-group population margin $d_0 - s_0$.**

| $\beta$ | $\kappa_\sigma$ | $m = 2$ | $m = 3$ | $m = 4$ | $m = 5$ | $m = 8$ | $m = 16$ |
|---|---|---|---|---|---|---|---|
| 0.25 | 0.9163 | 0.6331 | 0.6588 | 0.6649 | 0.6657 | 0.6660 | 0.6660 |
| 0.5 | 0.8763 | 0.6297 | 0.6664 | 0.6743 | 0.6755 | 0.6759 | 0.6759 |
| 1 | 0.7460 | 0.6185 | 0.6922 | 0.7057 | 0.7087 | 0.7096 | 0.7096 |
| 2 | 0.4677 | 0.5943 | 0.7518 | 0.7792 | 0.7863 | 0.7885 | 0.7885 |

Population values at $(\theta_S, \theta_N) = (0.05, 0.5)$, rounded to four decimals. $\kappa_\sigma = a_1^2/\lambda_{\text{base}}$ is the activation constant of (C.18). The $m = 2$ column is the two-element case in Table F.1. Values are quadrature evaluations.

Table F.3: **Population margin upper bound.**

| $\beta$ | $m = 2$ | $m = 3$ | $m = 4$ | $m = 5$ | $m = 8$ | $m = 16$ |
|---|---|---|---|---|---|---|
| 0.25 | 0.6931 | 1.0986 | 0.7255 | 0.8350 | 0.7266 | 0.7266 |
| 0.5 | 0.6931 | 1.0986 | 0.7385 | 0.8468 | 0.7401 | 0.7401 |
| 1 | 0.6931 | 1.0986 | 0.7818 | 0.8863 | 0.7857 | 0.7857 |
| 2 | 0.6931 | 1.0986 | 0.8811 | 0.9762 | 0.8904 | 0.8905 |
| $\log m$ | 0.6931 | 1.0986 | 1.3863 | 1.6094 | 2.0794 | 2.7726 |

Sharp upper bound $\log(1/\underline{q}_G)$, attained at $\theta_N = \theta^*$. It equals $\log m$ for $m \in \{2, 3\}$ and is below $\log m$ for $m \geq 4$. The similar values at $m = 8$ and $m = 16$ indicate saturation (Remark C.16).

The aligned-reflection construction provides a contrast: its $\mathcal{N}$ block is inactive, so the comparison uses the spectral-decay quantity in (B.28).

**B: Natural Finite-Cyclic Results.** Study B uses $G = C_m$ acting through $q$ regular-representation blocks, $d = qm$, $\operatorname{rank} \mathbf{P}_G = q$, and fixed-trace Gaussian covariance $\mathbf{\Sigma}(\chi) = \lambda_{\text{inv}}\mathbf{P}_G + \lambda_{\perp}(\mathbf{I} - \mathbf{P}_G)$, where $\lambda_{\perp} = m/(\chi + m - 1)$ and $\lambda_{\text{inv}} = \chi m/(\chi + m - 1)$.

The design fixes $q = 128$, $p = 512$, $n = 256$, and $t = 0.5$, and varies $m \in \{2, 4, 8, 16\}$ and $\chi \in \{1, 2, 4, 8\}$. Each cell uses 30 paired seeds. The designated cells are $(m, \chi) = (4, 8), (8, 8), (16, 8)$.

Table F.4 reports the Avg/Base comparisons on common reference subspaces. These subspaces define spectral-decay quantities for the selected feature directions. Since $\kappa = \chi/(\chi + m - 1)$ varies with $m$, the designated rows compare joint changes in group order and covariance contrast.

For each finite-cyclic run, we record target contribution, initial parameter-error share, commutator and cross-subspace measures, reference ranks, differences between independent estimates, active rank, trace ratios, reference-subspace crossing times, and numerical status. Quantities computed from compressed endpoints are reported as descriptive margins, separate from the theorem's interval separation margins.

Table F.4: **Finite-cyclic reference-subspace decay.**

| $m$ | $\kappa$ | $\widetilde{\Delta}_S^{\text{ref}}$ | $\widetilde{\Delta}_N^{\text{ref}}$ | $\widetilde{\Delta}_{\text{win}}^{\text{ref}}$ |
|---|---|---|---|---|
| 4 | 0.727 | $-0.109$ $[-0.112, -0.107]$ | 3.431 [3.429, 3.433] | 3.540 [3.537, 3.543] |
| 8 | 0.533 | 0.437 [0.433, 0.441] | 4.204 [4.202, 4.206] | 3.767 [3.763, 3.771] |
| 16 | 0.348 | 1.113 [1.108, 1.118] | 5.006 [5.004, 5.007] | 3.893 [3.888, 3.898] |

Paired means with 95% bias-corrected and accelerated bootstrap confidence intervals at $\rho = 0.1$ (30 pairs per row). Since $\kappa = \chi/(\chi + m - 1)$ changes with $m$, rows compare joint settings rather than an isolated group-order effect.

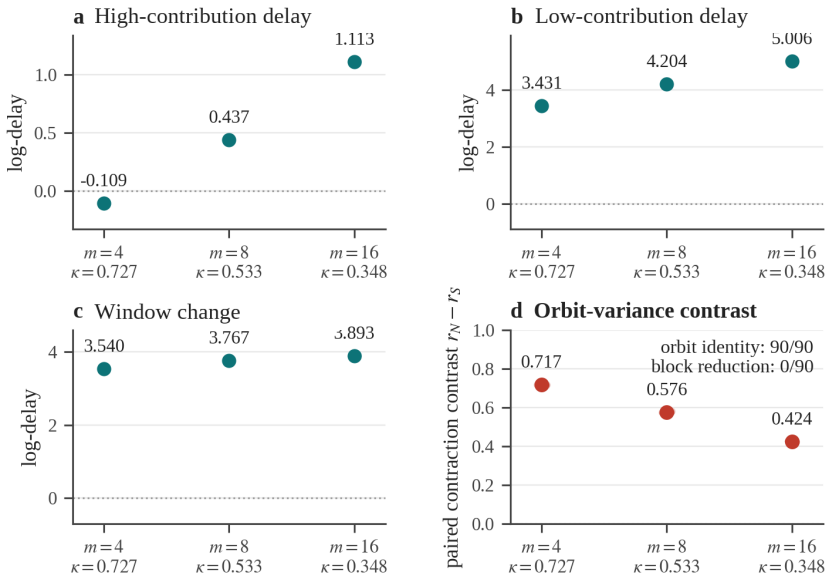


Figure F.2: **Natural finite-cyclic study.** The three designated values of $m$ show the spectral-decay contrasts on the selected high- and low-contribution references and the orbit-variance contrast.

**C: Orbit-Variance Results.** Study C estimates the following three matrices from the same $(\mathbf{X}, \boldsymbol{\xi}, g)$ draws:

$$\mathbf{U}_{\text{orb2}} = \mathbb{E}_{\boldsymbol{\xi},g}[\boldsymbol{\psi}_g \boldsymbol{\psi}_g^{\top}], \qquad \mathbf{U}_{\text{avg}}, \qquad \mathbf{V}_G = \mathbb{E}_{\boldsymbol{\xi},g}[(\boldsymbol{\psi}_g - \boldsymbol{\phi})(\boldsymbol{\psi}_g - \boldsymbol{\phi})^{\top}].$$

This common construction evaluates the positive-semidefinite (PSD) identity directly. On common references, the contractions are

$$r_S = \frac{\operatorname{Tr}(\mathbf{Q}_S \mathbf{V}_G)}{\operatorname{Tr}(\mathbf{Q}_S \mathbf{U}_{\text{base}})}, \qquad r_N = \frac{\operatorname{Tr}(\mathbf{Q}_N \mathbf{V}_G)}{\operatorname{Tr}(\mathbf{Q}_N \mathbf{U}_{\text{base}})}. \tag{F.4}$$

We report the relative orbit-identity error, a concentration bound, and independent estimates. We compare the scalar approximation $-\log(1-r_B)$ with a spectral-decay delay when the block-reduction error

$$E_{\mathrm{red}} = \frac{\|\mathbf{Q}_S \mathbf{U}_{\mathrm{base}} \mathbf{Q}_N\|_{\mathrm{op}} + \|\mathbf{Q}_S \mathbf{V}_G \mathbf{Q}_N\|_{\mathrm{op}}}{\|\mathbf{U}_{\mathrm{base}}\|_{\mathrm{op}}}$$

is below its prespecified threshold and $0 \le r_B < 1$. For other settings, we summarize the orbit-variance difference by the paired contrast $r_N - r_S$.

**D: Compact-Grid Approximation.** Study D uses three fixed torus representations indexed by $L \in \{7, 15, 23\}$, covariance scales $\lambda_0 \in \{4, 8, 16, 32\}$, raw and trace-normalized covariance modes, and 10 seeds. Across the 240 representation, covariance-mode, and seed combinations, it evaluates tensor-grid resolutions $m \in \{16, 32, 64, 128, 256, 512\}$ against a finite reference at $m_{\mathrm{ref}} = 1024$. It records relative discrepancies for the Gram, cross-moment matrix, target-contribution matrix, and target projector. Medians and interquartile ranges are computed over the four covariance scales and ten seeds for each representation and covariance mode.

Across the tested resolutions, discrepancies decrease toward the finite reference at $m_{\mathrm{ref}} = 1024$. The finite-grid error criterion is unmet in all 240 settings, and no aggregated cell satisfies the hypotheses of Corollary 5.5. These results quantify convergence relative to the finite reference; exact Haar matrices require a separate error analysis.

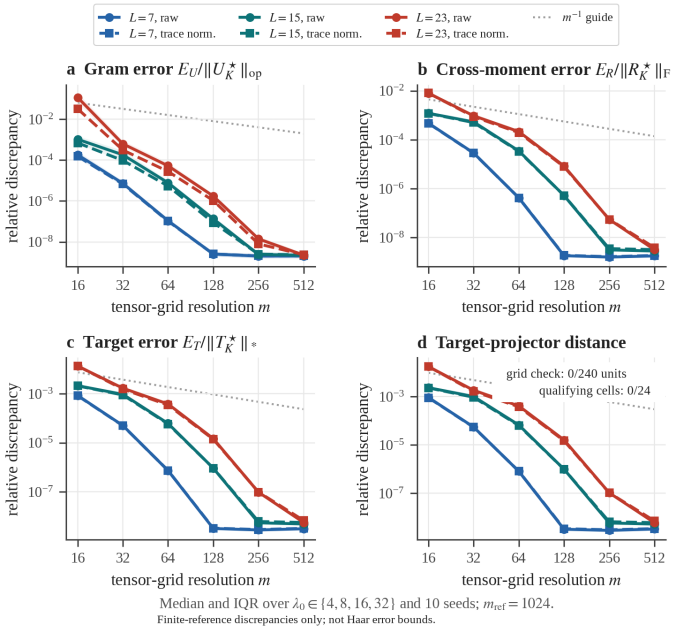


Figure F.3: **Compact tensor-grid approximation.** Gram, cross-moment, target-contribution, and projector discrepancies across grid resolutions, relative to the finite reference $m_{\mathrm{ref}} = 1024$.

**E: Structured Finite-Group Results.** Study E evaluates the theorem conditions in a structured construction for $C_4, C_8, C_{16}$, with $p = 12$, $n = 8192$, an independent held-out target sample of size 4096, and 30 seeds per group and regime. The realized Grams are exactly block diagonal with disjoint spectra. At each power-of-two resolution, we compute the Gram error and use the same exact spectral blocks for both models. The positive-margin and null-boundary regimes share the numerical, spectral, target-contribution, reference-overlap, and activity conditions; they differ in the constructed margin. Subspaces and thresholds are fixed before evaluating crossing times at $\rho = 0.1$. The independent target sample is used to evaluate the high- and low-contribution inequalities on those fixed subspaces.

In the positive-margin regime, all 90 runs satisfy the theorem conditions. The mean $\Delta_S$ is zero, while mean $\Delta_N = \Delta_{\mathrm{win}}$ is $1.626, 1.629, 1.635$ for $C_4, C_8, C_{16}$, respectively. The mean margins $M_{\mathrm{gap}} = d_0 - s_0$ are $0.706, 0.706, 0.704$. In the null-boundary regime, all 90 runs retain finite active crossing times and satisfy the other conditions, but have $M_{\mathrm{gap}} < 0$ and $\Delta_{\mathrm{win}} = 0$; none satisfies all

theorem conditions. The held-out target inequalities hold in all 180 runs. In this construction, the positive margin separates the regime with widening from the null-boundary control.

Table F.5: **Structured finite-group results.**

| Group | Regime | Conditions met | $\Delta_S$ | $\Delta_N$ | $\Delta_{\text{win}}$ | $M_{\text{gap}}$ |
|---|---|---|---|---|---|---|
| $C_4$ | positive | 30/30 | 0.000 | 1.626 | 1.626 | 0.706 |
| $C_8$ | positive | 30/30 | 0.000 | 1.629 | 1.629 | 0.706 |
| $C_{16}$ | positive | 30/30 | 0.000 | 1.635 | 1.635 | 0.704 |

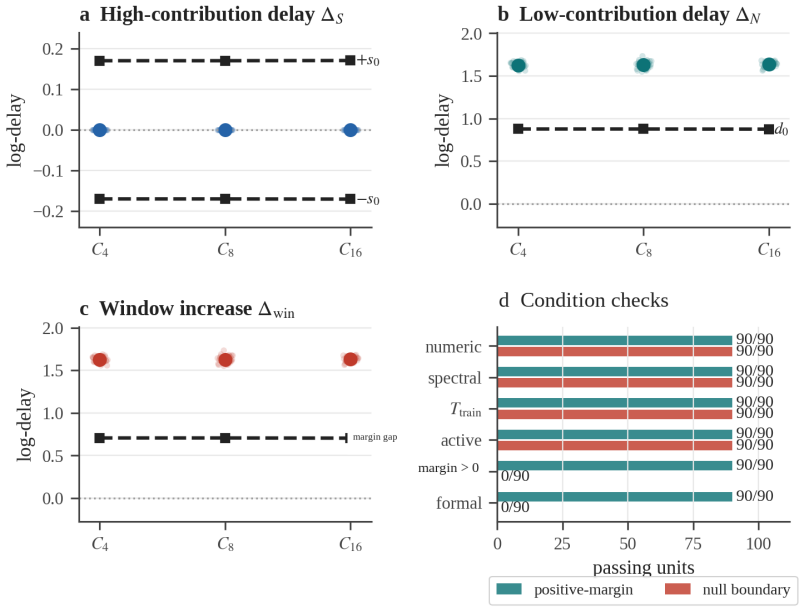


Figure F.4: **Structured finite-group results.** Results for the positive-margin and null-boundary regimes at $C_4, C_8, C_{16}$. Bars summarize the numerical, spectral, target, activity, and margin conditions; group means appear in Table F.5.

### F.4 QM9 Learned-Feature Results

The study uses 4,096 training molecules and fixed validation, test, and generation-reference sets of 10,000 molecules each, with three paired seeds. Each model receives atom types and masks from held-out composition templates and generates molecular coordinates. The four architectures are EGNN, raw-coordinate GNN, GNN-Dist, and EGNN-NoCoord. Let $\tau_{\text{gen}}$ denote the first generation-quality threshold crossing and $\tau_{\text{mem}}$ the first copying-detector trigger. Define $A_{\text{mol}} = \tau_{\text{mem}} - \tau_{\text{gen}}$ and $W_{\text{mol}} = \tau_{\text{mem}}/\tau_{\text{gen}}$. These clocks describe learned-feature training, while Section 5.1.3 defines times for frozen features. For the paired EGNN/raw-GNN comparison, the prespecified summary criterion requires: a majority of seeds with both larger $A_{\text{mol}}$ and larger $W_{\text{mol}}$; finite raw-GNN $\tau_{\text{mem}}$ in a majority of seeds; a median $\tau_{\text{gen}}$ ratio at most 1.25; and a mean best-quality ratio at most 1.1. Geometry quality is a composite of molecular distribution-distance metrics, with lower values indicating better agreement.

At each scheduled generation evaluation, $\tau_{\text{gen}}$ is the first of two consecutive valid evaluations with geometry quality at most 0.75. All observed generation-quality crossings occur at the first scheduled evaluation, 10,000 steps, so $\tau_{\text{gen}}$ is resolved at the schedule's first evaluation point for every model and seed. For the copying detector, let $d_1, d_2$ be the nearest and second-nearest training root-mean-square deviations (RMSDs) among molecules of the same composition. A generated sample is flagged when $d_1/d_2 < 1/3$ and $d_1$ is below the first percentile of the held-out-test nearest-distance calibration. The clock $\tau_{\text{mem}}$ is the first scheduled evaluation where the 95% bootstrap lower bound on the flagged

fraction is positive or its point estimate reaches 0.005. A trigger not observed by 500,000 steps is right-censored at that horizon.

Table F.6: **QM9 learned-feature comparison (4,096 training molecules).**

| Seed | Model | $\tau_{\text{gen}}$ | $\tau_{\text{mem}}$ | $A_{\text{mol}}$ | $W_{\text{mol}}$ | Best quality |
|---|---|---|---|---|---|---|
| 0 | EGNN | 10,000 | 420,000 | 410,000 | 42 | 0.0745 |
| 0 | raw GNN | 10,000 | 400,000 | 390,000 | 40 | 0.1001 |
| 1 | EGNN | 10,000 | 440,000 | 430,000 | 44 | 0.0789 |
| 1 | raw GNN | 10,000 | 360,000 | 350,000 | 36 | 0.1041 |
| 2 | EGNN | 10,000 | $\geq$ 500,000 | $\geq$ 490,000 | $\geq$ 50 | 0.0697 |
| 2 | raw GNN | 10,000 | 150,000 | 140,000 | 15 | 0.0785 |

Per-seed results. Best quality is the minimum composite geometry-distribution score; lower is better. Inequalities mark the right-censored EGNN seed.

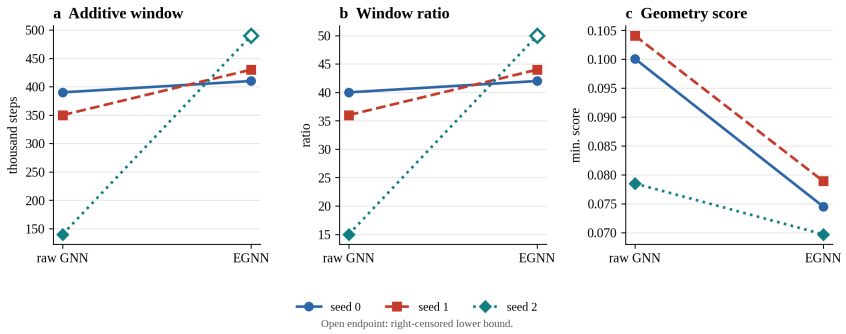


Figure F.5: **QM9 learned-feature comparison.** EGNN and raw-coordinate GNN results for the three paired seeds, showing duration, learning-time ratio, and geometry score. The open marker denotes the right-censored lower bound.

All three raw-GNN seeds have finite $\tau_{\text{mem}}$. EGNN has larger observed additive and multiplicative durations in all three paired runs; Figure F.5 shows the per-seed results. The median paired $A_{\text{mol}}^{\text{EGNN}} - A_{\text{mol}}^{\text{raw}}$ difference is 80,000 steps, the median paired $W_{\text{mol}}^{\text{EGNN}}/W_{\text{mol}}^{\text{raw}}$ ratio is 1.222, the median $\tau_{\text{gen}}^{\text{EGNN}}/\tau_{\text{gen}}^{\text{raw}}$ ratio is 1.000, and the mean EGNN/raw-GNN best-quality ratio is 0.797. These three paired runs satisfy the prespecified summary criterion, with the censored seed contributing its observed lower bound. Both control comparisons have positive median duration differences but exceed the quality-ratio threshold: EGNN/GNN-Dist has median additive difference 160,000, median learning-time ratio 1.471, and quality ratio 1.199, while EGNN-NoCoord/raw GNN has 120,000, 1.333, and 1.444, respectively. These comparisons serve as controls. All reported values summarize the three paired seeds.

### F.5 Experiment Summary

The table links each study to its measured quantity and interpretation.

### F.6 Statistical Summary

The seed is the independent statistical unit. Eigenvalues, time points, thresholds, models within a seed, and subspaces within a run are repeated measurements rather than independent samples. Avg/Base comparisons are paired.

For each designated finite-cyclic cell, Table F.4 reports paired means and 95% bias-corrected and accelerated bootstrap intervals. No equivalence margin was prespecified; results are therefore summarized by estimates and intervals. Because the fixed-trace design varies $m$ and $\kappa$ together, each designated setting is reported separately. Compact-study medians and interquartile ranges summarize variation over covariance scale and seed within each representation and covariance mode. Study

Table F.7: **Experimental quantities and interpretation.**

| Study | Reported quantity | Interpretation |
|---|---|---|
| Tilted reflection | Population moments, retained variance, and margin | Population formula evaluation and Monte Carlo agreement. |
| Cyclic population family | Margin and upper bound across group orders | Dependence of the population margin on group order, activation, and tilt. |
| Natural finite cyclic | $\widetilde{\Delta}_B^{\mathrm{ref}}$ on fixed common references | Spectral-decay contrasts for the designated joint settings. |
| Orbit variance | $r_S, r_N$ and block-reduction error | Orbit-variance contrast and conditions for its scalar timing approximation. |
| Compact tensor grid | Gram, cross-moment, target, and projector discrepancies | Convergence relative to the finite grid at $m_{\mathrm{ref}} = 1024$. |
| Structured finite group | Exact-subspace $\Delta_S, \Delta_N, \Delta_{\mathrm{win}}$ and $M_{\mathrm{gap}}$ | Theorem conditions evaluated in positive-margin and null-boundary regimes. |
| QM9 | Generation-quality and copying-detection times | Learned-feature comparison across three paired seeds. |

E reports all 30 independent seeds for each group–regime combination and evaluates the held-out target inequalities on their own sample. The three QM9 pairs are summarized with the right-censored endpoint retained at its observed lower bound.

Every run remains in its study denominator, with numerical, target, spectral, activity, and timing status recorded. Population calculations, spectral diagnostics, theorem-condition evaluations, and learned-feature comparisons are reported separately according to their estimands.

## AI USE STATEMENT

Large language model (LLM) assistants were used to support writing and polishing this manuscript, particularly to refine the language, improve readability, and clarify the exposition.

LLMs were not used to generate research ideas, develop theoretical models or conceptual frameworks, formulate mathematical claims or proof strategies, design or implement methods or experiments, generate or analyze data, correct or modify scientific figures. The authors take full responsibility for all content of this paper, including any text, code, proofs, or numerical results.