BLIND RECOVERY OF SIGNAL DOMAINS VIA UNSUPERVISED SYMMETRY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

A signal-processing perspective interprets data as discretized signals over geometric domains. When the underlying domain is inaccessible, we lose geometric priors, hindering interpretability and generalization. We propose an unsupervised framework for recovering signal domains by discovering symmetries of the data distribution. Our framework treats data samples as unknown linear measurements of continuous latent signals, and learns to recover them through a linear group-convolutional network. Assuming latent signals form a stationary Markov random field, we train the model by enforcing stationarity, minimizing total correlation, and maximizing information. We theoretically show that the proposed approach transforms an otherwise unidentifiable inverse problem into blind deconvolution, and we test the model on stochastic processes, Ising models, scrambled images, and biological neural recordings, revealing latent signals. Our results suggest a promising direction for unsupervised structure learning and blind inverse problems.

1 INTRODUCTION

A signal-processing perspective interprets data as discretized signals over continuous domains with geometric priors. When the underlying domain is inaccessible, these priors are lost, and we encounter fundamental difficulties. This setting arises in different fields, including imaging through unknown scattering media (Bertolotti et al., 2012; Katz et al., 2014), unsupervised neural decoding (Chaudhuri et al., 2019; Schneider et al., 2023), and data acquisition via uncalibrated and interfering sensors (Balzano & Nowak, 2007; Bravyi et al., 2021).

In this study, we propose an unsupervised framework to recover signal domains from unstructured observations. Assuming that the symmetries of the signal domain manifest in the data distribution, we use these symmetries as a surrogate for the inaccessible domain. Specifically, we leverage the free and transitive actions of Abelian symmetry groups by defining latent signals directly over group elements.

This setting naturally emerges in linear observations when the latent signals constitute a stationary random field, allowing signal recovery to be formulated as a group convolution (Cohen & Welling, 2016; Bekkers, 2021). While we utilize latent stationarity to learn the action of the latent symmetry, we introduce a complementary locality prior to reveal the appropriate convolutional filters. We formalize these priors by assuming the latent signal distribution resembles a stationary *Markov random field (MRF)* (Besag, 1974; Kindermann & Snell, 1980).

Our model consists of a recovery network, which maps observations to a symmetry-based representation, and a loss estimation network, which provides the information-theoretic quantities for training the former. The recovery network is trained under three objectives: information maximization prevents degenerate solutions, while stationarity and total correlation induce a stationary MRF structure at the model output to recover the signal domain.

Theoretically, we show that our method can recover signal domains up to axis permutations and reflections, and we empirically validate this on Generalized Shot Noise processes (Rice, 1944), Ising models (Ising, 1925; Külske, 2025), and pixel-shuffled and bit-scrambled MNIST datasets (Roux et al., 2007; Ye, 2010). Finally, on real-world mouse visual cortex recordings (Siegle et al., 2021), we show that the model successfully captures neural signal changes induced by stimulus rotations and reveals the underlying stimulus orientation.

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.

Our main contributions are:

- **Symmetry discovery for signal domain recovery:** We introduce an unsupervised approach to recover signal domains through leveraging regular (free and transitive) symmetry actions and distributional symmetries (Section 3).
- **Reducing blind linear inversion to blind deconvolution:** We analyze the mechanism behind signal recovery and show symmetry discovery reduces an otherwise unidentifiable inverse problem to blind deconvolution (Section 5).
- **Improving real-world applicability:** Leveraging complementary priors, we scale probabilistic symmetry discovery to higher-dimensional unstructured data, enabling signal recovery in challenging settings like bit-scrambled images and real-world biological neural recordings (Section 6).

Our implementation is publicly available at anonymous repository.

2 Related Work

Graph structure and manifold learning. Graph structure learning (GSL) infers a graph structure from statistically inferred variable dependencies (Dong et al., 2016; Kalofolias, 2016; Franceschi et al., 2019; Zhu et al., 2021). Inference requires accessing signal values at domain points, which are typically inaccessible in our setting. Alternatively, manifold learning seeks low-dimensional embeddings for high-dimensional observations (Tenenbaum et al., 2000; Belkin & Niyogi, 2003; McInnes et al., 2020). Rather than performing dimensionality reduction, our approach assumes the existence of a latent signal domain and aims to recover it.

Symmetry discovery. Existing methods learn function invariances in supervised settings (Romero & Lohit, 2021; Santos-Escriche & Jegelka, 2025; Forestano et al., 2023; Moskalev et al., 2022; Bhat et al., 2025; Zhou et al., 2020), model time-evolution from sequential data (Greydanus et al., 2019; Alet et al., 2021; Churchill & Xiu, 2023; Sohl-Dickstein et al., 2010; Champion et al., 2019; Dehmamy et al., 2021; Koyama et al., 2023; Miyato et al., 2022; Park et al., 2022; Yang et al., 2023a; Mitchel et al., 2024), or identify invariances in data distributions (Benton et al., 2020; Desai et al., 2022; Tombs & Lester, 2022; Yang et al., 2023b; Efe & Ozakin, 2025; Shaw et al., 2024). These methods primarily focus on discovering symmetries while we use them to reveal a latent signal domain.

Blind source separation and blind deconvolution. Blind source separation recovers latent signals from linear mixtures, whereas blind deconvolution inverts unknown convolutions (Comon, 1994; Bell & Sejnowski, 1995; Hyvärinen & Oja, 2000; Hyvarinen & Morioka, 2017; Kaaresen & Taxt, 1998; Levin et al., 2009; Corbetta et al., 2022). While both operate blindly via statistical assumptions, blind source separation recovers independent sources without relating them to a domain, and blind deconvolution presupposes the domain on which the convolution acts. Our framework instead recovers this domain from unstructured observations.

3 Problem Setting

We establish our problem setting below, leaving identifiability analysis to Section 5. For definitions used globally, we reserve letters f, x, y, and z (and their font variants) for the variables, spaces, and random vectors, while the remaining letters denote linear operators. Lightface font denotes scalars, functions, indexed components (c, f, x_i, X_i), operators producing them (U), and random fields (F). Bold font denotes vectors ($\mathbf{x}$), random vectors ($\mathbf{X}$), and operators producing them ($\mathbf{A}$). Calligraphic letters denote spaces: light for functions ($\mathcal{F}$) and bold for vector spaces ($\boldsymbol{\mathcal{X}}$). Greek letters τ and ρ denote symmetry representations. Vectors and linear operators that are indexed by grid sites are equivalently viewed as order-P tensors, with entries $Y_{\mathbf{n}}$ indexed by site.

Using a symmetry group as a domain surrogate. We assume the latent signal domain Ω is a homogeneous space equipped with the regular action of a connected, compact, P-dimensional Abelian Lie group $G \cong \mathbb{T}^P$. By fixing an origin $p_0 \in \Omega$, regularity ensures a bijection $p = \mathbf{t} \cdot p_0$

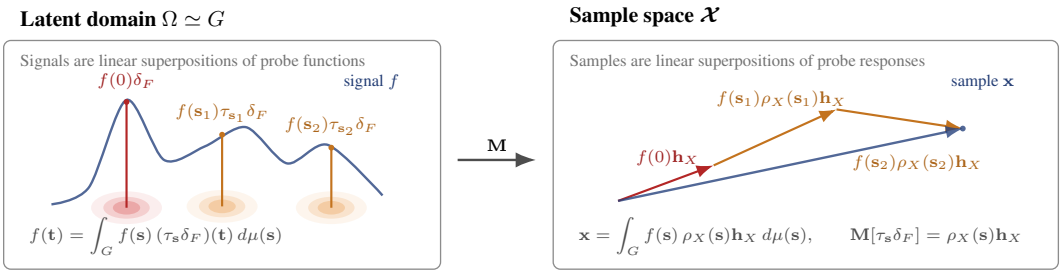


Figure 1: **Symmetry-based domain recovery.** At left, we illustrate the latent domain, where signals are superpositions of translated probe functions $\tau_{\mathbf{s}}\delta_F$. At right, linear superpositions of orbit elements $\rho_X(\mathbf{s})\mathbf{h}_X$ produce samples, where $\mathbf{h}_X$ and ρ_X are determined by the measurement operator $\mathbf{M}$.

between domain points and group elements. Thus, instead of using domain coordinates, we define latent signals over the group as $f \in L^2(G, \mathbb{R})$. [1]

Measurement operator. We assume sample vectors $\mathbf{x}_i \in \boldsymbol{\mathcal{X}} = \mathbb{R}^{d_X}$ are generated from latent signals $f_i \in \mathcal{F} := L^2(G, \mathbb{R})$ through an unknown bounded linear *measurement operator* $\mathbf{M} : \mathcal{F} \to \boldsymbol{\mathcal{X}}$. Because measurements are band-limited, we define the *reconstructible signal subspace* $\mathcal{F}_R := \ker(\mathbf{M})^\perp$, on which $\mathbf{M}$ is injective, and assume $\boldsymbol{\mathcal{X}} = \mathbf{M}(\mathcal{F}_R)$. Hereafter, we restrict our attention to projections onto $\mathcal{F}_R$, referring to them simply as latent signals.

The measurement operator can be factored into three sequential operations (Lemma B.1): a *function space projection operator* $\mathrm{P}_R : \mathcal{F} \to \mathcal{F}_R$, an *isometric sampling operator* $\mathbf{S} : \mathcal{F}_R \to \boldsymbol{\mathcal{F}}$ mapping functions to vectors, and an unknown square *mixing matrix* $\mathbf{A} : \boldsymbol{\mathcal{F}} \to \boldsymbol{\mathcal{X}}$, such that $\mathbf{M} = \mathbf{A}\mathbf{S}\mathrm{P}_R$.

Induced representation. We assume $\mathcal{F}_R$ is translation-invariant. As Appendix B.2 shows, this ensures latent translations admit a representation in the sample space. Thus, there exists a 2π-periodic representation $\rho_X : \mathbb{R}^P \to \mathrm{GL}(d_X)$ such that

$$\mathbf{M}[\tau_{\mathbf{t}} f] = \rho_X(\mathbf{t})\mathbf{M}[f], \qquad f \in \mathcal{F}_R, \mathbf{t} \in \mathbb{R}^P. \tag{1}$$

Probe function. Because $\mathcal{F}_R$ is translation-invariant, it admits a translation-invariant reproducing kernel k_R satisfying

$$f(\mathbf{t}) = \int_G f(\mathbf{s}) k_R(\mathbf{s}, \mathbf{t})\, d\mu(\mathbf{s}) = \int_G f(\mathbf{s}) k_R(\mathbf{0}, \mathbf{t} - \mathbf{s})\, d\mu(\mathbf{s}) = \int_G f(\mathbf{s}) \delta_F(\mathbf{t} - \mathbf{s})\, d\mu(\mathbf{s}), \tag{2}$$

where μ is the Haar measure on G. We define the *probe function* $\delta_F(\mathbf{r}) := k_R(\mathbf{0}, \mathbf{r})$, which acts as a Dirac delta within $\mathcal{F}_R$. Its image under the measurement operator is the *probe response* $\mathbf{h}_X := \mathbf{M}[\delta_F]$.

Data model. Under the translation convention $(\tau_{\mathbf{s}}\delta_F)(\mathbf{t}) = \delta_F(\mathbf{t} - \mathbf{s})$, the reproducing property gives $f = \int_G f(\mathbf{s})(\tau_{\mathbf{s}}\delta_F)\, d\mu(\mathbf{s})$ for any $f \in \mathcal{F}_R$. Applying the measurement operator yields the sample vector

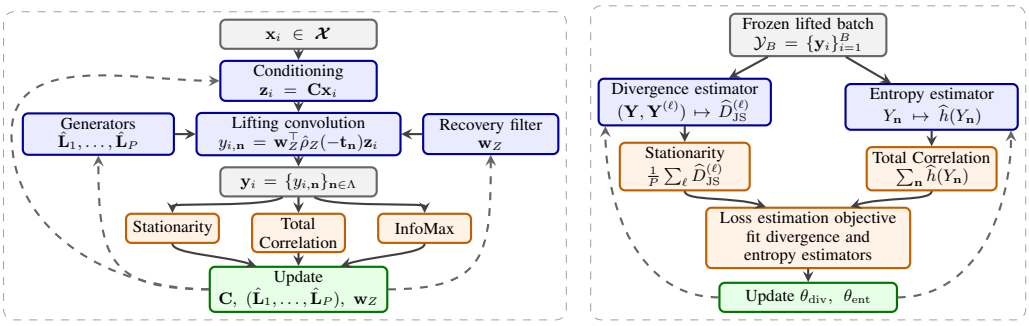
$$\mathbf{x} = \mathbf{M}[f] = \int_G f(\mathbf{s})\mathbf{M}[\tau_{\mathbf{s}}\delta_F]\, d\mu(\mathbf{s}) = \int_G f(\mathbf{s})\rho_X(\mathbf{s})\mathbf{h}_X\, d\mu(\mathbf{s}). \tag{3}$$

Consequently, $\boldsymbol{\mathcal{X}}$ is spanned by the transformed probe responses $\{\rho_X(\mathbf{s})\mathbf{h}_X \mid \mathbf{s} \in \mathbb{R}^P\}$. When latent signals are restricted to $f \in \{\tau_{\mathbf{s}}\delta_F \mid \mathbf{s} \in \mathbb{R}^P\}$, this reduces to the single-orbit setting (Anselmi et al., 2019), where data samples are generated by a symmetry group acting on a single vector.

Latent random field. We model latent signals $f \in \mathcal{F}$ as realizations of a continuous, stationary random field over G. Because we focus on reconstructible, band-limited signals, we define a *latent random vector* $\mathbf{F}$ with realizations $\mathbf{f} \in \boldsymbol{\mathcal{F}}$ obtained by projecting f onto a low-frequency subspace via $\mathrm{P}_R : \mathcal{F} \to \mathcal{F}_R$, followed by sampling $\mathbf{S} : \mathcal{F}_R \to \boldsymbol{\mathcal{F}}$:

$$\mathbf{f} = \mathbf{S}\mathrm{P}_R f.$$

[1] In practice, many signal domains are flat patches rather than tori. While these can be idealized as tori by identifying opposite boundaries (Grafakos, 2014), boundary effects disrupt latent statistics and preclude an exact theoretical treatment. Thus, our analysis focuses on the torus for mathematical tractability, while our experiments evaluate both settings for broader applicability.



(a) Recovery network update **(b)** Loss estimation network update

Figure 2: **Training scheme. (a)** The recovery network update optimizes $\mathbf{C}$, $(\hat{\mathbf{L}}_1, \ldots, \hat{\mathbf{L}}_P)$, and $\mathbf{w}_Z$ using stationarity, total correlation minimization, and information maximization losses. **(b)** The loss estimation network update fits statistical divergence and entropy estimators.

We assume the latent random vector $\mathbf{F}$ satisfies the local Markov property for an appropriate grid structure, i.e., $\mathbf{F}$ is an MRF.

4 Methodology

Our framework comprises two components: a *recovery network* mapping unstructured samples to latent signal representations, and a *loss estimation network* consisting of information-theoretic estimators. Both networks are trained jointly via simultaneous gradient updates, as visualized in Figure 2. Below, we describe the components of both networks and the training objective. Implementation details are in Appendix G.

4.1 Recovery Network

Parameterizing learned symmetry representations. Since G is compact, there exists a linear transformation on $\mathcal{X}$ which induces an equivalent orthogonal representation (Lemma B.8). We therefore introduce a learnable *conditioning matrix* $\mathbf{C} : \mathcal{X} \to \mathcal{Z}$, which maps samples $\mathbf{x}$ into *conditioned sample space* $\mathcal{Z} = \mathbb{R}^{d_Z}$, and we parametrize the learned symmetry representation $\hat{\rho}_Z \in \mathrm{SO}(d_Z)$ via P commuting skew-symmetric generators $(\hat{\mathbf{L}}_1, \ldots, \hat{\mathbf{L}}_P)$. For coordinates $\mathbf{t} \in \mathbb{R}^P$, we define:

$$\hat{\rho}_Z(\mathbf{t}) := \exp\left(\sum_{j=1}^{P} t_j \hat{\mathbf{L}}_j\right).$$

This parametrization is numerically stable and loses no generality: after successful training, the latent symmetry action ρ_Z on the conditioned sample space is orthogonal (Corollary B.9).

Lifting operation. We define a learnable *lifting operator* $\mathrm{U} : \mathcal{Z} \to L^2([0, 2\pi)^P)$ by

$$y(\mathbf{t}) = (\mathrm{U}\mathbf{z})(\mathbf{t}) := \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t})\mathbf{z}, \tag{4}$$

where $\mathbf{z} = \mathbf{C}\mathbf{x}$ for a given data sample $\mathbf{x}$. Here, $\hat{\rho}_Z$ is the learned representation and $\mathbf{w}_Z \in \mathcal{Z}$ is the learned *recovery filter*, normalized to $\|\mathbf{w}_Z\|_2 = 1$.

We evaluate lifting on regular grid points, the lattice Λ, indexed by the multi-index $\mathbf{n} \in \prod_{j=1}^{P}\{0, \ldots, N_j - 1\}$. Points in the parameter space are given by $\mathbf{t}_\mathbf{n} = 2\pi\mathbf{D}^{-1}\mathbf{n}$, where $\mathbf{D} := \mathrm{diag}(N_1, \ldots, N_P)$, and $h_\mathbf{n} := h(\mathbf{t}_\mathbf{n})$ discretizes any function h. This defines a *lifting matrix* $\mathbf{U} : \mathcal{Z} \to \mathcal{Y}$:

$$U_{ij} = \left[\mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t}_{\mathbf{n}(i)})\right]_j, \tag{5}$$

where $\mathbf{n}(i)$ is the multi-index for the flattened index i, yielding the final model output $\mathbf{y} = \mathbf{U}\mathbf{C}\mathbf{x}$.

4.2 Loss Estimation Network

The loss estimation network contains entropy and JS divergence estimators. We estimate component-wise entropies with trainable Gaussian mixture models (Pichler et al., 2022), and the joint entropy directly from the output covariance matrix (Appendix G.2). JS divergences are estimated by neural networks via the variational formula of Donsker & Varadhan (1975) (Appendix G.3).

For the joint entropy, we use a rank schedule: only the top-k eigenvalues of the output covariance are retained, and k increases gradually from 1 to its terminal value. In the single-orbit setting (Section 3), the DC and fundamental harmonics suffice to determine the latent state, so we compute the joint entropy in the subspace they span, with terminal value $k = 3$. Otherwise, $k = d_X$ unless stated.

4.3 Learning Objective

The learning objective combines stationarity, total correlation minimization, and information maximization terms on the model output $\mathbf{Y}$:

$$J(\mathbf{Y}) = J_{\text{stationarity}}(\mathbf{Y}) + \lambda_{\text{TC}} J_{\text{TC}}(\mathbf{Y}) + \lambda_{\text{infomax}} J_{\text{infomax}}(\mathbf{Y}). \tag{6}$$

Section 5 shows how these terms cooperate to discover the latent symmetry action and reveal the latent signals.

Stationarity. If the latent symmetry representation is learned, the model output is stationary (Lemma B.11). This also allows learning the symmetry action solely by enforcing stationarity of the model output. To accommodate open boundaries, we compare the output with its translation by one site only where both are defined, through a boundary projection $\mathbf{P}_\ell : \mathbb{R}^{d_Y} \to \mathbb{R}^{d_Y - (d_Y/N_\ell)}$:

$$J_{\text{stationarity}}(\mathbf{Y}) := \frac{1}{P} \sum_{\ell=1}^{P} D_{\text{JS}}(\mathbf{P}_\ell \mathbf{Y}, \mathbf{P}_\ell \mathbf{Y}^{(\ell)}),$$

where $Y^{(\ell)}_{\mathbf{n}} := Y_{\mathbf{n}+\mathbf{e}_\ell}$, $\mathbf{e}_\ell$ is the standard basis vector along the ℓ-th axis, and $\mathbf{P}_\ell$ discards the last slice along that axis, $n_\ell = N_\ell - 1$, where the translated output would wrap around.

Total correlation minimization. Total correlation regularizes high-order conditional dependencies (Lemma F.8) and resolves the latent dependency structure:

$$J_{\text{TC}}(\mathbf{Y}) := \sum_{\mathbf{n}} h(Y_{\mathbf{n}}) - h(\mathbf{Y}),$$

which vanishes only if the components $\{Y_{\mathbf{n}}\}$ are jointly independent. Except when the input is a zeroth-order MRF, this absolute minimum is never attained.

Information maximization. The sample space contains both high- and low-variance directions, and the preceding loss terms do not naturally bias the model toward informative directions. Information maximization encourages the model to prioritize high-variance directions and prevents dimensional collapse. We implement this by minimizing the negative joint entropy:

$$J_{\text{infomax}}(\mathbf{Y}) := -h(\mathbf{Y}).$$

5 Theoretical Interpretation and Identifiability

In this section, we outline the key elements of how the model learns latent symmetries and recovers latent signals, deferring the detailed formal treatment to Appendix A. We start by applying the lifting operation to the data model (Section 3):

$$y(\mathbf{t}) = \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t})\mathbf{z} = \int_G f(\mathbf{s})\, k(\mathbf{s}, \mathbf{t})\, d\mu(\mathbf{s}), \qquad k(\mathbf{s}, \mathbf{t}) := \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t}) \rho_Z(\mathbf{s}) \mathbf{h}_Z. \tag{7}$$

We refer to the right-hand side as the *recovery integral* and to $k(\mathbf{s}, \mathbf{t})$ as the *recovery kernel*, which will be central to our analysis.

Symmetry discovery. Non-degenerate information maximization ensures the lifting matrix $\mathbf{U}$ is injective (Lemmas C.4 and C.5). If the model output is stationary, the equivariance property $\mathbf{T}_{\mathbf{e}_i}\mathbf{U} = \mathbf{U}\hat{\rho}_Z(-\mathbf{t}_{\mathbf{e}_i})$ (Lemma B.11) yields:

$$\mathbf{T}_{\mathbf{e}_i}\mathbf{Y} \overset{d}{=} \mathbf{Y} \implies \mathbf{T}_{\mathbf{e}_i}\mathbf{U}\mathbf{Z} \overset{d}{=} \mathbf{U}\mathbf{Z} \implies \mathbf{U}\hat{\rho}_Z(-\mathbf{t}_{\mathbf{e}_i})\mathbf{Z} \overset{d}{=} \mathbf{U}\mathbf{Z} \implies \hat{\rho}_Z(-\mathbf{t}_{\mathbf{e}_i})\mathbf{Z} \overset{d}{=} \mathbf{Z}.$$

If all data distribution symmetries are group elements, this implies $\hat{\rho}_Z(\mathbf{s}) = \rho_Z(\mathbf{R}\mathbf{s})$ for a linear parametrization $\mathbf{R} : \mathbb{R}^P \to \mathbb{R}^P$ where $\mathbf{s} \in \Lambda$. In the absence of aliasing (Assumption 6), this is shown to be a signed permutation matrix (Theorem D.8). Defining $y^{(R)}(\mathbf{t}) := y(\mathbf{R}^\top \mathbf{t})$ and $\kappa(\mathbf{r}) := k(-\mathbf{r}, \mathbf{0}) = \mathbf{w}_Z^\top \rho_Z(-\mathbf{r})\mathbf{h}_Z$, the recovery integral reduces to a standard convolution which admits an exact discrete form with the kernel $\kappa_\mathbf{n} := \kappa(\mathbf{t}_\mathbf{n})$ (Lemma F.6).

Spectrum of the recovery kernel. Using the eigendecomposition $\rho_Z(\mathbf{r})\mathbf{u}_\mathbf{m} = e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{r}}\mathbf{u}_\mathbf{m}$, the shift-invariant kernel becomes:

$$\mathbf{w}_Z^\top \rho_Z(-\mathbf{r})\mathbf{h}_Z = \sum_\mathbf{m} e^{\mathrm{i}\mathbf{m}\cdot\mathbf{r}}\, \mathbf{w}_Z^\top \mathbf{u}_\mathbf{m}\mathbf{u}_\mathbf{m}^\dagger \mathbf{h}_Z = \sum_\mathbf{m} w_{Z,\mathbf{m}}^*\, h_{Z,\mathbf{m}}\, e^{\mathrm{i}\mathbf{m}\cdot\mathbf{r}},$$

where $w_{Z,\mathbf{m}} := \mathbf{u}_\mathbf{m}^\dagger \mathbf{w}_Z$ and $h_{Z,\mathbf{m}} := \mathbf{u}_\mathbf{m}^\dagger \mathbf{h}_Z$. Taking the lattice DFT yields $K_\mathbf{m} = w_{Z,\mathbf{m}}^*\, h_{Z,\mathbf{m}}$, where information maximization makes $|w_{Z,\mathbf{m}}|^2 = 1/d_X$ (Proposition C.13).

If the mixing matrix $\mathbf{A}$ is orthogonal, then $|h_{Z,\mathbf{m}}|^2 = 1$ (Lemma F.1 and Proposition F.2). For random mixing matrices, the relative variance of $|h_{Z,\mathbf{m}}|^2$ scales as $O(1/d_X)$, converging to a flat spectrum as $d_X \to \infty$ (Lemma F.4).

Emergence of a delta recovery kernel. Total correlation minimization regularizes conditional dependencies, thereby encouraging the model output to resemble a low-order MRF (Lemma F.8). Given that the latent field is a non-Gaussian finite-order MRF, this forces the recovery kernel κ to have compact support (Proposition F.9). Then, a flat kernel spectrum implies that it is a shifted delta.

To observe this, assume $P = 1$ (generalization in Proposition F.11) and suppose the support size q of κ satisfies $2 \le q \le N/2$. Aligning the support to $\{0, \dots, q-1\}$ ensures $\kappa_0, \kappa_{q-1} \neq 0$. The circular autocorrelation at shift $q-1$ is:

$$r(q-1) = \sum_{n=0}^{N-1} \kappa_n \kappa_{(n+q-1) \bmod N} = \kappa_0 \kappa_{q-1}.$$

However, a flat spectrum requires $r(j) = 0$ for all $j \in [1, N-1]$, contradicting $\kappa_0\kappa_{q-1} \neq 0$. Thus, if $q \le N/2$, we must have $q = 1$. Therefore, κ is a shifted and scaled delta, and signals are revealed up to scale, shift, and signed axis permutation $\mathbf{R}$.

6 Experiments

We evaluate the method using stochastic processes, Ising models, scrambled images, and neural recordings. We compare symmetry discovery performance against LieGAN (Yang et al., 2023b). For signal recovery, we adapt GLASSO (Friedman et al., 2008; Banerjee et al., 2008), Covariance Matching (CM) (Ottersten et al., 1998), ISOMAP (Tenenbaum et al., 2000), and UMAP (McInnes et al., 2020) selectively, depending on compatibility with the experiment. For neural recordings, we use manifold learning algorithms as a benchmark.

We tune baseline methods via grid search over multi-seed experiments (5 seeds), and report the configuration with the best average. For our method, we use $\lambda_{\mathrm{TC}} = \lambda_{\mathrm{infomax}} = 1$, unless otherwise stated. Complete experimental details (estimator architectures, hyperparameter values, and negative results) are available in Appendix H.

Evaluation protocol. To evaluate symmetry discovery performance, we obtain sample space minimal generators via $\hat{\rho}_X(\mathbf{t}_{\mathbf{e}_i}) = \mathbf{C}^\dagger \hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})\mathbf{C}$, where $\mathbf{C}^\dagger$ is the Moore-Penrose pseudo-inverse, and compute the cosine similarity between the actual and learned generators $\rho_X(\mathbf{t}_{\mathbf{e}_i})$ and $\hat{\rho}_X(\mathbf{t}_{\mathbf{e}_i})$ for each axis i. We compute cosine similarity over a subspace of frequencies up to 0.75 of the Nyquist limit which is referred to as $S_{0.75}$ (for the definition and full-band scores, see Appendix I.1.3). This filtering is applied for two reasons:

- Sampling continuous signals introduces aliasing artifacts at high frequencies, which violate the MRF structure.
- Frequencies beyond 0.75 of the Nyquist limit carry 0.5–2.7% of the signal energy (Appendix I.1.3), rendering these directions nearly empty.

For signal recovery, we report the maximum absolute Pearson correlation between the latent signal and the model output over all admissible axis permutations and reflections (i.e., $2^P P!$ alignments for a P-dimensional group).

6.1 WAVEFORM BENCHMARKS

Generalized Shot Noise experiments. We evaluate symmetry discovery against LieGAN (Yang et al., 2023b). To conform to LieGAN's validated operating region, we use 15-dimensional Gaussian *Generalized Shot Noise (GSN)* signals (Rice, 1944). We evaluate models on continuous, band-limited continuous, and discrete translations while training for the same number of steps. LieGAN performs perfectly for band-limited continuous translations ($S_{0.75} = 1.000$), but degrades under discrete or continuous translations, possibly due to aliasing in the latter. Our method remains robust in all cases ($S_{0.75} \geq 0.946$, Table 1).

We benchmark signal recovery against GLASSO-SPEQ and CM under no-mixing and DST-I mixing using 63-dimensional signals generated with Gaussian and Legendre bases. GLASSO-SPEQ estimates a sparse precision matrix with GLASSO (Friedman et al., 2008) and maps it to the 1D grid by spectral matching, whereas CM (Ottersten et al., 1998) aligns covariance eigenvectors with Fourier modes under an eigenvalue-decay prior. Without mixing, baseline correlations remain under $r = 0.67$ and degrade further under DST-I mixing ($r \leq 0.33$). Across settings, our method achieves $r \geq 0.93$ and identifies the symmetry action with $S_{0.75} \geq 0.91$ (Table 1).

Ising model experiments. Analog computing hardware, such as Ising machines (Mohseni et al., 2022), can suffer from calibration errors due to imperfect sensors (Maciejewski et al., 2020; Bravyi et al., 2021). We simulate a 33-node spin chain and apply permutation and random matrix mixing (condition number ≈ 1.77) to model uncalibrated sensors (Appendix H.1.2).

For random mixing, we use a learnable conditioning matrix. While scores for GLASSO-SPEQ and CM remain below $r = 0.72$, our approach achieves strong recovery for both cases ($r \geq 0.992$, Table 1). Figure 3a shows that the model's response $\mathbf{UC}$ inverts unknown mixing $\mathbf{A}$.

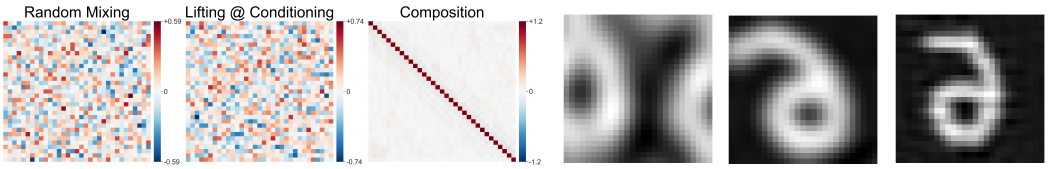
Table 1: **One-dimensional benchmarks.** We report generator similarity $S_{0.75}$ and signal recovery r.

Symmetry-generator discovery					
Dataset	d	Translation regime		LieGAN $S_{0.75}$	Ours $S_{0.75}$
GSN-G	15	Band-limited continuous		$\mathbf{1.000 \pm 0.000}$	0.946 ± 0.070
GSN-G	15	Discrete		0.218 ± 0.111	$\mathbf{0.962 \pm 0.010}$
GSN-G	15	Continuous		0.622 ± 0.460	$\mathbf{0.969 \pm 0.009}$
Signal recovery					
Dataset	d	Transform	Generator	Signal recovery r	
			$S_{0.75}$	Ours	GLASSO-SPEQ	CM
GSN-G	63	Identity	0.996 ± 0.002	$\mathbf{0.941 \pm 0.031}$	0.663 ± 0.076	0.470 ± 0.013
GSN-L	63	Identity	0.910 ± 0.022	$\mathbf{0.985 \pm 0.009}$	0.479 ± 0.007	0.403 ± 0.003
GSN-G	63	DST-I	0.994 ± 0.002	$\mathbf{0.931 \pm 0.015}$	0.240 ± 0.007	0.294 ± 0.010
GSN-L	63	DST-I	0.976 ± 0.007	$\mathbf{0.992 \pm 0.001}$	0.260 ± 0.007	0.330 ± 0.006
Ising	33	Permuted	0.999 ± 0.000	$\mathbf{0.999 \pm 0.000}$	0.715 ± 0.107	0.716 ± 0.042
Ising	33	Random	0.955 ± 0.026	$\mathbf{0.992 \pm 0.004}$	0.339 ± 0.016	0.654 ± 0.012

6.2 MNIST BENCHMARKS

Pixel shuffling. We randomly crop 15×15 and 27×27 MNIST patches, zero-padding pixels beyond the image boundary, and feed the methods 225- and 729-dimensional flat vectors with permuted components. For signal recovery, we estimate pixel connectivity with GLASSO and map the graph to the 2D grid by a permutation found with Fast Approximate QAP matching (GLASSO-FAQ; Vogelstein et al., 2015). We exclude spectral baselines as eigenvector ordering is ambiguous on 2D grids. Following Roux et al. (2007), we adapt manifold learning methods by estimating 2D pixel embeddings and reconstruct signals by interpolation. Improving on the best baseline, our method achieves $r = 0.99$ and $r = 0.95$ on 15×15 and 27×27 crop settings, although they search for $\mathcal{O}(d)$ pixel embeddings where our method learns linear transformations.

Symmetry discovery performance is stronger in the 15×15 setting, with $S_{0.75} = 0.900 \pm 0.034$ and 0.897 ± 0.039 along two axes. It decreases to $S_{0.75} \approx 0.72$ for 27×27 crops. The half-Nyquist



(a) Blind inversion of random matrix transformation. (b) Blind inversion of bit-scrambling transformation.

Figure 3: **(a)** Unknown random matrix mixing ($\mathbf{A}$), conditioning and lifting composed ($\mathbf{UC}$), and all three combined $\mathbf{UCA}$. **(b)** Training snapshots from the output of bit-scrambling experiment.

frequency band metric $S_{0.5}$ suggests that errors concentrate in higher-frequency modes ($S_{0.5} = 0.960$) for the latter.

Bit scrambling. Nonlinear problems can be addressed within a linear framework by utilizing high-dimensional representations. To test this setting, we apply bit-scrambling encryption to the cropped images (Ye, 2010). Scrambling acts as a permutation over 5832-dimensional ($27\times27\times8$) binary representations, although it is nonlinear in pixel intensities. For this experiment only, we use a dimensionality-reducing conditioning matrix $\mathbf{C}$. Operating on binary vectors, the model achieves strong recovery ($r = 0.941$, single run; Figure 3b) while it restores the domain structure by aggregating bits into pixels (Appendix, Figure 5a).

Table 2: **Two-dimensional benchmarks.** We report signal recovery performance r. †Single run.

MNIST crop	Ours	GLASSO-FAQ	UMAP	ISOMAP
15×15	$\mathbf{0.993 \pm 0.001}$	0.192 ± 0.016	0.846 ± 0.038	0.859 ± 0.064
27×27	$\mathbf{0.947}^{\dagger}$	0.153 ± 0.028	0.574 ± 0.028	0.731 ± 0.016

6.3 Learning Rotational Symmetry from Mouse Visual Cortex Signals

We next test our method on real mouse neural recordings from the Allen Brain Observatory (Siegle et al., 2021). In the selected experiment, visual patterns are shown in six random orientations ($\theta \in \{0°, 30°, \ldots, 150°\}$) while 110 visual cortex neurons are recorded. Extracting spike rates yields a 110-dimensional dataset with $\sim$ 1.2k samples (Appendix H.1.5).

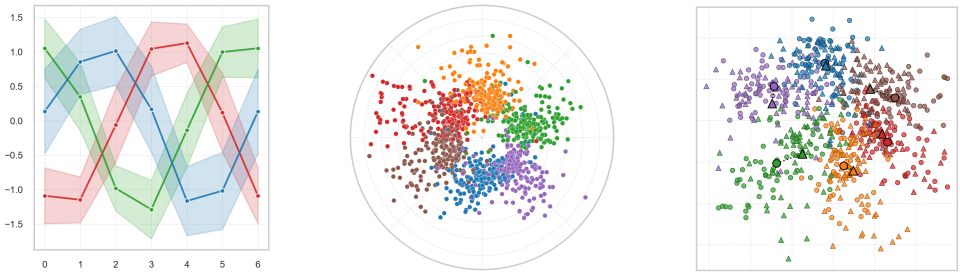
We use the single-orbit setting of the model (Section 4.2) and read the orientation from the phase of the fundamental harmonic; for the manifold learning baselines, we use the angular positions of 2D embeddings (Appendix H.3.5). As reference orientations are π-periodic, we double the angles and quantify alignment using the absolute circular correlation r_{cc} (Fisher & Lee, 1983; Jammalamadaka & Sengupta, 2001). We also report the mean absolute orientation error (MAE) for the best rotation and reflection. Since orientations take six discrete values, metrics are bounded by 0.912 and 7.5° for a model spreading its outputs uniformly. The model's output shows consistent orientation-based clustering (Figure 4b) and reaches $r_{\text{cc}} = 0.808$ and MAE $= 9.7°$, outperforming UMAP and ISOMAP (Table 3).

To investigate the learned symmetry, we apply the learned generator to spike rate vectors. The original and transformed samples reveal substantial overlap (Figure 4c), indicating that the learned minimal generator approximates the induced 30° stimulus rotation over the first two principal components[2].

Table 3: **Stimulus orientation recovery.** Limits assume uniformly spread outputs. †Deterministic.

Limit ($r_{\text{cc}} \uparrow$ / MAE $\downarrow$)	Ours ($r_{\text{cc}} \uparrow$ / MAE $\downarrow$)	UMAP ($r_{\text{cc}} \uparrow$ / MAE $\downarrow$)	ISOMAP ($r_{\text{cc}} \uparrow$ / MAE $\downarrow$)
0.912 / 7.5°	$\mathbf{0.808 \pm 0.013}$ / $\mathbf{9.7° \pm 0.3°}$	0.568 ± 0.025 / $16.0° \pm 0.4°$	0.076 / 27.5°†

[2]A successfully learned minimal generator corresponds to a $2\pi/N$ rotation in the latent signal domain, which is 60° for this setting ($N = 6$). However, as the stimulus is π-periodic, this corresponds to rotating it by 30°.



(a) Model output. (b) Polar code. (c) Generator action.

Figure 4: **Neural decoding.** Samples are color-coded by stimulus orientation. **(a–b)** Phase and amplitude of the output recover stimulus orientation and response strength. **(c)** Real (circles) and rotated (triangles) samples along the first two principal axes.

6.4 Model Analysis

Hyperparameter selection. All loss terms are per-dimension entropies, which eases hyperparameter selection, and a Morris sensitivity analysis shows stable performance under changes of loss weights and learning rates (Appendix I.1.2). We set $\lambda_{\text{TC}} = \lambda_{\text{infomax}} = 1$, except for 27×27 MNIST, where we set both to 0.25, to ensure stationarity dominates (Appendix H.2).

Computational complexity. Two eigendecompositions per update dominate the cost: of the $d_Y \times d_Y$ output covariance (joint entropy) and of the $d_Z \times d_Z$ generator parametrization (Appendix G.1). For batch size B, their per-sample costs are $\mathcal{O}(d_Y^2 + d_Y^3/B)$ and $\mathcal{O}(d_Y d_Z + d_Z^2 + d_Z^3/B)$, and the estimators scale almost linearly in d_Y. Large batches ($B = 2{,}500$) and dimensionality-reducing conditioning matrices reduce diagonalization costs.

Ablation study. Table 4 reports ablations on 63-dimensional GSN. Dropping total correlation causes complete failure ($S_{0.75} \leq 0.08$, $r \leq 0.06$), and dropping stationarity degrades performance severely ($S_{0.75} \leq 0.78$, $r \leq 0.25$). Without information maximization, Legendre recovery is largely intact ($r = 0.943$) but Gaussian recovery is unreliable ($r = 0.480 \pm 0.391$), suggesting that the joint-entropy term of total correlation alone can sometimes prevent dimensional collapse.

Table 4: **Ablation on GSN.** We report generator similarity $S_{0.75}$ and signal recovery r.

Dataset	Metric	Full	w/o Stationarity	w/o TC	w/o InfoMax
GSN-G	$S_{0.75}$	$\mathbf{0.996 \pm 0.002}$	0.772 ± 0.156	-0.011 ± 0.028	0.917 ± 0.106
	r	$\mathbf{0.941 \pm 0.031}$	0.241 ± 0.153	0.059 ± 0.035	0.480 ± 0.391
GSN-L	$S_{0.75}$	0.910 ± 0.022	0.581 ± 0.137	0.080 ± 0.021	$\mathbf{0.967 \pm 0.013}$
	r	$\mathbf{0.985 \pm 0.009}$	0.139 ± 0.052	0.059 ± 0.014	0.943 ± 0.042

7 Discussion

We introduced a method for a challenging setting frequently encountered across disciplines, built on a symmetry-based perspective. Our experiments and theoretical analysis suggest that the method can turn an otherwise unidentifiable linear inverse problem into blind deconvolution, which is still challenging but practically solvable. Moreover, by scaling up probabilistic symmetry discovery, it can offer a new perspective on exploratory data analysis.

However, our experiments are focused on controlled settings and lower dimensionalities. The bit-scrambling experiment suggests that dimensionality-reducing conditioning matrices can be a tractable route to scaling. Scaling further requires more efficient diagonalization algorithms. We also assumed that signal domains are tori or rectangular patches embeddable in a torus. The method can be adapted to other geometries by replacing the CNN-based stationarity estimators with group convolutional networks for the presumed group. A more challenging but promising avenue is to learn the group structure from data. We leave these for future work.

AI USE STATEMENT

In this work, we used generative AI tools, primarily Claude (Anthropic), for formulating mathematical claims in the appendix, including generalizations of existing results; for providing proof strategies and counterexamples and assisting in writing the appendix proofs; for refining hypotheses, such as the account of the training dynamics in Remark D.9; for feedback on the methodology and experiments; for interpreting results; for generating synthetic datasets; for implementing and evaluating the baselines; and for cleaning and reformatting datasets. We have not used generative AI tools for the core theoretical formulation in the main text, for implementing our method, or for translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for writing assistance, including proofreading, rephrasing, and checking the consistency of terminology and notation; for formatting tables and figures in LaTeX; for literature search; for modifying our training and evaluation pipelines; and for short scripts that numerically check theoretical statements. We have reviewed all AI-assisted work: the authors checked every AI-assisted claim, proof, and derivation, and tested AI-assisted code before use; suggestions that did not survive verification were discarded or corrected. The research questions and the method are the authors' own. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our implementation is available in an anonymous repository. The theoretical results are stated together with their assumptions and architectural constraints in Appendix A, and complete proofs are given in Appendices B–F. Appendix G describes the model as implemented. Appendix H specifies the generation of every dataset, with MNIST and the Allen Brain Observatory recordings being publicly available (Appendix H.1), the configuration of our model in each experiment, including seeds, training budgets, and run times (Appendix H.2), and the configurations, tuning grids, and evaluation protocols of all baselines (Appendix H.3). The full baseline sweeps are reported in Appendix I.

REFERENCES

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Ferran Alet, Dylan Doblar, Allan Zhou, Josh Tenenbaum, Kenji Kawaguchi, and Chelsea Finn. Noether networks: meta-learning useful conserved quantities. *Advances in Neural Information Processing Systems*, 34:16384–16397, 2021.

Fabio Anselmi, Georgios Evangelopoulos, Lorenzo Rosasco, and Tomaso Poggio. Symmetry-adapted representation learning. *Pattern Recognition*, 86:201–208, 2019.

Laura Balzano and Robert Nowak. Blind calibration of sensor networks. In *Proceedings of the 6th International Conference on Information Processing in Sensor Networks*, pp. 79–90. ACM, 2007.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Jan Beirlant, Edward J. Dudewicz, László Györfi, and Edward C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Erik J Bekkers. B-spline cnns on lie groups, 2021. URL https://arxiv.org/abs/1909.12057.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33:17605–17616, 2020.

Jacopo Bertolotti, Elbert G Van Putten, Christian Blum, Ad Lagendijk, Willem L Vos, and Allard P Mosk. Non-invasive imaging through opaque scattering layers. *Nature*, 491(7423):232–234, 2012.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–236, 1974.

Manu Bhat, Jonghyun Park, Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Atlasd: Automatic local symmetry discovery. *arXiv preprint arXiv:2504.10777*, 2025.

Sergey Bravyi, Sarah Sheldon, Abhinav Kandala, David C. Mckay, and Jay M. Gambetta. Mitigating measurement errors in multiqubit experiments. *Physical Review A*, 103(4):042605, 2021.

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116 (45):22445–22451, October 2019. ISSN 1091-6490. doi: 10.1073/pnas.1906995116. URL http://dx.doi.org/10.1073/pnas.1906995116.

Rishidev Chaudhuri, Berk Gerçek, Biraj Pandey, Adrien Peyrache, and Ila Fiete. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature Neuroscience*, 22(9):1512–1520, 2019.

Victor Churchill and Dongbin Xiu. Flow map learning for unknown dynamical systems: Overview, implementation, and benchmarks. *Journal of Machine Learning for Modeling and Computing*, 4 (2):173–201, 2023. ISSN 2689-3967.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. doi: 10.1016/0165-1684(94)90029-9.

S Corbetta, D Ancora, and A Bassi. Blind deconvolution in autocorrelation inversion for multiview light-sheet microscopy. *Microscopy Research and Technique*, 85(6):2349–2358, 2022.

T.M. Cover. *Elements of Information Theory*. Wiley series in telecommunications and signal processing. Wiley-India, 1999. ISBN 9788126508143. URL https://books.google.com.tr/books?id=3yGJrqyanyYC.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. *Advances in Neural Information Processing Systems*, 34:2503–2515, 2021.

Krish Desai, Benjamin Nachman, and Jesse Thaler. Symmetry discovery with deep learning. *Physical Review D*, 105(9):096031, 2022.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23): 6160–6173, 2016. doi: 10.1109/TSP.2016.2602809.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on pure and applied mathematics*, 28(1):1–47, 1975.

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

Onur Efe and Arkadas Ozakin. Symmetrylens: Unsupervised symmetry learning via locality and density preservation. *Symmetry*, 17(3), 2025. doi: 10.3390/sym17030425.

Nick I Fisher and Alan J Lee. A correlation coefficient for circular data. *Biometrika*, 70(2):327–332, 1983.

Roy T Forestano, Konstantin T Matchev, Katia Matcheva, Alexander Roman, Eyup B Unlu, and Sarunas Verner. Deep learning symmetries and their lie groups, algebras, and subalgebras from first principles. *Machine Learning: Science and Technology*, 4(2):025027, 2023.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *International conference on machine learning*, pp. 1972–1982. PMLR, 2019.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Loukas Grafakos. *Classical Fourier Analysis*, volume 249 of *Graduate Texts in Mathematics*. Springer, New York, 3 edition, 2014. doi: 10.1007/978-1-4939-1194-3.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1st edition, 1991.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469. PMLR, 2017.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000. doi: 10.1016/S0893-6080(00)00026-5.

Ernst Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.

S Rao Jammalamadaka and Ambar Sengupta. *Topics in circular statistics*, volume 5. world scientific, 2001.

Kjetil F Kaaresen and Torfinn Taxt. Multichannel blind deconvolution of seismic signals. *Geophysics*, 63(6):2093–2107, 1998.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 920–929. PMLR, 2016.

Ori Katz, Pierre Heidmann, Mathias Fink, and Sylvain Gigan. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature Photonics*, 8(10): 784–790, 2014.

Ross Kindermann and J. Laurie Snell. *Markov Random Fields and Their Applications*, volume 1 of *Contemporary Mathematics*. American Mathematical Society, Providence, RI, 1980.

Masanori Koyama, Kenji Fukumizu, Kohei Hayashi, and Takeru Miyato. Neural fourier transform: A general approach to equivariant representation learning. *arXiv preprint arXiv:2305.18484*, 2023.

Christof Külske. The ising model: highlights and perspectives: C. külske. *Mathematical Physics, Analysis and Geometry*, 28(3):20, 2025.

Anat Levin, Yair Weiss, Fredo Freeman, and Fredo Durand. Understanding and evaluating blind deconvolution algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1964–1971. IEEE, 2009.

Filip B. Maciejewski, Zoltán Zimborás, and Michał Oszmaniec. Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography. *Quantum*, 4:257, 2020.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884, 2020.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Thomas Mitchel, Michael J Taylor, and Vincent Sitzmann. Neural isometries: Taming transformations for equivariant ml. *Advances in Neural Information Processing Systems*, 37:7311–7338, 2024.

Takeru Miyato, Masanori Koyama, and Kenji Fukumizu. Unsupervised learning of equivariant structure from sequences. *Advances in Neural Information Processing Systems*, 35:768–781, 2022.

Naeimeh Mohseni, Peter L. McMahon, and Tim Byrnes. Ising machines as hardware solvers of combinatorial optimization problems, 2022. URL https://arxiv.org/abs/2204.00276.

Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. Liegg: Studying learned lie group generators. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25212–25223. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a120382cf4e2e06d94d7ae7ac96fbe25-Paper-Conference.pdf.

Björn Ottersten, Petre Stoica, and Richard Roy. Covariance matching estimation techniques for array signal processing applications. *Digital Signal Processing*, 8(3):185–210, 1998.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan Willem van de Meent, and Robin Walters. Learning symmetric embeddings for equivariant world models. *arXiv preprint arXiv:2204.11371*, 2022.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, pp. 17691–17715. PMLR, 2022.

Stephen O Rice. Mathematical analysis of random noise. *The Bell System Technical Journal*, 23(3): 282–332, 1944.

David W Romero and Suhas Lohit. Learning partial equivariances from data. *arXiv preprint arXiv:2110.10211*, 2021.

Nicolas Roux, Yoshua Bengio, Pascal Lamblin, Marc Joliveau, and Balázs Kégl. Learning the 2-d topology of images. *Advances in Neural Information Processing Systems*, 20, 2007.

Eduardo Santos-Escriche and Stefanie Jegelka. Learning equivariant models by discovering symmetries with learnable augmentations, 2025. URL https://arxiv.org/abs/2506.03914.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

Ben Shaw, Abram Magner, and Kevin R. Moon. Symmetry discovery beyond affine transformations, 2024. URL https://arxiv.org/abs/2406.03619.

Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.

Jascha Sohl-Dickstein, Ching Ming Wang, and Bruno A Olshausen. An unsupervised algorithm for learning lie group transformations. *arXiv preprint arXiv:1001.1027*, 2010.

Sreejith Sreekumar and Ziv Goldfeld. Neural estimation of statistical divergences, 2022. URL https://arxiv.org/abs/2110.03652.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Rupert Tombs and Christopher G Lester. A method to challenge symmetries in data with self-supervised learning. *Journal of Instrumentation*, 17(08):P08024, 2022.

Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Louis J. Podrazik, Steven G. Kratzer, Eric T. Harley, Donniell E. Fishkind, R. Jacob Vogelstein, and Carey E. Priebe. Fast approximate quadratic programming for graph matching. *PLOS ONE*, 10(4):e0121002, 2015.

Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery. *arXiv preprint arXiv:2310.00105*, 2023a.

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *International Conference on Machine Learning*, pp. 39488–39508. PMLR, 2023b.

Guodong Ye. Image scrambling encryption algorithm of pixel bit based on chaos map. *Pattern Recognition Letters*, 31(5):347–354, 2010.

Matthew D Zeiler and Rob Fergus. Stochastic pooling for regularization of deep convolutional neural networks. *arXiv preprint arXiv:1301.3557*, 2013.

Allan Zhou, Tom Knowles, and Chelsea Finn. Meta-learning symmetries by reparameterization. *arXiv preprint arXiv:2007.02933*, 2020.

Yanqiao Zhu, Weizhi Xu, Jinghao Zhang, Yuanqi Du, Jieyu Zhang, Qiang Liu, Carl Yang, and Shu Wu. A survey on graph structure learning: Progress and opportunities. *arXiv preprint arXiv:2103.03036*, 2021.

APPENDIX CONTENTS

THEORETICAL ANALYSIS 16

A Overview of the Analysis Framework **16**
A.1 Preliminaries ... 16
A.2 Foundational Assumptions ... 18
A.3 Architectural Constraints ... 18
A.4 Proof Convention ... 19

B Implications of Problem Setting and Architecture **19**
B.1 Measurement Operator Properties ... 19
B.2 Symmetry Representation and Properties ... 21
B.3 Probe Orbit Spans the Sample Space ... 23
B.4 Effect of Learnable Conditioning ... 24
B.5 Properties of Lifting Operation ... 25
B.6 Possibility of Signal Recovery ... 26

C Analysis of Information Maximization Loss **27**
C.1 Information Maximization and Singular Values of the Output Covariance ... 27
C.2 Implications of Positive Top-k Singular Value Product ... 29
C.3 Implications of Injectivity Arguments ... 29
C.4 The Lifting Matrix Maximizing Information ... 31
C.5 Emergent Periodicity of Learned Symmetry Representation ... 33

D Analysis of Stationarity Loss **35**
D.1 Assumptions ... 35
D.2 Invariance of Conditioned Sample Space Under Group Actions ... 35
D.3 Connection Between the Learned and True Group Actions ... 37

E Analysis of Total Correlation Minimization Loss **40**
E.1 Definitions ... 40
E.2 Total Correlation Surrogate ... 41
E.3 Reinterpreting Total Correlation ... 44
E.4 Total Correlation of a Stationary Field ... 46

F Analysis of Joint Loss **49**
F.1 Probe Response Spectrum in the Conditioned Sample Space ... 49
F.2 Recovery Integral on Lattice ... 54
F.3 Compactness of the Recovery Kernel ... 56
F.4 Signal Recovery ... 59

IMPLEMENTATION AND EXPERIMENTAL DETAILS 60

G Implementation Details **60**
G.1 Recovery Network ... 60
G.2 Entropy Estimation ... 61
G.3 JS-Divergence Estimation ... 62
G.4 Output Coarse-Graining and Multiscale Losses ... 64
G.5 Optimization ... 65

H Experimental Details **66**
H.1 Datasets ... 66
H.2 Configuration of Our Model ... 68
H.3 Details Related to Baselines ... 68

I Additional Results **74**
I.1 Our Model ... 75
I.2 Baselines ... 78

THEORETICAL ANALYSIS

A OVERVIEW OF THE ANALYSIS FRAMEWORK

In this section, we briefly summarize the analysis framework, and define the foundational assumptions and architectural constraints that are used throughout the derivations.

We investigate the architectural implications in Appendix B, while each loss function is analyzed within its own section in Appendix C, Appendix D, and Appendix E. We order the loss analysis according to the dependency structure within their arguments. First, we analyze information maximization, as the arguments proposed there do not assume optimality of the other loss terms. Then we analyze the stationarity loss, as it uses results from information maximization, followed by the analysis of the total correlation minimization term, which depends on results from the other two loss terms. Two results of the information maximization analysis, the flat filter spectrum and the periodic axis (Propositions C.13 and C.14), are stated under an invariance hypothesis (Hypothesis 1) that holds trivially when $d_Z = d_X$ and is established in the stationarity analysis otherwise (Corollary D.7). The identification of the symmetry (Theorem D.8) therefore holds where information maximization and stationarity are optimal jointly.

The analysis in Appendix D demonstrates that the latent symmetry representation can be discovered, while the signal recovery mechanism is investigated in Appendix F.

Remark A.1. *Throughout our analysis, we assume that the signal domain is toroidal, as flat domains can be made toroidal by stitching opposite boundaries (Section 3). This conforms to the geometric setting, and the information maximization and total correlation minimization losses do not depend directly on the wrap-around idealization. However, since boundary effects alter the statistics around the boundary, the stationarity term is affected by it. We handle this experimentally by selecting a sampling grid where boundaries are excluded and estimating the JS divergence between shifted versions of the grid. In our theoretical analysis, we will constrain ourselves to the case where toroidal idealization holds. There, translations of the model output are circular shifts of the lattice, and stationarity forces the learned symmetry representation to close up after N_i steps along each axis, so that a circular shift of the output is exactly a learned translation of the conditioned sample (Lemma D.1). Therefore, for flat domains, our treatment holds up to boundary effects.*

A.1 PRELIMINARIES

A.1.1 TENSOR SLICE NOTATION

Vectors indexed by lattice sites, such as the model output $\mathbf{Y} \in \mathbb{R}^{d_Y}$, are also viewed as order-P tensors of shape $N_1 \times \cdots \times N_P$, and we use whichever form is convenient: the matrix form in products such as $\mathbf{Y} = \mathbf{UCX}$, and the tensor form when indexing sites or slicing along axes.

We require establishing relations between groups of elements within the same tensor. For example, let $\mathbf{T}$ denote a tensor variable, and suppose we are given the following equality:

$$\mathbf{T}_{n_1,n_2,s_3,\ldots,s_P} = \mathbf{T}_{m_1,m_2,s_3,\ldots,s_P} \qquad \text{for all } s_i \in \{1,\ldots,N_i\}.$$

Here, $s_3,\ldots,s_P$ denote the indices common to both sides of the equality, whereas (n_1,n_2) and (m_1,m_2) denote distinct index sets.

Instead of using this fully explicit notation, we write only the indices that differ between the two sides of the equation, keeping the common, universally quantified indices implicit. Therefore, the above equality is compactly expressed as:

$$\mathbf{T}_{n_1,n_2} = \mathbf{T}_{m_1,m_2}.$$

We interpret this as an equality between the *tensor slices* $\mathbf{T}_{n_1,n_2}$ and $\mathbf{T}_{m_1,m_2}$. Specifically, for a tensor $\mathbf{T}$ of order P, we denote the slice of $\mathbf{T}$ where the first k indices ($k \leq P$) are fixed to specific values $i_1,\ldots,i_k$ by:

$$\mathbf{T}_{i_1,\ldots,i_k} := \mathbf{T}_{i_1,\ldots,i_k,:,\ldots,:}.$$

This operation results in a tensor of order $P-k$ for any $(i_1,\ldots,i_k)$.

For simplicity, we frequently express the fixed indices as a k-dimensional multi-index $\mathbf{i}^{(k)} := (i_1, \ldots, i_k)$. This allows us to denote this slice simply as $\mathbf{T}_{\mathbf{i}^{(k)}}$. In this notation, $\mathbf{T}_{\mathbf{i}^{(0)}}$ corresponds to the whole tensor, thus $\mathbf{T}_{\mathbf{i}^{(0)}} := \mathbf{T}$.

A.1.2 Random-Vector Formulation

We assume access to the idealized random vectors rather than the empirical distributions. Therefore, $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{Y}$ denote the random vectors associated with the sample space $\boldsymbol{\mathcal{X}}$, conditioned sample space $\boldsymbol{\mathcal{Z}}$, and model output space $\boldsymbol{\mathcal{Y}}$, respectively. Their distributions are implicit in the notation, and all expectations below are taken with respect to those distributions.

We write the covariance of a centered random variable $\mathbf{W}$ by

$$\boldsymbol{\Sigma}_W := \mathbb{E}[\mathbf{W}\mathbf{W}^\top].$$

For two random vectors $\mathbf{W}$ and $\mathbf{W}'$ taking values in a common domain $\boldsymbol{\mathcal{D}} \subseteq \mathbb{R}^d$, with probability density functions $p_W, p_{W'} : \boldsymbol{\mathcal{D}} \to \mathbb{R}_{\geq 0}$, we write $\mathbf{W} \overset{d}{=} \mathbf{W}'$ if

$$p_W(\mathbf{t}) = p_{W'}(\mathbf{t})$$

for all $\mathbf{t} \in \boldsymbol{\mathcal{D}}$.

A.1.3 Continuous Analysis

We base our analysis on the continuous forward model in Section 5. This relation connects the latent field to the model output representation through the *recovery integral* equation:

$$y(\mathbf{t}) = \int_G f(\mathbf{s}) k(\mathbf{s}, \mathbf{t})\, d\mu(\mathbf{s}), \qquad k(\mathbf{s}, \mathbf{t}) := \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t}) \rho_Z(\mathbf{s}) \mathbf{h}_Z, \tag{8}$$

where $k(\mathbf{s}, \mathbf{t})$ is referred to as the *recovery kernel*.

A.1.4 Discrete Analysis

The lifting admits the discretization of Section 4.1, which turns it into the lifting matrix $\mathbf{U} : \boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{Y}}$. The model is therefore the product of two matrices, the conditioning matrix $\mathbf{C} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Z}}$ and the lifting matrix $\mathbf{U}$, and the relation between the model output and input becomes

$$\mathbf{Y} = \mathbf{U}\mathbf{C}\mathbf{X}. \tag{9}$$

When the latent symmetry is discovered up to its unavoidable ambiguities, the recovery kernel becomes translation invariant, and the exact-discretization Lemma of the joint loss analysis (Lemma F.6) turns (8) into a circular convolution on the lattice. Writing $\mathbf{F}$ for the latent random vector of Section 3 and $\mathbf{K}$ for the circulant matrix of this convolution,

$$\mathbf{Y} = \mathbf{K}\mathbf{F}.$$

A.1.5 Low-rank Gaussian Entropy Formulation and Rank Scheduling

The loss estimation network needs to be robust in order to provide stable gradients to the recovery model. However, differential entropy estimation in high-dimensional spaces is often difficult and can be unreliable Beirlant et al. (1997); Paninski (2003); McAllester & Stratos (2020).

Therefore, we estimate the differential entropy under a Gaussian approximation. Furthermore, to ensure robustness against the small eigenvalues of the covariance matrix, we introduce a low-rank entropy, defined as follows.

Definition 1 (Low-rank entropy)**.** *Let $\boldsymbol{\Sigma}_W = \mathbb{E}[\mathbf{W}\mathbf{W}^\top]$ denote the covariance matrix of a centered random vector $\mathbf{W}$ taking values in $\mathbb{R}^d$, and let its eigenvalues $\lambda_i(\boldsymbol{\Sigma}_W)$ be ordered such that $\lambda_1(\boldsymbol{\Sigma}_W) \geq \lambda_2(\boldsymbol{\Sigma}_W) \geq \cdots \geq \lambda_d(\boldsymbol{\Sigma}_W) \geq 0$. Assuming $\lambda_k(\boldsymbol{\Sigma}_W) > 0$, we define the low-rank Gaussian entropy as*

$$h_k(\mathbf{W}) := \frac{k}{2} \log(2\pi e) + \frac{1}{2} \sum_{i=1}^{k} \log\left(\lambda_i(\boldsymbol{\Sigma}_W)\right).$$

Throughout the training, we increase the output rank from 1 to the dimension of the sample space, $\dim(\mathrm{span}(\mathbf{X})) = d_X$.

Remark A.2. *In Lemma E.3 (gradients of the smoothed joint entropy, total correlation minimization analysis), we show that the gradients obtained using this approximation are identical to those of the ideal entropy, due to the linearity of the model when $k = \dim(\boldsymbol{\mathcal{X}})$.*

A.2 Foundational Assumptions

Assumption 1 (Latent symmetry group)**.** *The latent symmetry group is a connected compact Abelian Lie group G acting freely and transitively on the signal domain.*

Assumption 2 (Symmetry maximality)**.** *With ρ_X denoting the action of latent translations on the sample space, every linear map $\mathbf{T}$ preserving the distribution of $\mathbf{X}$ can be written as $\mathbf{T} = \rho_X(\mathbf{t})$ for some $\mathbf{t} \in \mathbb{R}^P$.*

Assumption 3 (Support of the measurement operator)**.** *The G-invariant subspace $\mathcal{F}_R \subset L^2$ on which the bounded linear operator $\mathbf{M}$ is injective is spanned by the characters of a box-shaped set of integer frequencies,*

$$\boldsymbol{\mathcal{B}} = \left\{\mathbf{m} \in \mathbb{Z}^P \ : \ |m_j| \leq M_j,\ j = 1, \ldots, P\right\},$$

whose half-widths satisfy, with $H_j := \frac{N_j - 1}{2}$ the Nyquist limit of axis j,

(i) $M_j \leq H_j$: the band is resolved by the sampling lattice;

(ii) $M_j \geq 1$: the band contains the fundamental harmonics;

(iii) either $M_j = H_j$ for every j, so that $\boldsymbol{\mathcal{B}}$ is the full Nyquist box, or $2 \min_j M_j > \max_i H_i$, so that the band reaches beyond a quarter of the largest resolution along every axis.

For equal resolutions, the second alternative of (iii) says that the band extends beyond half of the Nyquist range.

Assumption 4 (Non-degenerate sample distribution)**.** *Let $\mathbf{X}$ be the centered random sample vector. The covariance $\boldsymbol{\Sigma}_X := \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]$ is non-degenerate*

$$\mathrm{rank}(\boldsymbol{\Sigma}_X) = d_X.$$

A.3 Architectural Constraints

Constraint 1 (Special orthogonal symmetry representation)**.** *The learned symmetry representation takes values in $\mathrm{SO}(d_Z)$ and is parameterized by P commuting skew-symmetric generators, $\hat{\rho}_Z(\mathbf{t}) := \exp\left(\sum_{j=1}^P t_j \hat{\mathbf{L}}_j\right)$.*

Constraint 2 (Eigenphase bound)**.** *The eigenphases of the learned symmetry representation are constrained to $(-\pi, \pi)$. Since the grid step along axis j is $2\pi/N_j$, a mode of angular frequency $\hat{k}^{(j)}$ advances $2\pi\hat{k}^{(j)}/N_j$ per step, and the constraint reads*

$$\left|\hat{k}^{(j)}\right| < \frac{N_j}{2}, \qquad j = 1, \ldots, P.$$

This is the Nyquist limit of the grid of Constraint 4.

Constraint 3 (L_2-normalized recovery filter)**.** *The recovery filter is normalized, $\|\mathbf{w}_Z\|_2 = 1$.*

Constraint 4 (Conditioned sample space and sampling lattice dimensions)**.** *The conditioned sample space and the sampling lattice are at least as large as the sample space:*

$$d_Z = \dim(\boldsymbol{\mathcal{Z}}) \ \geq \ \dim(\boldsymbol{\mathcal{X}}) = d_X, \qquad d_Y = \prod_{p=1}^{P} N_p \ \geq \ d_X,$$

where d_Y is the number of grid points. When $d_Z = d_X$, the conditioning matrix is square; zero padding and the dimension-increasing learnable conditioning matrix of the experiments give $d_Z > d_X$.

Constraint 5 (Grid resolution)**.** *Each axis resolution is odd and at least three,*

$$N_j \in \{3, 5, 7, \dots\}, \qquad j = 1, \dots, P.$$

Oddness makes the Nyquist band $\left\{-\frac{N_j-1}{2}, \dots, \frac{N_j-1}{2}\right\}$ *symmetric about zero and its corner frequency* $\frac{N_j-1}{2}$ *an integer mode, so that condition (i) of Assumption 3 can hold with equality.*

Constraint 6 (Rank schedule)**.** *The rank* k *used in the low-rank entropy estimates is scheduled between* 1 *and the rank* d_X *of the sample covariance during training, as described in Section 4.2. Unless stated otherwise, the analyses are carried out at the terminal rank* $k = d_X$*.*

Constraint 7 (Centered variables)**.** *Input samples* $\mathbf{X}$ *are centered such that* $\mathbb{E}[\mathbf{X}] = \mathbf{0}$*. Due to linearity of the model, this also implies* $\mathbb{E}[\mathbf{Z}] = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$*.*

Constraint 8 (Batch size)**.** *Mini-batches satisfy* $B \geq d_X$*. The population quantities of the analyses are estimated from a batch through the empirical covariance* $\hat{\mathbf{\Sigma}}_Y = \frac{1}{B}\sum_{b=1}^{B} \mathbf{y}_b \mathbf{y}_b^\top$ *of the batch samples* $\mathbf{y}_1, \dots, \mathbf{y}_B$*, whose rank is at most* B*; the bound is therefore what allows the estimate to attain the rank* $k = d_X$ *at which the analyses are carried out.*

A.4 Proof Convention

The analyses that follow draw on three kinds of statements, each cited in its own way. We refer to lemmas, propositions, theorems, corollaries, and remarks collectively as *results*, and distinguish them from the *assumptions* of Appendix A.2 and the *constraints* of Appendix A.3. The results and assumptions that a statement depends on are named in the statement itself, whereas constraints are invoked only in the proofs.

B Implications of Problem Setting and Architecture

In this section, we will analyze the implications of the problem setting and the proposed architecture under the assumptions and constraints acknowledged in Appendix A.2, without considering optimization aspects.

B.1 Measurement Operator Properties

Lemma B.1 (Decomposition of the measurement operator)**.** *Under Assumption 3 (support of the measurement operator), the measurement operator admits a decomposition*

$$\mathbf{M} = \mathbf{A}\mathbf{S}\mathrm{P}_R,$$

where $\mathrm{P}_R : \mathcal{F} \to \mathcal{F}_R$*,* $\mathbf{S} : \mathcal{F}_R \to \boldsymbol{\mathcal{F}}$*, and* $\mathbf{A} : \boldsymbol{\mathcal{F}} \to \boldsymbol{\mathcal{X}}$ *denote the orthogonal projector onto the reconstructible subspace of the latent signal space* $\mathcal{F} := L^2(G, \mathbb{R})$*, an isometric sampling operator mapping continuous functions to discrete vectors in* $\boldsymbol{\mathcal{F}}$*, and an invertible square mixing matrix, respectively.*

Proof. We need to show that there exist $\mathbf{A}$, $\mathbf{S}$, and P_R such that, for any $f \in \mathcal{F}$,

$$\mathbf{A}\mathbf{S}\mathrm{P}_R[f] - \mathbf{M}[f] = 0.$$

Taking P_R to project onto $\mathcal{F}_R$, and since $\mathbf{M}$ annihilates the complement of $\mathcal{F}_R$, this is equivalent to requiring, for any $f_R \in \mathcal{F}_R$,

$$\mathbf{A}\mathbf{S}[f_R] - \mathbf{M}[f_R] = 0.$$

Since the sampling operator $\mathbf{S}$ is invertible on $\mathcal{F}_R$, we can substitute $f_R = \mathbf{S}^{-1}[f_S]$ for any $f_S \in \boldsymbol{\mathcal{F}}$, which gives

$$\mathbf{A}[f_S] - \mathbf{M}\mathbf{S}^{-1}[f_S] = 0.$$

Here $\mathbf{M}\mathbf{S}^{-1}$ is a linear operator from $\boldsymbol{\mathcal{F}}$ to $\boldsymbol{\mathcal{X}}$. Since $\mathbf{A}$ may be any linear map from $\boldsymbol{\mathcal{F}}$ to $\boldsymbol{\mathcal{X}}$, we can choose

$$\mathbf{A} = \mathbf{M}|_{\mathcal{F}_R}\mathbf{S}^{-1}.$$

As $\mathbf{A}$ composes two invertible operators, $\mathbf{S}^{-1} : \boldsymbol{\mathcal{F}} \to \mathcal{F}_R$ and $\mathbf{M}|_{\mathcal{F}_R}$, it is also invertible. □

Lemma B.2 (Characters on the discretized latent space)**.** *Let the measurement operator be decomposed as* $\mathbf{M} = \mathbf{A}\mathbf{S}\mathrm{P}_R$ *by Lemma B.1 (decomposition of the measurement operator), and define the discretized latent space as*

$$\boldsymbol{\mathcal{F}} := \mathbf{S}\mathrm{P}_R\mathcal{F} = \mathbf{S}\mathcal{F}_R.$$

Since the sampling operator is isometric, the eigenvectors of the induced representation, i.e. the sampled characters $\mathbf{u}_\mathbf{m} := \mathbf{S}^C[\chi_\mathbf{m}]$*, are orthonormal.*

Proof. Since sampling is an invertible linear map, latent symmetry has an induced representation ρ_F on the discretized latent space

$$\rho_F(\mathbf{s})\mathbf{S} = \mathbf{S}\tau_\mathbf{s}.$$

Let us investigate the action of the sampling operator on characters. To this end, we define the complexification of the sampling operator as

$$\mathbf{S}^C[q] := \mathbf{S}[\mathrm{R}(q)] + \mathrm{i}\,\mathbf{S}[\mathrm{I}(q)],$$

for $q \in \mathcal{F}_R^C := \{f + \mathrm{i}g \mid f, g \in \mathcal{F}_R\}$, where the operators $\mathrm{R} : \mathcal{F}_R^C \to \mathcal{F}_R$ and $\mathrm{I} : \mathcal{F}_R^C \to \mathcal{F}_R$ take the real and imaginary parts of a complex function.

Investigating the action on character $\chi_\mathbf{m}$ gives

$$\begin{aligned}\rho_F(\mathbf{s})\mathbf{S}^C[\chi_\mathbf{m}] &= \mathbf{S}^C\tau_\mathbf{s}[\chi_\mathbf{m}] \\ &= e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{s}}\mathbf{S}^C[\chi_\mathbf{m}].\end{aligned}$$

Therefore, $\mathbf{u}_\mathbf{m} := \mathbf{S}^C[\chi_\mathbf{m}]$ is an eigenvector of ρ_F with the eigenvalue $e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{s}}$.

Due to orthogonality of characters,

$$\langle \chi_\mathbf{m}, \chi_{\boldsymbol{\ell}} \rangle_G = \int_G \overline{\chi_\mathbf{m}(\mathbf{s})}\chi_{\boldsymbol{\ell}}(\mathbf{s})d\mu(\mathbf{s}) = \delta_{\mathbf{m},\boldsymbol{\ell}}$$

As $\mathbf{S}^C$ is an isometry

$$\left(\mathbf{S}^C[\chi_\mathbf{m}]\right)^\dagger \left(\mathbf{S}^C[\chi_{\boldsymbol{\ell}}]\right) = \mathbf{u}_\mathbf{m}^\dagger \mathbf{u}_{\boldsymbol{\ell}} = \delta_{\mathbf{m},\boldsymbol{\ell}}$$

Therefore, each character is represented with an orthonormal eigenvector $\mathbf{u}_\mathbf{m}$. □

Lemma B.3 (Spectrum of probe response before mixing)**.** *Let the measurement operator be decomposed as* $\mathbf{M} = \mathbf{A}\mathbf{S}\mathrm{P}_R$ *by Lemma B.1 (decomposition of the measurement operator), and let* $\boldsymbol{\mathcal{F}} = \mathbf{S}\mathcal{F}_R$ *be the discretized latent space of Lemma B.2. Denote by* ρ_F *the symmetry representation induced on the discretized latent space, and by* $\mathbf{h}_F$ *the probe response there. Since the sampling operator is isometric, the component of the probe response along each unit eigenvector of* ρ_F *is*

$$h_{F,\mathbf{m}} = 1.$$

Proof. The probe response on the discretized latent space is

$$\mathbf{h}_F := \mathbf{S}\mathrm{P}_R[\delta_F] = \mathbf{S}[\delta_F],$$

where the second equality holds because $\delta_F \in \mathcal{F}_R$ and P_R acts as the identity there.

To investigate the spectral properties of the probe response on the discretized latent space, let us recall the reproducing property of δ_F: for any $f \in \mathcal{F}_R$,

$$\int_G f(\mathbf{s})\,(\tau_\mathbf{s}\delta_F)(\mathbf{t})\,d\mu(\mathbf{s}) = f(\mathbf{t}).$$

Applying the linear sampling operator to both sides gives

$$\int_G f(\mathbf{s})\,\mathbf{S}[\tau_\mathbf{s}\delta_F]\,d\mu(\mathbf{s}) = \mathbf{S}[f].$$

Since $\mathbf{S}$ is invertible on $\mathcal{F}_R$, the symmetry representation on the discretized latent space, $\rho_F(\mathbf{s}) := \mathbf{S}\tau_{\mathbf{s}}\mathbf{S}^{-1}$, is well defined and satisfies $\rho_F(\mathbf{s})\mathbf{S} = \mathbf{S}\tau_{\mathbf{s}}$. Hence

$$\int_G f(\mathbf{s})\, \rho_F(\mathbf{s})\mathbf{S}[\delta_F]\, d\mu(\mathbf{s}) = \mathbf{S}[f],$$

and since $\mathbf{h}_F = \mathbf{S}[\delta_F]$, this becomes

$$\int_G f(\mathbf{s})\, \rho_F(\mathbf{s})\mathbf{h}_F\, d\mu(\mathbf{s}) = \mathbf{S}[f]. \tag{10}$$

Now we investigate the action of the sampling operator on characters. To this end, we define the complexification of the sampling operator as

$$\mathbf{S}^C[q] := \mathbf{S}[\mathrm{R}(q)] + \mathrm{i}\,\mathbf{S}[\mathrm{I}(q)],$$

for $q \in \mathcal{F}_R^C := \{f + \mathrm{i}g \mid f, g \in \mathcal{F}_R\}$, where the operators $\mathrm{R} : \mathcal{F}_R^C \to \mathcal{F}_R$ and $\mathrm{I} : \mathcal{F}_R^C \to \mathcal{F}_R$ take the real and imaginary parts of a complex function.

Extending (10) linearly to $\mathcal{F}_R^C$ and setting $f = \chi_{\mathbf{m}}$ gives

$$\int_G \chi_{\mathbf{m}}(\mathbf{s})\, \rho_F(\mathbf{s})\mathbf{h}_F\, d\mu(\mathbf{s}) = \mathbf{S}^C[\chi_{\mathbf{m}}].$$

Writing $\mathbf{u}_{\mathbf{m}} := \mathbf{S}^C[\chi_{\mathbf{m}}]$ for the sampled character, this reads

$$\int_G \chi_{\mathbf{m}}(\mathbf{s})\, \rho_F(\mathbf{s})\mathbf{h}_F\, d\mu(\mathbf{s}) = \mathbf{u}_{\mathbf{m}}.$$

Each $\mathbf{u}_{\mathbf{m}}$ is an eigenvector of ρ_F: since $\tau_{\mathbf{s}}\chi_{\mathbf{m}} = e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{s}}\chi_{\mathbf{m}} = \overline{\chi_{\mathbf{m}}(\mathbf{s})}\,\chi_{\mathbf{m}}$, we have $\rho_F(\mathbf{s})\mathbf{u}_{\mathbf{m}} = \overline{\chi_{\mathbf{m}}(\mathbf{s})}\,\mathbf{u}_{\mathbf{m}}$. The characters are orthonormal under the normalized Haar measure, and the sampling operator is isometric by Lemma B.1 (decomposition of the measurement operator), so the eigenvectors $\{\mathbf{u}_{\mathbf{m}}\}$ are orthonormal as well. We can therefore expand $\rho_F(\mathbf{s}) = \sum_{\mathbf{k}} \overline{\chi_{\mathbf{k}}(\mathbf{s})}\,\mathbf{u}_{\mathbf{k}}\mathbf{u}_{\mathbf{k}}^\dagger$, resulting in

$$\int_G \sum_{\mathbf{k}} \chi_{\mathbf{m}}(\mathbf{s})\overline{\chi_{\mathbf{k}}(\mathbf{s})}\,\mathbf{u}_{\mathbf{k}}\mathbf{u}_{\mathbf{k}}^\dagger\mathbf{h}_F\, d\mu(\mathbf{s}) = \mathbf{u}_{\mathbf{m}}.$$

Taking the integral, the orthogonality of characters, $\int_G \chi_{\mathbf{m}}(\mathbf{s})\overline{\chi_{\mathbf{k}}(\mathbf{s})}\, d\mu(\mathbf{s}) = \delta_{\mathbf{mk}}$, leaves only the term $\mathbf{k} = \mathbf{m}$:

$$\mathbf{u}_{\mathbf{m}}\mathbf{u}_{\mathbf{m}}^\dagger\mathbf{h}_F = \mathbf{u}_{\mathbf{m}}.$$

Therefore, writing $h_{F,\mathbf{m}} := \mathbf{u}_{\mathbf{m}}^\dagger\mathbf{h}_F$ for the component of the probe response along the unit eigenvector $\mathbf{u}_{\mathbf{m}}$,

$$h_{F,\mathbf{m}} = 1.$$

□

B.2 Symmetry Representation and Properties

Lemma B.4 (Translation symmetry representation exists)**.** *Under Assumption 3 (support of the measurement operator), of which only the injectivity of $\mathbf{M}$ on the G-invariant subspace $\mathcal{F}_R$ is used, there exists a representation $\rho_X : \mathbb{R}^P \to \mathrm{GL}(\boldsymbol{\mathcal{X}})$ satisfying*

$$\mathbf{M}\tau_{\mathbf{s}}f = \rho_X(\mathbf{s})\mathbf{M}f$$

for any $\mathbf{s} \in \mathbb{R}^P$, $f \in \mathcal{F}_R$.

Proof. For any $f \in \mathcal{F}_R$,

$$\mathbf{M}\tau_{\mathbf{s}}f = \mathbf{M}\tau_{\mathbf{s}}\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1}\mathbf{M}f,$$

where $\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1}$ is well defined because $\mathbf{M}$ is injective on $\mathcal{F}_R$. The expression $\mathbf{M}\tau_{\mathbf{s}}\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1}$ is composed of linear maps and, since $\mathcal{F}_R$ is translation invariant, takes an element of $\boldsymbol{\mathcal{X}}$ and maps it back to $\boldsymbol{\mathcal{X}}$. Therefore, we define the linear map $\mathbf{R}_{\mathbf{s}} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$:

$$\mathbf{R}_{\mathbf{s}} := \mathbf{M}\tau_{\mathbf{s}}\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1}$$

Now, we can show that $\mathbf{R}_{\mathbf{s}}$ satisfies the group laws. For any $\mathbf{s}, \mathbf{t} \in \mathbb{R}^P$, using $\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1}\mathbf{M} = \mathrm{I}$ on $\mathcal{F}_R$, we have:

- $\mathbf{R_0} = \mathbf{M}\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1} = \mathbf{I}$
- $\mathbf{R_s}\mathbf{R_t} = \mathbf{M}\tau_\mathbf{s}\tau_\mathbf{t}\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1} = \mathbf{M}\tau_{\mathbf{s+t}}\left(\mathbf{M}|_{\mathcal{F}_R}\right)^{-1} = \mathbf{R_{s+t}}$
- $\mathbf{R_s}\mathbf{R_{-s}} = \mathbf{R_0} = \mathbf{I}$, so each $\mathbf{R_s}$ is invertible

Hence $\rho_X(\mathbf{s}) := \mathbf{R_s}$ is the required representation. It is 2π-periodic, since $\tau_\mathbf{s}$ depends on $\mathbf{s}$ only modulo 2π, and continuous, since $\mathbf{M}$ is bounded and $\mathbf{s} \mapsto \tau_\mathbf{s} f$ is continuous in L^2. □

Corollary B.5 (Spectrum of the translation symmetry representation)**.** *Under Assumption 3 (support of the measurement operator), of which only the injectivity of* $\mathbf{M}$ *on the G-invariant subspace* $\mathcal{F}_R$ *is used, the spectrum of* $\rho_X(\mathbf{s})$ *restricted to the image subspace* $\boldsymbol{\mathcal{X}} = \mathbf{M}(\mathcal{F}_R)$ *coincides with the spectrum of latent translations* $\tau_\mathbf{s}$ *on* $\mathcal{F}_R$*:*

$$\operatorname{spec}\left(\rho_X(\mathbf{s})\right) = \operatorname{spec}\left(\tau_\mathbf{s}|_{\mathcal{F}_R}\right).$$

Proof. Let $\mathcal{F}_R^C := \{f + ig \mid f, g \in \mathcal{F}_R\}$ denote the complexified signal subspace, and let $\boldsymbol{\mathcal{X}}^C := \{\mathbf{u} + i\mathbf{v} \mid \mathbf{u}, \mathbf{v} \in \boldsymbol{\mathcal{X}}\}$ be the complexified sample space. We extend $\mathbf{M}$ linearly via

$$\mathbf{M}^C(f + ig) := \mathbf{M}(f) + i\mathbf{M}(g).$$

The representations $\tau_\mathbf{s}$ and $\rho_X(\mathbf{s})$ are extended to $\mathcal{F}_R^C$ and $\boldsymbol{\mathcal{X}}^C$ analogously.

As $\mathcal{F}_R$ is translation invariant, it is spanned by Fourier basis elements $e^{i\boldsymbol{\omega}\cdot\mathbf{t}} \in \mathcal{F}_R^C$, where $\boldsymbol{\omega} \in \operatorname{supp}(\mathcal{F}_R)$ denotes the frequency. Applying the complexified intertwining relation yields:

$$\rho_X(\mathbf{s})\mathbf{M}^C\left(e^{i\boldsymbol{\omega}\cdot\mathbf{t}}\right) = \mathbf{M}^C\tau_\mathbf{s}\left(e^{i\boldsymbol{\omega}\cdot\mathbf{t}}\right) \tag{11}$$

$$\rho_X(\mathbf{s})\left(\mathbf{M}^C e^{i\boldsymbol{\omega}\cdot\mathbf{t}}\right) = e^{-i\boldsymbol{\omega}\cdot\mathbf{s}}\left(\mathbf{M}^C e^{i\boldsymbol{\omega}\cdot\mathbf{t}}\right). \tag{12}$$

Since $\mathbf{M}$ is injective on $\mathcal{F}_R$, its extension $\mathbf{M}^C$ is injective on $\mathcal{F}_R^C$. Because $e^{i\boldsymbol{\omega}\cdot\mathbf{t}} \neq 0$, it follows that $\mathbf{M}^C e^{i\boldsymbol{\omega}\cdot\mathbf{t}} \neq 0$.

Therefore, $\mathbf{M}^C e^{i\boldsymbol{\omega}\cdot\mathbf{t}}$ is a non-trivial eigenvector of $\rho_X(\mathbf{s})$ with eigenvalue $e^{-i\boldsymbol{\omega}\cdot\mathbf{s}}$. Varying over all frequencies spanning $\mathcal{F}_R^C$ establishes the identity of the spectra. □

Corollary B.6 (Faithfulness of the translation symmetry representation)**.** *Under Assumption 3 (support of the measurement operator), of which only the presence of the fundamental modes* $\mathbf{e}_i$ *in* $\boldsymbol{\mathcal{B}}$ *is used, the representation* ρ_X *is faithful.*

Proof. Let $\mathbf{u_m} \in \boldsymbol{\mathcal{X}}^C$ denote an eigenvector of ρ_X at the mode $\mathbf{m}$,

$$\rho_X(\mathbf{s})\,\mathbf{u_m} = e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{s}}\,\mathbf{u_m} \qquad \text{for all } \mathbf{s} \in \mathbb{R}^P,$$

which exists for every active mode $\mathbf{m} \in \boldsymbol{\mathcal{B}}$ because, by Corollary B.5 (spectrum of the translation symmetry representation), the spectrum of ρ_X is that of the latent translations on $\mathcal{F}_R$.

Suppose now that $\rho_X(\mathbf{s}) = \rho_X(\mathbf{t})$ while $\mathbf{s} \neq \mathbf{t}$. Then

$$\rho_X(\mathbf{s} - \mathbf{t}) = \mathbf{I} \implies (e^{-\mathrm{i}\mathbf{m}\cdot(\mathbf{s}-\mathbf{t})} - 1)\mathbf{u_m} = 0$$

for every active mode $\mathbf{m} \in \boldsymbol{\mathcal{B}}$.

By condition (ii) of Assumption 3, the band contains the fundamental harmonics, so taking $\mathbf{m} = \mathbf{e}_i$ for the i-th axis gives

$$e^{-\mathrm{i}(s_i - t_i)} = 1 \implies s_i - t_i \in 2\pi\mathbb{Z}.$$

Therefore $\mathbf{s} - \mathbf{t} \in 2\pi\mathbb{Z}^P$, so the representation is faithful. □

B.3 Probe Orbit Spans the Sample Space

Both the recovery construction below and the identifiability argument of Appendix D rely on the translated probe responses exhausting the sample space.

Lemma B.7 (Probe components on active frequencies are non-zero)**.** *Under Assumption 3 (support of the measurement operator), the spectral coefficients of the probe response are non-zero on every active mode, $h_{X,\mathbf{m}} \neq 0$ for all $\mathbf{m} \in \boldsymbol{\mathcal{B}}$.*

Proof. The probe δ_F is the reproducing kernel of the reconstructible signal subspace, so

$$\mathcal{F}_R = \operatorname{span}\{\tau_{\mathbf{s}}\delta_F \,|\, \mathbf{s} \in \mathbb{R}^P\}$$

At the same time, from the definition in Section 3, the sample space $\boldsymbol{\mathcal{X}}$ is the image of the reconstructible signal subspace $\mathcal{F}_R$ under the measurement operator $\mathbf{M}$. Therefore

$$\boldsymbol{\mathcal{X}} = \mathbf{M}[\mathcal{F}_R] = \mathbf{M}[\operatorname{span}\{\tau_{\mathbf{s}}\delta_F \,|\, \mathbf{s} \in \mathbb{R}^P\}]$$

Applying the intertwining relation of Lemma B.4 (translation symmetry representation exists) gives

$$\boldsymbol{\mathcal{X}} = \operatorname{span}\{\rho_X(\mathbf{s})\mathbf{M}[\delta_F] \,|\, \mathbf{s} \in \mathbb{R}^P\} = \operatorname{span}\{\rho_X(\mathbf{s})\mathbf{h}_X \,|\, \mathbf{s} \in \mathbb{R}^P\}$$

Therefore, the probe orbit $\{\rho_X(\mathbf{s})\mathbf{h}_X \,|\, \mathbf{s} \in \mathbb{R}^P\}$ spans the sample space $\boldsymbol{\mathcal{X}}$.

Now, let us identify the eigenvectors of ρ_X. Latent translation acts as a phase shift over its characters. Let us denote $\chi_{\mathbf{m}}(\mathbf{s}) := e^{\mathbf{i}\mathbf{m}\cdot\mathbf{s}}$. Extending the intertwining relation of Lemma B.4 (translation symmetry representation exists) linearly to the complexifications gives

$$\rho_X(\mathbf{t})\mathbf{M}^C[\chi_{\mathbf{m}}] = \mathbf{M}^C[\tau_{\mathbf{t}}\chi_{\mathbf{m}}].$$

Since $\tau_{\mathbf{t}}\chi_{\mathbf{m}} = e^{-\mathbf{i}\mathbf{m}\cdot\mathbf{t}}\chi_{\mathbf{m}}$, we have

$$\rho_X(\mathbf{t})\mathbf{M}^C[\chi_{\mathbf{m}}] = e^{-\mathbf{i}\mathbf{m}\cdot\mathbf{t}}\mathbf{M}^C[\chi_{\mathbf{m}}].$$

Consequently, $\mathbf{u}_{\mathbf{m}} := \mathbf{M}^C[\chi_{\mathbf{m}}]$ is an eigenvector of the symmetry representation ρ_X with eigenvalue $e^{-\mathbf{i}\mathbf{m}\cdot\mathbf{t}}$. By Assumption 3 the characters $\{\chi_{\mathbf{m}}\}_{\mathbf{m}\in\boldsymbol{\mathcal{B}}}$ form a basis of $\mathcal{F}_R^C$, and $\mathbf{M}^C$ is injective on $\mathcal{F}_R^C$ with image $\boldsymbol{\mathcal{X}}^C$, as in the proof of Corollary B.5 (spectrum of the translation symmetry representation). Hence $\{\mathbf{u}_{\mathbf{m}}\}_{\mathbf{m}\in\boldsymbol{\mathcal{B}}}$ is a basis of $\boldsymbol{\mathcal{X}}^C$ while every $\mathbf{u}_{\mathbf{m}}$ is non-zero.

Because ρ_X takes values in $\mathrm{GL}(\boldsymbol{\mathcal{X}})$ rather than in an orthogonal group, this basis need not be orthonormal, so we read off components with its dual basis $\{\tilde{\mathbf{u}}_{\mathbf{m}}\}_{\mathbf{m}\in\boldsymbol{\mathcal{B}}}$, defined by $\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{u}_{\mathbf{m}} = \delta_{\mathbf{k}\mathbf{m}}$. Each dual vector is a left eigenvector of ρ_X: for every $\mathbf{m} \in \boldsymbol{\mathcal{B}}$,

$$\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\rho_X(\mathbf{s})\,\mathbf{u}_{\mathbf{m}} = e^{-\mathbf{i}\mathbf{m}\cdot\mathbf{s}}\,\delta_{\mathbf{k}\mathbf{m}} = e^{-\mathbf{i}\mathbf{k}\cdot\mathbf{s}}\,\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{u}_{\mathbf{m}},$$

and since $\{\mathbf{u}_{\mathbf{m}}\}$ is a basis, $\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\rho_X(\mathbf{s}) = e^{-\mathbf{i}\mathbf{k}\cdot\mathbf{s}}\,\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger$. The spectral coefficients of the probe response are $h_{X,\mathbf{k}} := \tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{h}_X$.

For contradiction, now assume that $h_{X,\mathbf{k}} = 0$ for some active mode $\mathbf{k} \in \boldsymbol{\mathcal{B}}$. As $\operatorname{span}\{\rho_X(\mathbf{s})\mathbf{h}_X \,|\, \mathbf{s} \in \mathbb{R}^P\} = \boldsymbol{\mathcal{X}}$, their projections under the dual vector $\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger$ should span the same set,

$$\operatorname{span}\{\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\rho_X(\mathbf{s})\mathbf{h}_X \,|\, \mathbf{s} \in \mathbb{R}^P\} = \operatorname{span}\{\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{x} \,|\, \mathbf{x} \in \boldsymbol{\mathcal{X}}\}.$$

Applying the left eigenvalue relation gives

$$\operatorname{span}\{e^{-\mathbf{i}\mathbf{k}\cdot\mathbf{s}}\,\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{h}_X \,|\, \mathbf{s} \in \mathbb{R}^P\} = \operatorname{span}\{\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{x} \,|\, \mathbf{x} \in \boldsymbol{\mathcal{X}}\}.$$

The left side is $\{0\}$, since $\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{h}_X = h_{X,\mathbf{k}} = 0$. The right side is all of $\mathbb{C}$: the eigenvector $\mathbf{u}_{\mathbf{k}} = \mathbf{M}^C[\chi_{\mathbf{k}}]$ lies in $\boldsymbol{\mathcal{X}}^C$, which $\boldsymbol{\mathcal{X}}$ spans, and $\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger\mathbf{u}_{\mathbf{k}} = 1$ by duality, so $\tilde{\mathbf{u}}_{\mathbf{k}}^\dagger$ does not vanish on $\boldsymbol{\mathcal{X}}$. This contradiction shows that $h_{X,\mathbf{k}} \neq 0$ for every active mode $\mathbf{k} \in \boldsymbol{\mathcal{B}}$. □

B.4 Effect of Learnable Conditioning

To ensure numerical stability during training, we parameterize the learned latent group actions using the special orthogonal group $\mathrm{SO}(\boldsymbol{\mathcal{Z}})$ rather than the full general linear group $\mathrm{GL}(\boldsymbol{\mathcal{Z}})$. In this section, we show that this choice does not restrict the class of symmetries the model can discover to $\mathrm{SO}(\boldsymbol{\mathcal{Z}})$ when using a learnable conditioning matrix.

Lemma B.8 (General linear representations admit a special orthogonal form)**.** *Under Assumption 1 (latent symmetry group), of which only the compactness and connectedness of G are used, let $\rho : G \to \mathrm{GL}(\mathbb{R}^d)$ be a real representation of G. There exists a symmetric strictly positive-definite linear transformation $\mathbf{P} \in \mathrm{GL}(\mathbb{R}^d)$ such that the conjugate representation:*

$$\tilde{\rho}(\mathbf{t}) := \mathbf{P}^{-1}\rho(\mathbf{t})\mathbf{P}$$

satisfies $\tilde{\rho}(\mathbf{t}) \in \mathrm{SO}(d)$ for all $\mathbf{t} \in G$.

Proof. First, construct the group average operator on $\mathbb{R}^d$:

$$\mathbf{B} := \int_G \rho(\mathbf{s})\rho^\top(\mathbf{s})\, d\mu(\mathbf{s}),$$

where μ is the normalized Haar measure on G. For any $\mathbf{t} \in G$, using the right-invariance of the Haar measure $d\mu(\mathbf{st}) = d\mu(\mathbf{s})$, we have:

$$\rho(\mathbf{t})\mathbf{B}\rho^\top(\mathbf{t}) = \int_G \rho(\mathbf{ts})\rho^\top(\mathbf{ts})\, d\mu(\mathbf{s}) = \mathbf{B}.$$

For any non-zero $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$:

$$\mathbf{x}^\top\mathbf{B}\mathbf{x} = \int_G \mathbf{x}^\top\rho(\mathbf{s})\rho^\top(\mathbf{s})\mathbf{x}\, d\mu(\mathbf{s}) = \int_G \|\rho^\top(\mathbf{s})\mathbf{x}\|_2^2\, d\mu(\mathbf{s}) > 0,$$

since $\rho(\mathbf{s})$ is non-singular for all $\mathbf{s} \in G$, ensuring $\|\rho^\top(\mathbf{s})\mathbf{x}\|_2^2 > 0$ everywhere. Thus, $\mathbf{B}$ is symmetric and strictly positive-definite.

Consequently, $\mathbf{B}$ admits a unique symmetric positive-definite square root $\mathbf{P} := \mathbf{B}^{1/2}$ such that $\mathbf{B} = \mathbf{P}^\top\mathbf{P}$ with $\mathbf{P} = \mathbf{P}^\top$, where $\mathbf{P}$ is strictly invertible. Substituting $\mathbf{B} = \mathbf{P}^\top\mathbf{P}$ into the invariance relation gives:

$$\begin{aligned}
&\rho(\mathbf{t})\mathbf{P}^\top\mathbf{P}\rho^\top(\mathbf{t}) = \mathbf{P}^\top\mathbf{P} \\
&\left(\mathbf{P}^\top\right)^{-1}\rho(\mathbf{t})\mathbf{P}^\top\mathbf{P}\rho^\top(\mathbf{t})\mathbf{P}^{-1} = \mathbf{I} \\
&\left(\mathbf{P}\rho^\top(\mathbf{t})\mathbf{P}^{-1}\right)^\top\left(\mathbf{P}\rho^\top(\mathbf{t})\mathbf{P}^{-1}\right) = \mathbf{I}.
\end{aligned}$$

Defining the conjugate representation $\tilde{\rho}(\mathbf{t}) := \left(\mathbf{P}^{-1}\right)^\top\rho(\mathbf{t})\mathbf{P}^\top = \mathbf{P}^{-1}\rho(\mathbf{t})\mathbf{P}$, we obtain:

$$\tilde{\rho}(\mathbf{t})\tilde{\rho}^\top(\mathbf{t}) = \mathbf{I} \quad \Longrightarrow \quad \tilde{\rho}(\mathbf{t}) \in \mathrm{O}(d).$$

Finally, since G is connected, its continuous image under $\tilde{\rho}$ lies entirely within the identity component of $\mathrm{O}(d)$, which is $\mathrm{SO}(d)$. Thus, $\tilde{\rho}(\mathbf{t}) \in \mathrm{SO}(d)$ for all $\mathbf{t} \in G$. □

Corollary B.9 (A conditioning matrix exists under which the symmetry acts orthogonally)**.** *Under Assumption 1 (latent symmetry group), let $\rho_X : G \to \mathrm{GL}(\boldsymbol{\mathcal{X}})$ denote a representation on a sample space $\boldsymbol{\mathcal{X}} \cong \mathbb{R}^{d_X}$. If $\dim(\boldsymbol{\mathcal{Z}}) \geq \dim(\boldsymbol{\mathcal{X}})$, there exists a linear conditioning matrix $\mathbf{C} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Z}}$ and a special orthogonal representation $\rho_Z : G \to \mathrm{SO}(\boldsymbol{\mathcal{Z}})$ satisfying:*

$$\mathbf{C}\rho_X(\mathbf{t}) = \rho_Z(\mathbf{t})\mathbf{C} \quad \forall \mathbf{t} \in G.$$

Proof. By Lemma B.8, there exists a positive-definite linear operator $\mathbf{P} \in \mathrm{GL}(\boldsymbol{\mathcal{X}})$ such that $\tilde{\rho_X}(\mathbf{t}) := \mathbf{P}^{-1}\rho_X(\mathbf{t})\mathbf{P} \in \mathrm{SO}(d_X)$, which rearranges to:

$$\mathbf{P}^{-1}\rho_X(\mathbf{t}) = \tilde{\rho_X}(\mathbf{t})\mathbf{P}^{-1}.$$

For any orthogonal matrix $\mathbf{Q} \in \mathrm{O}(d_Z)$, we define $\rho_Z(\mathbf{t}) \in \mathrm{SO}(\boldsymbol{\mathcal{Z}})$ via the block-diagonal conjugation:

$$\rho_Z(\mathbf{t}) := \mathbf{Q}^\top \begin{pmatrix} \tilde{\rho_X}(\mathbf{t}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d_Z - d_X} \end{pmatrix} \mathbf{Q}.$$

Since $\tilde{\rho_X}(\mathbf{t}) \in \mathrm{SO}(d_X)$ and G is connected, $\rho_Z(\mathbf{t}) \in \mathrm{SO}(\boldsymbol{\mathcal{Z}})$ for all $\mathbf{t} \in G$.

Because the conditioning matrix sends samples to the conditioned sample space, $\mathbf{z} = \mathbf{C}\mathbf{x}$, it is an element of $\mathbb{R}^{d_Z \times d_X}$. We construct it as:

$$\mathbf{C} := \mathbf{Q}^\top \begin{pmatrix} \mathbf{P}^{-1} \\ \mathbf{0}_{(d_Z - d_X) \times d_X} \end{pmatrix} \in \mathbb{R}^{d_Z \times d_X}.$$

Evaluating both sides of the desired equivariance relation:

$$\mathbf{C}\rho_X(\mathbf{t}) = \mathbf{Q}^\top \begin{pmatrix} \mathbf{P}^{-1} \\ \mathbf{0} \end{pmatrix} \rho_X(\mathbf{t}) = \mathbf{Q}^\top \begin{pmatrix} \mathbf{P}^{-1}\rho_X(\mathbf{t}) \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}^\top \begin{pmatrix} \tilde{\rho_X}(\mathbf{t})\mathbf{P}^{-1} \\ \mathbf{0} \end{pmatrix},$$

$$\begin{aligned} \rho_Z(\mathbf{t})\mathbf{C} &= \left[\mathbf{Q}^\top \begin{pmatrix} \tilde{\rho_X}(\mathbf{t}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d_Z - d_X} \end{pmatrix} \mathbf{Q}\right] \mathbf{Q}^\top \begin{pmatrix} \mathbf{P}^{-1} \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{Q}^\top \begin{pmatrix} \tilde{\rho_X}(\mathbf{t}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d_Z - d_X} \end{pmatrix} \begin{pmatrix} \mathbf{P}^{-1} \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}^\top \begin{pmatrix} \tilde{\rho_X}(\mathbf{t})\mathbf{P}^{-1} \\ \mathbf{0} \end{pmatrix}. \end{aligned}$$

Thus, $\mathbf{C}\rho_X(\mathbf{t}) = \rho_Z(\mathbf{t})\mathbf{C}$ holds identically for all $\mathbf{t} \in G$. □

Remark B.10 (Orthogonal freedom of the conditioning matrix)**.** *The construction above fixes* $\mathbf{P} = \mathbf{B}^{1/2} \in \mathrm{GL}(d_X)$ *uniquely but leaves* $\mathbf{Q}$ *free. Since the conditioning matrix acts on the left of a sample,* $\mathbf{z} = \mathbf{C}\mathbf{x}$*, this freedom is a* left*-multiplication: the construction yields the family of conditioning matrices*

$$\mathbf{C} = \mathbf{Q}^\top \begin{pmatrix} \mathbf{P}^{-1} \\ \mathbf{0}_{(d_Z - d_X) \times d_X} \end{pmatrix}, \qquad \mathbf{Q} \in \mathrm{O}(d_Z),$$

all of which admit a special orthogonal latent representation. More generally, if $\mathbf{C}$ *admits one,* ρ_Z*, then so does* $\mathbf{Q}'\mathbf{C}$*, namely* $\mathbf{Q}'\rho_Z\mathbf{Q}'^\top$*, for every* $\mathbf{Q}' \in \mathrm{O}(d_Z)$*, so* $\mathbf{C}$ *and* $\mathbf{Q}'\mathbf{C}$ *are equally valid.*

B.5 Properties of Lifting Operation

Lemma B.11 (Lifting converts learned symmetry actions to translations)**.** *Lifting operator* $\mathrm{U} : \boldsymbol{\mathcal{Z}} \to L^2([0, 2\pi)^P)$ *converts the learned symmetry action* $\hat{\rho}_Z \in \mathrm{SO}(d_Z)$ *on* $\boldsymbol{\mathcal{Z}}$ *to translations on the lifted space* $\mathcal{Y} = L^2([0, 2\pi)^P)$*. With* $\tau_{\mathbf{s}}^{\mathcal{Y}}$ *denoting translations on the function space, it satisfies*

$$\mathrm{U}\hat{\rho}_Z(\mathbf{s}) = \tau_{\mathbf{s}}^{\mathcal{Y}}\mathrm{U}.$$

Proof. The lifting operation is defined as $(\mathrm{U}\mathbf{z})(\mathbf{t}) := \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t})\mathbf{z}$ for $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$ and $\mathbf{t} \in \mathbb{R}^P$. To observe equivariance, we plug $\hat{\rho}_Z(\mathbf{s})\mathbf{z} \to \mathbf{z}$, which gives

$$\begin{aligned} (\mathrm{U}\hat{\rho}_Z(\mathbf{s})\mathbf{z})(\mathbf{t}) &= \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t})\hat{\rho}_Z(\mathbf{s})\mathbf{z} && (13) \\ &= \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t} + \mathbf{s})\mathbf{z} && (14) \\ &= (\mathrm{U}\mathbf{z})(\mathbf{t} - \mathbf{s}). && (15) \end{aligned}$$

Consequently, under the translation convention $(\tau_{\mathbf{s}}^{\mathcal{Y}} f)(\mathbf{t}) = f(\mathbf{t} - \mathbf{s})$ the lifting operator satisfies

$$\mathrm{U}\hat{\rho}_Z(\mathbf{s}) = \tau_{\mathbf{s}}^{\mathcal{Y}}\mathrm{U}.$$

where $\tau_{\mathbf{s}}^{\mathcal{Y}}$ denotes translations in the continuum idealization of recovered domain. □

Lemma B.12 (Lifted samples are continuous functions)**.** *Let* $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$ *be a conditioned sample, and let* U *be the lifting operator defined by* $(\mathrm{U}\mathbf{z})(\mathbf{t}) = \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t})\mathbf{z}$ *for group parameters* $\mathbf{t} \in \mathbb{R}^P$*. Then, the lifted function* $(\mathrm{U}\mathbf{z})(\mathbf{t})$ *is continuous in* $\mathbf{t}$*.*

Proof. Let $\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{R}^P$ be two points in the group parameter space, and define $y_1 := (\mathrm{U}\mathbf{z})(\mathbf{t}_1)$ and $y_2 := (\mathrm{U}\mathbf{z})(\mathbf{t}_2)$. By Cauchy-Schwarz and operator norm inequalities, the absolute difference between y_1 and y_2 is bounded by:

$$|y_1 - y_2| = \left|\mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t}_1)\mathbf{z} - \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t}_2)\mathbf{z}\right| \leq \|\mathbf{w}_Z\| \, \|\hat{\rho}_Z(-\mathbf{t}_1) - \hat{\rho}_Z(-\mathbf{t}_2)\| \, \|\mathbf{z}\|,$$

where $\|\cdot\|$ denotes a compatible norm.

In the case of finite-dimensional vector spaces, $\hat{\rho}_Z$ is a matrix representation, and we may choose the Frobenius norm $\|\cdot\|_F$. The continuity of the matrix exponential map guarantees that

$$\|\hat{\rho}_Z(-\mathbf{t}_1) - \hat{\rho}_Z(-\mathbf{t}_2)\|_F \to 0 \quad \text{as } \mathbf{t}_1 \to \mathbf{t}_2.$$

Consequently, $|y_1 - y_2| \to 0$ as $\mathbf{t}_1 \to \mathbf{t}_2$, demonstrating that the lifting operator maps $\mathbf{z}$ to a continuous function indexed by $\mathbf{t}$. □

B.6 Possibility of Signal Recovery

In this section, we demonstrate that the continuous forward model (8) corresponds to the identity convolution for a specific filter when the actual symmetry representation is discovered and orthogonal. Therefore, latent signals are recovered in this case.

Proposition B.13 (Oracle deconvolution recovers the active-mode projection)**.** *Following the conditioning-flexibility Corollary (Corollary B.9), let the conditioning matrix be one under which the latent symmetry acts orthogonally on the conditioned sample space $\boldsymbol{\mathcal{Z}}$, and following the probe-orbit Corollary of the information maximization analysis (Corollary C.8), let the probe response be non-zero on every active mode. Under Assumption 3 (support of the measurement operator), the projection of the latent signals to the active frequency subspace can be recovered up to a scalar.*

Proof. Let the latent symmetry group be parameterized on the torus $\mathbb{T}^P = [0, 2\pi)^P$, and let $\rho_Z(\mathbf{t})$ denote its orthogonal representation on $\boldsymbol{\mathcal{Z}}$. Our goal is to recover the projection of a signal f onto the active mode set $\boldsymbol{\mathcal{B}}$ of Assumption 3 (support of the measurement operator), which is symmetric, $\boldsymbol{\mathcal{B}} = -\boldsymbol{\mathcal{B}}$, because $\mathcal{F}_R$ is a real subspace and its characters therefore come in conjugate pairs.

The ideal orthogonal projection operator $P_B : L^2(\mathbb{T}^P) \to \mathrm{span}\{e^{i\mathbf{m}\cdot\mathbf{t}}\}_{\mathbf{m}\in\boldsymbol{\mathcal{B}}}$ is defined as:

$$(P_B f)(\mathbf{t}) := \sum_{\mathbf{m}\in\boldsymbol{\mathcal{B}}} \hat{f}(\mathbf{m}) e^{i\mathbf{m}\cdot\mathbf{t}}, \quad \text{where } \hat{f}(\mathbf{m}) = \frac{1}{(2\pi)^P} \int_{\mathbb{T}^P} f(\mathbf{t}) e^{-i\mathbf{m}\cdot\mathbf{t}} \, d\mathbf{t}.$$

This projection corresponds to a convolution with the kernel $\delta_B(\mathbf{r}) = \sum_{\mathbf{m}\in\boldsymbol{\mathcal{B}}} e^{-i\mathbf{m}\cdot\mathbf{r}}$, where $\mathbf{r} = \mathbf{s} - \mathbf{t}$.

Since $\rho_Z(\mathbf{t})$ is an orthogonal representation of a commutative group, it can be diagonalized by an orthonormal basis $\{\mathbf{u}_\mathbf{m}\}_{\mathbf{m}\in\mathbb{Z}^P}$ such that:

$$\rho_Z(\mathbf{t})\mathbf{u}_\mathbf{m} = e^{-i\mathbf{m}\cdot\mathbf{t}}\mathbf{u}_\mathbf{m}.$$

Let $\mathbf{h}_Z \in \boldsymbol{\mathcal{Z}}$ be the real-valued probe vector associated with the signal. We define its spectral coefficients as $h_{Z,\mathbf{m}} := \mathbf{u}_\mathbf{m}^\dagger \mathbf{h}_Z$. By the probe-orbit Corollary of the information maximization analysis (Corollary C.8), $h_{Z,\mathbf{m}} \neq 0$ for all active modes $\mathbf{m} \in \boldsymbol{\mathcal{B}}$.

We seek a filter $\mathbf{w}_\star$ such that the reconstructed signal $f_{\text{rec}}(\mathbf{t}) := \mathbf{w}_\star^\top \rho_Z(-\mathbf{t})\mathbf{z}$ matches the projection $(P_B f)(\mathbf{t})$. We construct this filter explicitly in the eigenbasis as:

$$\mathbf{w}_\star := \sum_{\mathbf{m}\in\boldsymbol{\mathcal{B}}} \frac{1}{h_{Z,\mathbf{m}}^*} \mathbf{u}_\mathbf{m}.$$

First, we verify that $\mathbf{w}_\star$ is real-valued. Because the set of active modes is symmetric ($\boldsymbol{\mathcal{B}} = -\boldsymbol{\mathcal{B}}$) and the representation is orthogonal, the eigenvectors satisfy $\mathbf{u}_{-\mathbf{m}} = \mathbf{u}_\mathbf{m}^*$. Since $\mathbf{h}_Z$ is real, it follows that:

$$h_{Z,-\mathbf{m}} = \mathbf{u}_{-\mathbf{m}}^\dagger \mathbf{h}_Z = (\mathbf{u}_\mathbf{m}^\dagger \mathbf{h}_Z)^* = h_{Z,\mathbf{m}}^*.$$

The components of our filter, $w_{\star,\mathbf{m}} = 1/h_{Z,\mathbf{m}}^*$, therefore satisfy $w_{\star,-\mathbf{m}} = w_{\star,\mathbf{m}}^*$. This conjugate symmetry guarantees that $\mathbf{w}_\star$ is strictly real.

Finally, we verify that this filter produces the correct projection. Because $\mathbf{w}_\star$ is real, we have $\mathbf{w}_\star^\top = \mathbf{w}_\star^\dagger$, meaning its projection onto a basis vector is $\mathbf{w}_\star^\top \mathbf{u}_\mathbf{m} = \mathbf{w}_\star^\dagger \mathbf{u}_\mathbf{m} = w_{\star,\mathbf{m}}^*$.

The effective reconstruction kernel induced by $\mathbf{w}_\star$ is given by $k(\mathbf{r}) = \mathbf{w}_\star^\top \rho_Z(\mathbf{r}) \mathbf{h}_Z$. Expanding $\rho_Z(\mathbf{r})$ over the active eigenbasis gives:

$$k(\mathbf{r}) = \mathbf{w}_\star^\top \left(\sum_{\mathbf{m} \in \boldsymbol{\mathcal{B}}} \mathbf{u}_\mathbf{m} \mathbf{u}_\mathbf{m}^\dagger e^{-i\mathbf{m}\cdot\mathbf{r}} \right) \mathbf{h}_Z \tag{16}$$

$$= \sum_{\mathbf{m} \in \boldsymbol{\mathcal{B}}} (\mathbf{w}_\star^\top \mathbf{u}_\mathbf{m})(\mathbf{u}_\mathbf{m}^\dagger \mathbf{h}_Z) e^{-i\mathbf{m}\cdot\mathbf{r}} \tag{17}$$

$$= \sum_{\mathbf{m} \in \boldsymbol{\mathcal{B}}} w_{\star,\mathbf{m}}^* \, h_{Z,\mathbf{m}} \, e^{-i\mathbf{m}\cdot\mathbf{r}}. \tag{18}$$

Substituting $w_{\star,\mathbf{m}}^* = 1/h_{Z,\mathbf{m}}$, we obtain:

$$k(\mathbf{r}) = \sum_{\mathbf{m} \in \boldsymbol{\mathcal{B}}} \left(\frac{1}{h_{Z,\mathbf{m}}} \right) h_{Z,\mathbf{m}} \, e^{-i\mathbf{m}\cdot\mathbf{r}} = \sum_{\mathbf{m} \in \boldsymbol{\mathcal{B}}} e^{-i\mathbf{m}\cdot\mathbf{r}} = \delta_B(\mathbf{r}).$$

Since the effective kernel $k(\mathbf{r})$ perfectly matches the ideal projection kernel $\delta_B(\mathbf{r})$, the filter guarantees that $f_{\text{rec}}(\mathbf{t}) = (P_B f)(\mathbf{t})$. □

C Analysis of Information Maximization Loss

In this section, we will analyze information maximization loss, which provides the theoretical basis for other loss terms and eventually symmetry identification and recovery. First we will show that low-rank entropy maximization (Appendix A.1.5) can be formulated in terms of top-k singular products. Then, we will analyze implications of low-rank entropy maximization over the recovery network.

The linear model introduced in Appendix A will be central to our analysis:

$$\mathbf{Y} = \mathbf{UCX}.$$

We will heavily rely on the sample space and model output covariance matrices, which are related by $\mathbf{\Sigma}_Y = \mathbf{UC\Sigma}_X\mathbf{C}^\top\mathbf{U}^\top$. The sample covariance is positive definite (Assumption 4), so it has a positive definite square root, and the output covariance factors through it:

$$\mathbf{\Sigma}_Y = (\mathbf{UC\Sigma}_X^{1/2})(\mathbf{UC\Sigma}_X^{1/2})^\top.$$

C.1 Information Maximization and Singular Values of the Output Covariance

Definition 2 (Top-k singular value product)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and let $k \in \{1, \ldots, \min(m,n)\}$. We define the top-$k$ singular value product of $\mathbf{A}$ as:*

$$\pi_k(\mathbf{A}) := \prod_{i=1}^{k} \sigma_i(\mathbf{A}),$$

where $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_{\min(m,n)}(\mathbf{A}) \geq 0$ denote the ordered singular values of $\mathbf{A}$.

Lemma C.1 (Submultiplicativity of the top-k singular value product)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$. For any integer $k \leq \min(m,n,p)$, the top-k singular value product is submultiplicative:*

$$\pi_k(\mathbf{AB}) \leq \pi_k(\mathbf{A})\pi_k(\mathbf{B}).$$

Proof. By Horn's singular value inequality for matrix products (Horn & Johnson, 1991, Theorem 3.3.4), the product of the first k singular values of $\mathbf{AB}$ is bounded by the product of the first k singular values of $\mathbf{A}$ and $\mathbf{B}$:

$$\prod_{i=1}^{k} \sigma_i(\mathbf{AB}) \leq \prod_{i=1}^{k} \left(\sigma_i(\mathbf{A})\sigma_i(\mathbf{B})\right).$$

Rearranging the product on the right-hand side yields:

$$\prod_{i=1}^{k}(\sigma_i(\mathbf{A})\sigma_i(\mathbf{B})) = \left(\prod_{i=1}^{k}\sigma_i(\mathbf{A})\right)\left(\prod_{i=1}^{k}\sigma_i(\mathbf{B})\right) = \pi_k(\mathbf{A})\pi_k(\mathbf{B}).$$

Substituting the definition of $\pi_k(\mathbf{AB})$ completes the proof. □

Lemma C.2 (Equality of top-k products under singular vector alignment)**.** *For any matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$, let their singular value decompositions be denoted by $\mathbf{A} = \mathbf{U}_A\mathbf{D}_A\mathbf{V}_A^\top$ and $\mathbf{B} = \mathbf{U}_B\mathbf{D}_B\mathbf{V}_B^\top$. For any integer $k \le \min(m, n, p)$,*

$$\pi_k(\mathbf{AB}) = \pi_k(\mathbf{A})\pi_k(\mathbf{B})$$

when $\mathbf{V}_A = \mathbf{U}_B$.

Proof. Expanding the product using the singular value decompositions yields:

$$\pi_k(\mathbf{AB}) = \pi_k(\mathbf{U}_A\mathbf{D}_A\mathbf{V}_A^\top\mathbf{U}_B\mathbf{D}_B\mathbf{V}_B^\top).$$

Since $\mathbf{V}_A$ and $\mathbf{U}_B$ are orthogonal matrices and $\mathbf{V}_A = \mathbf{U}_B$, their product simplifies to the identity matrix ($\mathbf{V}_A^\top\mathbf{U}_B = \mathbf{I}$). Therefore:

$$\pi_k(\mathbf{AB}) = \pi_k(\mathbf{U}_A(\mathbf{D}_A\mathbf{D}_B)\mathbf{V}_B^\top).$$

Because the product of two descending sequences of non-negative numbers is also a descending sequence, $\mathbf{D}_A\mathbf{D}_B$ represents a valid singular value matrix for $\mathbf{AB}$. Thus, the singular values of the product matrix are exactly $\sigma_i(\mathbf{AB}) = \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$. Consequently:

$$\begin{aligned}\pi_k(\mathbf{AB}) &= \prod_{i=1}^{k}(\sigma_i(\mathbf{A})\sigma_i(\mathbf{B})) \\ &= \left(\prod_{i=1}^{k}\sigma_i(\mathbf{A})\right)\left(\prod_{i=1}^{k}\sigma_i(\mathbf{B})\right) \\ &= \pi_k(\mathbf{A})\pi_k(\mathbf{B}).\end{aligned}$$

□

Lemma C.3 (Maximizing low-rank entropy maximizes top-k singular value product of output covariance)**.** *For $1 \le k \le \min(d_X, d_Y)$, assume that the k largest eigenvalues of the output covariance matrix are positive. Then maximizing the low-rank entropy is equivalent to maximizing the top-k singular value product of $\mathbf{UC}\boldsymbol{\Sigma}_X^{1/2}$, where $\mathbf{U}$ and $\mathbf{C}$ correspond to the lifting and conditioning matrices.*

Proof. Information maximization is defined as a low rank entropy, which is

$$h_k(\mathbf{Y}) = \frac{k}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^{k}\log\left(\lambda_i(\boldsymbol{\Sigma}_Y)\right).$$

Substituting $\lambda_i(\boldsymbol{\Sigma}_Y) = \sigma_i(\mathbf{UC}\boldsymbol{\Sigma}_X^{1/2})^2$ which follows from $\boldsymbol{\Sigma}_Y = (\mathbf{UC}\boldsymbol{\Sigma}_X^{1/2})(\mathbf{UC}\boldsymbol{\Sigma}_X^{1/2})^\top$ and abbreviating $\sigma_i := \sigma_i(\mathbf{UC}\boldsymbol{\Sigma}_X^{1/2})$:

$$h_k(\mathbf{Y}) = \frac{k}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^{k}\log\left(\sigma_i^2\right).$$

By moving the $\frac{1}{2}$ coefficient inside the logarithm as a square root and converting the sum of logarithms into the logarithm of a product, we obtain the equality

$$h_k(\mathbf{Y}) = \frac{k}{2}\log(2\pi e) + \log\left(\prod_{i=1}^{k}\sigma_i\right).$$

In terms of the top-k singular value product, this reads

$$h_k(\mathbf{Y}) = \frac{k}{2}\log(2\pi e) + \log\left(\pi_k(\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2})\right).$$

Because the logarithm is a strictly increasing function, maximizing the low-rank entropy $h_k(\mathbf{Y})$ maximizes the top-k singular value product of $\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}$, whose squared singular values are the eigenvalues of the output covariance matrix. □

C.2 Implications of Positive Top-k Singular Value Product

In this section, we show that a positive top-k singular value product for the model output leads to several injectivity arguments.

Lemma C.4 (Injectivity of the conditioning matrix)**.** *Assume that the top-k product of the output covariance eigenvalues is strictly positive at the rank $k = d_X$. Then the conditioning matrix $\mathbf{C}$ is injective.*

Proof. Under the assumption $\pi_k(\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}) > 0$ and applying Lemma C.1 twice gives:

$$0 < \pi_k(\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}) \leq \pi_k(\mathbf{U})\pi_k(\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}) \leq \pi_k(\mathbf{U})\pi_k(\mathbf{C})\pi_k(\boldsymbol{\Sigma}_X^{1/2}). \tag{19}$$

The hypothesis $k \leq \min(d_Y, d_Z, d_X)$ of that lemma holds because $k = d_X \leq \min(d_Y, d_Z)$ by Constraint 4.

Since $\pi_k(\mathbf{C}) = \prod_{i=1}^{k}\sigma_i(\mathbf{C}) > 0$, the first k singular values of $\mathbf{C}$ are strictly positive, implying $\operatorname{rank}(\mathbf{C}) \geq k$. As $\mathbf{C} \in \mathbb{R}^{d_Z \times d_X}$ with $k = d_X$, this says $\mathbf{C}$ has full column rank, so

$$\ker(\mathbf{C}) = \{\mathbf{0}\}.$$

Consequently, $\mathbf{C}$ is injective on the sample space. The same chain gives $\pi_k(\boldsymbol{\Sigma}_X^{1/2}) > 0$, so the hypothesis already makes the sample covariance full rank; Assumption 4 (non-degenerate sample distribution) is what makes the hypothesis attainable. □

Lemma C.5 (Injectivity of the lifting matrix on the range of the conditioning matrix)**.** *Assume that the top-k product of the output covariance eigenvalues is strictly positive at the rank $k = d_X$. Then the lifting matrix $\mathbf{U}$ is injective on the range $\mathbf{C}\,\mathbb{R}^{d_X}$ of the conditioning matrix, which is the whole conditioned sample space $\boldsymbol{\mathcal{Z}}$ when $d_Z = d_X$.*

Proof. Under the assumption $\pi_k(\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}) > 0$ and applying Lemma C.1 gives:

$$0 < \pi_k(\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}) \leq \pi_k(\mathbf{U})\pi_k(\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}). \tag{20}$$

Hence the first k singular values of $\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2} \in \mathbb{R}^{d_Y \times d_X}$ are strictly positive, so this matrix has full column rank $k = d_X$. The same chain gives $\pi_k(\boldsymbol{\Sigma}_X^{1/2}) > 0$, so $\boldsymbol{\Sigma}_X^{1/2}$ is invertible and $\mathbf{U}\mathbf{C}$ has full column rank as well. Every element of the range of $\mathbf{C}$ is $\mathbf{C}\mathbf{x}$ for some $\mathbf{x}$, and $\mathbf{U}\mathbf{C}\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$, so

$$\ker(\mathbf{U}) \cap \mathbf{C}\,\mathbb{R}^{d_X} = \{\mathbf{0}\}.$$

Consequently, $\mathbf{U}$ is injective on the range of $\mathbf{C}$. Combined with Lemma C.4, the composite map $\mathbf{x} \mapsto \mathbf{U}\mathbf{C}\mathbf{x}$ is injective, so no information about the conditioned sample is lost by lifting. □

C.3 Implications of Injectivity Arguments

C.3.1 Injectivity of Conditioning Matrix

In Appendix B.2, we establish arguments about existence of a symmetry representation, and its properties. Besides, in Appendix B.3 we show that the probe orbit of $\mathbf{h}_X$ spans the sample space $\boldsymbol{\mathcal{X}}$. In this section we will show that these properties are transferred to the conditioned sample space $\boldsymbol{\mathcal{Z}}$ when $\mathbf{C}$ is injective.

Corollary C.6 (Existence of a conditioned sample space symmetry)**.** *Under Lemma C.4 (injectivity of the conditioning matrix) and the representation-existence Lemma of the problem setting (Lemma B.4), the latent symmetry group G admits a representation on the conditioned sample space,*

$$\rho_Z : \mathbb{R}^P \to \mathrm{GL}(\boldsymbol{\mathcal{Z}}), \qquad \rho_Z(\mathbf{s}) := \mathbf{C}\,\rho_X(\mathbf{s})\,\mathbf{C}^+ + (\mathbf{I} - \boldsymbol{\Pi}_C)\,,$$

where $\mathbf{C}^+$ is the Moore–Penrose pseudo-inverse of $\mathbf{C}$ and $\boldsymbol{\Pi}_C := \mathbf{C}\mathbf{C}^+$ is the orthogonal projector onto the range of $\mathbf{C}$. It intertwines the conditioning matrix, $\rho_Z(\mathbf{s})\,\mathbf{C} = \mathbf{C}\,\rho_X(\mathbf{s})$, and leaves the range of $\mathbf{C}$ invariant. When $d_Z = d_X$, $\mathbf{C}^+ = \mathbf{C}^{-1}$, $\boldsymbol{\Pi}_C = \mathbf{I}$, and $\rho_Z(\mathbf{s}) = \mathbf{C}\,\rho_X(\mathbf{s})\,\mathbf{C}^{-1}$.

Proof. By Lemma C.4 the conditioning matrix is injective, i.e. it has full column rank. Its pseudo-inverse $\mathbf{C}^+ = (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top$ is then a left inverse, and $\boldsymbol{\Pi}_C = \mathbf{C}\mathbf{C}^+$ is the orthogonal projector onto the range of $\mathbf{C}$:

$$\mathbf{C}^+\mathbf{C} = \mathbf{I}_{d_X}, \qquad \mathbf{C}^+\,(\mathbf{I} - \boldsymbol{\Pi}_C) = \mathbf{0}, \qquad (\mathbf{I} - \boldsymbol{\Pi}_C)\,\mathbf{C} = \mathbf{0}.$$

The representation ρ_X on the sample space is supplied by Lemma B.4.

We may therefore define, for every $\mathbf{s} \in \mathbb{R}^P$,

$$\rho_Z(\mathbf{s}) := \mathbf{C}\,\rho_X(\mathbf{s})\,\mathbf{C}^+ + (\mathbf{I} - \boldsymbol{\Pi}_C)\,,$$

which is a linear map of $\boldsymbol{\mathcal{Z}}$ to itself. It acts through ρ_X on the range of $\mathbf{C}$ and as the identity on its orthogonal complement, as in the block construction of the conditioning-flexibility Corollary of the problem setting (Corollary B.9). Multiplying on the right by $\mathbf{C}$, and using $\mathbf{C}^+\mathbf{C} = \mathbf{I}_{d_X}$ and $(\mathbf{I} - \boldsymbol{\Pi}_C)\mathbf{C} = \mathbf{0}$, gives the intertwining relation $\rho_Z(\mathbf{s})\mathbf{C} = \mathbf{C}\,\rho_X(\mathbf{s})$; in particular, $\rho_Z(\mathbf{s})$ maps the range of $\mathbf{C}$ into itself.

It remains to check the group laws. Since ρ_X is a representation,

- $\rho_Z(\mathbf{0}) = \mathbf{C}\mathbf{C}^+ + \mathbf{I} - \boldsymbol{\Pi}_C = \mathbf{I}$;
- the cross terms of $\rho_Z(\mathbf{s})\rho_Z(\mathbf{t})$ vanish by the identities above, and $(\mathbf{I} - \boldsymbol{\Pi}_C)^2 = \mathbf{I} - \boldsymbol{\Pi}_C$, so
$$\begin{aligned}\rho_Z(\mathbf{s})\rho_Z(\mathbf{t}) &= \mathbf{C}\,\rho_X(\mathbf{s})\,\mathbf{C}^+\mathbf{C}\,\rho_X(\mathbf{t})\,\mathbf{C}^+ + (\mathbf{I} - \boldsymbol{\Pi}_C) \\ &= \mathbf{C}\,\rho_X(\mathbf{s}+\mathbf{t})\,\mathbf{C}^+ + (\mathbf{I} - \boldsymbol{\Pi}_C) = \rho_Z(\mathbf{s}+\mathbf{t});\end{aligned}$$
- consequently $\rho_Z(\mathbf{s})\rho_Z(-\mathbf{s}) = \rho_Z(\mathbf{0}) = \mathbf{I}$, so each $\rho_Z(\mathbf{s})$ is invertible.

Hence ρ_Z is a representation of G on the conditioned sample space. The identity on the complement of the range is one admissible extension; the data, and hence the analysis that follows, involve only the range of $\mathbf{C}$. □

Corollary C.7 (Faithfulness transfers to the conditioned sample space)**.** *Under Lemma C.4 (injectivity of the conditioning matrix) and the faithfulness Corollary of the problem setting (Corollary B.6), the representation ρ_Z is faithful.*

Proof. By Lemma C.4 the conditioning matrix is injective, so its pseudo-inverse is a left inverse, $\mathbf{C}^+\mathbf{C} = \mathbf{I}_{d_X}$.

Suppose $\rho_Z(\mathbf{s}) = \rho_Z(\mathbf{t})$. Substituting the definition of Corollary C.6 (existence of a conditioned sample space symmetry) and canceling the common term $\mathbf{I} - \boldsymbol{\Pi}_C$ gives

$$\mathbf{C}\,\rho_X(\mathbf{s})\,\mathbf{C}^+ = \mathbf{C}\,\rho_X(\mathbf{t})\,\mathbf{C}^+,$$

and multiplying by $\mathbf{C}^+$ on the left and $\mathbf{C}$ on the right yields $\rho_X(\mathbf{s}) = \rho_X(\mathbf{t})$. By Corollary B.6 the representation ρ_X is faithful, so $\mathbf{s} - \mathbf{t} \in 2\pi\mathbb{Z}^P$. Therefore ρ_Z is faithful. □

Corollary C.8 (Probe orbit spans the range of the conditioning matrix)**.** *Under Lemma C.4 (injectivity of the conditioning matrix) and the probe-component Lemma of the problem setting (Lemma B.7), the probe response on the conditioned sample space, $\mathbf{h}_Z := \mathbf{C}\mathbf{h}_X$, has an orbit spanning the range of the conditioning matrix,*

$$\mathrm{span}\{\rho_Z(\mathbf{s})\mathbf{h}_Z \mid \mathbf{s} \in \mathbb{R}^P\} = \mathbf{C}\,\boldsymbol{\mathcal{X}},$$

and its spectral coefficients are non-zero on every active mode, $h_{Z,\mathbf{m}} \neq 0$ for all $\mathbf{m} \in \boldsymbol{\mathcal{B}}$. Consequently the spectrum of ρ_Z on the range $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ is exactly the active mode set $\boldsymbol{\mathcal{B}}$. When $d_Z = d_X$, $\mathbf{C}\,\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Z}}$.

Proof. Applying the intertwining relation of Corollary C.6 (existence of a conditioned sample space symmetry) to the probe response gives

$$\rho_Z(\mathbf{s})\mathbf{h}_Z = \rho_Z(\mathbf{s})\,\mathbf{C}\mathbf{h}_X = \mathbf{C}\,\rho_X(\mathbf{s})\mathbf{h}_X,$$

so the orbit in the conditioned sample space is the image under $\mathbf{C}$ of the orbit in the sample space. By Lemma B.7 the latter spans $\boldsymbol{\mathcal{X}}$, and therefore

$$\operatorname{span}\{\rho_Z(\mathbf{s})\mathbf{h}_Z\} = \mathbf{C}\operatorname{span}\{\rho_X(\mathbf{s})\mathbf{h}_X\} = \mathbf{C}\,\boldsymbol{\mathcal{X}}.$$

For the spectral coefficients, expand $\mathbf{h}_X = \sum_{\mathbf{m}} h_{X,\mathbf{m}}\mathbf{u}^X_{\mathbf{m}}$ in the eigenbasis of ρ_X. Conjugation carries eigenvectors to eigenvectors at the same eigenvalue, $\rho_Z(\mathbf{s})\mathbf{C}\mathbf{u}^X_{\mathbf{m}} = e^{-\mathbf{i}\mathbf{m}\cdot\mathbf{s}}\mathbf{C}\mathbf{u}^X_{\mathbf{m}}$, so $\{\mathbf{C}\mathbf{u}^X_{\mathbf{m}}\}$, which is linearly independent because $\mathbf{C}$ is injective, is an eigenbasis of ρ_Z on $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ and

$$\mathbf{h}_Z = \mathbf{C}\mathbf{h}_X = \sum_{\mathbf{m}} h_{X,\mathbf{m}}\,\mathbf{C}\mathbf{u}^X_{\mathbf{m}}.$$

The expansion coefficients of $\mathbf{h}_Z$ in this eigenbasis therefore coincide with those of $\mathbf{h}_X$, and Lemma B.7 gives $h_{X,\mathbf{m}} \neq 0$ on every active mode. Hence $h_{Z,\mathbf{m}} \neq 0$ for all $\mathbf{m} \in \boldsymbol{\mathcal{B}}$, and every active mode is present in $\mathbf{C}\,\boldsymbol{\mathcal{X}}$, so the spectrum of ρ_Z on $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ is exactly $\boldsymbol{\mathcal{B}}$. □

C.3.2 Injectivity of the Lifting Matrix on the Conditioned Sample Space

Here, we will show that injectivity of the lifting matrix on the conditioned sample space ensures that top-k entropy is well-defined and finite.

Corollary C.9 (Finiteness of the k-dimensional output entropy)**.** *By Assumption 4 (non-degenerate sample distribution) and the conditioning-matrix and lifting injectivity Lemmas of the information maximization analysis (Lemmas C.4 and C.5), the k-dimensional differential entropy of the output $h_k(\mathbf{Y})$ is finite.*

Proof. From Lemma C.4 (injectivity of the conditioning matrix), $\mathbf{C}$ is injective on $\operatorname{span}(\mathbf{X})$. And from Lemma C.5 (injectivity of the lifting matrix), $\mathbf{U}$ is injective on $\operatorname{span}(\mathbf{Z})$, where $\mathbf{Z} = \mathbf{C}\mathbf{X}$. Therefore

$$\dim(\operatorname{span}(\mathbf{Y})) = \dim(\operatorname{span}(\mathbf{C}\mathbf{X})) = \dim(\operatorname{span}(\mathbf{X})) = k.$$

The rank of the covariance matrix, i.e. its number of non-zero singular values, equals $\dim(\operatorname{span}(\mathbf{Y})) = k$. Therefore, the covariance matrix has k non-zero singular values. From Definition 1 (low-rank entropy), $h_k(\mathbf{Y})$ is finite. □

C.4 The Lifting Matrix Maximizing Information

In this section, we first show that the orthogonal degree of freedom of the conditioning matrix $\mathbf{C}$, proposed in Remark B.10 in Appendix B.4, enables the model to attain the top-k singular value product upper bound. Under Assumption 4, it is guaranteed that there exists a conditioning matrix $\mathbf{C}^*$ such that $\pi_k(\mathbf{U}\mathbf{C}^*\boldsymbol{\Sigma}_X^{1/2}) = \pi_k(\mathbf{U})\pi_k(\mathbf{C}^*\boldsymbol{\Sigma}_X^{1/2})$.

Subsequently, we show that for this optimal conditioning matrix, maximizing the output top-k singular value product requires maximizing both $\pi_k(\mathbf{U})$ and $\pi_k(\mathbf{C}^*\boldsymbol{\Sigma}_X^{1/2})$. Because the Frobenius norm of the lifting matrix is fixed by the architecture, maximizing $\pi_k(\mathbf{U})$ forces $\mathbf{U}$ to become a scaled isometry on the span of its top-k right singular vectors. This holds for all dimensions admitted by Constraint 4, $d_Z \geq d_X$ and $d_Y \geq d_X$.

Lemma C.10 (Existence of an aligning conditioning matrix)**.** *There exists a conditioning matrix $\mathbf{C}^*$ that simultaneously enables the latent symmetry representation to be orthogonal and the inequality $\pi_k(\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}) \leq \pi_k(\mathbf{U})\pi_k(\mathbf{C}\boldsymbol{\Sigma}_X^{1/2})$ to be exact. Thus, for this conditioning matrix:*

$$\pi_k(\mathbf{U}\mathbf{C}^*\boldsymbol{\Sigma}_X^{1/2}) = \pi_k(\mathbf{U})\pi_k(\mathbf{C}^*\boldsymbol{\Sigma}_X^{1/2}).$$

Proof. As established previously, for any matrices $\mathbf{A}$ and $\mathbf{B}$, $\pi_k(\mathbf{A}\mathbf{B}) = \pi_k(\mathbf{A})\pi_k(\mathbf{B})$ when the left singular vectors of $\mathbf{B}$ perfectly align with the right singular vectors of $\mathbf{A}$.

Let the singular value decomposition of the lifting matrix be $\mathbf{U} = \mathbf{U}_U \mathbf{D}_U \mathbf{V}_U^\top$, and let the decomposition of the whitened conditioned input be $\mathbf{C}\mathbf{\Sigma}_X^{1/2} = \mathbf{U}_C \mathbf{D}_C \mathbf{V}_C^\top$. Because both $\mathbf{V}_U$ and $\mathbf{U}_C$ are orthogonal matrices (or have orthonormal columns), there exists an orthogonal matrix $\mathbf{Q}$ such that:

$$\mathbf{V}_U = \mathbf{Q}\mathbf{U}_C.$$

As stated in Remark B.10, the conditioning matrix is free up to a left-orthogonal transformation without breaking the orthogonality of the latent symmetry representation. Therefore, there exists an equally valid conditioning matrix in the solution space defined by:

$$\mathbf{C}^* = \mathbf{Q}\mathbf{C}.$$

Substituting $\mathbf{C}^*$ replaces the left singular vectors of $\mathbf{C}\mathbf{\Sigma}_X^{1/2}$ with $\mathbf{Q}\mathbf{U}_C = \mathbf{V}_U$, aligning them with the right singular vectors of $\mathbf{U}$ and making the top-k singular value product exactly multiplicative. □

Proposition C.11 (Optimal lifting matrix is a scaled isometry)**.** *Suppose that the conditioning matrix is the one supplied by Lemma C.10 (existence of an aligning conditioning matrix), so that maximizing $\pi_k(\mathbf{U}\mathbf{C}^*\mathbf{\Sigma}_X^{1/2})$ amounts to maximizing $\pi_k(\mathbf{U})$ and $\pi_k(\mathbf{C}^*\mathbf{\Sigma}_X^{1/2})$ separately. Let the rank be terminal, $k = d_X$ (Constraint 6). Then information maximization forces the lifting matrix to be a scaled isometry on the span of its top-k right singular vectors,*

$$\mathbf{U}^\top \mathbf{U} = \frac{d_Y}{d_X} \mathbf{\Pi}_U,$$

where $\mathbf{\Pi}_U$ is the orthogonal projector onto that span. When $d_Z = d_X$, $\mathbf{\Pi}_U = \mathbf{I}$ and $\mathbf{U}^\top \mathbf{U} = \frac{d_Y}{d_Z}\mathbf{I}$.

Proof. The lifting matrix elements are defined as $\mathbf{U}_{ij} := \left[\mathbf{w}_Z^\top \hat{\rho_Z}(-\mathbf{t}_i)\right]_j$, so row i of $\mathbf{U}$ is the vector $\hat{\rho_Z}(-\mathbf{t}_i)^\top \mathbf{w}_Z \in \mathbb{R}^{d_Z}$. By Constraint 1 the matrix $\hat{\rho_Z}(-\mathbf{t}_i)$ is orthogonal, and by Constraint 3 we have $\|\mathbf{w}_Z\|_2 = 1$; hence every row has unit norm,

$$\left\|\hat{\rho_Z}(-\mathbf{t}_i)^\top \mathbf{w}_Z\right\|_2 = \|\mathbf{w}_Z\|_2 = 1.$$

By Constraint 4 there are d_Y rows, one per grid point. The squared Frobenius norm of the lifting matrix is therefore not an imposed budget but is fixed identically:

$$\|\mathbf{U}\|_F^2 = \sum_{i=1}^{d_Y} \left\|\hat{\rho_Z}(-\mathbf{t}_i)^\top \mathbf{w}_Z\right\|_2^2 = d_Y.$$

Because the squared Frobenius norm equals the sum of the squared singular values of $\mathbf{U} \in \mathbb{R}^{d_Y \times d_Z}$,

$$\sum_{i=1}^{\min(d_Y, d_Z)} \sigma_i^2(\mathbf{U}) = d_Y.$$

The objective involves only the top $k = d_X \leq \min(d_Y, d_Z)$ singular values. Any squared singular value beyond the top k uses part of the fixed budget without contributing to the objective, so at the maximum $\sigma_i(\mathbf{U}) = 0$ for $i > k$ and

$$\sum_{i=1}^{k} \sigma_i^2(\mathbf{U}) = d_Y.$$

Simultaneously, the objective maximizes the top-k singular value product of the lifting matrix:

$$\pi_k(\mathbf{U}) = \prod_{i=1}^{k} \sigma_i(\mathbf{U}).$$

By the Arithmetic Mean-Geometric Mean (AM-GM) inequality, the product of a set of non-negative numbers with a fixed sum is maximized when all the numbers are equal. Therefore, over all matrices with squared Frobenius norm d_Y, the product π_k is uniquely maximized when

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \frac{d_Y}{k} = \frac{d_Y}{d_X}, \qquad \sigma_i = 0 \quad (i > k).$$

This bound is taken over all matrices of the given size and norm, whereas $\mathbf{U}$ is not free: it is determined by the pair $(\mathbf{w}_Z, \hat{\rho_Z})$ through $\mathbf{U}_{ij} = [\mathbf{w}_Z^\top \hat{\rho_Z}(-\mathbf{t}_i)]_j$. The argument therefore establishes the conclusion only if the unconstrained maximizer is attainable within that parameterized family. It is. Let $\{\mathbf{v_m}\}$ be the eigenvectors of $\hat{\rho_Z}$ with eigenphases $\hat{\mathbf{k}}_\mathbf{m}$, and write $w_{Z,\mathbf{m}} := \mathbf{v}_\mathbf{m}^\dagger \mathbf{w}_Z$. Expanding the inner product $\langle \mathbf{U}\mathbf{v}_{\boldsymbol{\ell}}, \mathbf{U}\mathbf{v_m}\rangle$ in this basis gives

$$\langle \mathbf{U}\mathbf{v}_{\boldsymbol{\ell}}, \mathbf{U}\mathbf{v_m}\rangle = w_{Z,\boldsymbol{\ell}} w^*_{Z,\mathbf{m}} \sum_{i=1}^{d_Y} \exp\left(\mathrm{i}(\hat{\mathbf{k}}_{\boldsymbol{\ell}} - \hat{\mathbf{k}}_\mathbf{m}) \cdot \mathbf{t}_i\right).$$

Select k of the d_Z eigenvectors of $\hat{\rho_Z}$, choose their eigenphases to be integers differing, for each selected pair $\boldsymbol{\ell} \neq \mathbf{m}$, by a non-multiple of N_p along at least one axis p, and choose $\mathbf{w}_Z$ with $|w_{Z,\mathbf{m}}|^2 = 1/k$ on the selected modes and $w_{Z,\mathbf{m}} = 0$ on the others. For selected modes, the lattice sum then vanishes for $\boldsymbol{\ell} \neq \mathbf{m}$ and equals d_Y for $\boldsymbol{\ell} = \mathbf{m}$; for the others, $\mathbf{U}\mathbf{v_m} = \mathbf{0}$. Hence $\mathbf{U}^\top\mathbf{U} = \frac{d_Y}{k}\boldsymbol{\Pi}$, with $\boldsymbol{\Pi}$ the projector onto the span of the selected eigenvectors. Such a choice exists whenever $d_Y \geq k$, since k distinct residues modulo the grid are then available. The maximum over the parameterized family therefore coincides with the unconstrained maximum.

Substituting these singular values back into the singular value decomposition of $\mathbf{U}$, and keeping the top-k left and right singular vectors $\mathbf{U}_k$ and $\mathbf{V}_k$, yields

$$\mathbf{U} = \sqrt{\tfrac{d_Y}{d_X}}\, \mathbf{U}_k \mathbf{V}_k^\top,$$

and hence, using orthonormality of the columns of $\mathbf{U}_k$,

$$\mathbf{U}^\top\mathbf{U} = \frac{d_Y}{d_X}\mathbf{V}_k\mathbf{V}_k^\top = \frac{d_Y}{d_X}\boldsymbol{\Pi}_U, \qquad \boldsymbol{\Pi}_U := \mathbf{V}_k\mathbf{V}_k^\top.$$

Therefore the optimal lifting matrix $\mathbf{U}$ is a scaled isometry on the span of its top-k right singular vectors, and a scaled isometry on all of $\boldsymbol{\mathcal{Z}}$ when $d_Z = d_X$. □

Remark C.12 (Band-pass character of the optimal lifting matrix)**.** *By Proposition C.11 (the optimal lifting matrix is a scaled isometry), the Frobenius norm of the lifting matrix is fixed, so optimal information maximization drives every singular value outside the top-k subspace to zero, i.e. $\sigma_i = 0$ for $i > k$.*

C.5 Emergent Periodicity of Learned Symmetry Representation

C.5.1 Definitions

We denote the orthonormal eigenvectors of the learned symmetry representation $\hat{\rho_Z}$ by $\{\mathbf{v_m}\}$. The results of this subsection that concern individual modes rely on the following hypothesis on the range of the projector $\boldsymbol{\Pi}_U$ of Proposition C.11.

Hypothesis 1 (Invariance of the lifting range)**.** *The range of $\boldsymbol{\Pi}_U$ is invariant under $\hat{\rho_Z}(\mathbf{s})$ for every $\mathbf{s} \in \mathbb{R}^P$.*

Since $\{\hat{\rho_Z}(\mathbf{s})\}$ is a commuting family of orthogonal matrices, an invariant subspace is spanned by joint eigenvectors, and the eigenbasis can be chosen with d_X eigenvectors in the range of $\boldsymbol{\Pi}_U$ and the remaining $d_Z - d_X$ orthogonal to it. Under Hypothesis 1, a mode is *active* if its eigenvector lies in that range, and we index the active modes by $\mathbf{m} \in \boldsymbol{\mathcal{B}}$; the remaining modes are *inactive*. When $d_Z = d_X$, $\boldsymbol{\Pi}_U = \mathbf{I}$, so the hypothesis holds trivially and every mode is active. When $d_Z > d_X$, it does not follow from information maximization alone; it holds at the joint optimum of information maximization and stationarity, which Corollary D.7 of the stationarity analysis establishes. For every eigenvector we have:

$$\hat{\rho_Z}(\mathbf{s})\mathbf{v_m} = \mathbf{v_m} \exp\left(\mathrm{i}\hat{\mathbf{k}}_\mathbf{m} \cdot \mathbf{s}\right), \qquad \hat{\mathbf{k}}_\mathbf{m} = (\hat{k}^{(1)}_\mathbf{m}, \ldots, \hat{k}^{(P)}_\mathbf{m}) \in \mathbb{R}^P$$

where $(\hat{k}^{(1)}_\mathbf{m}, \ldots, \hat{k}^{(P)}_\mathbf{m})$ denote the learned frequencies, which are not necessarily integer.

C.5.2 Implications of the Scaled-Isometric Lifting Matrix

Throughout this section, we will take the lifting matrix as a scaled isometry with the constant established in Proposition C.11, namely $\mathbf{U}^\top\mathbf{U} = \frac{d_Y}{d_X}\mathbf{\Pi}_U$.

We start by investigating the action of $\mathbf{U}$ on the eigenvectors $\{\mathbf{v_m}\}$ for each row of $\mathbf{U}$ separately:

$$[\mathbf{U}\mathbf{v_m}]_i = \mathbf{w}_Z^\top \hat{\rho_Z}(-\mathbf{t}_i)\mathbf{v_m} = w^*_{Z,\mathbf{m}} \exp(-\mathrm{i}\hat{\mathbf{k}}_\mathbf{m} \cdot \mathbf{t}_i). \tag{21}$$

Here, $w_{Z,\mathbf{m}} = \mathbf{v}_\mathbf{m}^\dagger \mathbf{w}_Z$.

Under the scaled-isometry condition $\mathbf{U}^\top\mathbf{U} = \frac{d_Y}{d_X}\mathbf{\Pi}_U$, active eigenvectors satisfy $\mathbf{\Pi}_U\mathbf{v_m} = \mathbf{v_m}$, so for active modes $\boldsymbol{\ell}, \mathbf{m} \in \boldsymbol{\mathcal{B}}$:

$$\mathbf{v}_\boldsymbol{\ell}^\dagger \mathbf{U}^\top \mathbf{U}\mathbf{v_m} = \frac{d_Y}{d_X}\delta_{\boldsymbol{\ell},\mathbf{m}}. \tag{22}$$

Expanding the inner product $\langle \mathbf{U}\mathbf{v}_\boldsymbol{\ell}, \mathbf{U}\mathbf{v_m}\rangle$ explicitly yields:

$$\langle \mathbf{U}\mathbf{v}_\boldsymbol{\ell}, \mathbf{U}\mathbf{v_m}\rangle = \sum_{i=1}^{d_Y} w_{Z,\boldsymbol{\ell}} w^*_{Z,\mathbf{m}} \exp\left(\mathrm{i}(\hat{\mathbf{k}}_\boldsymbol{\ell} - \hat{\mathbf{k}}_\mathbf{m}) \cdot \mathbf{t}_i\right) = \frac{d_Y}{d_X}\delta_{\boldsymbol{\ell},\mathbf{m}}. \tag{23}$$

Proposition C.13 (Recovery filter has a flat spectrum over active modes)**.** *By Proposition C.11 (the optimal lifting matrix is a scaled isometry), and under Hypothesis 1 (invariance of the lifting range), the recovery filter has a flat spectrum over the active modes in the eigenbasis of $\hat{\rho_Z}$, and vanishes on the inactive ones:*

$$|w_{Z,\mathbf{m}}|^2 = \frac{1}{d_X} \ \ (\mathbf{m} \in \boldsymbol{\mathcal{B}}), \qquad w_{Z,\mathbf{m}} = 0 \ \ (\mathbf{m} \textit{ inactive}).$$

When $d_Z = d_X$, every mode is active and $1/d_X = 1/d_Z$.

Proof. Setting $\boldsymbol{\ell} = \mathbf{m}$ in Equation (23) simplifies the phase exponent to $\exp(0) = 1$:

$$\sum_{i=1}^{d_Y} w_{Z,\mathbf{m}} w^*_{Z,\mathbf{m}} = d_Y |w_{Z,\mathbf{m}}|^2 = \frac{d_Y}{d_X}. \tag{24}$$

Dividing both sides by d_Y, we obtain for every active mode:

$$|w_{Z,\mathbf{m}}|^2 = \frac{1}{d_X}.$$

For an inactive mode, $\mathbf{\Pi}_U\mathbf{v_m} = \mathbf{0}$, so $\|\mathbf{U}\mathbf{v_m}\|_2^2 = \mathbf{v}_\mathbf{m}^\dagger \mathbf{U}^\top \mathbf{U}\mathbf{v_m} = 0$. Since every entry of $\mathbf{U}\mathbf{v_m}$ has modulus $|w_{Z,\mathbf{m}}|$, this gives $w_{Z,\mathbf{m}} = 0$. □

Proposition C.14 (Existence of a periodic axis)**.** *Let $\hat{\mathbf{k}}_\boldsymbol{\ell}$ and $\hat{\mathbf{k}}_\mathbf{m}$ be two distinct active frequencies. By Proposition C.11 (the optimal lifting matrix is a scaled isometry), and under Hypothesis 1 (invariance of the lifting range), there is an axis $p \in \{1, \dots, P\}$ such that*

$$\hat{\mathbf{k}}_\boldsymbol{\ell}^{(p)} - \hat{\mathbf{k}}_\mathbf{m}^{(p)} \in \mathbb{Z} \setminus N_p\mathbb{Z}.$$

Proof. From Equation (23), for any distinct modes $\boldsymbol{\ell} \neq \mathbf{m}$:

$$\sum_{i=1}^{d_Y} w_{Z,\boldsymbol{\ell}} w^*_{Z,\mathbf{m}} \exp\left(\mathrm{i}(\hat{\mathbf{k}}_\boldsymbol{\ell} - \hat{\mathbf{k}}_\mathbf{m}) \cdot \mathbf{t}_i\right) = 0. \tag{25}$$

Factoring out $w_{Z,\boldsymbol{\ell}} w^*_{Z,\mathbf{m}} \neq 0$ from the sum yields:

$$\sum_{i=1}^{d_Y} \exp\left(\mathrm{i}(\hat{\mathbf{k}}_\boldsymbol{\ell} - \hat{\mathbf{k}}_\mathbf{m}) \cdot \mathbf{t}_i\right) = 0. \tag{26}$$

Writing the sum over the P-dimensional spatial grid indices $(n_1, \ldots, n_P)$ instead of the flattened index $i \in \{1, \ldots, d_Y\}$ (where $d_Y = \prod_{p=1}^{P} N_p$):

$$\begin{aligned}\sum_{i=1}^{d_Y} \exp\left(\mathrm{i}(\hat{\mathbf{k}}_{\boldsymbol{\ell}} - \hat{\mathbf{k}}_{\mathbf{m}}) \cdot \mathbf{t}_i\right) &= \sum_{n_1=0}^{N_1-1} \cdots \sum_{n_P=0}^{N_P-1} \exp\left(\mathrm{i}\sum_{p=1}^{P}(\hat{\mathbf{k}}_{\boldsymbol{\ell}}^{(p)} - \hat{\mathbf{k}}_{\mathbf{m}}^{(p)})\frac{2\pi n_p}{N_p}\right) \\ &= \prod_{p=1}^{P}\sum_{n_p=0}^{N_p-1} \exp\left(\mathrm{i}(\hat{\mathbf{k}}_{\boldsymbol{\ell}}^{(p)} - \hat{\mathbf{k}}_{\mathbf{m}}^{(p)})\frac{2\pi n_p}{N_p}\right) = 0.\end{aligned} \tag{27}$$

For any $r \in \mathbb{R}$, the geometric series evaluates to:

$$\sum_{n=0}^{N_p-1} \exp\left(\mathrm{i}r\frac{2\pi n}{N_p}\right) = \begin{cases} N_p, & r \in N_p\mathbb{Z}, \\ 0, & r \in \mathbb{Z} \setminus N_p\mathbb{Z}, \\ \dfrac{\exp(\mathrm{i}2\pi r) - 1}{\exp\left(\mathrm{i}2\pi r/N_p\right) - 1}, & r \notin \mathbb{Z}. \end{cases} \tag{28}$$

Because the product of factors equals zero, at least one factor in the product over $p \in \{1, \ldots, P\}$ must vanish. A factor vanishes if and only if:

$$\hat{\mathbf{k}}_{\boldsymbol{\ell}}^{(p)} - \hat{\mathbf{k}}_{\mathbf{m}}^{(p)} \in \mathbb{Z} \setminus N_p\mathbb{Z}$$

for that corresponding axis p. Thus, such a periodic axis p must exist between any pair of distinct active frequencies. □

D Analysis of Stationarity Loss

The stationarity loss term is composed of the JS divergence between the output $\mathbf{Y}$ and its single-axis translated versions along each axis p. In this section, we will show that this enforces discovering symmetry representations.

D.1 Assumptions

We estimate the JS divergence by combining two Donsker–Varadhan bounds, each maximized over a critic network (Appendix G.3). Each bound lies between zero, attained by a constant critic, and the true KL divergence, so the estimate vanishes whenever the two distributions are equal. The converse depends on the critic, and we assume it.

Assumption 5 (Zero-consistency of the JS estimator)**.** *Let $\mathbf{W}$ and $\mathbf{W}'$ be two random vectors taking values in the same space, and let their JS divergence be estimated with critics at their optimum. The critic family is rich enough to detect every difference between the two distributions, so the estimate is zero exactly when the true JS divergence is zero:*

$$\hat{D}_{\mathrm{JS}}(\mathbf{W}, \mathbf{W}') = 0 \quad \Longleftrightarrow \quad D_{\mathrm{JS}}(P_W, P_{W'}) = 0.$$

Equivalently, the estimate is zero exactly when $\mathbf{W}$ and $\mathbf{W}'$ have the same distribution.

D.2 Invariance of Conditioned Sample Space Under Group Actions

In this section, we will show that optimal stationarity loss ensures that the distribution of the conditioned sample space $\mathbf{Z}$ becomes invariant under the learned symmetry actions $\hat{\rho}_Z(\mathbf{t}_i)$ where $\mathbf{t}_i \in \Lambda$.

On the torus, translating the output is a circular shift of its sites, which also moves the last site to the first. The following lemma shows that, in one dimension, stationarity forces the learned representation to agree with this shift on every site, including the wrapped one, and hence to close up after N steps.

Lemma D.1 (Stationarity closes the learned representation on the torus)**.** *Let $P = 1$, so that the lattice has $d_Y = N$ sites, let $d_Z = d_X$, and let $\mathbf{T}$ be the circular shift, $[\mathbf{T}\mathbf{y}]_n = y_{n-1}$ with indices*

modulo N. Let the conditioning matrix be injective (information maximization analysis, Lemma C.4) and the lifting matrix a scaled isometry, $\mathbf{U}^\top\mathbf{U} = c\mathbf{I}$ with $c = d_Y/d_X$ (Proposition C.11). If $\mathbf{Y} \overset{d}{=} \mathbf{T}\mathbf{Y}$, then, with $\hat{\rho} := \hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_1})$,

$$\mathbf{T}\mathbf{U} = \mathbf{U}\,\hat{\rho} \qquad \textit{and} \qquad \hat{\rho}^N = \mathbf{I}.$$

Proof. The shift preserves the range of the lifting matrix. Equality in distribution gives equal covariances, $\mathbf{T}\mathbf{\Sigma}_Y\mathbf{T}^\top = \mathbf{\Sigma}_Y$. Since $\mathbf{C}$ is injective and $\mathbf{\Sigma}_X$ is positive definite (Assumption 4), the range of $\mathbf{\Sigma}_Y = \mathbf{U}\mathbf{C}\mathbf{\Sigma}_X\mathbf{C}^\top\mathbf{U}^\top$ is the range of $\mathbf{U}$, so $\mathbf{T}$ maps the range of $\mathbf{U}$ onto itself.

The shift pulls back to an orthogonal map of order N with determinant one. Let $\tilde{\rho} := c^{-1}\mathbf{U}^\top\mathbf{T}\mathbf{U}$. Since $c^{-1}\mathbf{U}\mathbf{U}^\top$ is the orthogonal projector onto the range of $\mathbf{U}$, which contains the range of $\mathbf{T}\mathbf{U}$,

$$\mathbf{U}\tilde{\rho} = \mathbf{T}\mathbf{U}, \qquad \tilde{\rho}^\top\tilde{\rho} = c^{-1}\mathbf{U}^\top\mathbf{T}^\top\mathbf{T}\mathbf{U} = \mathbf{I}.$$

Iterating, $\mathbf{U}\tilde{\rho}^N = \mathbf{T}^N\mathbf{U} = \mathbf{U}$, and the injectivity of $\mathbf{U}$ gives $\tilde{\rho}^N = \mathbf{I}$. Through the scaled isometry $\mathbf{U}$, $\tilde{\rho}$ is similar to the restriction of $\mathbf{T}$ to the range of $\mathbf{U}$, whose eigenvalues are N-th roots of unity. Since N is odd (Constraint 5), -1 is not among them, and the non-real ones come in conjugate pairs, so $\det\tilde{\rho} = 1$.

The pulled-back shift and the learned translation differ by at most rank one. The lifting matrix evaluates the learned representation at the lattice points, $\mathbf{U}_{n,:} = \mathbf{w}_Z^\top\hat{\rho}_Z(-\mathbf{t}_n)$ for the row of site n. For every site except the wrapped one, $n = 1, \ldots, N-1$,

$$[\mathbf{T}\mathbf{U}]_{n,:} = \mathbf{U}_{n-1,:} = \mathbf{w}_Z^\top\hat{\rho}_Z(-\mathbf{t}_n)\,\hat{\rho} = [\mathbf{U}\hat{\rho}]_{n,:}.$$

Hence $\mathbf{U}(\tilde{\rho} - \hat{\rho}) = \mathbf{e}_0\mathbf{b}^\top$ for some vector $\mathbf{b}$, and multiplying by $c^{-1}\mathbf{U}^\top$ shows that $\tilde{\rho} - \hat{\rho}$ has rank at most one.

Two rotations differing by rank one are equal. The matrix $\mathbf{O} := \hat{\rho}^\top\tilde{\rho}$ is orthogonal, and $\mathbf{O} - \mathbf{I}$ has rank at most one, so all eigenvalues of $\mathbf{O}$ but one equal 1. Being real and orthogonal, $\mathbf{O}$ is diagonalizable with the remaining eigenvalue ± 1: either $\mathbf{O} = \mathbf{I}$, or $\mathbf{O}$ is a reflection with determinant -1. Since $\hat{\rho} \in \mathrm{SO}(d_Z)$ (Constraint 1) and $\det\tilde{\rho} = 1$, $\det\mathbf{O} = 1$, so $\mathbf{O} = \mathbf{I}$ and $\tilde{\rho} = \hat{\rho}$. Therefore $\mathbf{T}\mathbf{U} = \mathbf{U}\hat{\rho}$ and $\hat{\rho}^N = \tilde{\rho}^N = \mathbf{I}$. □

Remark D.2. *For $P > 1$, the wrapped sites along axis i form a slice of d_Y/N_i rows, and the same argument bounds the rank of $\tilde{\rho} - \hat{\rho}$ by d_Y/N_i. When the remaining sites still determine the conditioned sample, i.e. the lifting matrix restricted to them is injective, which requires $d_Y(1 - 1/N_i) \geq d_X$, the two maps coincide directly. Otherwise, we take the identity $\mathbf{T}_{\mathbf{e}_i}\mathbf{U} = \mathbf{U}\hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})$ as part of the toroidal idealization.*

Lemma D.3 (Output stationarity transfers to the conditioned sample space)**.** *By the lifting-injectivity Lemma of the information maximization analysis (Lemma C.5), let the lifting matrix $\mathbf{U}$ be injective on the range of the conditioning matrix, which contains $\mathbf{Z}$. Under Assumption 5 (zero-consistency of the JS estimator), suppose that $\hat{D}_{\mathrm{JS}}(\mathbf{Y}, \mathbf{T}_{\mathbf{e}_i}\mathbf{Y}) = 0$ for every $i \in \{1, 2, \ldots, P\}$. Then the distribution of $\mathbf{Z}$ is invariant under $\hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})$ for every i.*

Proof. By Assumption 5, $\hat{D}_{\mathrm{JS}}(\mathbf{Y}, \mathbf{T}_{\mathbf{e}_i}\mathbf{Y}) = 0$ gives $\mathbf{Y} \overset{d}{=} \mathbf{T}_{\mathbf{e}_i}\mathbf{Y}$ for each i.

The lifting matrix evaluates the learned representation at the lattice points, $Y_{\mathbf{n}} = \mathbf{w}_Z^\top\hat{\rho}_Z(-\mathbf{t}_{\mathbf{n}})\mathbf{Z}$. In the toroidal idealization, translating the output by one site along axis i therefore translates the argument of the representation (Lemma D.1 and the subsequent remark), and by the homomorphism property

$$[\mathbf{T}_{\mathbf{e}_i}\mathbf{Y}]_{\mathbf{n}} = Y_{\mathbf{n}-\mathbf{e}_i} = \mathbf{w}_Z^\top\hat{\rho}_Z(-\mathbf{t}_{\mathbf{n}})\,\hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})\mathbf{Z}, \qquad \text{i.e.} \qquad \mathbf{T}_{\mathbf{e}_i}\mathbf{Y} = \mathbf{U}\hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})\mathbf{Z}.$$

Hence

$$\mathbf{U}\hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})\mathbf{Z} \overset{d}{=} \mathbf{U}\mathbf{Z}.$$

Since $\mathbf{U}$ is injective on the range of the conditioning matrix, it has a left inverse there, and applying it to both sides gives

$$\hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})\mathbf{Z} \overset{d}{=} \mathbf{Z}.$$

□

Corollary D.4 (Invariance under integer translations)**.** *Following Lemma D.3 (output stationarity transfers to the conditioned sample space), assume that the distribution of* $\mathbf{Z}$ *is invariant under* $\hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})$ *for every* i*. Then the distribution of* $\mathbf{Z}$ *is invariant under* $\hat{\rho}_Z(\mathbf{t}_\mathbf{s})$ *for every integer translation vector* $\mathbf{s} \in \mathbb{Z}^P$*, where* $\mathbf{t}_\mathbf{s} = 2\pi\mathbf{D}^{-1}\mathbf{s}$ *is the corresponding lattice point.*

Proof. Consider an arbitrary $\mathbf{s} = (s_1, \ldots, s_P) \in \mathbb{Z}^P$. Because $\hat{\rho}_Z$ is a group homomorphism, the action decomposes into single-axis steps, each of which preserves the distribution:

$$\hat{\rho}_Z(\mathbf{t}_\mathbf{s})\mathbf{Z} = \left(\prod_{i=1}^{P} \hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})^{s_i}\right) \mathbf{Z} \overset{d}{=} \mathbf{Z}.$$

□

D.3 Connection Between the Learned and True Group Actions

In this section, we show that invariance of $\mathbf{Z}$ under the integer translations $\hat{\rho_Z}(\mathbf{t}_\mathbf{s})$, $\mathbf{s} \in \mathbb{Z}^P$, identifies $\hat{\rho_Z}$ with ρ_Z on the lattice up to a linear reparametrization of the group coordinates, and that the identification extends to all $\mathbf{s} \in \mathbb{R}^P$ in the absence of aliasing.

Lemma D.5 (Symmetry maximality transfers to the conditioned sample space)**.** *Under Assumption 2 (symmetry maximality), and given that the conditioning matrix is injective (information maximization analysis, Lemma C.4), every linear map* $\mathbf{T}$ *satisfying distribution invariance* $\mathbf{TZ} \overset{d}{=} \mathbf{Z}$ *maps the range* $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ *of the conditioning matrix onto itself and agrees there with the representation of Corollary C.6: there is a* $\mathbf{t} \in \mathbb{R}^P$ *with*

$$\mathbf{TC} = \rho_Z(\mathbf{t})\,\mathbf{C}.$$

When $d_Z = d_X$, $\mathbf{C}\,\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Z}}$ *and* $\mathbf{T} = \rho_Z(\mathbf{t})$.

Proof. Based on the distribution invariance property, we have the following relation:

$$\mathbf{TZ} \overset{d}{=} \mathbf{Z} \implies \mathbf{TCX} \overset{d}{=} \mathbf{CX}$$

First, $\mathbf{T}$ maps the range of $\mathbf{C}$ onto itself. Equality in distribution implies equal covariances,

$$\mathbf{TC}\,\boldsymbol{\Sigma}_X\,\mathbf{C}^\top\mathbf{T}^\top = \mathbf{C}\,\boldsymbol{\Sigma}_X\,\mathbf{C}^\top,$$

and since $\boldsymbol{\Sigma}_X$ is positive definite (Assumption 4), the two sides have ranges $\mathbf{TC}\,\boldsymbol{\mathcal{X}}$ and $\mathbf{C}\,\boldsymbol{\mathcal{X}}$, which therefore coincide.

By Lemma C.4 (injectivity of the conditioning matrix), the pseudo-inverse $\mathbf{C}^+$ is a left inverse, $\mathbf{C}^+\mathbf{C} = \mathbf{I}_{d_X}$. Multiplying both sides of the distributional identity by $\mathbf{C}^+$ gives

$$\mathbf{C}^+\mathbf{TCX} \overset{d}{=} \mathbf{C}^+\mathbf{CX} = \mathbf{X}.$$

Therefore, $\mathbf{T}_X := \mathbf{C}^+\mathbf{TC}$ is a distribution invariance of $\boldsymbol{\mathcal{X}}$. By Assumption 2, every invariance is an element of the underlying symmetry. Therefore, there must be a $\mathbf{t}$ satisfying $\mathbf{T}_X = \rho_X(\mathbf{t})$. Since $\mathbf{TC}$ maps into the range of $\mathbf{C}$, $\boldsymbol{\Pi}_C\mathbf{TC} = \mathbf{TC}$ with $\boldsymbol{\Pi}_C = \mathbf{CC}^+$, and the intertwining relation of Corollary C.6 gives

$$\mathbf{TC} = \mathbf{C}\,\mathbf{C}^+\mathbf{TC} = \mathbf{C}\,\rho_X(\mathbf{t}) = \rho_Z(\mathbf{t})\,\mathbf{C}.$$

□

Lemma D.6 (Additivity of the parametrization)**.** *By Corollary D.4 (invariance under integer translations),* $\mathbf{Z}$ *is invariant under* $\hat{\rho_Z}(\mathbf{t}_\mathbf{s})$ *for every* $\mathbf{s} \in \mathbb{Z}^P$*. Under Lemma D.5 (symmetry maximality transfers to the conditioned sample space) and the faithfulness-transfer Corollary of the information maximization analysis (Corollary C.7), there is a matrix* $\mathbf{R} \in \mathbb{R}^{P\times P}$ *with*

$$\hat{\rho_Z}(\mathbf{t}_\mathbf{s})\,\mathbf{C} = \rho_Z(\mathbf{R}\mathbf{t}_\mathbf{s})\,\mathbf{C} \qquad \textit{for every } \mathbf{s} \in \mathbb{Z}^P,$$

that is, $\hat{\rho_Z} = \rho_Z \circ \mathbf{R}$ *at the lattice points, on the range* $\mathbf{C}\,\boldsymbol{\mathcal{X}}$*, which carries the data. When* $d_Z = d_X$*, this range is* $\boldsymbol{\mathcal{Z}}$*.*

Proof. By Lemma D.5, for every $\mathbf{s} \in \mathbb{Z}^P$ there is an $\mathbf{a}(\mathbf{s}) \in \mathbb{R}^P$ with

$$\hat{\rho_Z}(\mathbf{t_s})\,\mathbf{C} = \rho_Z(\mathbf{a}(\mathbf{s}))\,\mathbf{C} = \mathbf{C}\,\rho_X(\mathbf{a}(\mathbf{s})).$$

Since $\mathbf{C}$ is injective, $\rho_Z(\mathbf{a})\mathbf{C} = \rho_Z(\mathbf{b})\mathbf{C}$ implies $\rho_X(\mathbf{a}) = \rho_X(\mathbf{b})$, and by the faithfulness-transfer Corollary of the information maximization analysis (Corollary C.7), $\mathbf{a}(\mathbf{s})$ is unique up to the periods of ρ_X.

For $\mathbf{s}, \mathbf{s}' \in \mathbb{Z}^P$, $\hat{\rho_Z}$ is a homomorphism and $\mathbf{t_{s+s'}} = \mathbf{t_s} + \mathbf{t_{s'}}$; therefore

$$\begin{aligned}\rho_Z(\mathbf{a}(\mathbf{s}+\mathbf{s}'))\,\mathbf{C} &= \hat{\rho_Z}(\mathbf{t_s})\,\hat{\rho_Z}(\mathbf{t_{s'}})\,\mathbf{C} = \hat{\rho_Z}(\mathbf{t_s})\,\mathbf{C}\,\rho_X(\mathbf{a}(\mathbf{s}')) \\ &= \mathbf{C}\,\rho_X(\mathbf{a}(\mathbf{s}))\,\rho_X(\mathbf{a}(\mathbf{s}')) = \rho_Z(\mathbf{a}(\mathbf{s})+\mathbf{a}(\mathbf{s}'))\,\mathbf{C},\end{aligned}$$

so $\mathbf{a}$ is additive up to the periods of ρ_X. An additive map on $\mathbb{Z}^P$ is determined by its values on the unit vectors, $\mathbf{a}(\mathbf{s}) = \sum_i s_i\,\mathbf{a}(\mathbf{e}_i)$. Defining $\mathbf{R}$ by $\mathbf{R}\mathbf{t}_{\mathbf{e}_i} := \mathbf{a}(\mathbf{e}_i)$, that is, $\mathbf{R} := \frac{1}{2\pi}[\mathbf{a}(\mathbf{e}_1), \ldots, \mathbf{a}(\mathbf{e}_P)]\,\mathbf{D}$, gives $\mathbf{a}(\mathbf{s}) = \mathbf{R}\mathbf{t_s}$ up to the periods of ρ_X, which ρ_Z does not see. □

Between the lattice points, integer-translation invariance does not constrain the learned representation. Two representations that agree on the lattice can differ there by *aliasing*: along axis j, the learned frequency of an active mode may exceed its latent-induced frequency by a multiple of N_j. We exclude this.

Assumption 6 (No aliasing)**.** *The identification of Lemma D.6 (additivity of the parametrization) extends to all translations,*

$$\hat{\rho_Z}(\mathbf{t})\,\mathbf{C} = \rho_Z(\mathbf{R}\mathbf{t})\,\mathbf{C} \qquad \textit{for every } \mathbf{t} \in \mathbb{R}^P.$$

It holds, for instance, whenever the latent-induced frequencies of the active modes lie within the eigenphase bound of Constraint 2, so that the learned generators, whose eigenphases lie there too, cannot differ from them by a multiple of N_j. Remark D.9 discusses the ambiguity it excludes and why training avoids it.

Corollary D.7 (The invariance hypothesis holds at the joint optimum)**.** *Let the conditioning matrix be aligned with the lifting matrix as in the aligning-conditioning Lemma of the information maximization analysis (Lemma C.10), and let the lifting matrix be a scaled isometry on the span of its top-k right singular vectors (Proposition C.11). Under Lemma D.6 (additivity of the parametrization) and Assumption 6 (no aliasing), the range of the projector $\mathbf{\Pi}_U$ is the range $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ of the conditioning matrix, and it is invariant under $\hat{\rho_Z}(\mathbf{s})$ for every $\mathbf{s} \in \mathbb{R}^P$. Hence the invariance hypothesis of the information maximization analysis (Hypothesis 1) holds, and Propositions C.13 and C.14 apply for every $d_Z \geq d_X$.*

Proof. By the aligning-conditioning Lemma, the top-k right singular vectors of $\mathbf{U}$ coincide with the left singular vectors of $\mathbf{C}\mathbf{\Sigma}_X^{1/2}$, so the range of $\mathbf{\Pi}_U$ is the range of $\mathbf{C}\mathbf{\Sigma}_X^{1/2}$. The sample covariance is full rank (proof of Lemma C.4), so $\mathbf{\Sigma}_X^{1/2}$ is invertible and this range is $\mathbf{C}\,\boldsymbol{\mathcal{X}}$. By Assumption 6, $\hat{\rho_Z}(\mathbf{s})\,\mathbf{C} = \rho_Z(\mathbf{R}\mathbf{s})\,\mathbf{C} = \mathbf{C}\,\rho_X(\mathbf{R}\mathbf{s})$, so $\hat{\rho_Z}(\mathbf{s})$ maps $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ into itself for every $\mathbf{s}$. □

Theorem D.8 (Symmetry identification up to signed axis permutations)**.** *By Lemma D.6 (additivity of the parametrization) and Assumption 6 (no aliasing), let $\mathbf{R} \in \mathbb{R}^{P\times P}$ be the matrix satisfying $\hat{\rho}_Z(\mathbf{s}) = \rho_Z(\mathbf{R}\mathbf{s})$ on the range $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ for any $\mathbf{s} \in \mathbb{R}^P$. In the toroidal idealization (Lemma D.1 and the subsequent remark), and under Assumption 3 (support of the measurement operator), $\mathbf{R}$ is a signed permutation matrix.*

Proof. Learned frequencies. By the probe-orbit Corollary of the information maximization analysis (Corollary C.8), the translated probe responses $\{\rho_Z(\mathbf{s})\mathbf{h}_Z\}$ span the range $\mathbf{C}\,\boldsymbol{\mathcal{X}}$, which carries the data, so the spectrum of ρ_Z on $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ is exactly the active mode set $\boldsymbol{\mathcal{B}}$. Since $\rho_Z(\mathbf{R}\mathbf{s})$ multiplies the mode $\mathbf{m}$ by $e^{-\mathrm{i}\,\mathbf{m}\cdot\mathbf{R}\mathbf{s}}$, and the learned representation multiplies its eigenvector of learned frequency $\hat{\mathbf{k}}$ by $e^{\mathrm{i}\,\hat{\mathbf{k}}\cdot\mathbf{s}}$ (Appendix C.5.1), the learned frequency of the active mode $\mathbf{m}$ is

$$\hat{\mathbf{k}}_{\mathbf{m}} = \mathbf{\Phi}\,\mathbf{m}, \qquad \mathbf{\Phi} := -\mathbf{R}^\top,$$

and it suffices to show that $\mathbf{\Phi}$ is a signed permutation matrix.

Integrality. In the toroidal idealization, the learned representation closes up after N_i steps along axis i (Lemma D.1), $\hat{\rho}_Z(\mathbf{t}_{N_i\mathbf{e}_i}) = \mathbf{I}$ on $\mathbf{C}\,\boldsymbol{\mathcal{X}}$. Since $\mathbf{t}_{N_i\mathbf{e}_i} = 2\pi\mathbf{e}_i$, this gives $e^{2\pi\mathrm{i}\,[\hat{\mathbf{k}}_\mathbf{m}]_i} = 1$, so every learned frequency is an integer vector. The fundamental harmonics lie in $\boldsymbol{\mathcal{B}}$ by condition (ii) of Assumption 3, and taking $\mathbf{m} = \mathbf{e}_j$ shows that the j-th column of $\boldsymbol{\Phi}$ is an integer vector. Hence $\boldsymbol{\Phi}$ is an integer matrix.

Every column is non-zero. On the lattice, the lifting matrix maps the eigenvector of the active mode $\mathbf{m}$ to a multiple of the lattice harmonic with frequency $\hat{\mathbf{k}}_\mathbf{m}$ (Appendix C.5.2). Two distinct active modes with the same learned frequency modulo N_i along every axis would therefore be lifted to parallel vectors, contradicting the injectivity of the lifting matrix on the conditioned sample space (information maximization analysis, Lemma C.5). Applied to $\mathbf{e}_j$ and $\mathbf{0}$, this gives $\boldsymbol{\Phi}\mathbf{e}_j \neq \mathbf{0}$, and since $\boldsymbol{\Phi}$ is an integer matrix, every column j has an entry with

$$|\Phi_{\sigma(j),j}| \geq 1. \tag{29}$$

This lower bound also follows, without periodicity, from the periodic-axis Proposition of the information maximization analysis (Proposition C.14).

Nyquist bound. The learned frequencies are integers, and the eigenphase bound of Constraint 2 confines them to the Nyquist limit of the grid, $|[\hat{\mathbf{k}}_\mathbf{m}]_i| \leq H_i := \frac{N_i-1}{2}$, the grid resolutions being odd (Constraint 5). For each axis i, the sign-pattern corner

$$\mathbf{q}_j^{(i)} := \operatorname{sgn}(\Phi_{ij})\, M_j, \qquad j = 1, \ldots, P,$$

with $\operatorname{sgn}(0) := 0$, lies in $\boldsymbol{\mathcal{B}}$ because $\boldsymbol{\mathcal{B}}$ is box-shaped. It is the active mode whose learned frequency along axis i is largest, and the Nyquist limit gives

$$\sum_j |\Phi_{ij}|\, M_j \leq H_i \qquad \text{for every } i. \tag{30}$$

Case of a band reaching beyond a quarter of the largest resolution, $2\min_j M_j > \max_i H_i$. Bounding every M_j from below by $\min_j M_j$ in (30) gives $\big(\sum_j |\Phi_{ij}|\big)\min_j M_j \leq H_i < 2\min_j M_j$, so $\sum_j |\Phi_{ij}| \leq 1$: every row of the integer matrix $\boldsymbol{\Phi}$ has at most one non-zero entry, of magnitude one. By (29), each of the P columns has at least one. Hence every row and every column has exactly one entry ± 1, and $\boldsymbol{\Phi}$ is a signed permutation matrix.

Case of the full Nyquist box, $M_j = H_j$ for every j. Then (30) reads

$$\sum_j |\Phi_{ij}| \frac{N_j - 1}{N_i - 1} \leq 1. \tag{31}$$

We group the axes by their lengths into $G_1, \ldots, G_R$, placing axes with equal N_i in the same group and sorting the groups in descending order, so that $N_i > N_j$ for $i \in G_p$ and $j \in G_{p+1}$. For $i \in G_p$ and $j \in G_q$ with $p > q$, $\frac{N_j-1}{N_i-1} > 1$, and (31) implies

$$\Phi_{ij} = 0 \qquad \text{whenever } N_j > N_i, \tag{32}$$

so $\boldsymbol{\Phi}$ is block-triangular with respect to the partition $G_1, \ldots, G_R$.

Consider G_1, which contains the longest axes. By (32), every column $j \in G_1$ can have non-zero entries only in rows $i \in G_1$, and by (29) each such column contains an entry of magnitude at least 1:

$$\sum_{i \in G_1} \sum_{j \in G_1} |\Phi_{ij}| \geq \operatorname{size}(G_1). \tag{33}$$

On the other hand, summing (31) over $i \in G_1$ and using $N_i = N_j$ for $i, j \in G_1$,

$$\sum_{i \in G_1} \sum_{j \in G_1} |\Phi_{ij}| + \sum_{i \in G_1} \sum_{j \notin G_1} |\Phi_{ij}| \frac{N_j - 1}{N_i - 1} \leq \operatorname{size}(G_1).$$

By (33), the second sum vanishes and

$$\sum_{i \in G_1} \sum_{j \in G_1} |\Phi_{ij}| = \operatorname{size}(G_1), \tag{34}$$

so $\Phi_{ij} = 0$ for $i \in G_1$, $j \notin G_1$, and the first block decouples from all other groups. Every column in G_1 contains an entry of magnitude at least 1, while the total absolute mass of the block is exactly $\mathrm{size}(G_1)$, so every column contains exactly one non-zero entry, of magnitude 1; by (31), every row has absolute sum 1. Therefore $\mathbf{\Phi}_{G_1,G_1}$ is a signed permutation matrix.

The same argument applies successively to $G_2, \ldots, G_R$: at each step the preceding groups are already decoupled, and (32) prevents entries from a smaller-N column from appearing in a larger-N row. Hence $\mathbf{\Phi}$ is block-diagonal with signed permutation matrices as its diagonal blocks, and it is a signed permutation matrix.

In both cases $\mathbf{R} = -\mathbf{\Phi}^\top$ is a signed permutation matrix as well. □

Remark D.9 (Integer translations and the no-aliasing assumption)**.** Why integer translations. *The stationarity loss compares the output only with its integer translations. This conforms to the practical setting: many datasets are sampled on a fixed grid, fractional shifts of the data must then be synthesized by interpolation, and aliasing corrupts such interpolation. This may explain why LieGAN succeeds on band-limited translations but degrades on non-band-limited continuous ones (Section 6.1). Testing fractional shifts would nevertheless be cheap in our framework: the lifting evaluates the learned representation at any group element, so a fractionally translated output, $\mathbf{w}_Z^\top \hat{\rho_Z}(-\mathbf{t_n} - \boldsymbol{\delta})\mathbf{Z}$, requires no interpolation. Enforcing it would give the identification of Assumption 6 directly.*

What integer translations leave open. *On the torus, for any integer a coprime to N, the output $Y_n = F_{an \bmod N}$ is jointly stationary: a translation of the output by one site is a translation of the latent field by a sites. Total correlation and information maximization do not depend on the order of the sites, so no term of the loss distinguishes these* stride solutions *from the recovered signal. Reflection is the stride $a = -1$. For $|a| \geq 2$, the latent-induced frequency aM of the corner mode exceeds the Nyquist limit $\frac{N-1}{2}$ whenever the band reaches beyond half of it (condition (iii) of Assumption 3), and aliases back into it, which Assumption 6 excludes. For a narrower band, strides are not aliased, and Theorem D.8 can fail.*

Why training avoids stride solutions. *On the torus, the stationary generators are isolated. For a single active mode, a circular translation carries the last site to the first, where the lifted sinusoid jumps in phase by $2\pi\hat{k}$, so the output is stationary exactly at integer learned frequencies $\hat{k}$ and the stationarity loss is positive between them. With several active modes, stationarity restricts the learned frequencies to the stride solutions, a discrete set, so every continuous path between two of them crosses positive stationarity loss. During training, the learned frequencies grow continuously from zero, and the rank schedule activates the modes in order of output power. The first active mode settles at the first nonzero integer frequency, and since one output step corresponds to a single latent translation, this fixes the stride of all later modes through $a\, m_0 \equiv \pm 1 \pmod{N}$, where m_0 is the latent frequency of the first mode. When the fundamental carries the most power, $m_0 = \pm 1$ and the minimal generator, $a = \pm 1$, is reached; the barriers then retain it. This is a statement about the dynamics rather than a proof, and it predicts stride solutions for latent fields whose power peaks at a higher harmonic.*

E Analysis of Total Correlation Minimization Loss

In this section, we analyze the effects of total correlation. The analysis is carried out at the terminal rank $k = \dim(\boldsymbol{\mathcal{X}})$, in the regime where information maximization is non-degenerate and stationarity is optimal.

E.1 Definitions

Let $\mathbf{Q}_k$ correspond to the top-k eigenvectors of the covariance matrix $\mathbf{\Sigma}_Y := \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\top\right]$. We use centered variables due to Constraint 7.

We define the projector to the top-k subspace as

$$\mathbf{P}_k := \mathbf{Q}_k \mathbf{Q}_k^\top,$$

and the smoothed random vector as

$$\tilde{\mathbf{Y}} := \mathbf{P}_k \mathbf{Y} + \epsilon \left(\mathbf{I} - \mathbf{P}_k\right) \mathbf{N},$$

where $\epsilon > 0$ and $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is independent of $\mathbf{Y}$.

This definition also has the following implication.

Corollary E.1 (Top-k projection acts as the identity on the model output)**.** *Suppose the conditioning matrix $\mathbf{C}$ is injective and the lifting matrix $\mathbf{U}$ has full column rank on the conditioned sample space, by the conditioning-matrix and lifting injectivity Lemmas of the information maximization analysis (Lemmas C.4 and C.5). Then the output random vector $\mathbf{Y}$ lies entirely within a k-dimensional subspace. Consequently, projecting it onto its top-k principal subspace acts as the identity operation:*

$$\mathbf{P}_k \mathbf{Y} = \mathbf{Y}.$$

Proof. Let us apply the resolution of the identity to $\mathbf{Y}$. Because $\mathbf{P}_k$ and $\mathbf{I} - \mathbf{P}_k$ project onto orthogonal subspaces, we can write the dimension equality for the subspaces spanned by these projections:

$$\dim(\text{span}(\mathbf{P}_k \mathbf{Y})) + \dim(\text{span}((\mathbf{I} - \mathbf{P}_k)\mathbf{Y})) = \dim(\text{span}(\mathbf{Y})).$$

Since $\mathbf{P}_k$ projects onto the top-k subspace, $\dim(\text{span}(\mathbf{P}_k \mathbf{Y})) = k$. By our premises, $\dim(\text{span}(\mathbf{Y})) = k$ as well. Substituting these yields:

$$\dim(\text{span}((\mathbf{I} - \mathbf{P}_k)\mathbf{Y})) = 0.$$

A subspace of dimension zero contains only the zero vector. Therefore, almost surely:

$$(\mathbf{I} - \mathbf{P}_k)\mathbf{Y} = \mathbf{0} \implies \mathbf{P}_k \mathbf{Y} = \mathbf{Y}.$$

□

E.2 Total Correlation Surrogate

Total correlation estimation requires estimating the component-wise entropies and the joint entropy of the output random vector $\mathbf{Y}$. Because estimating the joint entropy is challenging in high-dimensional spaces, we use the low-rank entropy surrogate (defined in Appendix A.1.5) instead of a direct estimation. In Lemma E.3 (gradients of the smoothed joint entropy), we show that this yields identical gradients when $k = \dim(\boldsymbol{\mathcal{X}})$ due to the linearity of the model.

The component-wise entropies are much more tractable, since they are one-dimensional. Each estimate is the cross-entropy of Y_i under a fitted mixture density $\hat{p}_i$ (Appendix G.2), so it exceeds the true entropy by a Kullback–Leibler gap,

$$\hat{h}(Y_i) = h(Y_i) + D_{KL}\big(p_{Y_i} \,\|\, \hat{p}_i\big) \geq h(Y_i).$$

In practice, owing to optimization dynamics and the limited capacity of the mixture, the estimates are slightly higher than the true entropies. We neglect this gap.

Assumption 7 (Exact component-wise entropy estimates)**.** *For every i, the density of Y_i belongs to the family of the component-wise estimator, and the estimator is at its optimum, so that $\hat{h}(Y_i) = h(Y_i)$. Since only gradients enter the analysis, it suffices that the gap $D_{KL}(p_{Y_i} \| \hat{p}_i)$ does not depend on the model parameters.*

E.2.1 Gradient Relations Between Entropy Estimates

Lemma E.2 (Gradients of the smoothed component-wise entropy)**.** *By Corollary E.1 (the top-k projection acts as the identity), $\mathbf{P}_k \mathbf{Y} = \mathbf{Y}$. Let the differential entropy $h(Y_i)$ exist and assume that the density of $[\mathbf{P}_k \mathbf{Y}]_i$ is twice continuously differentiable with respect to its domain. Furthermore, assume the density is sufficiently regular with respect to the parameters θ such that the gradient of its Fisher information, $\nabla_\theta J([\mathbf{P}_k \mathbf{Y}]_i)$, exists and is bounded. Then the gradient of the marginal entropy of the smoothed output random vector component $h(\tilde{Y}_i)$ corresponds to*

$$\nabla_\theta h(\tilde{Y}_i) = \nabla_\theta h(Y_i) + \mathcal{O}(\epsilon^2).$$

Proof. We decompose the smoothed component i as:

$$\tilde{Y}_i = [\mathbf{P}_k \mathbf{Y}]_i + \epsilon[(\mathbf{I} - \mathbf{P}_k)\mathbf{N}]_i.$$

Let $S_i = [\mathbf{P}_k\mathbf{Y}]_i$ with density $p_S(y)$, and $W_i = [(\mathbf{I} - \mathbf{P}_k)\mathbf{N}]_i$. Since $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$ is independent of $\mathbf{Y}$, W_i is independent of S_i. Because $\mathbf{P}_k$ is an orthogonal projection, $(\mathbf{I} - \mathbf{P}_k)$ is symmetric and idempotent, meaning $W_i \sim \mathcal{N}(0, \sigma_i^2)$ with $\sigma_i^2 = [\mathbf{I} - \mathbf{P}_k]_{ii}$.

The density of the smoothed variable $\tilde{Y}_i = S_i + \epsilon W_i$ is given by the expectation:

$$p_{\tilde{Y}_i}(y) = \mathbb{E}_{W_i}[p_S(y - \epsilon W_i)].$$

By our assumption that $p_S(y)$ is twice continuously differentiable with respect to y, we apply a Taylor expansion with respect to ϵ:

$$p_S(y - \epsilon W_i) = p_S(y) - \epsilon W_i p_S'(y) + \frac{1}{2}\epsilon^2 W_i^2 p_S''(y) + \mathcal{O}(\epsilon^3).$$

Taking the expectation over W_i, and noting that $\mathbb{E}[W_i] = 0$ and $\mathbb{E}[W_i^2] = \sigma_i^2$:

$$p_{\tilde{Y}_i}(y) = p_S(y) + \frac{1}{2}\epsilon^2\sigma_i^2 p_S''(y) + \mathcal{O}(\epsilon^3).$$

We now expand the differential entropy $h(\tilde{Y}_i) = -\int p_{\tilde{Y}_i}(y) \log p_{\tilde{Y}_i}(y) dy$ using the first-order approximation $\log(p_S + \delta) \approx \log p_S + \frac{\delta}{p_S}$:

$$\begin{aligned} h(\tilde{Y}_i) &\approx -\int \left(p_S(y) + \frac{1}{2}\epsilon^2\sigma_i^2 p_S''(y)\right)\left(\log p_S(y) + \frac{\frac{1}{2}\epsilon^2\sigma_i^2 p_S''(y)}{p_S(y)}\right) dy \\ &\approx h(S_i) - \frac{1}{2}\epsilon^2\sigma_i^2 \int p_S''(y) \log p_S(y) dy - \frac{1}{2}\epsilon^2\sigma_i^2 \int p_S''(y) dy. \end{aligned}$$

The last integral vanishes because $\int p_S''(y)dy = p_S'(\infty) - p_S'(-\infty) = 0$. Applying integration by parts to the middle term yields $\int p_S''(y) \log p_S(y) dy = -\int \frac{(p_S'(y))^2}{p_S(y)} dy$.

Substituting this back, we obtain the explicit dependence of the entropy on ϵ:

$$h(\tilde{Y}_i) = h(S_i) + \frac{1}{2}\epsilon^2\sigma_i^2 \int \frac{(p_S'(y))^2}{p_S(y)} dy + \mathcal{O}(\epsilon^3).$$

Notice that the integral term is exactly the Fisher information, $J(S_i)$, of the variable S_i. Taking the gradient with respect to the parameters θ on both sides yields:

$$\nabla_\theta h(\tilde{Y}_i) = \nabla_\theta h(S_i) + \frac{1}{2}\epsilon^2\sigma_i^2 \nabla_\theta J(S_i) + \nabla_\theta \mathcal{O}(\epsilon^3).$$

By our regularity assumption, the gradient of the Fisher information $\nabla_\theta J(S_i)$ and the higher-order terms are bounded, allowing us to absorb them into the asymptotic error term:

$$\nabla_\theta h(\tilde{Y}_i) = \nabla_\theta h([\mathbf{P}_k\mathbf{Y}]_i) + \mathcal{O}(\epsilon^2).$$

Finally, by Corollary E.1 (the top-k projection acts as the identity), the projection perfectly captures the data subspace, $\mathbf{P}_k\mathbf{Y} = \mathbf{Y}$. Then $h([\mathbf{P}_k\mathbf{Y}]_i) = h(Y_i)$, leading directly to the simplified gradient relation:

$$\nabla_\theta h(\tilde{Y}_i) = \nabla_\theta h(Y_i) + \mathcal{O}(\epsilon^2).$$

□

Lemma E.3 (Gradients of the smoothed joint entropy)**.** *Under the conditioning-matrix and lifting injectivity Lemmas of the information maximization analysis (Lemmas C.4 and C.5), so that* $\mathbf{W} := \mathbf{UC}$ *has full column rank k, the gradients of the joint entropy of the smoothed random vector* $\nabla_\theta h(\tilde{\mathbf{Y}})$ *are equivalent to the gradients of the k-dimensional surrogate joint entropy* $\nabla_\theta h_{surr}(\mathbf{Y})$.

Proof. Recall the definition of the smoothed random vector:

$$\tilde{\mathbf{Y}} = \mathbf{P}_k\mathbf{Y} + \epsilon(\mathbf{I} - \mathbf{P}_k)\mathbf{N},$$

where $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$ is independent of $\mathbf{Y}$, and $\mathbf{P}_k = \mathbf{Q}_k\mathbf{Q}_k^\top$ is the orthogonal projection onto the top-k principal subspace.

Let $\mathbf{Q} = [\mathbf{Q}_k \quad \mathbf{Q}_\perp]$ be an orthogonal matrix where $\mathbf{Q}_\perp$ spans the null space of $\mathbf{P}_k$. Because differential entropy is invariant under orthogonal transformations, applying this rotation decouples the random vector:

$$h(\tilde{\mathbf{Y}}) = h(\mathbf{Q}^\top \tilde{\mathbf{Y}}) = h(\mathbf{Q}_k^\top \mathbf{Y}) + h(\epsilon \mathbf{Q}_\perp^\top \mathbf{N}).$$

The second term is the entropy of isotropic Gaussian noise in $D-k$ dimensions. As it evaluates to the analytic constant $\frac{D-k}{2}\log(2\pi e\epsilon^2)$, its gradient with respect to the model parameters θ vanishes:

$$\nabla_\theta h(\tilde{\mathbf{Y}}) = \nabla_\theta h(\mathbf{Q}_k^\top \mathbf{Y}) + 0.$$

We must now relate the true joint entropy on the subspace, $h(\mathbf{Q}_k^\top \mathbf{Y})$, to the surrogate entropy. Recalling the model relation (9), the output random vector corresponds to

$$\mathbf{Y} = \mathbf{U}\mathbf{C}\mathbf{X}.$$

Let $\mathbf{W} = \mathbf{U}\mathbf{C}$ represent the model. Since $\mathbf{X}$ is full rank and $\dim(\boldsymbol{\mathcal{X}}) = k$, the matrix $\mathbf{W}$ has full column rank k. Using the QR decomposition, we can express $\mathbf{W} = \mathbf{Q}_k\mathbf{R}$, where $\mathbf{R}$ is a $k \times k$ invertible matrix.

The true entropy on the principal subspace evaluates to:

$$h(\mathbf{Q}_k^\top \mathbf{Y}) = h(\mathbf{Q}_k^\top \mathbf{W}\mathbf{X}) = h(\mathbf{R}\mathbf{X}) = h(\mathbf{X}) + \log|\det(\mathbf{R})|.$$

Noting that $\mathbf{W}^\top\mathbf{W} = \mathbf{R}^\top\mathbf{R}$, we can rewrite $\log|\det(\mathbf{R})| = \frac{1}{2}\log\det(\mathbf{W}^\top\mathbf{W})$, yielding:

$$h(\mathbf{Q}_k^\top \mathbf{Y}) = h(\mathbf{X}) + \frac{1}{2}\log\det(\mathbf{W}^\top\mathbf{W}).$$

Next, consider our surrogate joint entropy, defined via the top-k eigenvalues of the covariance matrix $\boldsymbol{\Sigma}_Y = \mathbf{W}\boldsymbol{\Sigma}_X\mathbf{W}^\top$:

$$h_{\text{surr}}(\mathbf{Y}) = \frac{k}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^{k}\log\left(\lambda_i(\boldsymbol{\Sigma}_Y)\right).$$

The product of the k non-zero eigenvalues of $\boldsymbol{\Sigma}_Y$ corresponds to the determinant of $\mathbf{W}^\top\mathbf{W}\boldsymbol{\Sigma}_X$, which factors as $\det(\mathbf{W}^\top\mathbf{W})\det(\boldsymbol{\Sigma}_X)$. Substituting this into the surrogate definition gives:

$$h_{\text{surr}}(\mathbf{Y}) = \underbrace{\frac{k}{2}\log(2\pi e) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}_X)}_{\text{Gaussian Entropy of } \mathbf{X}} + \frac{1}{2}\log\det(\mathbf{W}^\top\mathbf{W}).$$

Crucially, while the true entropy $h(\mathbf{Q}_k^\top\mathbf{Y})$ relies on the actual differential entropy $h(\mathbf{X})$ and the surrogate $h_{\text{surr}}(\mathbf{Y})$ relies on the Gaussian entropy of $\mathbf{X}$, both of these baseline terms are properties of the inherent data distribution and are entirely independent of the network parameters θ.

Taking the gradient with respect to θ annihilates the data-dependent constants in both equations, isolating the volume-expansion term:

$$\nabla_\theta h(\mathbf{Q}_k^\top\mathbf{Y}) = \nabla_\theta\left(\frac{1}{2}\log\det(\mathbf{W}^\top\mathbf{W})\right) = \nabla_\theta h_{\text{surr}}(\mathbf{Y}).$$

Therefore, $\nabla_\theta h(\tilde{\mathbf{Y}}) = \nabla_\theta h_{\text{surr}}(\mathbf{Y})$, proving that their gradients are strictly equivalent. □

E.2.2 Effect of Minimizing Total Correlation Surrogate

In this section, we show that minimizing the total correlation surrogate corresponds to minimizing the true total correlation of the smoothed output random vector $\tilde{\mathbf{Y}}$.

Proposition E.4 (Gradients of the surrogate total correlation)**.** *Define the surrogate total correlation as*

$$\text{TC}_{\text{sur}}(\mathbf{Y}) := \sum_i h(Y_i) - h_{surr}(\mathbf{Y}).$$

By the smoothed component-wise and joint entropy gradient Lemmas (Lemmas E.2 and E.3), the gradient of the surrogate total correlation matches the gradient of the total correlation of the smoothed random vector $\tilde{\mathbf{Y}}$, up to an $\mathcal{O}(\epsilon^2)$ error term:

$$\nabla_\theta \operatorname{TC}(\tilde{\mathbf{Y}}) = \nabla_\theta \operatorname{TC}_{\text{sur}}(\mathbf{Y}) + \mathcal{O}(\epsilon^2).$$

Therefore, minimizing total correlation surrogate is equivalent to minimizing actual total correlation in the $\epsilon \to 0$ limit.

Proof. By definition, the total correlation of the smoothed random vector $\tilde{\mathbf{Y}}$ is:

$$\operatorname{TC}(\tilde{\mathbf{Y}}) = \sum_i h(\tilde{Y}_i) - h(\tilde{\mathbf{Y}}).$$

Applying the gradient operator ∇_θ to both sides, and using the linearity of the gradient over the finite sum, yields:

$$\nabla_\theta \operatorname{TC}(\tilde{\mathbf{Y}}) = \sum_i \nabla_\theta h(\tilde{Y}_i) - \nabla_\theta h(\tilde{\mathbf{Y}}).$$

We now substitute the results from our previous lemmas. From Lemma E.2, we replace the gradient of each smoothed marginal entropy:

$$\nabla_\theta \operatorname{TC}(\tilde{\mathbf{Y}}) = \sum_i \big(\nabla_\theta h(Y_i) + \mathcal{O}(\epsilon^2)\big) - \nabla_\theta h(\tilde{\mathbf{Y}}).$$

Since the dimensionality $D = \dim(\boldsymbol{\mathcal{Y}})$ is finite and independent of ϵ, summing the asymptotic error term D times remains bounded as $\mathcal{O}(\epsilon^2)$. Next, we apply the result from the previous lemma regarding the joint entropy, substituting $\nabla_\theta h(\tilde{\mathbf{Y}}) = \nabla_\theta h_{\text{surr}}(\mathbf{Y})$:

$$\nabla_\theta \operatorname{TC}(\tilde{\mathbf{Y}}) = \sum_i \nabla_\theta h(Y_i) - \nabla_\theta h_{\text{surr}}(\mathbf{Y}) + \mathcal{O}(\epsilon^2).$$

Recognizing that the first two terms form exactly the gradient of the surrogate total correlation, we conclude:

$$\nabla_\theta \operatorname{TC}(\tilde{\mathbf{Y}}) = \nabla_\theta \operatorname{TC}_{\text{sur}}(\mathbf{Y}) + \mathcal{O}(\epsilon^2).$$

□

E.3 Reinterpreting Total Correlation

In this section, we will reinterpret total correlation in terms of conditional mutual informations, and show that the covariance structure of the model output places a lower bound on total correlation. For notational simplicity, we will take $\mathbf{Y}$ as the smoothed output random vector $\tilde{\mathbf{Y}}$, and following Proposition E.4, we will assume that minimizing total correlation surrogate leads to minimizing total correlation of the smoothed output random vector. Throughout our derivations, we will use the tensor slice notation introduced in Appendix A.1.1.

E.3.1 Multiscale Decomposition for Total Correlation

Under the introduced notation, the total correlation is written as

$$\operatorname{TC}(\mathbf{Y}) = \left[\sum_{i_1=1}^{N_1} \cdots \sum_{i_P=1}^{N_P} h(\mathbf{Y}_{i_1,\ldots,i_P})\right] - h(\mathbf{Y}) \tag{35}$$

$$= \left[\sum_{\mathbf{i}^{(P)}} h(\mathbf{Y}_{\mathbf{i}^{(P)}})\right] - h(\mathbf{Y}). \tag{36}$$

In our proofs, we use a decomposed form of total correlation. Each term corresponds to the total correlation for a different scale. For $s = 0, \ldots, P-1$, we define

$$\operatorname{TC}^{(s)}(\mathbf{Y}) := \left[\sum_{n_{s+1}=1}^{N_{s+1}} h\big(\mathbf{Y}_{\mathbf{n}^{(s)}, n_{s+1}}\big) - h(\mathbf{Y}_{\mathbf{n}^{(s)}})\right],$$

the common value of this bracket across the slices of scale s. The total correlation is recovered by summing these contributions over the slices of every scale. For $s = 0$ the scale contains a single slice, $\mathbf{Y}_{\mathbf{n}^{(0)}} := \mathbf{Y}$.

We write $d_s := \dim(\mathbf{Y}_{\mathbf{n}^{(s)}}) = \prod_{q=s+1}^{P} N_q$ for the dimension of a slice at scale s.

Remark E.5 (Total correlation dominates each scale contribution)**.** *Notably, as any* $\mathrm{TC}^{(s)}(\mathbf{Y})$ *is non-negative, this also demonstrates that total correlation* $\mathrm{TC}(\mathbf{Y})$ *acts like an upper bound to the total correlation contribution of each scale.*

E.3.2 Total Correlation as Sum of Conditional Mutual Informations

In this section, we will show that any contribution $\mathrm{TC}^{(s)}$ can be written as a sum of conditional mutual information terms.

Lemma E.6 (Multiscale decomposition into conditional mutual informations)**.** *Assume that the differential entropies of* $\mathbf{Y}$ *and of all its tensor slices are finite, so that every conditional entropy below is well defined. For the smoothed output of Appendix E.1, the added noise makes the covariance full rank, and the top-k part has finite entropy by Corollary C.9 (finiteness of the k-dimensional output entropy). Define*

$$\mathbf{Y}_{\mathbf{n}^{(s)},a:b} := (\mathbf{Y}_{\mathbf{n}^{(s)},a}, \mathbf{Y}_{\mathbf{n}^{(s)},a+1}, \dots, \mathbf{Y}_{\mathbf{n}^{(s)},b}),$$

where $\mathbf{n}^{(s)} = (n_1, n_2, \dots, n_s)$*.* $\mathbf{Y}_{\mathbf{n}^{(s)},a:b}$ *will be non-existent whenever* $a > b$*, thus conditioning for this case will be equivalent to the unconditioned case.*

Then the s-th contribution to total correlation can be written as the sum of pair-wise and conditional mutual information terms:

$$\begin{aligned}\mathrm{TC}^{(s)}(\mathbf{Y}) = &\sum_{i=2}^{N_{s+1}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},i}; \mathbf{Y}_{\mathbf{n}^{(s)},i-1}\big) \\ &+ \sum_{m=1}^{N_{s+1}-2} \sum_{i=m+2}^{N_{s+1}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},i}; \mathbf{Y}_{\mathbf{n}^{(s)},i-m-1} \mid \mathbf{Y}_{\mathbf{n}^{(s)},i-m:i-1}\big)\,.\end{aligned}$$

Proof. For $m = 0, \dots, N_{s+1} - 1$, define the m-th order chain-rule approximation

$$h^{(m)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) := \sum_{i=1}^{N_{s+1}} h\big(\mathbf{Y}_{\mathbf{n}^{(s)},i} \mid \mathbf{Y}_{\mathbf{n}^{(s)},\max(1,i-m):i-1}\big)\,.$$

With this convention,

$$h^{(0)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) = \sum_{i=1}^{N_{s+1}} h(\mathbf{Y}_{\mathbf{n}^{(s)},i}),$$

while

$$h^{(N_{s+1}-1)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) = \sum_{i=1}^{N_{s+1}} h(\mathbf{Y}_{\mathbf{n}^{(s)},i} \mid \mathbf{Y}_{\mathbf{n}^{(s)},1:i-1}) = h(\mathbf{Y}_{\mathbf{n}^{(s)}})$$

by the chain rule (Cover, 1999). Therefore,

$$\begin{aligned}\mathrm{TC}^{(s)}(\mathbf{Y}) &= h^{(0)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) - h^{(N_{s+1}-1)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) \\ &= \sum_{m=0}^{N_{s+1}-2} \Big[h^{(m)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) - h^{(m+1)}(\mathbf{Y}_{\mathbf{n}^{(s)}})\Big]\,.\end{aligned}$$

For each m, the terms for $i \le m+1$ cancel out perfectly. The difference is therefore:

$$h^{(m)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) - h^{(m+1)}(\mathbf{Y}_{\mathbf{n}^{(s)}}) = \sum_{i=m+2}^{N_{s+1}} \Big[h(\mathbf{Y}_{\mathbf{n}^{(s)},\,i} \mid \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m:i-1}) - h(\mathbf{Y}_{\mathbf{n}^{(s)},\,i} \mid \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m-1:i-1}) \Big]$$
$$= \sum_{i=m+2}^{N_{s+1}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},\,i}; \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m-1} \mid \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m:i-1}\big)\,.$$

Substituting this into the sum gives

$$\mathrm{TC}^{(s)}(\mathbf{Y}) = \sum_{m=0}^{N_{s+1}-2} \sum_{i=m+2}^{N_{s+1}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},\,i}; \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m-1} \mid \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m:i-1}\big)\,.$$

If we isolate the term $m = 0$, we have:

$$\mathrm{TC}^{(s)}(\mathbf{Y}) = \sum_{i=2}^{N_{s+1}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},\,i}; \mathbf{Y}_{\mathbf{n}^{(s)},\,i-1}\big) + \sum_{m=1}^{N_{s+1}-2} \sum_{i=m+2}^{N_{s+1}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},\,i}; \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m-1} \mid \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m:i-1}\big)\,.$$

□

E.4 Total Correlation of a Stationary Field

Optimal stationarity makes the slices at a given scale identically distributed. This allows us to focus on scale-level relations rather than individual slices by aggregating metrics over the slices. For a functional m evaluated on identically distributed slices $\{\mathbf{Y}_{\mathbf{n}^{(s)}}\}$, we denote its aggregate by

$$m^{(s)}_{\mathrm{agg}}(\mathbf{Y}) := \sum_{\mathbf{n}^{(s)}} m(\mathbf{Y}_{\mathbf{n}^{(s)}})\,. \tag{37}$$

Since the summation is taken over all slices, the slice index no longer needs to be carried explicitly.

In this notation, the total correlation decomposition reads

$$\mathrm{TC}(\mathbf{Y}) = \sum_{s=0}^{P-1} \mathrm{TC}^{(s)}_{\mathrm{agg}}(\mathbf{Y}). \tag{38}$$

E.4.1 Mutual Information Lower Bound from Linear Prediction Error

Here, we show that the residual of the best linear predictor of one random vector from the other sets a mutual information lower bound.

Lemma E.7 (Mutual information lower bound from linear prediction error)**.** *Let $\mathbf{X} \in \mathbb{R}^{d'}$ and $\mathbf{Y} \in \mathbb{R}^{d}$ denote two zero-mean multi-variate random vectors, and write*

$$\mathrm{Cov}(\mathbf{W}, \mathbf{W}') := \mathbb{E}\big[\mathbf{W}\mathbf{W}'^{\top}\big]$$

for their cross-covariance, so that $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$ is the covariance of $\mathbf{X}$. Let

$$\mathbf{\Sigma}(\mathbf{Y}, \mathbf{X}) := \mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) - \mathrm{Cov}(\mathbf{Y}, \mathbf{X})\ \mathrm{Cov}(\mathbf{X}, \mathbf{X})^{-1}\,\mathrm{Cov}(\mathbf{Y}, \mathbf{X})^{\top}$$

denote the error covariance of the best linear predictor of $\mathbf{Y}$ from $\mathbf{X}$. Assume that the differential entropy of $\mathbf{Y}$ is finite and that $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$ and $\mathbf{\Sigma}(\mathbf{Y}, \mathbf{X})$ have full rank. Then, their mutual information is lower bounded by

$$I(\mathbf{X}; \mathbf{Y}) \ge h(\mathbf{Y}) - \frac{1}{2} \log\big((2\pi e)^d \det \mathbf{\Sigma}(\mathbf{Y}, \mathbf{X})\big),$$

that is, by the entropy of a Gaussian carrying the prediction error covariance.

Proof. To find a lower bound for mutual information, we first find an upper bound for the conditional entropy. Consider a predictor function $f : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$, and the prediction error $\mathbf{E} = \mathbf{Y} - f(\mathbf{X})$. Because the Gaussian distribution maximizes entropy for a given covariance, the conditional entropy is upper bounded by the entropy of a Gaussian with covariance $\mathbb{E}[\mathbf{E}\mathbf{E}^\top]$:

$$h(\mathbf{Y}|\mathbf{X}) \leq h(\mathbf{E}) \leq \frac{1}{2} \log \left((2\pi e)^d \det \mathbb{E} \left[\mathbf{E}\mathbf{E}^\top \right] \right).$$

This can be expanded to:

$$h(\mathbf{Y}|\mathbf{X}) \leq \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log \det \mathbb{E} \left[\mathbf{E}\mathbf{E}^\top \right].$$

The above inequality is satisfied for any f, so we are free to choose the one that makes the right-hand side smallest. Among linear predictors $f(\mathbf{X}) = \mathbf{B}\mathbf{X}$, the error covariance $\mathbb{E}[\mathbf{E}\mathbf{E}^\top]$ is minimized by $\mathbf{B} = \mathrm{Cov}(\mathbf{Y}, \mathbf{X})\ \mathrm{Cov}(\mathbf{X}, \mathbf{X})^{-1}$. Substituting it gives

$$h(\mathbf{Y}|\mathbf{X}) \leq \frac{1}{2} \log \left((2\pi e)^d \det \boldsymbol{\Sigma}(\mathbf{Y}, \mathbf{X}) \right),$$

and subtracting from $h(\mathbf{Y})$ yields the claim. When $\mathbf{X}$ and $\mathbf{Y}$ have the same dimension, the identity predictor $f(\mathbf{X}) = \mathbf{X}$ is the special case $\mathbf{B} = \mathbf{I}$, so the bound is never weaker than the one it yields. □

E.4.2 Mutual Information Lower Bound for Model Output Slices

The stationarity of the model output fixes the covariance between neighboring slices. This allows us to define lower bounds for the total correlation at each scale by using Lemma E.7.

Lemma E.8 (Covariance mutual information lower bound for output slices)**.** *For a scale s, let the linear-prediction lower bound of Lemma E.7 apply to the model output slices. By the integer-translation Corollary of the stationarity analysis (Corollary D.4), joint stationarity holds under circulant translations along every axis, so the lag-r cross-covariance of sub-slices along axis $s+1$,*

$$\mathbf{C}^{(s)}(\mathbf{Y}; r) := \mathbf{C}\big(\mathbf{Y}_{\mathbf{n}^{(s)},i+r},\ \mathbf{Y}_{\mathbf{n}^{(s)},i}\big),$$

does not depend on i or on $\mathbf{n}^{(s)}$; as with the other scale-indexed quantities, the superscript records the scale and the slice index is left implicit. Then, for any lag $r \neq 0$, the aggregated mutual information between sub-slices separated by r,

$$I^{(s)}_{\mathrm{agg}}(\mathbf{Y}; r) := \sum_{\mathbf{n}^{(s)}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},i}; \mathbf{Y}_{\mathbf{n}^{(s)},i+r}\big),$$

is lower bounded by

$$I^{(s)}_{\mathrm{agg}}(\mathbf{Y}; r) \geq \frac{1}{N_{s+1}} h^{(s+1)}_{\mathrm{agg}}(\mathbf{Y}) - h^{\mathrm{cond},(s)}_{\mathrm{agg}}(\mathbf{Y}; r), \tag{39}$$

where the aggregated conditional term is the scale aggregate of

$$\begin{aligned} h^{\mathrm{cond}}(\mathbf{Y}_{\mathbf{n}^{(s)}}; r) &:= \frac{1}{2} \log \Big((2\pi e)^{d_{s+1}} \det \boldsymbol{\Sigma}^{(s)}(\mathbf{Y}; r) \Big), \\ \boldsymbol{\Sigma}^{(s)}(\mathbf{Y}; r) &:= \mathbf{C}^{(s)}(\mathbf{Y}; 0) - \mathbf{C}^{(s)}(\mathbf{Y}; r)\, \mathbf{C}^{(s)}(\mathbf{Y}; 0)^{-1} \mathbf{C}^{(s)}(\mathbf{Y}; r)^\top, \end{aligned} \tag{40}$$

that is, $\boldsymbol{\Sigma}^{(s)}(\mathbf{Y}; r)$ is the error covariance of the best linear predictor of a sub-slice from the one r steps away along axis $s+1$.

Proof. Fix a scale s and a lag $r \neq 0$, and apply Lemma E.7 to the pair

$$\mathbf{X} := \mathbf{Y}_{\mathbf{n}^{(s)},i}, \qquad \mathbf{Y}' := \mathbf{Y}_{\mathbf{n}^{(s)},i+r},$$

both of which live in $\mathbb{R}^{d_{s+1}}$ and are centered by Constraint 7. By stationarity their covariances and cross-covariance are $\mathbf{C}^{(s)}(\mathbf{Y}; 0)$ and $\mathbf{C}^{(s)}(\mathbf{Y}; r)$ irrespective of i, so the error covariance of the best linear predictor is exactly the $\boldsymbol{\Sigma}^{(s)}(\mathbf{Y}; r)$ of (40) and the lemma gives

$$\begin{aligned} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},i}; \mathbf{Y}_{\mathbf{n}^{(s)},i+r}\big) &\geq h\big(\mathbf{Y}_{\mathbf{n}^{(s)},i+r}\big) - \frac{1}{2} \log \Big((2\pi e)^{d_{s+1}} \det \boldsymbol{\Sigma}^{(s)}(\mathbf{Y}; r) \Big) \\ &= h\big(\mathbf{Y}_{\mathbf{n}^{(s)},i+r}\big) - h^{\mathrm{cond}}(\mathbf{Y}_{\mathbf{n}^{(s)}}; r). \end{aligned}$$

Summing both sides over all slices gives

$$\sum_{\mathbf{n}^{(s)}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},i};\mathbf{Y}_{\mathbf{n}^{(s)},i+r}\big) \geq \sum_{\mathbf{n}^{(s)}} h\big(\mathbf{Y}_{\mathbf{n}^{(s)},i+r}\big) - \sum_{\mathbf{n}^{(s)}} h^{\mathrm{cond}}(\mathbf{Y}_{\mathbf{n}^{(s)}};r)\,.$$

By the aggregate notation (37),

$$I^{(s)}_{\mathrm{agg}}(\mathbf{Y};r) = \sum_{\mathbf{n}^{(s)}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},i};\mathbf{Y}_{\mathbf{n}^{(s)},i+r}\big)$$
$$h^{\mathrm{cond},(s)}_{\mathrm{agg}}(\mathbf{Y};r) = \sum_{\mathbf{n}^{(s)}} h^{\mathrm{cond}}(\mathbf{Y}_{\mathbf{n}^{(s)}};r)\,.$$

For the remaining entropy term, we use that slices are identically distributed to convert the sum. For $j := i + r$, we have

$$\sum_{\mathbf{n}^{(s)}} h\big(\mathbf{Y}_{\mathbf{n}^{(s)},j}\big) = \frac{1}{N_{s+1}} \sum_{\mathbf{n}^{(s)}} \sum_{n_{s+1}=1}^{N_{s+1}} h\big(\mathbf{Y}_{\mathbf{n}^{(s)},n_{s+1}}\big) = \frac{1}{N_{s+1}} h^{(s+1)}_{\mathrm{agg}}(\mathbf{Y})\,.$$

Collecting the terms gives the stated bound (39).

$$I^{(s)}_{\mathrm{agg}}(\mathbf{Y};r) \geq \frac{1}{N_{s+1}} h^{(s+1)}_{\mathrm{agg}}(\mathbf{Y}) - h^{\mathrm{cond},(s)}_{\mathrm{agg}}(\mathbf{Y};r)$$

□

E.4.3 Lower Bound to Total Correlation

In this section, we obtain a lower bound for total correlation, which also places an upper bound for conditional dependencies.

First, let us define the sum of conditional mutual informations for pairs of tensor slices where the index gap is ≥ 2 as

$$I(\mathbf{Y}_{\mathbf{n}^{(s)}};\geq 2) = \sum_{m=1}^{N_{s+1}-2} \sum_{i=m+2}^{N_{s+1}} I\big(\mathbf{Y}_{\mathbf{n}^{(s)},\,i};\mathbf{Y}_{\mathbf{n}^{(s)},\,i-m-1} \mid \mathbf{Y}_{\mathbf{n}^{(s)},\,i-m:i-1}\big)\,.$$

To omit the slice index, we aggregate this metric over all slices, and define

$$I^{(s)}_{\mathrm{agg}}(\mathbf{Y};\geq 2) := \sum_{\mathbf{n}^{(s)}} I(\mathbf{Y}_{\mathbf{n}^{(s)}};\geq 2)\,.$$

Proposition E.9 (Total correlation bounds the higher-order conditional dependencies)**.** *Let the multiscale decomposition of Lemma E.6 and the covariance lower bound of Lemma E.8 apply, and let* $\mathbf{Y}$ *denote the smoothed output of Proposition E.4 (gradients of the surrogate total correlation), so that* $h(\mathbf{Y})$ *carries the additive offset of the smoothing and the magnitude of the bound depends on* ϵ*. Then the total correlation is lower bounded by*

$$\mathrm{TC}(\mathbf{Y}) \geq \sum_{s=0}^{P-1} \left(\left(1 - \frac{1}{N_{s+1}}\right) h(\mathbf{Y}) - (N_{s+1}-1)\, h^{\mathrm{cond},(s)}_{\mathrm{agg}}(\mathbf{Y};1) + I^{(s)}_{\mathrm{agg}}(\mathbf{Y};\geq 2) \right),$$

where $h^{\mathrm{cond},(s)}_{\mathrm{agg}}$ *and* $I^{(s)}_{\mathrm{agg}}(\,\cdot\,;\geq 2)$ *are the scale aggregates of* (37) *and* h^{cond} *is the per-slice conditional term* (40)*.*

Proof. By Lemma E.6, the s-th scale contribution $\mathrm{TC}^{(s)}(\mathbf{Y})$ decomposes into adjacent pairwise dependencies and higher-order conditional dependencies:

$$\mathrm{TC}^{(s)}(\mathbf{Y}) = \sum_{i=2}^{N_{s+1}} I(\mathbf{Y}_{\mathbf{n}^{(s)},\,i};\mathbf{Y}_{\mathbf{n}^{(s)},\,i-1}) + I(\mathbf{Y}_{\mathbf{n}^{(s)}};\geq 2)\,.$$

By stationarity along axis $s+1$, each of the $N_{s+1}-1$ adjacent mutual information terms is the same. Aggregating the decomposition over the slices of scale s in the sense of (37) therefore gives

$$\mathrm{TC}^{(s)}_{\mathrm{agg}}(\mathbf{Y}) = (N_{s+1}-1)\, I^{(s)}_{\mathrm{agg}}(\mathbf{Y};1) + I^{(s)}_{\mathrm{agg}}(\mathbf{Y};\geq 2).$$

Applying the aggregated covariance lower bound (39) at lag $r=1$,

$$\mathrm{TC}^{(s)}_{\mathrm{agg}}(\mathbf{Y}) \geq (N_{s+1}-1)\left(\frac{1}{N_{s+1}} h^{(s+1)}_{\mathrm{agg}}(\mathbf{Y}) - h^{\mathrm{cond},(s)}_{\mathrm{agg}}(\mathbf{Y};1)\right) + I^{(s)}_{\mathrm{agg}}(\mathbf{Y};\geq 2).$$

The scale-$(s+1)$ slices partition the coordinates of $\mathbf{Y}$, so subadditivity of differential entropy is exactly the statement that the whole is dominated by its aggregate,

$$h(\mathbf{Y}) \leq h^{(s+1)}_{\mathrm{agg}}(\mathbf{Y}),$$

whence $\left(1-\frac{1}{N_{s+1}}\right) h^{(s+1)}_{\mathrm{agg}}(\mathbf{Y}) \geq \left(1-\frac{1}{N_{s+1}}\right) h(\mathbf{Y})$. Substituting this and summing over scales using (38) results in the final lower bound. □

F Analysis of Joint Loss

In this section, we will investigate the combined effect of the loss terms, and show how these lead to signal recovery. For this, first we will show that the recovery kernel is translation invariant, and then that the recovery admits an exact discrete convolutional form (Appendix F.2).

F.1 Probe Response Spectrum in the Conditioned Sample Space

The recovery kernel is built from the recovery filter and the probe response, so its spectrum is set jointly by the two. Information maximization already fixes the filter spectrum (Proposition C.13). Here we determine the probe response spectrum. It is set by the energy that the conditioning matrix places on each eigenvector of ρ_Z, together with the mixing matrix of the measurement operator. Information maximization drives every such energy to the bound imposed by the architecture, and for a randomly drawn mixing matrix the resulting spectrum concentrates as the dimension grows.

Constraint 9 (Mode-energy bound of the conditioning matrix)**.** *Let $\mathbf{u_m}$ denote the eigenvector of ρ_Z carrying the mode $\mathbf{m}$. The energy that the conditioning matrix places on it, which we call the* mode energy*,*

$$e_{\mathbf{m}} := \|\mathbf{C}^\top \mathbf{u_m}\|_2^2 = \mathbf{u_m}^\dagger \mathbf{C}\mathbf{C}^\top \mathbf{u_m},$$

is bounded by one, $e_{\mathbf{m}} \leq 1$, for every mode. The implementation enforces a soft version of this bound after every gradient update, by rescaling $\mathbf{C}$ along the eigenvectors of ρ_Z (Appendix G.1).

Lemma F.1 (Probe response spectrum is set by the mode energy and the mixing matrix)**.** *Following the conditioning-injectivity Lemma of the information maximization analysis (Lemma C.4), let the conditioning matrix be injective, and following the symmetry-identification Theorem of the stationarity analysis (Theorem D.8), let the learned symmetry representation carry the same eigenvectors as the representation induced on the conditioned sample space by the latent symmetry. Let $\mathbf{A}$ be the invertible mixing matrix of the decomposition Lemma of the problem setting (Lemma B.1), let $\mathbf{q_k} := \mathbf{S}^C[\chi_{\mathbf{k}}]$ be the sampled character of mode $\mathbf{k}$, and define*

$$\beta_{\mathbf{k}} := \|\mathbf{A}^{-\top} \mathbf{q_k}\|_2^2.$$

Then the spectral coefficients of the probe response are determined by the mode energy of Constraint 9 and by the mixing matrix,

$$h_{Z,\mathbf{k}} = \alpha_{\mathbf{k}} = \sqrt{\frac{e_{\mathbf{k}}}{\beta_{\mathbf{k}}}}.$$

Proof. By the symmetry-transfer Corollary of the information maximization analysis (Corollary C.6), the conditioning matrix intertwines the two representations, $\rho_Z(\mathbf{s})\mathbf{C} = \mathbf{C}\rho_X(\mathbf{s})$, and by the representation-existence Lemma of the problem setting (Lemma B.4), the measurement operator intertwines the latent symmetry, $\rho_X(\mathbf{s})\mathbf{M} = \mathbf{M}\tau_{\mathbf{s}}$. Composing the two, the latent symmetry induces the representation ρ_Z on the conditioned sample space through the composite map,

$$\rho_Z(\mathbf{s})\mathbf{C}\mathbf{M} = \mathbf{C}\mathbf{M}\tau_{\mathbf{s}}.$$

Characters are complex valued, so they are eigenfunctions of the translation only after the operators are extended to complex arguments. We therefore complexify the measurement operator exactly as the sampling operator is complexified in the proof of the before-mixing Lemma of the problem setting (Lemma B.3),

$$\mathbf{M}^C[q] := \mathbf{M}[\mathrm{R}(q)] + \mathrm{i}\,\mathbf{M}[\mathrm{I}(q)], \qquad q \in \mathcal{F}^C := \{f + \mathrm{i}g \mid f, g \in \mathcal{F}\},$$

and extend $\tau_{\mathbf{s}}$, $\mathbf{C}$ and $\rho_Z(\mathbf{s})$ to complex arguments in the same way. Each extension is $\mathbb{C}$-linear and agrees with the original operator on real arguments, so an identity between real-linear operators holds between their complexifications as well, applied separately to the real and imaginary parts. Hence

$$\rho_Z(\mathbf{s})\mathbf{C}\mathbf{M}^C = \mathbf{C}\mathbf{M}^C\tau_{\mathbf{s}}.$$

Investigating the action of both sides on the character $\chi_{\mathbf{m}}$, and using $\tau_{\mathbf{s}}\chi_{\mathbf{m}} = e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{s}}\chi_{\mathbf{m}}$, gives

$$\rho_Z(\mathbf{s})\mathbf{C}\mathbf{M}^C[\chi_{\mathbf{m}}] = e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{s}}\mathbf{C}\mathbf{M}^C[\chi_{\mathbf{m}}].$$

Since $\mathbf{C}$ is injective and $\mathbf{M}^C[\chi_{\mathbf{m}}] \neq \mathbf{0}$, the vector $\mathbf{C}\mathbf{M}^C[\chi_{\mathbf{m}}]$ is an eigenvector of ρ_Z with the eigenvalue $e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{s}}$. This eigenvector lies in the range $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ of the conditioning matrix, which ρ_Z leaves invariant. These eigenvalues are pairwise distinct as functions of $\mathbf{s}$, and by the probe-orbit Corollary of the information maximization analysis (Corollary C.8) the spectrum of ρ_Z on $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ is exactly the active mode set, so each joint eigenspace of the family $\{\rho_Z(\mathbf{s})\}$ restricted to $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ is one-dimensional and is labeled by a single mode. Consequently

$$\mathbf{C}\mathbf{M}^C[\chi_{\mathbf{m}}] = \alpha_{\mathbf{m}}\mathbf{u}_{\mathbf{m}}, \qquad \alpha_{\mathbf{m}} > 0,$$

where $\mathbf{u}_{\mathbf{m}}$ is the eigenvector of the learned symmetry representation carrying the mode $\mathbf{m}$, whose free phase we fix so that $\alpha_{\mathbf{m}}$ is positive.

By the decomposition Lemma of the problem setting (Lemma B.1), $\mathbf{M} = \mathbf{A}\mathbf{S}\mathrm{P}_R$, and the projector acts as the identity on $\mathcal{F}_R$, which contains every active character. Therefore

$$\mathbf{M}^C[\chi_{\mathbf{m}}] = \mathbf{A}\mathbf{q}_{\mathbf{m}}, \qquad \mathbf{C}\mathbf{A}\mathbf{q}_{\mathbf{m}} = \alpha_{\mathbf{m}}\mathbf{u}_{\mathbf{m}}, \tag{41}$$

where $\mathbf{q}_{\mathbf{m}} := \mathbf{S}^C[\chi_{\mathbf{m}}]$ is the sampled character, written $\mathbf{u}_{\mathbf{m}}$ in that analysis and renamed here because $\mathbf{u}_{\mathbf{m}}$ denotes the eigenvectors of ρ_Z throughout this section. The character Lemma of the problem setting (Lemma B.2) shows the sampled characters to be orthonormal across modes, and there are d_X active modes, so $\{\mathbf{q}_{\mathbf{m}}\}$ is an orthonormal basis of $\mathbb{C}^{d_X}$.

In the before-mixing Lemma of the problem setting (Lemma B.3), the probe response at the discretized latent space is given by $h_{F,\mathbf{m}} = 1$. Expanding it over the orthonormal sampled characters and acting with the remaining factors of the composite map therefore gives

$$\mathbf{h}_Z = \mathbf{C}\mathbf{A}\sum_{\mathbf{m}} h_{F,\mathbf{m}}\mathbf{q}_{\mathbf{m}} = \sum_{\mathbf{m}} \mathbf{C}\mathbf{A}\mathbf{q}_{\mathbf{m}} = \sum_{\mathbf{m}} \alpha_{\mathbf{m}}\mathbf{u}_{\mathbf{m}}.$$

Constraint 1 (special orthogonal symmetry representation) makes ρ_Z orthogonal, hence normal, so the eigenvectors $\{\mathbf{u}_{\mathbf{m}}\}$ of the active modes form an orthonormal basis of $\mathbf{C}\,\boldsymbol{\mathcal{X}}$ and

$$h_{Z,\mathbf{k}} = \mathbf{u}_{\mathbf{k}}^\dagger \sum_{\mathbf{m}} \alpha_{\mathbf{m}}\mathbf{u}_{\mathbf{m}} = \alpha_{\mathbf{k}}.$$

It remains to express $\alpha_{\mathbf{k}}$ through the mode energy. By (41) and the orthonormality of the eigenvectors, $\mathbf{u}_{\mathbf{k}}^\dagger\mathbf{C}\mathbf{A}\mathbf{q}_{\mathbf{m}} = \alpha_{\mathbf{m}}\,\mathbf{u}_{\mathbf{k}}^\dagger\mathbf{u}_{\mathbf{m}} = \alpha_{\mathbf{k}}\delta_{\mathbf{k},\mathbf{m}}$. Expanding the row vector $\mathbf{u}_{\mathbf{k}}^\dagger\mathbf{C}\mathbf{A}$ in the orthonormal basis $\{\mathbf{q}_{\mathbf{m}}^\dagger\}$ therefore leaves a single term,

$$\mathbf{u}_{\mathbf{k}}^\dagger\mathbf{C}\mathbf{A} = \sum_{\mathbf{m}} \left(\mathbf{u}_{\mathbf{k}}^\dagger\mathbf{C}\mathbf{A}\mathbf{q}_{\mathbf{m}}\right)\mathbf{q}_{\mathbf{m}}^\dagger = \alpha_{\mathbf{k}}\,\mathbf{q}_{\mathbf{k}}^\dagger.$$

Multiplying from the right by $\mathbf{A}^{-1}$ and taking the conjugate transpose, with $\mathbf{C}$ and $\mathbf{A}$ real,

$$\mathbf{C}^\top\mathbf{u}_{\mathbf{k}} = \alpha_{\mathbf{k}}\,\mathbf{A}^{-\top}\mathbf{q}_{\mathbf{k}}.$$

Taking squared norms gives $e_{\mathbf{k}} = \alpha_{\mathbf{k}}^2\,\beta_{\mathbf{k}}$, and since $\alpha_{\mathbf{k}} > 0$,

$$h_{Z,\mathbf{k}} = \alpha_{\mathbf{k}} = \sqrt{\frac{e_{\mathbf{k}}}{\beta_{\mathbf{k}}}}.$$

□

The lemma separates the two factors of the probe response spectrum: the mode energy $e_{\mathbf{k}}$ is controlled by the conditioning matrix, whereas $\beta_{\mathbf{k}}$ depends on the mixing matrix alone. The next proposition shows that information maximization fixes the first factor.

Proposition F.2 (Information maximization saturates the mode-energy bound)**.** *Under the assumptions of the probe-spectrum Lemma (Lemma F.1), let the rank be terminal, $k = d_X$ (Constraint 6), and let the conditioning matrix be aligned with the lifting matrix as in the aligning-conditioning Lemma of the information maximization analysis (Lemma C.10). Then, at the maximum of information over the conditioning matrices that satisfy Constraint 9, the mode energy of every active mode attains its bound, $e_{\mathbf{m}} = 1$, and the probe response spectrum is*

$$|h_{Z,\mathbf{m}}|^2 = \frac{1}{\beta_{\mathbf{m}}}.$$

In particular, if the mixing matrix is orthogonal, then $\beta_{\mathbf{m}} = \|\mathbf{q}_{\mathbf{m}}\|_2^2 = 1$ and $|h_{Z,\mathbf{m}}| = 1$ for every mode.

Proof. By the aligning-conditioning Lemma, $\pi_k(\mathbf{U}\mathbf{C}\boldsymbol{\Sigma}_X^{1/2}) = \pi_k(\mathbf{U})\,\pi_k(\mathbf{C}\boldsymbol{\Sigma}_X^{1/2})$, so the conditioning matrix enters the information only through $\pi_k(\mathbf{C}\boldsymbol{\Sigma}_X^{1/2})$. At the terminal rank $k = d_X$, the injective conditioning matrix factors as $\mathbf{C} = \mathbf{W}\mathbf{R}$, with $\mathbf{W} \in \mathbb{R}^{d_Z \times d_X}$ having orthonormal columns and $\mathbf{R} \in \mathbb{R}^{d_X \times d_X}$ invertible. Left multiplication by $\mathbf{W}$ preserves singular values, so for every invertible $\mathbf{B} \in \mathbb{R}^{d_X \times d_X}$

$$\pi_k(\mathbf{C}\mathbf{B}) = |\det \mathbf{R}|\,|\det \mathbf{B}| = \pi_k(\mathbf{C})\,|\det \mathbf{B}|,$$

where $\pi_k(\mathbf{C})$ is the product of the d_X singular values of $\mathbf{C}$, equal to $|\det \mathbf{C}|$ when $d_Z = d_X$. With $\mathbf{B} = \boldsymbol{\Sigma}_X^{1/2}$, only $\pi_k(\mathbf{C})$ depends on the conditioning matrix. By (41), $\mathbf{C}\mathbf{A} = \sum_{\mathbf{m}} \alpha_{\mathbf{m}} \mathbf{u}_{\mathbf{m}} \mathbf{q}_{\mathbf{m}}^\dagger$, the product of the matrix of active eigenvectors (orthonormal columns), the diagonal matrix of the $\alpha_{\mathbf{m}}$, and a unitary matrix. Hence $\pi_k(\mathbf{C}\mathbf{A}) = \prod_{\mathbf{m}} \alpha_{\mathbf{m}}$, and with $\mathbf{B} = \mathbf{A}$ and the probe-spectrum Lemma

$$\pi_k(\mathbf{C}) = \frac{1}{|\det \mathbf{A}|} \prod_{\mathbf{m}} \sqrt{\frac{e_{\mathbf{m}}}{\beta_{\mathbf{m}}}}.$$

The $\beta_{\mathbf{m}}$ depend on the mixing matrix alone, so $\pi_k(\mathbf{C})$ is increasing in every active mode energy.

The mode energies can be changed one by one without leaving the admissible set. Since ρ_Z is real, $\overline{\mathbf{u}_{\mathbf{m}}}$ is an eigenvector for the mode $-\mathbf{m}$, so $\mathbf{u}_{-\mathbf{m}}$ equals $\overline{\mathbf{u}_{\mathbf{m}}}$ up to a unit phase, and since $\mathbf{C}$ is real, $e_{-\mathbf{m}} = e_{\mathbf{m}}$. For positive scales with $s_{-\mathbf{m}} = s_{\mathbf{m}}$, the matrix $\mathbf{D} := \sum_{\mathbf{m}} s_{\mathbf{m}} \mathbf{u}_{\mathbf{m}} \mathbf{u}_{\mathbf{m}}^\dagger + (\mathbf{I} - \boldsymbol{\Pi}_C)$, with the sum over the active modes and $\boldsymbol{\Pi}_C$ the projector onto the range of $\mathbf{C}$, is therefore real and symmetric. It acts on the range through the eigenvectors of $\rho_Z(\mathbf{s})$ and as the identity on its complement, as $\rho_Z(\mathbf{s})$ does (Corollary C.6), so it commutes with $\rho_Z(\mathbf{s})$, and $\mathbf{D}\mathbf{C}$ satisfies the same intertwining relation as $\mathbf{C}$. Its mode energies are $\|(\mathbf{D}\mathbf{C})^\top \mathbf{u}_{\mathbf{m}}\|_2^2 = \|\mathbf{C}^\top \mathbf{D}\mathbf{u}_{\mathbf{m}}\|_2^2 = s_{\mathbf{m}}^2 e_{\mathbf{m}}$. If some mode energy were below its bound, $e_{\mathbf{m}} < 1$, the choice $s_{\pm\mathbf{m}} = e_{\mathbf{m}}^{-1/2}$, with all other scales equal to one, would keep Constraint 9 satisfied and strictly increase $\pi_k(\mathbf{C})$. Therefore $e_{\mathbf{m}} = 1$ for every active mode at the maximum; the inactive eigenvectors are orthogonal to the range of $\mathbf{C}$, so their mode energy vanishes, and the probe-spectrum Lemma gives $|h_{Z,\mathbf{m}}|^2 = 1/\beta_{\mathbf{m}}$. If $\mathbf{A}$ is orthogonal, $\mathbf{A}^{-\top} = \mathbf{A}$ and $\beta_{\mathbf{m}} = \|\mathbf{A}\mathbf{q}_{\mathbf{m}}\|_2^2 = \|\mathbf{q}_{\mathbf{m}}\|_2^2 = 1$. □

Remark F.3. *The rescaling $\mathbf{C} \leftarrow \mathbf{D}\mathbf{C}$ used in the proof is the family through which the implementation enforces Constraint 9 (Appendix G.1). The analysis therefore uses the same normalization of the conditioning matrix as the implementation, and requires no normalization of $\mathbf{C}^{-1}$.*

For a general mixing matrix, $\beta_{\mathbf{m}}$ varies across modes. The following lemma shows that, when the mixing matrix is drawn at random, this variation vanishes as the dimension grows.

Lemma F.4 (Probe response spectrum for random matrix mixing)**.** *Following the mode-energy saturation Proposition (Proposition F.2), let the mixing matrix $\mathbf{A}$ admit the singular value decomposition*

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Sigma}\mathbf{V}^\top, \qquad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_d), \qquad d := \dim(\boldsymbol{\mathcal{X}}),$$

with singular values $\sigma_{\max} = \sigma_1 \geq \cdots \geq \sigma_d = \sigma_{\min} > 0$, condition number $\kappa := \sigma_{\max}/\sigma_{\min}$, and inverse moments $m'_p := \mathrm{tr}(\boldsymbol{\Sigma}^{-p})/d$. Let $\mathbf{V}$ be a random matrix sampled from the orthogonal group

uniformly on the Haar measure, with $\mathbf{Q}$ *and* $\mathbf{\Sigma}$ *fixed. Then* $\beta_{\mathbf{k}}$ *of the probe-spectrum Lemma has the mode-independent mean*

$$\mu := \mathbb{E}[\beta_{\mathbf{k}}] = m_2',$$

and the variance

$$\mathrm{Var}(\beta_{\mathbf{k}}) = \gamma_{\mathbf{k}} \frac{2\left(m_4' - m_2'^2\right)}{d+2}, \qquad \gamma_{\mathbf{k}} = \begin{cases} 1, & \mathbf{k} = \mathbf{0}, \\ \dfrac{d-2}{2(d-1)}, & \mathbf{k} \neq \mathbf{0}. \end{cases}$$

Consequently, the relative fluctuations of the probe response spectrum vanish as the dimension grows:

$$\frac{\mathrm{Var}\left(|h_{Z,\mathbf{k}}|^2\right)}{\mathbb{E}\left[|h_{Z,\mathbf{k}}|^2\right]^2} \;\le\; \kappa^4 \frac{\mathrm{Var}(\beta_{\mathbf{k}})}{\mu^2} \;=\; \kappa^4 \gamma_{\mathbf{k}} \frac{2\left(m_4'/m_2'^2 - 1\right)}{d+2},$$

which is of order $1/d$ *whenever* κ *stays bounded. By Chebyshev's inequality, for every* $\varepsilon > 0$*, the probability that* $|h_{Z,\mathbf{k}}|^2$ *deviates from its mean by more than a fraction* ε *is at most this ratio divided by* ε^2*, so a randomly drawn mixing matrix yields an approximately flat probe spectrum in high dimension.*

Proof. Let us first simplify $\beta_{\mathbf{k}} = \mathbf{q}_{\mathbf{k}}^\dagger \mathbf{A}^{-1}\mathbf{A}^{-\top}\mathbf{q}_{\mathbf{k}}$ by plugging in $\mathbf{A}^{-\top} = \mathbf{Q}\mathbf{\Sigma}^{-1}\mathbf{V}^\top$, under which the factor $\mathbf{Q}$ cancels,

$$\beta_{\mathbf{k}} = \mathbf{q}_{\mathbf{k}}^\dagger \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \mathbf{q}_{\mathbf{k}}, \qquad \mathbf{\Lambda} := \mathbf{\Sigma}^{-2} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \quad \lambda_i := \sigma_i^{-2}.$$

Defining $\tilde{\mathbf{q}}_{\mathbf{k}} := \mathbf{V}^\top \mathbf{q}_{\mathbf{k}}$ we obtain

$$\beta_{\mathbf{k}} = \tilde{\mathbf{q}}_{\mathbf{k}}^\dagger \mathbf{\Lambda} \tilde{\mathbf{q}}_{\mathbf{k}}.$$

We first collect the moments of the quadratic form $Z := \mathbf{g}^\top \mathbf{\Lambda} \mathbf{g}$ taken along a single direction $\mathbf{g}$ that is uniformly distributed on the unit sphere, which is what $\mathbf{V}^\top$ produces from any fixed real unit vector. Instead of taking the expectation on $\mathbf{V}$, we take it on $\mathbf{g}$, and leverage the isotropy of the Gaussian distribution to compute it.

Setting $\mathbf{g} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ while $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ realizes the uniform direction, and because the norm and the direction of an isotropic Gaussian vector are independent, the moments factorize, $\mathbb{E}\left[(\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x})^p\right] = \mathbb{E}\left[\|\mathbf{x}\|_2^{2p}\right] \mathbb{E}\left[Z^p\right]$. For the mean this gives

$$\mathbb{E}\left[Z\right] = \mathbb{E}\left[\frac{\mathbf{x}^\top}{\|\mathbf{x}\|_2} \mathbf{\Lambda} \frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right] = \frac{\mathbb{E}\left[\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}\right]}{\mathbb{E}\left[\|\mathbf{x}\|_2^2\right]}.$$

Working on the term $\mathbb{E}\left[\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}\right]$ gives

$$\mathbb{E}\left[\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}\right] = \sum_i \lambda_i \mathbb{E}[x_i^2] = \sum_i \lambda_i.$$

$\|\mathbf{x}\|_2^2$ follows a chi-squared distribution with d degrees of freedom, and its mean is given as

$$\mathbb{E}\left[\|\mathbf{x}\|_2^2\right] = d.$$

Therefore the mean of the quadratic form along a uniform direction is

$$\mathbb{E}\left[Z\right] = \frac{\mathrm{tr}(\mathbf{\Lambda})}{d} = m_2'.$$

To obtain its variance, we investigate the term $\mathbb{E}\left[Z^2\right]$,

$$\mathbb{E}\left[Z^2\right] = \mathbb{E}\left[\left(\frac{\mathbf{x}^\top}{\|\mathbf{x}\|_2} \mathbf{\Lambda} \frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right)^2\right] = \frac{\mathbb{E}\left[\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}\, \mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}\right]}{\mathbb{E}\left[\|\mathbf{x}\|_2^4\right]}.$$

$\mathbb{E}\left[\|\mathbf{x}\|_2^4\right]$ is the second raw moment of a chi-squared distribution, and equals $\mathbb{E}\left[\|\mathbf{x}\|_2^4\right] = d(d+2)$.

Investigating the term $\mathbb{E}\left[\mathbf{x}^\top \mathbf{\Lambda x}\, \mathbf{x}^\top \mathbf{\Lambda x}\right]$, and using the Gaussian moments $\mathbb{E}[x_i^4] = 3$ and $\mathbb{E}[x_i^2 x_j^2] = 1$ for $i \neq j$,

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}^\top \mathbf{\Lambda x}\, \mathbf{x}^\top \mathbf{\Lambda x}\right] &= \mathbb{E}\left[\sum_{i,j} x_i \lambda_i x_i x_j \lambda_j x_j\right] \\ &= \sum_{i,j} \lambda_i \lambda_j \mathbb{E}[x_i^2 x_j^2] \\ &= \sum_i \lambda_i^2 \mathbb{E}[x_i^4] + \sum_{\substack{i,j \\ i \neq j}} \lambda_i \lambda_j \mathbb{E}[x_i^2]\mathbb{E}[x_j^2] \\ &= 2\,\mathrm{tr}(\mathbf{\Lambda}^2) + (\mathrm{tr}(\mathbf{\Lambda}))^2 .\end{aligned}$$

Consequently

$$\mathbb{E}\left[Z^2\right] = \frac{2\,\mathrm{tr}(\mathbf{\Lambda}^2) + (\mathrm{tr}(\mathbf{\Lambda}))^2}{d(d+2)}.$$

Previously, we calculated the mean as $\mathrm{tr}(\mathbf{\Lambda})/d$. This gives us the variance expression

$$\begin{aligned}\mathrm{Var}(Z) &= \frac{2\,\mathrm{tr}(\mathbf{\Lambda}^2) + (\mathrm{tr}(\mathbf{\Lambda}))^2}{d(d+2)} - \frac{(\mathrm{tr}(\mathbf{\Lambda}))^2}{d^2} \\ &= \frac{2\left(d\,\mathrm{tr}(\mathbf{\Lambda}^2) - (\mathrm{tr}(\mathbf{\Lambda}))^2\right)}{d^2(d+2)} \\ &= \frac{2\left(\mathrm{tr}(\mathbf{\Lambda}^2)/d - (\mathrm{tr}(\mathbf{\Lambda})/d)^2\right)}{d+2} = \frac{2\left(m_4' - m_2'^2\right)}{d+2}.\end{aligned} \tag{42}$$

It remains to express $\tilde{\mathbf{q}}_\mathbf{k}^\dagger \mathbf{\Lambda} \tilde{\mathbf{q}}_\mathbf{k}$ through such directions. Splitting the sampled character into its real and imaginary parts, $\mathbf{q_k} = \mathbf{a_k} + \mathrm{i}\mathbf{b_k}$, and using that $\mathbf{\Lambda}$ is symmetric so that the cross terms cancel,

$$\tilde{\mathbf{q}}_\mathbf{k}^\dagger \mathbf{\Lambda} \tilde{\mathbf{q}}_\mathbf{k} = \tilde{\mathbf{a}}_\mathbf{k}^\top \mathbf{\Lambda} \tilde{\mathbf{a}}_\mathbf{k} + \tilde{\mathbf{b}}_\mathbf{k}^\top \mathbf{\Lambda} \tilde{\mathbf{b}}_\mathbf{k}, \qquad \tilde{\mathbf{a}}_\mathbf{k} := \mathbf{V}^\top \mathbf{a_k}, \quad \tilde{\mathbf{b}}_\mathbf{k} := \mathbf{V}^\top \mathbf{b_k}.$$

Since the sampling operator is real, conjugating a character conjugates its sample, $\mathbf{q}_{-\mathbf{k}} = \overline{\mathbf{q_k}}$. The sampled characters are orthonormal by the character Lemma of the problem setting (Lemma B.2), so $\mathbf{q}_{-\mathbf{k}}^\dagger \mathbf{q_k} = \|\mathbf{a_k}\|_2^2 - \|\mathbf{b_k}\|_2^2 + 2\mathrm{i}\,\mathbf{a}_\mathbf{k}^\top \mathbf{b_k}$ vanishes whenever $-\mathbf{k} \neq \mathbf{k}$, which Constraint 5 (grid resolution) guarantees for every mode but $\mathbf{k} = \mathbf{0}$, oddness of each axis resolution leaving no other self-conjugate mode. Hence the real and imaginary parts are orthogonal and of equal length,

$$\|\mathbf{a_k}\|_2^2 = \|\mathbf{b_k}\|_2^2 = \tfrac{1}{2}, \qquad \mathbf{a}_\mathbf{k}^\top \mathbf{b_k} = 0, \qquad \mathbf{k} \neq \mathbf{0},$$

while $\mathbf{q_0}$ is real of unit length.

For $\mathbf{k} = \mathbf{0}$ the direction $\tilde{\mathbf{q}}_\mathbf{0} = \mathbf{V}^\top \mathbf{q_0}$ is a single uniform direction, so (42) applies as it stands and $\gamma_\mathbf{0} = 1$.

For $\mathbf{k} \neq \mathbf{0}$, write $\mathbf{g}_1 := \sqrt{2}\tilde{\mathbf{a}}_\mathbf{k}$ and $\mathbf{g}_2 := \sqrt{2}\tilde{\mathbf{b}}_\mathbf{k}$, which are orthonormal, and, being the images of a fixed orthonormal pair under a Haar-distributed orthogonal matrix, are distributed as the first two vectors of a uniformly random orthonormal basis. Then

$$\beta_\mathbf{k} = \tfrac{1}{2}\left(Z_1 + Z_2\right), \qquad Z_i := \mathbf{g}_i^\top \mathbf{\Lambda} \mathbf{g}_i,$$

where each Z_i is marginally the single-direction quadratic form treated above. Their covariance follows from completing the pair to a uniformly random orthonormal basis $\{\mathbf{g}_i\}_{i=1}^d$: the sum $\sum_i Z_i = \mathrm{tr}(\mathbf{\Lambda})$ is deterministic, and the pairs are exchangeable, so

$$0 = \mathrm{Var}\left(\sum_i Z_i\right) = d\,\mathrm{Var}(Z_1) + d(d-1)\,\mathrm{Cov}(Z_1, Z_2), \qquad \mathrm{Cov}(Z_1, Z_2) = -\frac{\mathrm{Var}(Z_1)}{d-1}.$$

Therefore

$$\mathrm{Var}(\beta_\mathbf{k}) = \tfrac{1}{4}\left(2\,\mathrm{Var}(Z_1) + 2\,\mathrm{Cov}(Z_1, Z_2)\right) = \frac{d-2}{2(d-1)}\,\mathrm{Var}(Z_1),$$

which is the stated $\gamma_{\mathbf{k}}$ for $\mathbf{k} \neq \mathbf{0}$. The mean is unaffected by the splitting, since $\mathbb{E}[\beta_{\mathbf{k}}] = \frac{1}{2}(\mathbb{E}[Z_1] + \mathbb{E}[Z_2]) = m_2'$ holds in both cases.

Finally, we transfer the concentration of $\beta_{\mathbf{k}}$ to the probe response spectrum $|h_{Z,\mathbf{k}}|^2 = 1/\beta_{\mathbf{k}}$ of the mode-energy saturation Proposition. Since $\tilde{\mathbf{q}}_{\mathbf{k}}$ is a unit vector, $\beta_{\mathbf{k}}$ is a Rayleigh quotient of $\mathbf{\Lambda}$ and lies in $[\sigma_{\max}^{-2}, \sigma_{\min}^{-2}]$; the same interval contains μ. Hence

$$\left|\frac{1}{\beta_{\mathbf{k}}} - \frac{1}{\mu}\right| = \frac{|\beta_{\mathbf{k}} - \mu|}{\beta_{\mathbf{k}}\,\mu} \leq \frac{\sigma_{\max}^2}{\mu}\,|\beta_{\mathbf{k}} - \mu|,$$

and therefore

$$\operatorname{Var}\left(\frac{1}{\beta_{\mathbf{k}}}\right) \leq \mathbb{E}\left[\left(\frac{1}{\beta_{\mathbf{k}}} - \frac{1}{\mu}\right)^2\right] \leq \frac{\sigma_{\max}^4}{\mu^2}\operatorname{Var}(\beta_{\mathbf{k}}).$$

By Jensen's inequality, $\mathbb{E}[1/\beta_{\mathbf{k}}] \geq 1/\mu$. Dividing, and using $\mu \leq \sigma_{\min}^{-2}$,

$$\frac{\operatorname{Var}\left(|h_{Z,\mathbf{k}}|^2\right)}{\mathbb{E}\left[|h_{Z,\mathbf{k}}|^2\right]^2} \leq \sigma_{\max}^4 \operatorname{Var}(\beta_{\mathbf{k}}) = \sigma_{\max}^4\mu^2\,\frac{\operatorname{Var}(\beta_{\mathbf{k}})}{\mu^2} \leq \kappa^4\,\frac{\operatorname{Var}(\beta_{\mathbf{k}})}{\mu^2}.$$

Substituting the variance and mean of $\beta_{\mathbf{k}}$ gives the stated bound, and Chebyshev's inequality gives the probability statement. □

F.2 Recovery Integral on Lattice

We will first demonstrate that the recovery integral (8) admits a discrete convolutional form.

Lemma F.5 (Translation invariance of the recovery kernel)**.** *Following the symmetry-identification Theorem of the stationarity analysis (Theorem D.8), let the symmetry be identified up to a signed axis permutation. Then the recovery kernel $k(\mathbf{s}, \mathbf{t}) := \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t})\rho_Z(\mathbf{s})\mathbf{h}_Z$ is translation invariant: it depends only on $\mathbf{r} := \mathbf{s} - \mathbf{t}$,*

$$k(\mathbf{s}, \mathbf{t}) = k(\mathbf{s} - \mathbf{t}), \qquad k(\mathbf{r}) := \mathbf{w}_Z^\top \rho_Z(\mathbf{r})\mathbf{h}_Z.$$

Proof. In the recovery integral (8), the recovery kernel is given as

$$k(\mathbf{s}, \mathbf{t}) := \mathbf{w}_Z^\top \hat{\rho}_Z(-\mathbf{t})\rho_Z(\mathbf{s})\mathbf{h}_Z.$$

By the symmetry-identification Theorem, the symmetry has been identified up to a signed axis permutation on the range $\mathbf{C}\,\boldsymbol{\mathcal{X}}$, which contains the probe orbit $\{\rho_Z(\mathbf{s})\mathbf{h}_Z\}$. Up to this ambiguity, which only relabels the axes, we write

$$\hat{\rho}_Z(\mathbf{t}) = \rho_Z(\mathbf{t}) \quad \text{on } \mathbf{C}\,\boldsymbol{\mathcal{X}}.$$

Since $\rho_Z(\mathbf{s})\mathbf{h}_Z$ lies in that range and ρ_Z is a representation, $\hat{\rho}_Z(-\mathbf{t})\rho_Z(\mathbf{s})\mathbf{h}_Z = \rho_Z(-\mathbf{t})\rho_Z(\mathbf{s})\mathbf{h}_Z = \rho_Z(\mathbf{s} - \mathbf{t})\mathbf{h}_Z$. Defining $\mathbf{r} := \mathbf{s} - \mathbf{t}$,

$$k(\mathbf{s}, \mathbf{t}) = \mathbf{w}_Z^\top \rho_Z(\mathbf{s} - \mathbf{t})\mathbf{h}_Z = \mathbf{w}_Z^\top \rho_Z(\mathbf{r})\mathbf{h}_Z = k(\mathbf{r}).$$

□

F.2.1 Discretizing Recovery Integral

By the translation-invariance Lemma (Lemma F.5), the recovery integral (8) takes the form

$$y(\mathbf{t}) = \int_G f(\mathbf{s})\,k(\mathbf{s} - \mathbf{t})\,d\mu(\mathbf{s}), \tag{43}$$

so the recovery operation is a convolution, with the reflected kernel. Since f is band-limited, this convolution can also be discretized exactly, as the following lemma shows.

Lemma F.6 (Exact discretization of the recovery integral)**.** *Following the translation-invariance Lemma (Lemma F.5) and Assumption 3 (support of the measurement operator), let $f \in \mathcal{F}_R$. Because G is represented on the discrete lattice Λ, the recovery integral $y(\mathbf{t}) = \int_G f(\mathbf{s})\, k(\mathbf{s}-\mathbf{t})\, d\mu(\mathbf{s})$ admits an exact discrete form,*

$$y_{\boldsymbol{\ell}} = \frac{1}{|\Lambda|} \sum_{\mathbf{n}\in\Lambda} f_{\mathbf{n}}\, k_{\mathbf{n}-\boldsymbol{\ell}},$$

where $y_{\boldsymbol{\ell}} := y(\mathbf{t}_{\boldsymbol{\ell}})$, $f_{\mathbf{n}} := f(\mathbf{t}_{\mathbf{n}})$, and $k_{\mathbf{n}} := k(\mathbf{t}_{\mathbf{n}})$. In the frequency space of the symmetry representation,

$$Y_{\mathbf{m}} = K_{\mathbf{m}} F_{\mathbf{m}}.$$

Proof. First, let us see that k is a band-limited function

$$k(\mathbf{r}) = \mathbf{w}_Z^\top \rho_Z(\mathbf{r}) \mathbf{h}_Z = \sum_{\mathbf{m}} \mathbf{w}_Z^\top \mathbf{u}_{\mathbf{m}} \mathbf{u}_{\mathbf{m}}^\dagger \mathbf{h}_Z \exp(-\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{r}) = \sum_{\mathbf{m}} w_{Z,\mathbf{m}}^* h_{Z,\mathbf{m}} \exp(-\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{r}). \quad (44)$$

Because it contains only the Fourier modes of the symmetry representation, it is band-limited.

As both f and k are band-limited on the group, we can write them as

$$f(\mathbf{r}) = \sum_{\mathbf{m}} F_{\mathbf{m}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{r}\right) \quad (45)$$

$$k(\mathbf{r}) = \sum_{\mathbf{m}} K_{\mathbf{m}} \exp\left(-\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{r}\right), \qquad K_{\mathbf{m}} := w_{Z,\mathbf{m}}^* h_{Z,\mathbf{m}}. \quad (46)$$

Putting them inside the integral gives

$$y(\mathbf{t}) = \int_G \sum_{\mathbf{m},\boldsymbol{\ell}} F_{\mathbf{m}} K_{\boldsymbol{\ell}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{s}\right) \exp\left(-\mathrm{i}\boldsymbol{\theta}_{\boldsymbol{\ell}} \cdot \mathbf{s}\right) \exp\left(\mathrm{i}\boldsymbol{\theta}_{\boldsymbol{\ell}} \cdot \mathbf{t}\right) d\mu(\mathbf{s}) \quad (47)$$

$$y(\mathbf{t}) = \sum_{\mathbf{m}} F_{\mathbf{m}} K_{\mathbf{m}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{t}\right). \quad (48)$$

Here we used the orthogonality of the Fourier modes,

$$\int_G \exp\big(\mathrm{i}(\boldsymbol{\theta}_{\mathbf{m}} - \boldsymbol{\theta}_{\boldsymbol{\ell}}) \cdot \mathbf{s}\big)\, d\mu(\mathbf{s}) = \delta_{\mathbf{m},\boldsymbol{\ell}},$$

with μ normalized Haar measure. Therefore $Y_{\mathbf{m}} = K_{\mathbf{m}} F_{\mathbf{m}}$.

On the lattice this gives

$$y(\mathbf{t}_{\mathbf{n}}) = \sum_{\mathbf{m}} F_{\mathbf{m}} K_{\mathbf{m}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{t}_{\mathbf{n}}\right).$$

Now, let us take the discrete elements and write the discrete convolution explicitly:

$$y(\mathbf{t}_{\mathbf{n}}) = \frac{1}{|\Lambda|} \sum_{\mathbf{s}\in\Lambda} f(\mathbf{t}_{\mathbf{s}}) k(\mathbf{t}_{\mathbf{s}} - \mathbf{t}_{\mathbf{n}}) \quad (49)$$

$$= \frac{1}{|\Lambda|} \sum_{\mathbf{s}\in\Lambda} \sum_{\mathbf{m}} F_{\mathbf{m}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{t}_{\mathbf{s}}\right) \sum_{\boldsymbol{\ell}} K_{\boldsymbol{\ell}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\boldsymbol{\ell}} \cdot (\mathbf{t}_{\mathbf{n}} - \mathbf{t}_{\mathbf{s}})\right) \quad (50)$$

$$= \sum_{\mathbf{m},\boldsymbol{\ell}} F_{\mathbf{m}} K_{\boldsymbol{\ell}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\boldsymbol{\ell}} \cdot \mathbf{t}_{\mathbf{n}}\right) \frac{1}{|\Lambda|} \sum_{\mathbf{s}\in\Lambda} \exp\left(\mathrm{i}(\boldsymbol{\theta}_{\mathbf{m}} - \boldsymbol{\theta}_{\boldsymbol{\ell}}) \cdot \mathbf{t}_{\mathbf{s}}\right) \quad (51)$$

$$= \sum_{\mathbf{m}} F_{\mathbf{m}} K_{\mathbf{m}} \exp\left(\mathrm{i}\boldsymbol{\theta}_{\mathbf{m}} \cdot \mathbf{t}_{\mathbf{n}}\right) \quad (52)$$

where the last equality uses the discrete orthogonality relation

$$\frac{1}{|\Lambda|} \sum_{\mathbf{s}\in\Lambda} \exp\big(\mathrm{i}(\boldsymbol{\theta}_{\mathbf{m}} - \boldsymbol{\theta}_{\boldsymbol{\ell}}) \cdot \mathbf{t}_{\mathbf{s}}\big) = \delta_{\mathbf{m},\boldsymbol{\ell}}.$$

This relation holds because no two active modes alias on the lattice: by condition (i) of Assumption 3 (support of the measurement operator), $\boldsymbol{\mathcal{B}}$ lies within the Nyquist box, so two distinct active modes differ by less than N_j along every axis j.

Consequently, we see that both the discrete convolution and the continuous convolution evaluated at discrete indices give the same result. □

Corollary F.7 (Spectrum of the discretized recovery kernel)**.** *Following the symmetry-identification Theorem of the stationarity analysis (Theorem D.8), let the learned symmetry representation coincide with the one induced by the latent symmetry, and following the flat-filter Proposition of the information maximization analysis (Proposition C.13), let the recovery filter have a flat spectrum. Then the magnitude spectrum of the discretized recovery kernel is that of the probe response,*

$$|K_{\mathbf{m}}| = \frac{|h_{Z,\mathbf{m}}|}{\sqrt{d_X}}.$$

In particular, by the probe-spectrum Lemma (Lemma F.1) and the mode-energy saturation Proposition (Proposition F.2), $|K_{\mathbf{m}}| = 1/\sqrt{d_X\,\beta_{\mathbf{m}}}$ is set by the mixing matrix.

Proof. By the translation-invariance Lemma (Lemma F.5), the recovery kernel is

$$k(\mathbf{r}) := \mathbf{w}_Z^\top \rho_Z(\mathbf{r}) \mathbf{h}_Z.$$

Constraint 1 (special orthogonal symmetry representation) makes $\rho_Z(\mathbf{r})$ orthogonal, hence normal, so it diagonalizes in an orthonormal eigenbasis $\{\mathbf{u}_\mathbf{m}\}$. By the symmetry-identification Theorem, this is also the eigenbasis of the learned representation, in which the flat-filter Proposition is stated. With the eigenvalues $e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{r}}$,

$$\rho_Z(\mathbf{r}) = \sum_{\mathbf{m}} e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{r}}\, \mathbf{u}_\mathbf{m}\mathbf{u}_\mathbf{m}^\dagger.$$

Writing $w_{Z,\mathbf{m}} := \mathbf{u}_\mathbf{m}^\dagger \mathbf{w}_Z$ and $h_{Z,\mathbf{m}} := \mathbf{u}_\mathbf{m}^\dagger \mathbf{h}_Z$ for the spectral coefficients, and using that $\mathbf{w}_Z$ is real so that $\mathbf{w}_Z^\top \mathbf{u}_\mathbf{m} = w_{Z,\mathbf{m}}^*$, the kernel becomes

$$k(\mathbf{r}) = \sum_{\mathbf{m}} w_{Z,\mathbf{m}}^*\, h_{Z,\mathbf{m}}\, e^{-\mathrm{i}\mathbf{m}\cdot\mathbf{r}}.$$

To find its coefficients, we take the DFT,

$$K_\mathbf{m} = w_{Z,\mathbf{m}}^*\, h_{Z,\mathbf{m}}.$$

This is the coefficient convention of the exact-discretization Lemma (Lemma F.6).

The spectral magnitude of the kernel is therefore

$$K_\mathbf{m} K_\mathbf{m}^* = w_{Z,\mathbf{m}}^* w_{Z,\mathbf{m}}\, h_{Z,\mathbf{m}} h_{Z,\mathbf{m}}^*.$$

From the flat-filter Proposition, $w_{Z,\mathbf{m}}^* w_{Z,\mathbf{m}} = 1/d_X$ on every active mode, giving us

$$|K_\mathbf{m}| = \frac{|h_{Z,\mathbf{m}}|}{\sqrt{d_X}}.$$

Therefore, the spectrum of the discretized recovery kernel is completely determined by the spectrum of the probe response, which the probe-spectrum Lemma and the mode-energy saturation Proposition give as $|h_{Z,\mathbf{m}}| = \alpha_\mathbf{m} = \beta_\mathbf{m}^{-1/2}$. □

F.3 Compactness of the Recovery Kernel

In this section, we will show that total correlation minimization ensures compactness of recovery for first-order MRFs. For higher orders, it acts as a regularizer over high order conditional dependencies.

F.3.1 Total Correlation Minimization as Regularization to High-Order Conditional Dependencies

Lemma F.8 (Total correlation regularizes higher-order dependencies)**.** *Following the exact-discretization Lemma (Lemma F.6), let the recovery act in the frequency space as*

$$\mathbf{Y_m} = K_{\mathbf{m}}\mathbf{F_m},$$

where $\mathbf{Y_m}$*,* $K_{\mathbf{m}}$*, and* $\mathbf{F_m}$ *are the* $\mathbf{m}$*-th Fourier coefficients, and following the kernel-spectrum Corollary (Corollary F.7), let the magnitude* $|K_{\mathbf{m}}|$ *of the recovery kernel be fixed, so that only its phase changes during optimization. If the total-correlation bound of the total correlation minimization analysis (Proposition E.9),*

$$\mathrm{TC}(\mathbf{Y}) \geq \sum_{s=0}^{P-1} \left[\left(1 - \frac{1}{N_{s+1}}\right) h(\mathbf{Y}) - (N_{s+1} - 1)\, h_{\mathrm{agg}}^{\mathrm{cond},(s)}(\mathbf{Y}; 1) + I_{\mathrm{agg}}^{(s)}(\mathbf{Y}; \geq 2) \right],$$

is tight, then minimizing total correlation penalizes the higher-order conditional dependencies $I_{\mathrm{agg}}^{(s)}(\mathbf{Y}; \geq 2)$.

Proof. By the total-correlation bound Proposition of the total correlation minimization analysis (Proposition E.9),

$$\mathrm{TC}(\mathbf{Y}) \geq \sum_{s=0}^{P-1} \left[\left(1 - \frac{1}{N_{s+1}}\right) h(\mathbf{Y}) - (N_{s+1} - 1)\, h_{\mathrm{agg}}^{\mathrm{cond},(s)}(\mathbf{Y}; 1) + I_{\mathrm{agg}}^{(s)}(\mathbf{Y}; \geq 2) \right].$$

From the covariance lower bound Lemma of the total correlation minimization analysis (Lemma E.8), the conditional entropy term is

$$h^{\mathrm{cond}}(\mathbf{Y}_{\mathbf{n}^{(s)}}; 1) = \frac{1}{2} \log \Big((2\pi e)^{d_{s+1}} \det \mathbf{\Sigma}^{(s)}(\mathbf{Y}; 1) \Big),$$
$$\mathbf{\Sigma}^{(s)}(\mathbf{Y}; 1) = \mathbf{C}^{(s)}(\mathbf{Y}; 0) - \mathbf{C}^{(s)}(\mathbf{Y}; 1)\, \mathbf{C}^{(s)}(\mathbf{Y}; 0)^{-1} \mathbf{C}^{(s)}(\mathbf{Y}; 1)^{\top},$$

so it is determined entirely by the cross-covariances of adjacent output slices. We now show that these covariances do not depend on the phase of the kernel, which is the only part of the kernel that the optimization can still change.

Every slice cross-covariance $\mathbf{C}^{(s)}(\mathbf{Y}; r)$ is assembled from values of the output autocovariance $c_Y(\mathbf{j}) := \mathbb{E}[\mathbf{Y_n} \mathbf{Y_{n+j}}]$. Writing the recovery as $\mathbf{Y_n} = \sum_{\mathbf{a}} k_{\mathbf{a}} \mathbf{F_{n-a}}$ on the lattice, with the latent field stationary under the latent symmetry and c_F its autocovariance,

$$c_Y(\mathbf{j}) = \sum_{\mathbf{a},\mathbf{b}} k_{\mathbf{a}} k_{\mathbf{b}}\, c_F(\mathbf{j} + \mathbf{a} - \mathbf{b}) = \sum_{\mathbf{i}} \gamma_k(\mathbf{i})\, c_F(\mathbf{j} - \mathbf{i}), \qquad \gamma_k(\mathbf{i}) := \sum_{\mathbf{a}} k_{\mathbf{a}} k_{\mathbf{a+i}},$$

where the second equality groups the terms by $\mathbf{i} := \mathbf{b} - \mathbf{a}$. The kernel enters only through its autocorrelation γ_k, whose discrete Fourier transform is proportional to $|K_{\mathbf{m}}|^2$. Equivalently, the output power spectrum is $|K_{\mathbf{m}}|^2$ times the latent one. Since the magnitude $|K_{\mathbf{m}}|$ is fixed by hypothesis, every $\mathbf{C}^{(s)}(\mathbf{Y}; r)$, and with them $\mathbf{\Sigma}^{(s)}(\mathbf{Y}; 1)$, is independent of the phase of the kernel. Consequently $h^{\mathrm{cond}}(\mathbf{Y}_{\mathbf{n}^{(s)}}; 1)$ is constant with respect to the optimization objective.

In the special case of a flat spectrum, $|K_{\mathbf{m}}| = K_0$, this reduces to $\mathbf{C}^{(s)}(\mathbf{Y}; r) = K_0^2\, \mathbf{C}^{(s)}(\mathbf{F}; r)$ and $\mathbf{\Sigma}^{(s)}(\mathbf{Y}; 1) = K_0^2\, \mathbf{\Sigma}^{(s)}(\mathbf{F}; 1)$: the prediction error of the output is that of the latent field, scaled by the kernel's overall scale, which the normalization of the recovery filter, Constraint 3 (L_2-normalized recovery filter), fixes.

Thus, we can rewrite the inequality as

$$\mathrm{TC}(\mathbf{Y}) - \sum_{s=0}^{P-1} \left(1 - \frac{1}{N_{s+1}}\right) h(\mathbf{Y}) + \mathrm{const} \geq \sum_{s=0}^{P-1} I_{\mathrm{agg}}^{(s)}(\mathbf{Y}; \geq 2).$$

Here, the left-hand side is minimized through minimizing total correlation, while information maximization is optimal by assumption. Consequently, if the inequality is tight, the aggregated conditional mutual information terms $I_{\mathrm{agg}}^{(s)}(\mathbf{Y}; \geq 2)$ get regularized. □

F.3.2 Conditional Independence Leads to Compact Recovery Kernel

Here, we show how conditional independence enforces the recovery kernel to be compact.

Proposition F.9 (Banded Hessians imply compact kernels)**.** *Under the hypotheses of the TC-regularization Lemma (Lemma F.8), let the recovery kernel have a flat spectrum, $|K_{\mathbf{m}}| = K_0$, and let the conditional mutual information terms beyond order p vanish, as total-correlation minimization encourages by that lemma. Write $E_F := -\log p_F$ and $E_Y := -\log p_Y$ for the energies, i.e. the negative log-densities, of the latent field and the model output, and let $d(i, j)$ be the distance between lattice sites i and j on the periodic lattice, the largest circular index difference over the axes. Assume further that*

- *the latent field is a vector-valued Markov random field of order at most $q \leq p$;*
- *the latent energy Hessians have enough curvature variation that their span contains a matrix with a single non-zero entry on the diagonal: with $D := \prod_{r=1}^{P} N_r$, there is an index r such that*

$$\mathbf{e}_r \mathbf{e}_r^\top \in \operatorname{span}\{\nabla^2 E_F(\mathbf{f}) : \mathbf{f} \in \mathbb{R}^D\}.$$

Then each row of $\mathbf{K}$ has compact support of diameter at most p.

Proof. Let

$$\mathbf{H}_F(\mathbf{f}) := \nabla^2 E_F(\mathbf{f}), \qquad \mathbf{H}_Y(\mathbf{f}) := \nabla^2 E_Y(\mathbf{K}\mathbf{f})$$

denote the latent energy Hessian and the output energy Hessian at $\mathbf{y} = \mathbf{K}\mathbf{f}$. Since $\mathbf{Y} = \mathbf{K}\mathbf{F}$ with $\mathbf{K}$ invertible, $E_F(\mathbf{f}) = E_Y(\mathbf{K}\mathbf{f}) + \text{const}$, and the chain rule gives

$$\mathbf{H}_F(\mathbf{f}) = \mathbf{K}^\top \mathbf{H}_Y(\mathbf{f}) \mathbf{K}.$$

The flat spectrum $|K_{\mathbf{m}}| = K_0$ makes the circulant matrix orthogonal up to scale, $\mathbf{K}^\top \mathbf{K} = K_0^2 \mathbf{I}$, so this is equivalently

$$\mathbf{H}_Y(\mathbf{f}) = K_0^{-4}\, \mathbf{K} \mathbf{H}_F(\mathbf{f}) \mathbf{K}^\top.$$

The latent field is a Markov random field of order at most $q \leq p$, so its energy Hessian is banded:

$$[\mathbf{H}_F(\mathbf{f})]_{ij} = 0 \qquad \text{for } d(i, j) > p.$$

In the vector-valued case, this is understood as the corresponding block-banded condition.

Moreover, the conditional mutual information terms beyond order p vanish, so the output field is also a Markov random field of order at most p. Hence the output energy Hessian is banded as well:

$$[\mathbf{H}_Y(\mathbf{f})]_{ij} = 0 \qquad \text{for } d(i, j) > p.$$

Take the index r of the curvature-variation assumption, so that the single-entry diagonal matrix $\mathbf{e}_r \mathbf{e}_r^\top$ is a linear combination of latent Hessians,

$$\mathbf{e}_r \mathbf{e}_r^\top = \sum_i c_i\, \mathbf{H}_F(\mathbf{f}_i).$$

Applying the linear map $\mathbf{X} \mapsto K_0^{-4} \mathbf{K}\mathbf{X}\mathbf{K}^\top$ to both sides turns each latent Hessian into the corresponding output Hessian:

$$\sum_i c_i\, \mathbf{H}_Y(\mathbf{f}_i) = K_0^{-4}\, (\mathbf{K}\mathbf{e}_r)(\mathbf{K}\mathbf{e}_r)^\top.$$

The left-hand side is a linear combination of banded matrices, hence banded, so the right-hand side vanishes outside the band:

$$K_{ir} K_{jr} = 0 \qquad \text{for all } d(i, j) > p.$$

Therefore, if $K_{ir} \neq 0$ and $K_{jr} \neq 0$, then necessarily $d(i, j) \leq p$. So the support of the r-th column of $\mathbf{K}$ has diameter at most p.

Since $\mathbf{K}$ is circulant, all columns have the same support pattern up to a cyclic shift, and so do all rows; one column therefore suffices. Hence each row of $\mathbf{K}$ has compact support of diameter at most p.

The orthogonality is what turns the conclusion into one about $\mathbf{K}$. Without the flat spectrum, the same argument applies to $\mathbf{H}_Y = \mathbf{K}^{-\top}\mathbf{H}_F\mathbf{K}^{-1}$ and bounds the support of $\mathbf{K}^{-1}$ instead, which does not make $\mathbf{K}$ compact: an inverse with the two taps $(1, a)$ gives a kernel with a geometric tail, although the latent Hessian is diagonal and the output Hessian is banded. □

Remark F.10 (Curvature variation as a Markov generalization of non-Gaussianity)**.** *The assumption that the diagonal entries of $\nabla^2 E_F(\mathbf{f})$ can be varied independently by changing $\mathbf{f}$ corresponds to non-Gaussianity for zeroth-order MRFs, which is the principle behind Independent Component Analysis (ICA). For higher-order Markov fields, the non-Gaussianity condition shifts to possessing local curvature variations. Therefore, this can be viewed as a finite-range Markov generalization of the ICA principle.*

F.4 Signal Recovery

The results so far give the machinery for the signal recovery claim. Signal recovery is possible in two cases:

- the mixing matrix $\mathbf{A}$ is orthogonal;
- the mixing matrix $\mathbf{A}$ is random and sufficiently high dimensional that the spectrum of the probe response $|h_{Z,\mathbf{m}}|$ is almost flat (Lemma F.4).

Below, we will focus on the idealized setting where the spectrum is exactly flat.

Appendix F.3.2 shows that total-correlation minimization encourages compact kernels by minimizing conditional dependencies. When the mixing matrix is orthogonal, $\beta_{\mathbf{m}} = 1$ and hence $|h_{Z,\mathbf{m}}| = 1$ for every mode (Proposition F.2), so the kernel-spectrum Corollary (Corollary F.7) makes the spectrum of the recovery kernel flat. Combined, compactness and flatness force a shifted delta, provided the Markov order is small compared with the lattice, $p + 1 \le N_\ell/2$ along every axis ℓ.

Proposition F.11 (Compact support and flat spectrum imply a delta kernel)**.** *Let the discretized recovery kernel k have a flat spectrum, $|K_{\mathbf{m}}|$ constant across modes, and let its support lie in a box whose side along every axis ℓ is $Q_\ell \le N_\ell/2$. Then k is a shifted and scaled Kronecker delta.*

Proof. The circular autocorrelation of the kernel, $r(\mathbf{j}) := \sum_{\mathbf{n}\in\Lambda} k(\mathbf{s_n})\, k(\mathbf{s_{n+j}})$, has discrete Fourier transform proportional to $|K_{\mathbf{m}}|^2$. A flat spectrum therefore requires

$$r(\mathbf{j}) = 0 \qquad \text{for every lag } \mathbf{j} \neq \mathbf{0}.$$

Suppose the support S of k contains at least two lattice points. Choose a linear functional $\mathbf{n} \mapsto \boldsymbol{\lambda} \cdot \mathbf{n}$, with $\boldsymbol{\lambda}$ generic so that it takes distinct values on S, and let $\mathbf{a}$ and $\mathbf{b}$ be the points of S where it is smallest and largest. Since every side of the support box is at most $N_\ell/2$, differences of support points do not wrap around the cyclic lattice. The lag $\mathbf{b} - \mathbf{a}$ is then realized by exactly one pair of support points: any other pair $(\mathbf{n}, \mathbf{n} + \mathbf{b} - \mathbf{a})$ in S with $\mathbf{n} \neq \mathbf{a}$ would have $\boldsymbol{\lambda} \cdot \mathbf{n} > \boldsymbol{\lambda} \cdot \mathbf{a}$, hence $\boldsymbol{\lambda} \cdot (\mathbf{n} + \mathbf{b} - \mathbf{a}) > \boldsymbol{\lambda} \cdot \mathbf{b}$, contradicting the maximality of $\mathbf{b}$. Therefore

$$r(\mathbf{b} - \mathbf{a}) = k(\mathbf{s_a})\, k(\mathbf{s_b}) \neq 0,$$

which contradicts the flat spectrum. In one dimension, $\mathbf{a}$ and $\mathbf{b}$ are the two ends of the support and this is the extreme-lag argument; the generic functional extends it to several axes, where a single axis alone would not suffice, since the products along one axis could cancel across the other axes.

Hence S is a single point, and k is a shifted and scaled Kronecker delta. □

IMPLEMENTATION AND EXPERIMENTAL DETAILS

G IMPLEMENTATION DETAILS

This appendix describes our model as implemented, together with the settings shared by all experiments. Settings that differ between experiments are listed in Table 6 (Appendix H.2).

G.1 RECOVERY NETWORK

Input normalization. Before the conditioning matrix, samples are centered and scaled with a mean and a variance shared by all coordinates, estimated over the batch during training and by running averages afterwards. This enforces Constraint 7 (centered variables), and sharing the statistics across coordinates keeps the normalization independent of the coordinate order.

Conditioning matrix. We choose the conditioning matrix as follows:

- Where the symmetry acts periodically and orthogonally on the sample space, we use the identity, $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}}$. This is the case in the symmetry-discovery benchmarks and the neural recordings.
- For signal-recovery benchmarks with open boundaries and an orthogonal mixing matrix, we use a zero-padding operator.
- For signal-recovery benchmarks with a random mixing matrix, we use a learnable linear operator into a space of higher dimension than the sample space, $d_Z > d_X$.
- For bit scrambling, we use a learnable linear operator into a space of lower dimension than the sample space, $d_Z < d_X$.

Zero padding. The zero-padding operator appends $d_Z - d_X$ zeros to the sample, split evenly between its two ends. Zero padding and the dimension-increasing learnable conditioning matrix serve two purposes: they relieve the gradient stiffness of optimizing on orthogonal (Stiefel) manifolds Edelman et al. (1998); Absil et al. (2008), and they give the model enough capacity to accommodate open boundary conditions.

Reconditioning of the conditioning matrix. After every gradient update, the learned conditioning matrix is reconditioned so that the energy it places on each eigenmode of the learned generators stays bounded. With $\mathbf{u}_j$ the columns of $\mathbf{V}$ (Appendix G.1) and $e_j := \mathbf{u}_j^\dagger \mathbf{C}\mathbf{C}^\top \mathbf{u}_j$ the energy along $\mathbf{u}_j$, we set

$$\mathbf{C} \leftarrow \mathbf{V}\,\mathrm{diag}(\mathbf{s})\,\mathbf{V}^\dagger \mathbf{C}, \qquad s_j = \sqrt{\frac{g(e_j)}{e_j}}, \qquad g(e) = \begin{cases} e, & e \le 1, \\ 1 + 0.2\left(1 - e^{-(e-1)/0.2}\right), & e > 1, \end{cases}$$

so that energies up to one are left unchanged and larger ones are softly capped below 1.2.

Efficient and stable lifting. To implement the lifting stably and efficiently, we use the basis transformation $\mathbf{V}$ that simultaneously diagonalizes the learned generators $\hat{\mathbf{L}}_i$. For a conditioned sample $\mathbf{z} = \mathbf{C}\mathbf{x}$, the lifted output is

$$\begin{aligned} y_{\mathbf{n}} &= \mathbf{w}_Z^\top \exp\left(-\sum_{i=1}^{P} t_{n_i} \hat{\mathbf{L}}_i\right) \mathbf{z} \\ &= \mathbf{w}_Z^\top \mathbf{V} \exp\left(-\sum_{i=1}^{P} t_{n_i} \mathbf{D}_i\right) \mathbf{V}^\dagger \mathbf{z} \\ &= \tilde{\mathbf{w}}_Z^\dagger \exp\left(-\sum_{i=1}^{P} t_{n_i} \mathbf{D}_i\right) \tilde{\mathbf{z}} \end{aligned}$$

where we defined $\tilde{\mathbf{w}}_Z := \mathbf{V}^\dagger \mathbf{w}_Z$ and $\tilde{\mathbf{z}} := \mathbf{V}^\dagger \mathbf{z}$ in the last expression, and $\mathbf{D}_i \in \mathbb{C}^{d_Z \times d_Z}$ are the diagonal eigenvalue matrices of $\hat{\mathbf{L}}_i$. We use this spectral implementation in all experiments.

Parametrization of Lie basis. For Abelian Lie groups, the Lie basis elements must share their eigenvectors. To ensure this, we parametrize all eigenvectors with a single learnable antisymmetric matrix $\mathbf{\Gamma} \in \mathbb{R}^{d_Z \times d_Z}$ and obtain its eigenvectors $\mathbf{V} \in \mathbb{C}^{d_Z \times d_Z}$, with columns $\mathbf{u}_j$, by diagonalization.

The first Lie basis element is parameterized by the eigenvalues of $\mathbf{\Gamma}$, while the others are parameterized separately with real vectors, $\{\boldsymbol{\alpha}_i \in \mathbb{R}^{d_Z} : i = 2, \ldots, P\}$. The eigenphases $\{\boldsymbol{\phi}_i\}$ of each basis element are obtained from the parameters $\{\boldsymbol{\alpha}_i\}$.

To ensure that each Lie basis element $\hat{\mathbf{L}}_i := \mathbf{V}\,\mathrm{diag}(\mathrm{i}\boldsymbol{\phi}_i)\,\mathbf{V}^\dagger$ is a real antisymmetric matrix, we align parametrizations of each Lie basis with the eigenvectors of $\mathbf{\Gamma}$. Specifically, we define the permutation matrix $\mathbf{P}_{ij} := \mathbf{u}_i^\top \mathbf{u}_j$, which maps each component j to the index i satisfying $\mathbf{u}_i = \mathbf{u}_j^*$. The conjugate-paired eigenphases are then

$$\boldsymbol{\phi}_i = \boldsymbol{\alpha}_i - \mathbf{P}\boldsymbol{\alpha}_i, \tag{53}$$

where $\boldsymbol{\alpha}_i$ is the parameter vector of the i-th Lie basis element.

Bounding the eigenphases. We constrain the eigenphases to $(-\pi, \pi)$, so that no mode advances by more than half a turn between two adjacent grid points. Since the grid step along axis j is $2\pi/N_j$, a mode of angular frequency $\hat{k}^{(j)}$ advances $2\pi\hat{k}^{(j)}/N_j$ per step, and the constraint reads $|\hat{k}^{(j)}| < N_j/2$. This is the no-aliasing normalization recorded as Constraint 2: frequencies differing by a multiple of N_j are indistinguishable from the grid alone, so bounding the eigenphases selects one representative per class rather than restricting the model.

Sampling lattice. The sampling lattice is centered on the origin and shifted along each axis by a learnable offset. Along axis i, the lattice coordinates of Section 4.1 are translated by $\frac{2\pi}{N_i}\left(o_i - \frac{N_i-1}{2}\right)$, where $o_i = \frac{N_i-1}{2}\tanh(2\omega_i)$ and the parameter ω_i is initialized at zero.

Initialization. The antisymmetric matrix $\mathbf{\Gamma}$ is the antisymmetric part of a parameter matrix whose entries are initialized from a normal distribution with $\mu = 0$ and $\sigma = 10^{-3}$. We initialize the eigenphase parameters $\boldsymbol{\alpha}_i$ from a normal distribution with $\mu = 0$ and $\sigma = 10^{-3}$. Smaller standard deviations did not affect performance, whereas substantially larger ones can derail the optimization.

We initialize the parameters $\boldsymbol{\psi}$ of the recovery filter with zeros and, early in training, add Gaussian noise to them before computing the loss of each batch,

$$\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} + \mathcal{N}(0, 0.1^2)\, e^{-t/\tau}, \tag{54}$$

where t is the training time in epochs. The noise amplitude decays exponentially with the short time constant $\tau = 10$ epochs. To compute the model output, the recovery filter is normalized at every step, $\mathbf{w}_Z = \boldsymbol{\psi}/\|\boldsymbol{\psi}\|_2$, as Constraint 3 ($L_2$-normalized recovery filter) requires.

G.2 Entropy Estimation

Component-wise entropy estimation. We estimate component-wise entropies using trainable Gaussian mixture models for each component, following Pichler et al. (2022). This approach models the density $\hat{p}_{\mathbf{n}}$ of each component of the model output as

$$\hat{p}_{\mathbf{n}}(y_{\mathbf{n}}) = \sum_{m=1}^{M} \frac{w_{\mathbf{n};m}}{\sigma_{\mathbf{n};m}\sqrt{2\pi}} \exp\left(-\frac{(y_{\mathbf{n}} - \mu_{\mathbf{n};m})^2}{2\sigma_{\mathbf{n};m}^2}\right) \tag{55}$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_P)$, with mixture weights satisfying $\sum_m w_{\mathbf{n};m} = 1$ and $w_{\mathbf{n};m} \geq 0$. We use $M = 4$ Gaussian components per mixture. The means are initialized evenly on $[-2.5, 2.5]$, the log-variances at 1, and the weight logits at a common value, so the initial weights are equal. Before use, the log-variances are clipped to $[-2.5, 2.5]$ and the weight logits to $[-3.5, 3.5]$.

The component-wise entropy $\hat{h}(Y_{\mathbf{n}})$ is computed via:

$$\hat{h}(Y_{\mathbf{n}}) = -\mathbb{E}\left[\log \hat{p}_{\mathbf{n}}(y_{\mathbf{n}})\right]$$

where the expectation is taken over a batch. For training probability estimators, we minimize the average expected entropy $\frac{1}{d_Y}\sum_{\mathbf{n}} \hat{h}(Y_{\mathbf{n}})$.

Joint entropy estimation. To estimate the joint differential entropy of the model output $\mathbf{Y}$, we use a multivariate Gaussian approximation. For each batch, we flatten the model output, compute its sample covariance matrix $[\mathbf{\Sigma}_Y]_{pr} = \mathrm{Cov}(y_{\mathbf{n}(p)}, y_{\mathbf{n}(r)})$, where p and r are flattened indices, load it with $\epsilon\mathbf{I}$, and diagonalize it with a symmetric eigendecomposition. Its eigenvalues, sorted in descending order, clipped at zero, and raised by a floor of $\epsilon = 10^{-5}$, give the Gaussian entropy $\frac{1}{2}(\log(2\pi\lambda_l) + 1)$ of each mode. The rank-k approximation of the entropy, a weighted average of the per-mode contributions, is

$$\bar{h}_k := \frac{\sum_{l=1}^{d_Y} w_{kl} \, \frac{1}{2}(\log(2\pi\lambda_l) + 1)}{\sum_{l=1}^{d_Y} w_{kl}}, \tag{56}$$

where the weights provide a soft threshold at $l = k$, a smooth version of the low-rank entropy of Definition 1 (low-rank entropy). We compute them with a sigmoid on the normalized eigenvalue index,

$$w_{kl} = \frac{1}{e^{\beta(l-k)/(d_Y-1)} + 1}, \tag{57}$$

with sharpness $\beta = 75$ in units of the normalized index $(l-1)/(d_Y-1) \in [0, 1]$.

Due to the normalization, (56) should be thought of as a per-rank version of the low-rank entropy. In particular, when $\bar{h}_k$ is combined with the component-wise entropies $\hat{h}(Y_\mathbf{n})$ in the total correlation loss, it is combined with their average rather than their sum.

The derivative of (56) with respect to λ_l is proportional to w_{kl}/λ_l. It diverges for vanishing eigenvalues, so the weakest modes would dominate the update while carrying almost no variance. In the backward pass we therefore evaluate it at $\max(\lambda_l, \lambda_1/\kappa)$, where λ_1 is the largest eigenvalue. The forward pass is unchanged. Since the weights w_{kl} do not increase with l, the gradient with respect to any eigenvalue is then at most κ times that with respect to λ_1, so κ bounds the condition number seen by the gradient: eigenvalues above λ_1/κ receive their exact gradient, and smaller ones are treated as if they were equal to λ_1/κ. The bound κ is a fixed hyperparameter, not tuned per run: $\kappa = 100$ in all experiments except the Ising chain with random-matrix mixing, where $\kappa = 1000$.

Controlling the rank of the joint entropy estimator. To facilitate optimization by first fitting the gross features of the data and refining them over time, we use a time-dependent rank k in the rank weights (57). To adjust k during training, we use a normalized training time $\bar{t}_n$ that measures the amount of gradient flow,

$$\bar{t}_n = \frac{\sum_{s=1}^{n} \eta(s)}{\sum_{s=1}^{T} \eta(s)}, \tag{58}$$

where $\eta(s)$ is the learning rate at training step (batch) s, and n and T are the current and the total number of training steps. We place the threshold at $k = 1 + \bar{t}_n(d_Y - 1)$, so that the rank sweeps the whole spectrum of the output covariance during training and reaches d_Y at its end. In all our experiments except bit scrambling and the neural recordings, the lattice has $d_Y = d_X$ sites (Table 6), so this is the terminal rank of Constraint 6 (rank schedule); with zero padding, $d_Z > d_X$, and the rank of the model output is still at most d_X. In bit scrambling, the learned conditioning matrix reduces the dimension, $d_Z < d_X$, and the lattice has $d_Y < d_X$ sites, so the terminal rank is d_Y. For the neural recordings, $d_Y = 6 < d_X$, and the joint entropy is computed in the subspace of the DC and fundamental harmonics with terminal rank $k = 3$ (Section 4.2). Both experiments therefore lie outside Constraint 4 (dimensions), and bit scrambling also outside Constraint 8 (batch size), as they lie outside the setting of our analysis: bit scrambling is nonlinear in the pixel intensities, and the neural recordings follow the single-orbit setting. The learning rate in $\bar{t}_n$ is that of the recovery network.

G.3 JS-Divergence Estimation

G.3.1 Overview

We probe shift invariance by estimating the Jensen–Shannon (JS) divergence between the distribution of the model output and its shifted versions along each axis.

To estimate a JS divergence, we first estimate the two KL divergences it is built from, for each group axis, using the dual representation (59), and then combine them as

$$D_{\mathrm{JS}}\big(\mathbf{Y}, \mathbf{Y}^{(\ell)}\big) = \tfrac{1}{2} D_{KL}\big(P_Y \,||\, M_\ell\big) + \tfrac{1}{2} D_{KL}\big(P_{Y^{(\ell)}} \,||\, M_\ell\big), \qquad M_\ell := \tfrac{1}{2}\big(P_Y + P_{Y^{(\ell)}}\big),$$

where $\mathbf{Y}^{(\ell)}$ is the model output shifted along axis ℓ (Section 4.3). This requires two networks per group axis, and it turns KL-divergence estimation into an efficient downstream task.

We use CNNs for their stability and convergence speed, with position embeddings to improve their expressive capacity. In some settings, we also coarse-grain the model output by stochastic pooling and estimate the JS divergence at two scales; the procedure is given in Appendix G.4.

G.3.2 Building Blocks: KL-Divergence Estimators

The stationarity loss is built from KL divergences between pairs of distributions. To estimate KL divergence, we use the duality formula Sreekumar & Goldfeld (2022); Donsker & Varadhan (1975) giving the KL divergence $D_{KL}(R||S)$ between two probability distributions R, S as a solution to the optimization problem

$$D_{KL}(R||S) = \sup_{\theta \in \Theta} \big\{ \mathbb{E}_{\mathbf{x}\sim R}[f_\theta(\mathbf{x})] \;-\; \log \mathbb{E}_{\mathbf{x}\sim S}[\exp f_\theta(\mathbf{x})] \big\} \tag{59}$$

where $f_\theta : \mathbb{R}^{d_Y} \to \mathbb{R}$ and θ ranges over a family rich enough to contain all measurable functions Donsker & Varadhan (1975).

We approximate the search over all f_θ using a neural network, with θ describing the network parameters, and using (59) to define the loss. To approximate the expectation values in (59), we use averaging over batches. For the network architecture, we use a CNN with additional learnable position embeddings (Appendix G.3.2), which turns out to work rather efficiently.

Since each JS divergence involves two KL-divergence estimates, we use two identical networks with separate weights for each term in the loss, and train these networks with the loss estimation network update of Figure 2(b).

Position embeddings. The dual representation of the KL divergence requires solving the optimization problem (59) over all functions $f : \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_P} \to \mathbb{R}$. In principle, deep multilayer perceptrons (MLPs) appear suitable for this task as they lack inductive biases. However, our experiments reveal that MLPs exhibit slow convergence, inducing instabilities in our method.

While Convolutional Neural Networks (CNNs) converge faster and more stably than MLPs, they can only represent translation-equivariant functions. To improve expressive capacity while preserving convergence efficiency, we introduce learnable position embeddings. The augmented model output becomes

$$y^{\text{embed}}_{n_1,n_2,\ldots,n_P;e} := y_{n_1,n_2,\ldots,n_P} + p_{n_1,n_2,\ldots,n_P;e}, \tag{60}$$

where $p \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_P \times d_e}$ denotes the learnable position embedding tensor and $d_e = 4$ is the embedding dimension used throughout our experiments.

G.3.3 Estimator Details

One-component shifts. Instead of circular shifts, the estimators compare cropped patches offset by one component, in order to conform with open boundaries. Along each axis ℓ, the estimator compares the model output restricted to $n_\ell \in \{0, \ldots, N_\ell - 2\}$ with the output restricted to $n_\ell \in \{1, \ldots, N_\ell - 1\}$, a one-component shift without wrapping around the boundary.

Mixture samples and estimates. Samples of the mixture M_ℓ are formed from the first half of a batch of $\mathbf{Y}$ and the second half of a batch of $\mathbf{Y}^{(\ell)}$ (see Appendix G.3.1 for definitions). Each of the two KL divergences that make up the JS divergence has its own network and position embedding. For each network, we maximize the objective (59) in both directions between the two distributions and use the larger value, clipped at zero. Each KL-divergence estimator network ends with global average pooling over the grid, and each of its output channels acts as a separate function f_θ, whose estimates are averaged. The network outputs are soft-clipped with $\gamma \tanh(f_\theta/\gamma)$ and the position embeddings as $\zeta \tanh(p/\zeta)$ where typically we set $\gamma = 25$, $\zeta = 5$.

Architecture. Each KL-divergence estimator is a convolutional network over the output grid. The single-channel model output is repeated to $d_e = 4$ channels, and a learnable position embedding is added. For the 27×27 MNIST crops, a learnable input embedding layer precedes the network (Table 7); like the position embeddings, it lets the translation-equivariant CNN adapt to settings that violate shift invariance. The network then applies a sequence of blocks, each consisting of a convolution with zero padding, a nonlinear activation (except the final layer), and a downsampling step, either a stride or stochastic pooling (Zeiler & Fergus, 2013). We use strides in the first layers, which are the most compute-intensive, and stochastic pooling in the remaining ones to avoid overfitting. At the last layer we apply global average pooling over the grid. The spectral norm of each layer is capped at $10^{(1/L)}$ for a network of L layers. Table 7 lists the layers of each estimator.

G.4 Output Coarse-Graining and Multiscale Losses

Experiments indicate that a lack of smoothness, as in the Ising data, or a high dimensionality along an axis, as in the 27×27 MNIST crops, can cause optimization difficulties, such as the optimizer becoming trapped in local minima. To mitigate this issue, we introduce a coarse-grained version of the model output and an additional JS-divergence estimator that runs on it.

The coarse-graining procedure consists of two steps:

1. Partition the model output into patches along each axis.

2. Randomly select one component within each patch.

This operation is formalized in (61). For each sample, we independently draw random integer offsets r_i uniformly from $\{0, 1, \ldots, q_i - 1\}$, where q_i is the dilation factor along axis $i = 1, \ldots, P$. The indices of the coarse-grained output y^{coarse} range over $\{0, 1, \ldots, \lfloor N_1/q_1 \rfloor - 1\} \times \cdots \times \{0, 1, \ldots, \lfloor N_P/q_P \rfloor - 1\}$.

$$y^{\text{coarse}}_{n_1,\ldots,n_P} := y_{n_1 q_1 + r_1,\, \ldots,\, n_P q_P + r_P} \tag{61}$$

Multiscale losses. At the coarse grained scale, we apply identical loss with identical loss coefficients. Before estimation, the coarse-grained output is rescaled by its root-mean-square standard deviation $\bar{\sigma}$ to unit scale, computed as $\mathbf{y}\,\big(1 + \epsilon/\operatorname{sg}(\bar{\sigma})\big)/(\bar{\sigma} + \epsilon)$, where $\epsilon = 0.1$. The operator sg stops the gradient. The value equals $\mathbf{y}/\bar{\sigma}$, while the offset ϵ damps the gradient through $\bar{\sigma}$.

G.4.1 Weighting JS-Divergence Estimates Across Different Scales

Scaling properties of the JS divergence. To quantify shift invariance, we estimate the JS divergence between the distribution of the model output and its shifted versions along each axis. When the axes have substantially different dimensionalities, averaging these divergences without accounting for dimensionality can degrade performance. Throughout this subsection, we write $p := P_Y$ for the distribution of the model output and $p_\ell := P_{Y^{(\ell)}}$ for the distribution of its one-component shift along axis ℓ, so that $\mathrm{JS}(p \parallel p_\ell) = D_{\mathrm{JS}}\big(\mathbf{Y}, \mathbf{Y}^{(\ell)}\big)$, and $h(p)$ denotes the differential entropy of a random vector with distribution p.

To address this, we weight JS divergence estimates based on their scaling behavior with respect to dimensionality. We assume the model output is the discrete version of a rectangular region of the signal domain. Under this assumption, increasing the dimensionality along an axis corresponds to increasing resolution, allowing us to estimate the scaling behavior of JS divergence.

Formally, we can parametrize p with real parameters θ such that a small shift in model output corresponds to a perturbation $\delta\theta$ in parameter space. Expanding the JS divergence in a Taylor series

to second order at θ_0 we have

$$\mathrm{JS}\big(p(\theta_0) \parallel p(\theta_0 + \delta\theta)\big) = \mathrm{JS}\big(p(\theta_0) \parallel p(\theta_0)\big) \tag{62}$$

$$+ \sum_i \frac{\partial}{\partial \theta_i} \mathrm{JS}\big(p(\theta_0) \parallel p(\theta)\big)\Big|_{\theta=\theta_0} \delta\theta_i \tag{63}$$

$$+ \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \mathrm{JS}\big(p(\theta_0) \parallel p(\theta)\big)\Big|_{\theta=\theta_0} \delta\theta_i \delta\theta_j \tag{64}$$

$$+ \mathcal{O}(\|\delta\theta\|^3). \tag{65}$$

Since $\mathrm{JS}\big(p(\theta_0) \parallel p(\theta_0)\big) = 0$ and the first derivative vanishes at the global minimum $\theta = \theta_0$, we obtain the second-order approximation:

$$\mathrm{JS}\big(p(\theta_0) \parallel p(\theta_0 + \delta\theta)\big) \approx \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \mathrm{JS}\big(p(\theta_0) \parallel p(\theta)\big)\Big|_{\theta=\theta_0} \delta\theta_i \delta\theta_j \tag{66}$$

Under (66), scaling the shift magnitude by $s > 0$ (which scales $\delta\theta \to s\delta\theta$ for small continuous shifts) yields:

$$\mathrm{JS}\big(p(\theta_0) \parallel p(\theta_0 + s\delta\theta)\big) \approx s^2 \cdot \mathrm{JS}\big(p(\theta_0) \parallel p(\theta_0 + \delta\theta)\big), \tag{67}$$

demonstrating quadratic scaling of JS divergence with shift magnitude.

Since we estimate $\mathrm{JS}(p \parallel p_\ell)$ for each axis ℓ where p_ℓ is the one-component shifted version of p along the ℓ-th axis, and we assume that increasing dimensionality merely changes the resolution, magnitude of one-component shift should be proportional to $1/N_\ell$. Then (67) gives that $\mathrm{JS}(p \parallel p_\ell) \propto 1/N_\ell^2$ while everything else is kept the same.

To stay consistent with the per-dimension entropies of the total correlation minimization and information maximization losses, we further divide the JS divergences by the total dimension $d_Y = \prod_{i=1}^{P} N_i$. This ensures that each loss term has a similar scale, because the JS divergence is itself a difference of entropies. Therefore

$$\mathrm{JS}(p \parallel p_\ell) = h\left(\frac{p + p_\ell}{2}\right) - \frac{h(p) + h(p_\ell)}{2} = h\left(\frac{p + p_\ell}{2}\right) - h(p), \tag{68}$$

and $\frac{1}{d_Y}\mathrm{JS}(p \parallel p_\ell)$ is a difference of per-dimension entropies.

Thus, we assign the weight $w_\ell := N_\ell^2/d_Y$ to the JS divergence along axis ℓ. For coarse-grained representations, we replace N_ℓ with N_ℓ/q_ℓ and d_Y with $\prod_{i=1}^{P} N_i/q_i$, the dimensions after coarse-graining. This scheme ensures consistent hyperparameter settings across experiments when the stated assumptions hold.

Consequently, we rescale each JS-divergence term as $D_{\mathrm{JS}}\big(\mathbf{Y}, \mathbf{Y}^{(\ell)}\big) \to w_\ell\, D_{\mathrm{JS}}\big(\mathbf{Y}, \mathbf{Y}^{(\ell)}\big)$. Since the estimators compare cropped outputs (Appendix G.3.3), the estimate along axis ℓ is in practice multiplied by N_ℓ^2 and divided by the number $d_Y(N_\ell - 1)/N_\ell$ of compared components, which is this weighting applied to the cropped comparison. The stationarity loss averages these weighted estimates over the axes.

G.5 Optimization

The recovery and loss estimation networks are trained with simultaneous updates (Section 4). Each step makes one forward pass through both networks, computes the gradients of both from the same parameters, and then applies them. Both networks use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$, and gradients are clipped to a global norm of 0.5. The learning rates decay geometrically, once per epoch, from the initial values of Table 6 to a tenth of them, and the normalized training time $\bar{t}_n$ of the rank schedule is computed from the learning rate of the recovery network.

Each training set is generated once, and every epoch is one pass over it in a new order, drawn with a shuffle buffer of $20{,}000$ samples. An epoch therefore consists of as many update steps as the training set has batches, 100 in all experiments except the 27×27 MNIST runs (Table 6).

H Experimental Details

H.1 Datasets

H.1.1 Generalized Shot Noise Process

Each sample is generated by Algorithm 1, following Rice (1944). The grid covers $[-(N-1)/2, (N-1)/2]$, while pulse centers are drawn from $[-N, N)$, beyond the grid on both sides, so that boundary effects do not distort translation invariance. The 15-dimensional datasets instead use periodic boundaries, and their translation regimes modify how centers are drawn and applied (Appendix H.1.1).

Algorithm 1 Generalized shot noise sampling

Require: pulse families $h^{(1)}, \dots, h^{(F)}$ with parameter sets $\Theta_1, \dots, \Theta_F$, $C = 5$ candidate pulses per family, lattice size N, grid points $t_i = i - (N+1)/2$ for $i = 1, \dots, N$, relative noise level $\nu = 0.05$
Ensure: sample $\mathbf{x} \in \mathbb{R}^N$
1: $\mathbf{x} \leftarrow \mathbf{0}$
2: **for** $f = 1, \dots, F$ **do**
3: **for** $c = 1, \dots, C$ **do**
4: With probability $1/(2F)$, draw $\theta \sim \mathcal{U}(\Theta_f)$ and a center $s \sim \mathcal{U}[-N, N)$, and set $x_i \leftarrow x_i + h_\theta^{(f)}(t_i - s)$ for all i
5: **end for**
6: **end for**
7: Draw $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and set $\mathbf{x} \leftarrow \mathbf{x} + \nu\hat{\sigma}\,\boldsymbol{\varepsilon}$, where $\hat{\sigma}$ is the standard deviation of the noiseless batch
8: **return** $\mathbf{x}$

Data-generation procedure. Table 5 lists the generation parameters of all one-dimensional datasets.

Gaussian pulses. Gaussian pulses are unit-area Gaussians scaled by the amplitude A,

$$h_{\text{gauss}}(n; A, \mu, \sigma) = \frac{A}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(n-\mu)^2}{2\sigma^2}\right),$$

evaluated at the integer grid points n from $-(N-1)/2$ to $(N-1)/2$, where N is the lattice size. Since the centers μ are drawn from $[-N, N)$, beyond the grid on both sides, a sample may contain only the tail of a pulse.

Legendre pulses. Legendre pulses are built from the associated Legendre functions and produce localized waveforms that can change sign. With center c, scale s, amplitude A, and orders (l, m),

$$h_{\text{leg}}^{(l,m)}(n; A, c, s) = \begin{cases} A\,P_l^m\left(\sin\frac{\pi(n-c)}{s}\right), & |n-c| \le s/2, \\ 0, & \text{otherwise}, \end{cases}$$

so that each pulse spans s grid units and vanishes at its ends. Every GSN-L sample superposes pulses of both orders, $(l, m) = (2, 1)$ and $(3, 1)$, which act as two pulse families in Algorithm 1. Centers are drawn as for the Gaussian pulses, and the remaining parameters from the ranges in Table 5.

Datasets and translation regimes. We refer to the datasets with Gaussian and Legendre bases as GSN-G and GSN-L. The 63-dimensional datasets are observed either directly (identity) or through a discrete sine transform (DST-I). The symmetry-discovery comparison with LieGAN uses 15-dimensional GSN-G signals with periodic boundary conditions, under the three translation regimes of Section 6.1. In the *band-limited continuous* regime, each pulse is discretized at the origin and then shifted cyclically by a continuous amount with an FFT-based translation operator, so that the translation acts exactly as a linear operator on the grid. In the *continuous* regime, each pulse is evaluated directly at a continuous center, with distances wrapped around the period before discretization. The *discrete* regime is the continuous one restricted to integer centers.

H.1.2 ISING MODEL

To generate spin signals, we initialize a chain of $3N$ spins uniformly at random in $\{-1, +1\}$ and perform 10 sequential heat-bath (Gibbs) sweeps with nearest-neighbor coupling and open boundaries, at an inverse temperature β drawn from $\mathcal{U}[1.0, 2.5)$ independently for each sample. We keep the central $N = 33$ spins, which removes the boundary effects of the finite chain. No noise is added.

After generating the stochastic samples following this procedure, we obscure the domain either by a fixed random permutation of the nodes or by a fixed random mixing matrix $\mathbf{A}$, both drawn once per seed. The linear transformation acts as an unknown, imperfect sensor response, effectively turning the task into a blind inverse problem.

Singular value distribution of the mixing matrix. To construct $\mathbf{A}$, we draw a 33×33 matrix with independent standard normal entries and keep its left and right singular vectors. Its singular values are mapped affinely onto $[0.75, 1.33]$, so that the smallest becomes 0.75 and the largest 1.33, which fixes the condition number of $\mathbf{A}$ at $1.33/0.75 \approx 1.77$. Since the map is affine, the rescaled singular values retain the distribution of the singular values of the Gaussian matrix and are not evenly spaced: they are skewed towards the lower end of the interval. Across the five seeds, 55–64% of them lie in the lower half of the interval, and their median lies between 0.97 and 1.01, below the midpoint 1.04.

Table 5: **Synthetic datasets.** Generation parameters of the one-dimensional datasets. Uniform ranges are half-open. Each of the F pulse families contributes five candidate pulses, each present independently with probability $1/(2F)$ (Algorithm 1). Noise is Gaussian with standard deviation 0.05 times that of the noiseless batch, and is added before mixing. The centers of the periodic datasets depend on the translation regime (Appendix H.1.1).

	GSN-G	GSN-L	GSN-G (periodic)	Ising
Lattice size N	63	63	15	33
Boundary	open	open	periodic	open
Pulse	Gaussian	P_2^1, P_3^1	Gaussian	—
Families F	1	2	1	—
Amplitude A	$[0.5, 1.5)$	$[0.5, 1.5)$	$[0.5, 1.5)$	—
Scale	$\sigma \in [0.5, 2.5)$	$s \in [6, 15)$	$\sigma \in [0.5, 2.5)$	—
Centers	$[-N, N)$	$[-N, N)$	regime-dependent	—
Inverse temperature β	—	—	—	$[1.0, 2.5)$
Gibbs sweeps	—	—	—	10
Noise level ν	0.05	0.05	0.05	none
Mixing	identity DST-I	identity DST-I	identity	permutation random matrix

H.1.3 MNIST PIXEL SHUFFLING

Samples are generated from the 70,000 MNIST images of the training and test sets, drawn uniformly with replacement. To remove the positional bias of the centered digits, each image is zero-padded by H pixels on every side, where $H \in \{15, 27\}$ is the crop size, and an $H \times H$ crop is taken at a uniformly random, continuous position inside the padded image. A crop can therefore cover only part of a digit, and crops without any nonzero pixel are discarded. Since the crop position is continuous, pixel values are obtained by bilinear interpolation, at a sampling pitch of $\frac{H}{28+2H} \cdot \frac{27+2H}{H-1}$ original pixels, which is 1.05 for $H = 15$ and 1.03 for $H = 27$. Intensities are kept on their original scale $[0, 255]$, and no noise is added.

Each crop is flattened in row-major order into a vector of dimension H^2, and a fixed random permutation, drawn once per seed, is applied to the coordinates of every sample. The observations are thus 225- and 729-dimensional vectors whose coordinate order carries no information about the pixel layout.

H.1.4 MNIST BIT SCRAMBLING

The bit-scrambling datasets start from the 27×27 crops of Appendix H.1.3. Each interpolated intensity is truncated to an 8-bit integer $v \in \{0, \dots, 255\}$ and replaced by its binary expansion, most significant bit first,

$$v = \sum_{b=0}^{7} 2^{7-b}\, \beta_b, \qquad \beta_b \in \{0, 1\},$$

so each crop becomes a binary tensor of shape $H \times H \times 8$. The tensor is flattened, and a fixed random permutation, drawn once per seed, is applied to all $8H^2$ bits jointly. This hides the order of the bits within a pixel together with the pixel layout, a bit-level scrambling similar to Ye (2010).

H.1.5 NEURAL CODING

To isolate orientation geometry from confounding factors such as contrast and speed, we restrict our analysis to a fixed spatial frequency (0.04 cpd), which leaves $1{,}162$ trials. For each trial and neuron, we count the spikes in a $250\,\mathrm{ms}$ window starting $80\,\mathrm{ms}$ after stimulus onset, to account for cortical processing time, and convert the count to the Gaussian quantile of its Poisson cumulative probability under the neuron's spontaneous firing rate, capped at 6. Each neuron is then standardized across trials. The selected recording session contains 110 units in primary visual cortex that pass the inter-spike-interval violation criterion of the Allen quality metrics (< 0.75), yielding a 110-dimensional sample space.

The data are not augmented. The training set of $250{,}000$ samples is drawn uniformly with replacement from the $1{,}162$ trials, and every method is evaluated on each trial once.

H.2 CONFIGURATION OF OUR MODEL

Our model uses the implementation of Appendix G in all experiments. Settings that differ between experiments are listed in Table 6, and the layers of the divergence estimators in Table 7. Every experiment is repeated with five seeds, 0 to 4. The seed sets the training data, the permutation or transformation of the observations and the initialization, and the configurations are otherwise identical across seeds. Reported numbers are computed at the end of training. For the neural recordings, the training samples are drawn in a fixed order, so the seed sets only the initialization, and the joint entropy is computed in the subspace of the DC and fundamental harmonics of the output, i.e. $k = 3$ (Section 4.2). The 27×27 pixel-shuffling and bit-scrambling experiments are single runs. The former trains for $35{,}000$ epochs, which takes about $132\,\mathrm{h}$, and the latter is early stopped at epoch $7{,}500$, after about $28\,\mathrm{h}$.

H.3 DETAILS RELATED TO BASELINES

This section documents how the baselines of Section 6 were configured, tuned and evaluated. For each baseline we give the method as implemented and the hyperparameter grid together with the results that motivated its range. The full sweeps are reported in Appendix I.

H.3.1 COMMON PROTOCOL

Model selection. Every configuration of a grid is evaluated under 5 seeds, and configurations are ranked by their average. We report that mean together with the standard deviation over seeds. Single-seed scans outside a grid, used below to justify its range, are diagnostics only and never enter a reported number.

Standardized observations. All second-order baselines operate on standardized observations. Stacking the N observations as the rows of $\mathbf{X} \in \mathbb{R}^{N \times d_X}$, we center and scale each coordinate by its sample mean $\bar{x}_j$ and sample standard deviation s_j,

$$\widetilde{X}_{ij} = \frac{X_{ij} - \bar{x}_j}{s_j},$$

and compute the covariance $\mathbf{\Sigma}_X$ from $\widetilde{\mathbf{X}}$.

Table 6: **Configuration of our model.** Settings that differ between experiments. The $d = 15$ column covers the three translation regimes of the LieGAN comparison, and the $d = 63$ column both the identity and DST-I observations. d_X and d_Z are the dimensions of the observations and of the conditioned sample space, and the lattice is $N_1 \times \cdots \times N_P$. An epoch is one pass over the training set. The initial learning rates η_{rec} and η_{loss} of the recovery and loss estimation networks decay tenfold over training. Estimators gives the number of JS-divergence estimators per axis, with the dilation q of the coarse-grained one. Time is the approximate wall-clock time per run on a single NVIDIA RTX 3090 GPU, estimated from the training logs, except for the 27×27 runs, which used an NVIDIA A100. *Early stopped.

	GSN-G $d = 15$	GSN-G, GSN-L $d = 63$	Ising permuted	Ising random	MNIST 15×15 permuted	MNIST 27×27 permuted	MNIST 27×27 bit-scrambled	Neural recordings
d_X	15	63	33	33	225	729	5,832	110
d_Z	15	99	51	51	325	825	825	110
Lattice	15	63	33	33	15×15	27×27	27×27	6
$\mathbf{C}$	identity	zero padding	zero padding	learned	zero padding	zero padding	learned	identity
Batch	500	2,500	2,500	2,500	2,500	2,500	2,500	2,500
Training set	50k	250k	250k	250k	250k	500k	2,500	250k
Epochs	1,500	10,000	10,000	10,000	15,000	35,000	7,500*	500
Updates	1.5×10^5	10^6	10^6	10^6	1.5×10^6	7×10^6	7,500	5×10^4
η_{rec}	5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}	10^{-4}	10^{-4}	10^{-4}
η_{loss}	5×10^{-3}	2.5×10^{-3}	2.5×10^{-3}	2.5×10^{-3}	2.5×10^{-3}	10^{-3}	10^{-3}	10^{-3}
Estimators	1	1	2 ($q = 3$)	2 ($q = 3$)	1	2 ($q = 3$)	2 ($q = 3$)	1
Time	1.3 h	11 h	15 h	15 h	28 h	132 h	28 h	0.6 h

One-dimensional recovery benchmarks. The signal-recovery benchmarks of Table 1 use the 63-dimensional GSN datasets of Appendix H.1.1 and the 33-node Ising chains of Appendix H.1.2, the latter under both permutation and random-matrix mixing. Each seed draws $50{,}000$ samples, and the seeds are 0 to 4. For GLASSO and Covariance Matching, the covariance is diagonally loaded with 10^{-2}. The recovery correlation is the Pearson correlation between a latent signal and its reconstruction, maximized over the ambiguities that one-dimensional signal recovery cannot resolve (cyclic translation, reflection, and global sign), and averaged over 500 samples. These samples are drawn from the pool used to estimate the covariance. Since the baselines never observe the latent signals, this evaluation measures the quality of the recovered map rather than generalization.

Two-dimensional recovery benchmarks. The pixel-shuffling benchmarks of Table 2 use the permuted 15×15 and 27×27 MNIST crops of Appendix H.1.3, with $50{,}000$ samples per seed and seeds 42 to 46. All baselines on these benchmarks use the same data generator and seeds, and therefore see identical data. The recovery correlation is the absolute Pearson correlation between an image and its reconstruction, maximized over the eight dihedral symmetries of the square,

$$r_i = \max_{g \in D_4} \left| \operatorname{corr}\left(\mathbf{f}_i,\, g \cdot \widehat{\mathbf{f}}_i \right) \right|,$$

averaged over 100 reconstructed images.

H.3.2 Covariance Matching

Method. Covariance Matching is a second-order baseline that aligns the eigenbasis of the observed covariance with the eigenbasis of the assumed signal domain, without estimating a graph. Let $\mathbf{\Sigma}_X = \mathbf{V}_{\text{obs}} \mathbf{\Lambda}_{\text{obs}} \mathbf{V}_{\text{obs}}^\top$ be the eigendecomposition of the loaded covariance of the standardized observations $\widetilde{\mathbf{X}}$ (Appendix H.3.1), with eigenvalues λ_i in descending order. Let $\mathbf{V}_{\text{ref}}$ contain the eigenvectors of the adjacency matrix of the path graph, i.e. the one-dimensional grid, also in descending eigenvalue order. Every configuration recovers the signals as

$$\widehat{\mathbf{F}} = \widetilde{\mathbf{X}} \mathbf{W}, \qquad \mathbf{W} = \mathbf{V}_{\text{obs}} \operatorname{diag}(\mathbf{s})\, \mathbf{V}_{\text{ref}}^\top,$$

and configurations differ only in the spectral gain $\mathbf{s}$:

- *rotate*, $s_i = 1$: a change of basis that keeps the empirical spectrum;
- *whiten*, $s_i = \lambda_i^{-1/2}$: principal-component whitening before the change of basis;

Table 7: **Architecture of the divergence estimators.** Layers of each KL-divergence estimator network (Appendix G.3.3). Filters is the number of output channels, and Pool marks stochastic pooling with window and stride 2 after the layer.

Experiment	Estimator	Layer	Filters	Kernel	Stride	Pool
One-dimensional grids						
GSN-G $d = 15$	Full ($N = 15$)	1	4	5	2	None
		2	6	3	1	Stoch
		3	8	3	1	Stoch
GSN-G, GSN-L $d = 63$	Full ($N = 63$)	1	4	5	2	None
		2	4	5	2	None
		3	6	3	2	None
		4	6	3	1	Stoch
		5	8	3	1	Stoch
Ising (both mixings)	Full ($N = 33$)	1	4	7	2	None
		2	6	5	2	None
		3	8	3	1	Stoch
		4	8	3	1	Stoch
	Coarse ($q = 3$, $N = 11$)	1	4	5	2	None
		2	6	3	1	Stoch
		3	8	3	1	Stoch
Neural recordings	Full ($N = 6$)	1	4	5	1	Stoch
		2	6	3	1	Stoch
Two-dimensional grids						
MNIST 15×15 (permuted)	Full (15×15)	1	4	5×5	2	None
		2	6	3×3	1	Stoch
		3	8	3×3	1	Stoch
MNIST 27×27 (both)	Input embedding	1	4	3×3	1	None
	Full (27×27)	1	4	7×7	1	Stoch
		2	6	5×5	1	Stoch
		3	8	3×3	1	Stoch
		4	8	3×3	1	Stoch
	Coarse ($q = 3$, 9×9)	1	4	5×5	1	Stoch
		2	6	3×3	1	Stoch
		3	8	3×3	1	Stoch

- *recolor*(γ, p), $s_i = (\sigma_i/\lambda_i)^{1/2}$: whitening followed by recoloring with the low-pass reference spectrum

$$\sigma_i \propto \left(1 + \gamma\, \omega_i^2\right)^{-p}, \qquad \omega_i = \frac{i-1}{d_X - 1} \in [0, 1],$$

normalized to unit mean. This spectrum encodes only the assumption that signals on a one-dimensional domain are smooth.

Grid. The two parameter-free configurations represent the two extremes of spectral gain: *rotate* keeps the empirical spectrum, and *whiten* discards it. The *recolor* configurations whiten the observations and then impose the low-pass reference spectrum σ_i; as $\gamma \to 0$ this spectrum becomes flat and *recolor* reduces to *whiten*. The strength of the resulting smoothness prior grows with both the decay rate γ and the exponent p. We sweep $\gamma \in \{0.5, 1, 2\}$ and $p \in \{1, 2\}$, from a weak to a strong prior, which together with *rotate* and *whiten* gives eight configurations. Each configuration was run on the six one-dimensional datasets (four GSN and two Ising) with five seeds, $8 \times 6 \times 5 = 240$ runs in total, all of which completed.

The full sweeps are reported in Tables 12 and 14 (Appendices I.2.1 and I.2.1).

H.3.3 Graphical Lasso

Method. The graphical-lasso (GLASSO) baseline first estimates a graph over the observed coordinates and then finds the transformation that aligns this graph with the target domain. GLASSO estimates a sparse precision matrix $\widehat{\mathbf{\Theta}}$ from the loaded covariance $\mathbf{\Sigma}_X$ by maximizing the Gaussian log-likelihood with an ℓ_1 penalty of weight α on the entries of the precision matrix, so that a larger α yields a sparser estimate (Friedman et al., 2008); the solver runs for at most 300 iterations. The off-diagonal magnitudes of $\widehat{\mathbf{\Theta}}$, with entries below 10^{-3} discarded, form the weighted adjacency matrix $\mathbf{A}_{\text{glasso}}$ of the estimated graph. The target adjacency matrix $\mathbf{A}_{\text{target}}$ is that of the path graph for the one-dimensional datasets and of the four-neighbor lattice for MNIST. With $\mathbf{V}_{\text{g}}$ and $\mathbf{V}_{\text{t}}$ the eigenvectors of $\mathbf{A}_{\text{glasso}}$ and $\mathbf{A}_{\text{target}}$ in descending eigenvalue order, the estimated graph is matched to the target in one of two ways.

- *Spectral matching* aligns the two eigenbases, $\mathbf{W} = \mathbf{V}_{\text{g}}\mathbf{V}_{\text{t}}^{\top}$, and recovers the signals as $\widehat{\mathbf{F}} = \widetilde{\mathbf{X}}\mathbf{W}$. We refer to this variant as GLASSO-SPEQ and use it on the one-dimensional datasets. Since $\mathbf{W}$ is dense, it is not restricted to permutations.
- *FAQ matching* (GLASSO-FAQ, MNIST only) binarizes $\mathbf{A}_{\text{glasso}}$ and finds the permutation that best aligns the two graphs by solving the quadratic assignment problem $\max_{\mathbf{P}\in\Pi} \operatorname{tr}(\mathbf{P}^{\top}\mathbf{A}_{\text{glasso}}\mathbf{P}\mathbf{A}_{\text{target}})$, where Π is the set of permutation matrices, with the Fast Approximate QAP solver (Vogelstein et al., 2015). Its output is therefore a permutation, the transformation class of the pixel-shuffling benchmark. On MNIST we use only this variant: the adjacency matrix of the lattice has highly degenerate eigenspaces, within which the descending-order alignment of spectral matching is arbitrary.

Empty-graph degeneracy. For large α, the estimated precision matrix $\widehat{\mathbf{\Theta}}$ has no off-diagonal entries. The estimated graph is then empty, $\mathbf{V}_{\text{g}}$ is the identity, and spectral matching reduces to the fixed map $\mathbf{W} = \mathbf{V}_{\text{t}}^{\top}$, which does not depend on the data. We exclude a configuration from selection if its graph is empty for every seed.

One-dimensional grid. We sweep the penalty over the five values $\alpha \in \{0.01, 0.05, 0.25, 1.25, 6.25\}$. For every seed, the graph is empty for $\alpha \geq 1.25$ on every GSN dataset, and already at $\alpha = 0.25$ for GSN-G under DST-I mixing. On both Ising datasets, $\alpha = 1.25$ and $\alpha = 6.25$ give identical scores on every seed, which is the signature of the data-independent map, and we exclude them likewise. The sweep comprises 5 penalties on six datasets (four GSN and two Ising) with five seeds, $5 \times 6 \times 5 = 150$ runs in total, all of which completed. The sweeps are reported in Tables 13 and 14 (Appendices I.2.1 and I.2.1).

MNIST grid. On 15×15 crops, $\alpha = 0.25$ makes the solver return a non-positive-definite result, and $\alpha \geq 1.25$ gives the empty graph. To locate the penalties for which the solver succeeds and the graph is non-empty, we counted the edges of the estimated graph at seed 42 over a range of penalties:

α	0.005	0.01	0.02	0.05	0.10	0.125	0.15	0.20	0.25	≥ 1.25
edges	fails	8070	6544	4696	3054	2734	fails	fails	fails	empty

The solver succeeds on the contiguous interval $0.01 \leq \alpha \leq 0.125$ and becomes unstable at its neighbors, $\alpha = 0.005$ and $0.15 \leq \alpha \leq 0.25$. We sweep nine penalties inside it, $\alpha \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.10, 0.125\}$, the same number of configurations as for the manifold-learning baselines. None of these configurations yields an empty graph: even the sparsest has about six times as many edges as the 15×15 lattice, which has 420.

At 27×27 we reuse the same grid without retuning, and the solver fails for some penalties on some seeds. We select only among configurations that converged on all five seeds, because the mean of a partially converged configuration is taken over a different subset of seeds and is therefore not comparable. The sweeps at both resolutions are reported in Table 15 (Appendix I.2.2).

H.3.4 Manifold Learning: Pixel Shuffling

Method. Following Roux et al. (2007), ISOMAP and UMAP are adapted to the permuted-pixel task by embedding pixels rather than images, using their intensities over the dataset (Algorithm 2). The pseudo-distance treats strongly correlated pixels as close. Both methods run the same pipeline and differ only in the embedding step (line 3): ISOMAP builds a k-nearest-neighbor graph and uses geodesic distances, while UMAP optimizes a fuzzy simplicial set whose local scale is set by k. Differences between the two are therefore attributable to that step.

Algorithm 2 Pixel-layout recovery with manifold learning

Require: permuted images as the rows of $\mathbf{X} \in \mathbb{R}^{n \times d}$, grid size $H \times W$, embedding method (ISOMAP or UMAP) with neighborhood size k, interpolation radius r, variance threshold τ, distance floor ϵ, RBF width c

Ensure: linear operator $\mathbf{W} \in \mathbb{R}^{d \times HW}$ and reconstructed images $\widehat{\mathbf{F}} = \mathbf{XW}$

1: Discard the pixels whose standard deviation over the dataset is below τ times the largest

2: For every pair of remaining pixels, set $d_{ii} \leftarrow 0$ and
$d_{ij} \leftarrow -\frac{1}{2}\left[\log\max(|\operatorname{corr}(i,j)|, \epsilon) + \log\max(|\operatorname{corr}(j,i)|, \epsilon)\right]$

3: Embed the pixels in two dimensions from the distance matrix (d_{ij}) with neighborhood size k, giving points $\mathbf{z}_p \in \mathbb{R}^2$

4: Rotate the embedding onto its principal axes and rescale it to span $[1, H] \times [1, W]$

5: **for** each grid point $\mathbf{g} \in \{1, \ldots, H\} \times \{1, \ldots, W\}$ **do**

6: $\mathcal{N}(\mathbf{g}) \leftarrow \{p : \|\mathbf{z}_p - \mathbf{g}\|_1 < r\}$, widening r only if the set is empty

7: $W_{p\mathbf{g}} \leftarrow \dfrac{\exp(-c\,\|\mathbf{z}_p - \mathbf{g}\|_2)}{\sum_{q \in \mathcal{N}(\mathbf{g})} \exp(-c\,\|\mathbf{z}_q - \mathbf{g}\|_2)}$ for $p \in \mathcal{N}(\mathbf{g})$, and $W_{p\mathbf{g}} \leftarrow 0$ otherwise

8: **end for**

9: **return** $\mathbf{W}$ and $\widehat{\mathbf{F}} = \mathbf{XW}$

Fixed parameters. Of the five parameters of the pipeline, three are held fixed because they are inert or saturated on this data.

- *Variance threshold,* $\tau = 0.05$. On 15×15 crops, the ratio of every pixel's standard deviation to the largest is at least 0.60, so any threshold below 0.60 removes no pixel. An earlier grid $\tau \in \{0.05, 0.10, 0.15\}$ gave identical per-seed scores for all three values.
- *Distance floor,* $\epsilon = 10^{-8}$. The smallest observed $|\operatorname{corr}(i,j)|$ is far above the floor (first percentile 0.0019), so no entry is clipped.
- *RBF width,* $c = 3$. Together with the radius, c sets the reach of the interpolation. At $c = 3$, the weight e^{-3d} is 0.0025 at a distance of $d = 2$ grid units, so pixels beyond that distance carry less than 2% of the total weight even when a larger radius admits them. Accordingly, the score saturates as the radius grows (ISOMAP, $k = 3$, 15×15):

r	0.5	1	2	4	6
correlation	0.841	0.844	0.859	0.861	0.861

Grid. The swept parameters are the neighborhood size k and the interpolation radius $r \in \{0.5, 1, 2\}$, which covers the radii below the saturation shown above. The useful range of k is narrow and differs between the methods, because k plays a different role in each: it is a hard neighbor count in ISOMAP and the scale of a fuzzy neighborhood in UMAP. Single-seed scans at 15×15 locate the best k of each method. ISOMAP scores 0.55, 0.31, 0.92, 0.78, and 0.47 for $k = 1, \ldots, 5$, and UMAP, at $r = 1$, scores as follows:

k	2	3	4	5	6	7	8	12	20	30	50
correlation	0.274	0.274	0.423	0.727	0.874	0.687	0.576	0.442	0.453	0.461	0.462

For both methods, the best k lies strictly inside the scanned range, and the score drops sharply one step away from it, so a coarse grid, or a grid shared by both methods, would miss it. We therefore center each grid on the best value of its own method: $k \in \{2, 3, 4\}$ for ISOMAP and $k \in \{5, 6, 7\}$ for UMAP. Reusing the ISOMAP range for UMAP would place all UMAP configurations at $k \leq 4$, where UMAP scores at most 0.423. The optimization of UMAP is seeded with the trial seed, so its runs are reproducible.

At 27×27 we reuse the same grids without retuning. A best configuration at the edge of a grid would suggest that a wider grid could improve the baseline, so we scanned k beyond the grid edges to verify that the best configuration of each method still lies strictly inside its grid. For ISOMAP at $r = 2$, k from 3 to 10 peaks at $k = 4$ (0.749), and every other value scores at most 0.619. For UMAP at $r = 2$, $k \in \{8, 10, 12, 15, 20, 30\}$ scores at most 0.547, and $k \in \{2, 3, 4\}$ at most 0.447, both below the best in-grid score of 0.574. Each grid has nine configurations (three values of k and three radii), run with five seeds, so that $9 \times 5 = 45$ runs were completed per method and resolution.

The full sweep at both resolutions is reported in Table 16 (Appendix I.2.2).

H.3.5 Manifold Learning: Neural Recordings

For the experiment of Section 6.3, ISOMAP and UMAP embed the 110-dimensional response vectors $\mathbf{x}_i$ of Appendix H.1.5 directly into two dimensions, $\mathbf{z}_i \in \mathbb{R}^2$. We read an angle from each embedded point relative to the centroid $\bar{\mathbf{z}}$ of the embedding,

$$\widehat{\theta}_i = \operatorname{atan2}(z_{i,2} - \bar{z}_2,\ z_{i,1} - \bar{z}_1) \bmod 2\pi,$$

and compare it with the grating orientation exactly as the output phase of our model, over all $1{,}162$ trials. Since the orientation is π-periodic, the orientation $\theta_i^{\deg}$ in degrees is doubled to cover the full circle, $\theta_i = 2\theta_i^{\deg} \cdot \pi/180^\circ$. The rotation and handedness of an embedding are arbitrary, so both scores are taken at the best alignment $s\widehat{\theta}_i - \delta$ with $s \in \{1, -1\}$: the circular correlation is invariant to δ and we report $\max_s r_{\mathrm{cc}}(s\widehat{\theta}, \theta)$, while the mean absolute orientation error

$$\mathrm{MAE} = \min_{s,\,\delta}\ \frac{1}{N} \sum_{i=1}^{N} \tfrac{1}{2} \left| \mathrm{wrap}\Big(s\widehat{\theta}_i - \delta - \theta_i \Big) \right|,$$

with wrap mapping to $(-\pi, \pi]$, is expressed in degrees of orientation. With six equally frequent orientations, outputs spread uniformly over the circle occupy a 60° sector per orientation even when perfectly ordered, which limits the scores to $r_{\mathrm{cc}}^{\max} = 0.912$ and $\mathrm{MAE}^{\min} = 7.5^\circ$; Table 3 reports these limits alongside the scores.

For UMAP, we sweep the neighborhood size $k \in \{5, 15, 30, 50\}$ and the minimum distance over $\{0.01, 0.1, 0.5\}$; for ISOMAP, which has no minimum-distance parameter, we sweep $k \in \{5, 15, 30, 50\}$. Both grids contain the default settings of the respective libraries ($k = 15$ and minimum distance 0.1 for UMAP, $k = 5$ for ISOMAP) and extend to neighborhoods of 50 points, about 4% of the trials. We report the configuration with the highest mean circular correlation over five seeds, $k = 30$ with minimum distance 0.1 for UMAP and $k = 5$ for ISOMAP; ISOMAP is deterministic, so its seeds coincide.

The full sweep is reported in Table 17 (Appendix I.2.3).

H.3.6 SYMMETRY DISCOVERY: LIEGAN

LieGAN (Yang et al., 2023b) learns a symmetry adversarially: a generator applies group transformations to the data, and a discriminator is trained to distinguish transformed from untransformed samples. Unlike Covariance Matching and GLASSO, which recover a map from second-order statistics, LieGAN parametrizes the symmetry as a Lie algebra element, so its output is directly comparable with the generators learned by our model. We compare the two methods on the 15-dimensional GSN-G datasets under the three translation regimes of Appendix H.1.1.

Method. We use LieGAN with a single channel, so that the generator learns one matrix $\mathbf{L} \in \mathbb{R}^{15\times 15}$, and in its Lie algebra mode, in which only $\mathbf{L}$ is learned and each observation is transformed as $\mathbf{x}' = \exp(z\mathbf{L})\,\mathbf{x}$ with $z \sim \mathcal{N}(0,1)$. The discriminator is a multilayer perceptron with two hidden layers of width 512 and LeakyReLU activations. The generator loss includes the regularizer of LieGAN, weighted by λ, which penalizes transformations that leave $\mathbf{x}$ unchanged and thereby excludes the trivial solution $\mathbf{L} = \mathbf{0}$. The channel-orthogonality regularizer vanishes for a single channel, and the sparsity regularizer is disabled, as in the defaults of LieGAN. Both networks are trained with Adam, alternating one generator step and one discriminator step per minibatch.

Data. LieGAN is trained on the same data as our model. For each regime and seed, we regenerate the dataset of the corresponding run of our model from its stored specification, which yields the same 50,000 samples (100 batches of 500). All nine configurations share this dataset.

Optimization budget. Every run trains for 200 epochs with minibatches of 64, which amounts to about 1.5×10^5 generator updates, the number of updates of our model (Table 6). Since our model uses larger batches to obtain low-variance estimates of its information-theoretic terms, we match the number of gradient updates rather than the number of epochs. This budget is sufficient for LieGAN to recover the generator: in the band-limited continuous regime, four of the nine configurations reach $S_{0.75} = 1.000$ (Table 18).

Grid. We sweep the two hyperparameters that govern the balance between the generator and the discriminator: the discriminator learning rate $\eta_D \in \{2\cdot 10^{-5}, 2\cdot 10^{-4}, 2\cdot 10^{-3}\}$ and the regularizer weight $\lambda \in \{0.1, 1, 10\}$, one value per decade, which gives nine configurations. The generator learning rate is fixed at $\eta_G = 10^{-3}$.

Evaluation. Since LieGAN learns a Lie algebra element, this comparison scores Lie generators rather than group elements, and both methods are scored by the same implementation. The reference generator is the spectral derivative $\mathbf{F}^\dagger \operatorname{diag}(2\pi \mathrm{i}k/N)_k \mathbf{F}$ in the band-limited continuous regime, where $\mathbf{F}$ is the unitary DFT matrix, and the central difference $\frac{1}{2}(\mathbf{S} - \mathbf{S}^\top)$ of the cyclic shift $\mathbf{S}$ in the continuous and discrete regimes. Since $\mathbf{L}$ and $-\mathbf{L}$ generate the same group, the sign of the reference is chosen to agree with the learned generator. The primary score is the similarity $S_{0.75}$ of Section 6, restricted to frequencies up to 0.75 of the Nyquist limit, which at $N = 15$ keeps the 11 Fourier modes with $|k| \le 5$. We also report $S_{1.0}$, the same similarity computed over all frequencies up to the Nyquist limit, i.e. without discarding any mode.

Model selection. Each run is scored on the learned Lie generator after its final epoch. We do not select a checkpoint by the minimum generator loss, because in adversarial training a low generator loss reflects a weak discriminator rather than a better symmetry. The full grid of nine configurations was trained in the three regimes with five seeds, $9 \times 3 \times 5 = 135$ runs in total, all of which completed. In each regime, the configuration with the highest mean $S_{0.75}$ over seeds is selected. Selection and reporting use the same five seeds, so the reported LieGAN score is the maximum of nine noisy means. The full sweep is reported in Table 18 (Appendix I.2.4).

I ADDITIONAL RESULTS

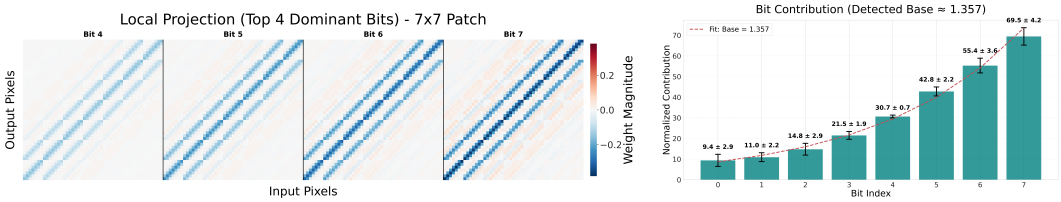
This appendix reports additional results of our model and the full hyperparameter sweeps of the baselines, whose configurations and grids are described in Appendix H.3. The baseline scores are defined for each family of datasets in Appendix H.3.1.

I.1 Our Model

I.1.1 Bit Scrambling Experiment

The bit-scrambling benchmark follows the image encryption scheme of Ye (2010), which represents each pixel by its binary expansion and permutes the positions of the resulting bits with a permutation derived from a chaotic map; for a fixed key, this permutation is the same for every image. Our variant differs in two respects (Appendix H.1.4). First, we draw the permutation with a pseudorandom number generator instead of a chaotic map. Our method makes no assumption on how the permutation is generated, only that it is fixed across samples, so this choice does not affect the recovery problem. Second, whereas the scheme of Ye (2010) regroups the scrambled bits into bytes to form an encrypted image, we present the scrambled bits to the model directly, on which the scrambling acts as a permutation (Section 6.2).

Figure 5 shows the bit weights learned by the composition $\mathbf{UC}$ of the conditioning and lifting matrices, which forms the response of the model.



(a) Learned bit-to-byte aggregation.

(b) Bit weights.

Figure 5: **Matrix representation of the model and bitstream recomposition. (a)** The matrix representation of the model shows parallel diagonal lines in every bit channel. Since vertical and horizontal pixel shifts correspond to index increments of 7 and 1 in a flattened 7×7 patch, this pattern forms a banded Toeplitz structure, which acts as a local averaging operator with a receptive field of approximately 3 pixels. **(b)** The standard deviation of the bit weights across pixels is small. The weights decay approximately exponentially from the most significant bit, so that the model learns an approximate number system with base 1.357.

I.1.2 Morris Sensitivity Analysis

We assess the robustness of our model to hyperparameter variations with a Morris elementary-effects analysis, which screens the influence of each parameter at a small computational cost and indicates nonlinear effects and interactions between parameters.

Experimental setup. We vary five hyperparameters that govern the optimization dynamics and the balance between the loss terms:

- η_{loss}, the learning rate of the loss estimation network, in $[0.0025, 0.0075]$;
- η_{rec}, the learning rate of the recovery network, in $[0.00025, 0.00075]$;
- λ_{stat}, a weight on the stationarity term, which is unweighted in (6), in $[1.0, 1.66]$;
- λ_{TC}, the weight of the total correlation term, in $[0.75, 1.25]$;
- λ_{infomax}, the weight of the information maximization term, in $[0.5625, 0.9375]$.

With 6 trajectories on an 8-level grid, the design comprises 36 configurations, each trained for $1{,}500$ epochs with a batch size of 500. We report the signal recovery r and the generator similarity $S_{0.75}$, both as defined in Section 6.

Global robustness. Across the 36 configurations (Table 8), the signal recovery lies between $r = 0.968$ and $r = 0.988$, and the generator similarity is at least $S_{0.75} = 0.990$, exceeding 0.995 in 33 of the 36 runs.

Table 8: **Morris sensitivity runs.** Hyperparameters, signal recovery r, and generator similarity $S_{0.75}$ of all 36 configurations.

Run	η_{loss}	η_{rec}	λ_{stat}	λ_{TC}	λ_{infomax}	r	$S_{0.75}$
1	0.0039	0.0007	1.5657	1.1786	0.7232	0.985	0.998
2	0.0039	0.0007	1.5657	0.8929	0.7232	0.987	0.999
3	0.0068	0.0007	1.5657	0.8929	0.7232	0.987	0.995
4	0.0068	0.0007	1.1886	0.8929	0.7232	0.987	0.999
5	0.0068	0.0007	1.1886	0.8929	0.9375	0.986	0.997
6	0.0068	0.0004	1.1886	0.8929	0.9375	0.986	0.999
7	0.0061	0.0003	1.2829	0.9643	0.5625	0.985	0.996
8	0.0032	0.0003	1.2829	0.9643	0.5625	0.985	0.999
9	0.0032	0.0006	1.2829	0.9643	0.5625	0.985	0.999
10	0.0032	0.0006	1.6600	0.9643	0.5625	0.985	0.999
11	0.0032	0.0006	1.6600	1.2500	0.5625	0.984	0.999
12	0.0032	0.0006	1.6600	1.2500	0.7768	0.985	0.997
13	0.0068	0.0004	1.4714	1.2500	0.9375	0.985	0.998
14	0.0068	0.0004	1.4714	1.2500	0.7232	0.985	0.990
15	0.0068	0.0004	1.0943	1.2500	0.7232	0.985	0.999
16	0.0068	0.0007	1.0943	1.2500	0.7232	0.985	0.997
17	0.0068	0.0007	1.0943	0.9643	0.7232	0.987	0.995
18	0.0039	0.0007	1.0943	0.9643	0.7232	0.986	0.999
19	0.0039	0.0008	1.2829	1.0357	0.6696	0.985	0.999
20	0.0068	0.0008	1.2829	1.0357	0.6696	0.985	0.995
21	0.0068	0.0005	1.2829	1.0357	0.6696	0.986	0.998
22	0.0068	0.0005	1.6600	1.0357	0.6696	0.985	0.999
23	0.0068	0.0005	1.6600	1.0357	0.8839	0.983	0.999
24	0.0068	0.0005	1.6600	0.7500	0.8839	0.968	0.997
25	0.0075	0.0006	1.0943	0.8214	0.7768	0.988	0.999
26	0.0075	0.0003	1.0943	0.8214	0.7768	0.984	0.997
27	0.0046	0.0003	1.0943	0.8214	0.7768	0.988	0.999
28	0.0046	0.0003	1.0943	1.1071	0.7768	0.985	0.997
29	0.0046	0.0003	1.4714	1.1071	0.7768	0.985	0.998
30	0.0046	0.0003	1.4714	1.1071	0.5625	0.983	0.997
31	0.0025	0.0008	1.1886	1.1786	0.6161	0.984	0.999
32	0.0025	0.0008	1.1886	0.8929	0.6161	0.985	0.997
33	0.0025	0.0008	1.1886	0.8929	0.8304	0.987	0.999
34	0.0025	0.0005	1.1886	0.8929	0.8304	0.984	0.997
35	0.0054	0.0005	1.1886	0.8929	0.8304	0.987	0.994
36	0.0054	0.0005	1.5657	0.8929	0.8304	0.985	0.999

Elementary effects. For each hyperparameter, we compute the mean elementary effect μ, the mean absolute effect μ^*, and the standard deviation of the effects σ. The value of μ^* measures the overall influence of a parameter, while σ indicates nonlinear effects and interactions with other parameters.

Table 9 shows that λ_{TC} has the largest influence on the signal recovery ($\mu^* = 0.0059$, $\sigma = 0.0102$), whereas λ_{stat} ($\mu^* = 0.0052$) and η_{loss} ($\mu^* = 0.0051$) have the largest influence on the generator similarity. These effects are small in absolute terms: across all runs, r and $S_{0.75}$ vary by at most 0.020 and 0.010, respectively.

Table 9: **Morris elementary effects.** Mean effect μ, mean absolute effect μ^*, and standard deviation σ of each hyperparameter on the signal recovery r and the generator similarity $S_{0.75}$.

	Signal recovery r			Generator similarity $S_{0.75}$		
Parameter	μ	μ^*	σ	μ	μ^*	σ
η_{loss}	-0.0003	0.0021	0.0034	-0.0051	0.0051	0.0012
η_{rec}	0.0017	0.0022	0.0031	-0.0005	0.0025	0.0029
λ_{stat}	-0.0008	0.0009	0.0014	-0.0012	0.0052	0.0077
λ_{TC}	0.0015	0.0059	0.0102	0.0003	0.0024	0.0029
$\lambda_{\mathrm{infomax}}$	0.0005	0.0023	0.0028	0.0018	0.0038	0.0056

I.1.3 Generator Similarity over the Full Band

Definition. Let $\mathbf{P}_q$ be the orthogonal projector onto the discrete Fourier modes of the lattice whose frequency satisfies $|f| \le q/2$ cycles per sample along every axis, i.e. up to the fraction $q \in (0, 1]$ of the Nyquist limit; on the two-dimensional lattices this band is a box. The projector is carried into the coordinates of the observations by the mixing matrix of the dataset, $\mathbf{\Pi}_q = \mathbf{A}\mathbf{P}_q\mathbf{A}^{-1}$. For a learned operator $\hat{\mathbf{G}}$ and its reference $\mathbf{G}^\star$, the generator similarity is the cosine similarity of the two operators restricted to this band,

$$S_q = \frac{\langle \mathbf{\Pi}_q \hat{\mathbf{G}} \mathbf{\Pi}_q,\ \mathbf{\Pi}_q \mathbf{G}^\star \mathbf{\Pi}_q \rangle_F}{\|\mathbf{\Pi}_q \hat{\mathbf{G}} \mathbf{\Pi}_q\|_F \, \|\mathbf{\Pi}_q \mathbf{G}^\star \mathbf{\Pi}_q\|_F},$$

where $\langle \cdot, \cdot \rangle_F$ and $\| \cdot \|_F$ are the Frobenius inner product and norm. The learned operator is the sample-space minimal generator $\hat{\rho}_X(\mathbf{t}_{\mathbf{e}_i}) = \mathbf{C}^\dagger \hat{\rho}_Z(\mathbf{t}_{\mathbf{e}_i})\mathbf{C}$ of each axis i (Section 6), and the reference is the translation by one lattice site along the axis and in the direction that match it best, carried into the observations as $\mathbf{A} \cdot \mathbf{A}^{-1}$. The 15-dimensional GSN datasets are the exception: as in the comparison with LieGAN, which learns a Lie algebra element, the Lie generators are compared instead, against the references of Appendix H.3.6. The main text reports $S_{0.75}$, and $S_{1.0}$ keeps every mode, so that it is the plain cosine similarity of the two operators.

Results. Table 10 reports both similarities for every experiment run with five seeds. $S_{1.0}$ is lower than $S_{0.75}$ in every experiment, although only in the third decimal for the permuted Ising chain. The difference comes from the directions that $S_{0.75}$ drops: although they make up 24–46% of the sample space, they carry only 0.5–2.7% of the signal energy (Table 11). The training objectives see the generator only through its action on the data, so these directions constrain it only weakly, and the learned operator is left nearly free there. In addition, sampling continuous signals aliases their highest frequencies, which violates the translation structure in exactly these modes (Section 6).

Signal energy beyond the evaluation band. Table 11 reports, for each dataset, the share of Fourier directions outside this band and the share of the signal energy they carry. The energy of a frequency is the mean squared magnitude of the discrete Fourier transform of the samples, including their mean, taken in the natural coordinates before any permutation or mixing. It is estimated from 10 training batches for each of the seeds 0 to 4. Although the excluded directions make up 24–46% of the sample space, they carry only 0.5–2.7% of the energy. The neural recordings and the bit-scrambling data are not included, since no generator similarity is reported for them.

Table 10: **Generator similarity over the band and the full spectrum.** $S_{0.75}$ and $S_{1.0}$ of our model: mean $\pm$ standard deviation over the seeds 0 to 4. For MNIST, the two values are the two group axes. †Lie generators are compared, as in the LieGAN comparison.

Dataset	Transform	$S_{0.75}$	$S_{1.0}$
GSN-G, $d = 15^{\dagger}$	Band-limited continuous	0.946 ± 0.070	0.718 ± 0.115
GSN-G, $d = 15^{\dagger}$	Discrete	0.962 ± 0.010	0.909 ± 0.052
GSN-G, $d = 15^{\dagger}$	Continuous	0.969 ± 0.009	0.926 ± 0.038
GSN-G, $d = 63$	Identity	0.996 ± 0.002	0.700 ± 0.064
GSN-L, $d = 63$	Identity	0.910 ± 0.022	0.618 ± 0.044
GSN-G, $d = 63$	DST-I	0.994 ± 0.002	0.842 ± 0.078
GSN-L, $d = 63$	DST-I	0.976 ± 0.007	0.914 ± 0.030
Ising, $d = 33$	Permuted	0.999 ± 0.000	0.998 ± 0.000
Ising, $d = 33$	Random	0.955 ± 0.026	0.744 ± 0.138
MNIST 15×15	Permuted	0.900 ± 0.034 0.897 ± 0.039	0.549 ± 0.049 0.532 ± 0.075

Table 11: **Signal energy beyond the evaluation band.** Share of the Fourier directions outside the band of $S_{0.75}$, and share of the signal energy they carry: mean $\pm$ standard deviation over the seeds 0 to 4.

Dataset	Lattice	Directions outside	Energy outside
GSN-G, band-limited continuous	15	$4/15$ (27%)	$1.43 \pm 0.03\%$
GSN-G, discrete	15	$4/15$ (27%)	$1.43 \pm 0.04\%$
GSN-G, continuous	15	$4/15$ (27%)	$0.91 \pm 0.02\%$
GSN-G	63	$16/63$ (25%)	$1.46 \pm 0.04\%$
GSN-L	63	$16/63$ (25%)	$0.51 \pm 0.01\%$
Ising	33	$8/33$ (24%)	$2.41 \pm 0.01\%$
MNIST	15×15	$104/225$ (46%)	$2.72 \pm 0.01\%$
MNIST	27×27	$288/729$ (40%)	$1.47 \pm 0.01\%$

I.2 Baselines

We report the full hyperparameter sweep of every baseline, grouped by the main-text results they support: signal recovery on the one-dimensional benchmarks (Table 1, Appendix I.2.1), signal recovery on the two-dimensional benchmarks (Table 2, Appendix I.2.2), and symmetry discovery (Table 1, Appendix I.2.4).

I.2.1 One-Dimensional Benchmarks

Covariance matching. Table 12 reports the recovery correlation of every covariance-matching configuration.

Graphical lasso. Table 13 reports the sweep over the penalty α.

Ising chains. Table 14 reports both sweeps on the 33-node Ising chains.

I.2.2 Two-Dimensional Benchmarks

Graphical lasso. Table 15 reports the sweeps at both crop sizes. At 15×15, all $9 \times 5 = 45$ runs converged, i.e. nine penalties α with five seeds each. At 27×27, some configurations converged only on a subset of the seeds; these configurations are excluded from selection (Appendix H.3.3).

Manifold learning. Table 16 reports the ISOMAP and UMAP sweeps on permuted MNIST.

Table 12: **Covariance matching sweep.** Recovery correlation on the GSN datasets: mean ± standard deviation over 5 seeds. Best configuration per dataset in bold.

	Identity		DST-I	
Configuration	GSN-G	GSN-L	GSN-G	GSN-L
rotate	$\mathbf{0.470 \pm 0.013}$	$\mathbf{0.403 \pm 0.003}$	0.254 ± 0.010	0.285 ± 0.005
whiten	0.407 ± 0.011	0.332 ± 0.004	0.243 ± 0.010	0.274 ± 0.005
recolor$(0.5, 1)$	0.412 ± 0.011	0.341 ± 0.004	0.252 ± 0.010	0.286 ± 0.006
recolor$(0.5, 2)$	0.417 ± 0.011	0.349 ± 0.004	0.261 ± 0.010	0.297 ± 0.006
recolor$(1, 1)$	0.416 ± 0.011	0.348 ± 0.004	0.259 ± 0.010	0.294 ± 0.006
recolor$(1, 2)$	0.424 ± 0.011	0.361 ± 0.004	0.274 ± 0.010	0.311 ± 0.006
recolor$(2, 1)$	0.422 ± 0.011	0.357 ± 0.004	0.270 ± 0.010	0.306 ± 0.006
recolor$(2, 2)$	0.435 ± 0.011	0.374 ± 0.004	$\mathbf{0.294 \pm 0.010}$	$\mathbf{0.330 \pm 0.006}$

Table 13: **GLASSO-SPEQ sweep on the GSN datasets.** Recovery correlation: mean ± standard deviation over 5 seeds. Gray cells are empty-graph degeneracies, excluded from selection. Best non-degenerate configuration per dataset in bold.

	Identity		DST-I	
α	GSN-G	GSN-L	GSN-G	GSN-L
0.01	0.460 ± 0.006	0.452 ± 0.009	$\mathbf{0.240 \pm 0.007}$	0.253 ± 0.006
0.05	0.519 ± 0.068	$\mathbf{0.479 \pm 0.007}$	0.235 ± 0.010	0.248 ± 0.006
0.25	$\mathbf{0.663 \pm 0.076}$	0.462 ± 0.009	0.336 ± 0.010	$\mathbf{0.260 \pm 0.007}$
1.25	0.264 ± 0.007	0.198 ± 0.008	0.336 ± 0.010	0.274 ± 0.004
6.25	0.264 ± 0.007	0.198 ± 0.008	0.336 ± 0.010	0.274 ± 0.004

I.2.3 Neural Recordings

Table 17 reports the ISOMAP and UMAP sweeps on the mouse visual cortex recordings of Section 6.3, scored as in Appendix H.3.5.

I.2.4 Symmetry Discovery

Table 18 reports the full LieGAN sweep, scored by the band-limited similarity $S_{0.75}$. To check that the comparison does not depend on the band restriction, we also compute $S_{1.0}$, the same similarity evaluated over all frequencies up to the Nyquist limit, so that no Fourier mode is discarded (Appendix H.3.6). Under $S_{1.0}$, the selected LieGAN configurations score 1.000 ± 0.000, 0.192 ± 0.096, and 0.521 ± 0.381 in the band-limited continuous, discrete, and continuous regimes, respectively, against 0.718 ± 0.115, 0.909 ± 0.052, and 0.926 ± 0.038 for our model on the same data. The ranking of the two methods is therefore the same as under $S_{0.75}$ (Table 1).

Table 14: **Covariance matching and GLASSO-SPEQ sweeps on the Ising chains.** Recovery correlation: mean $\pm$ standard deviation over 5 seeds. Gray cells are empty-graph degeneracies, excluded from selection. Best configuration per method and mixing in bold.

Method	Configuration	Permuted	Random
CM	*rotate*	$\mathbf{0.716 \pm 0.042}$	$\mathbf{0.654 \pm 0.012}$
	whiten	0.461 ± 0.025	0.419 ± 0.007
	recolor$(0.5, 1)$	0.482 ± 0.028	0.437 ± 0.007
	recolor$(0.5, 2)$	0.502 ± 0.029	0.455 ± 0.008
	recolor$(1, 1)$	0.498 ± 0.029	0.451 ± 0.008
	recolor$(1, 2)$	0.530 ± 0.033	0.481 ± 0.008
	recolor$(2, 1)$	0.521 ± 0.032	0.472 ± 0.008
	recolor$(2, 2)$	0.568 ± 0.037	0.516 ± 0.009
GLASSO-SPEQ	$\alpha = 0.01$	$\mathbf{0.715 \pm 0.107}$	0.332 ± 0.020
	$\alpha = 0.05$	0.689 ± 0.066	$\mathbf{0.339 \pm 0.016}$
	$\alpha = 0.25$	0.662 ± 0.009	0.308 ± 0.036
	$\alpha = 1.25$	0.395 ± 0.011	0.388 ± 0.010
	$\alpha = 6.25$	0.395 ± 0.011	0.388 ± 0.010

Table 15: **GLASSO-FAQ sweep on permuted MNIST.** Recovery correlation: mean $\pm$ standard deviation over the converged seeds, with the number of converged seeds at 27×27 in parentheses where it is below 5. For reference, a uniformly random permutation of the pixels yields a recovery correlation of 0.116 at 15×15 and 0.061 at 27×27; this value is nonzero because the recovery correlation takes the absolute value and the best of eight dihedral alignments. Best configuration per resolution in bold, selected among configurations that converged on all seeds.

α	15×15	27×27
0.01	0.176 ± 0.007	failed $(0/5)$
0.02	0.177 ± 0.012	0.140 ± 0.018
0.03	0.187 ± 0.025	0.125 ± 0.013
0.04	0.179 ± 0.007	0.128 ± 0.015
0.05	0.191 ± 0.016	$\mathbf{0.153 \pm 0.028}$
0.06	0.187 ± 0.010	0.146 ± 0.022
0.08	0.185 ± 0.028	0.132 ± 0.012 $(3/5)$
0.10	$\mathbf{0.192 \pm 0.016}$	0.138 $(1/5)$
0.125	0.190 ± 0.012	failed $(0/5)$

Table 16: **Manifold-learning sweep on permuted MNIST.** Recovery correlation for each neighborhood size k and interpolation radius: mean $\pm$ standard deviation over 5 seeds. Best configuration per method and resolution in bold.

Method	k	Interpolation radius 0.5	1	2
		$\mathbf{15 \times 15}$		
ISOMAP	2	0.340 ± 0.091	0.337 ± 0.087	0.350 ± 0.091
	3	0.841 ± 0.066	0.844 ± 0.067	$\mathbf{0.859 \pm 0.064}$
	4	0.771 ± 0.013	0.780 ± 0.012	0.793 ± 0.012
UMAP	5	0.787 ± 0.033	0.788 ± 0.033	0.806 ± 0.033
	6	0.827 ± 0.040	0.833 ± 0.040	$\mathbf{0.846 \pm 0.038}$
	7	0.739 ± 0.105	0.741 ± 0.104	0.756 ± 0.102
		$\mathbf{27 \times 27}$		
ISOMAP	2	0.291 ± 0.034	0.302 ± 0.037	0.310 ± 0.039
	3	0.587 ± 0.037	0.589 ± 0.033	0.610 ± 0.035
	4	0.709 ± 0.017	0.710 ± 0.016	$\mathbf{0.731 \pm 0.016}$
UMAP	5	0.544 ± 0.050	0.543 ± 0.050	0.558 ± 0.051
	6	0.560 ± 0.028	0.558 ± 0.027	$\mathbf{0.574 \pm 0.028}$
	7	0.520 ± 0.013	0.517 ± 0.012	0.531 ± 0.012

Table 17: **Manifold-learning sweep on neural recordings.** Circular correlation r_{cc} and mean absolute orientation error (MAE) for each neighborhood size k and minimum distance: mean $\pm$ standard deviation over 5 seeds. ISOMAP is deterministic and has no minimum distance. The configuration with the highest mean r_{cc} per method, reported in Table 3, is in bold.

Method	k	Min. distance	$r_{\mathrm{cc}} \uparrow$	MAE$\downarrow$
ISOMAP	5	–	$\mathbf{0.076}$	$\mathbf{27.5}^\circ$
	15	–	0.039	28.7°
	30	–	0.052	28.3°
	50	–	0.062	28.5°
UMAP	5	0.01	0.491 ± 0.017	$18.2^\circ \pm 0.4^\circ$
	5	0.1	0.498 ± 0.040	$17.9^\circ \pm 1.0^\circ$
	5	0.5	0.487 ± 0.058	$18.0^\circ \pm 1.8^\circ$
	15	0.01	0.517 ± 0.012	$17.3^\circ \pm 0.3^\circ$
	15	0.1	0.532 ± 0.010	$16.8^\circ \pm 0.2^\circ$
	15	0.5	0.514 ± 0.007	$16.9^\circ \pm 0.2^\circ$
	30	0.01	0.559 ± 0.020	$16.4^\circ \pm 0.4^\circ$
	30	0.1	$\mathbf{0.568 \pm 0.025}$	$\mathbf{16.0}^\circ \pm \mathbf{0.4}^\circ$
	30	0.5	0.516 ± 0.029	$16.8^\circ \pm 0.7^\circ$
	50	0.01	0.178 ± 0.020	$22.1^\circ \pm 0.4^\circ$
	50	0.1	0.163 ± 0.027	$22.3^\circ \pm 0.6^\circ$
	50	0.5	0.065 ± 0.010	$24.8^\circ \pm 0.3^\circ$

Table 18: **LieGAN sweep.** Generator similarity $S_{0.75}$ of every LieGAN configuration in the three translation regimes: mean $\pm$ standard deviation over 5 seeds. Selected configuration per regime in bold.

η_D	λ	Band-limited continuous	Discrete	Continuous
$2 \cdot 10^{-5}$	0.1	0.830 ± 0.379	0.059 ± 0.037	$\mathbf{0.622 \pm 0.460}$
$2 \cdot 10^{-5}$	1	0.430 ± 0.094	0.204 ± 0.222	0.444 ± 0.255
$2 \cdot 10^{-5}$	10	0.214 ± 0.125	0.212 ± 0.155	0.171 ± 0.165
$2 \cdot 10^{-4}$	0.1	1.000 ± 0.000	0.094 ± 0.073	0.016 ± 0.024
$2 \cdot 10^{-4}$	1	$\mathbf{1.000 \pm 0.000}$	0.016 ± 0.005	0.425 ± 0.489
$2 \cdot 10^{-4}$	10	0.002 ± 0.166	0.091 ± 0.085	0.134 ± 0.106
$2 \cdot 10^{-3}$	0.1	1.000 ± 0.000	0.056 ± 0.031	0.113 ± 0.021
$2 \cdot 10^{-3}$	1	1.000 ± 0.000	0.014 ± 0.012	0.024 ± 0.018
$2 \cdot 10^{-3}$	10	0.459 ± 0.360	$\mathbf{0.218 \pm 0.111}$	0.284 ± 0.111