Manifold-tiling Localized Receptive Fields are Optimal in Similarity-preserving Neural Networks

Anirvan M. Sengupta
Rutgers University
Flatiron Institute
anirvans@physics.rutgers.edu

Mariano Tepper*
Flatiron Institute
mtepper@flatironinstitute.org

Cengiz Pehlevan*
Flatiron Institute
cpehlevan@flatironinstitute.org

Alexander Genkin
NYU Langone Medical Center
alexander.genkin@gmail.com

Dmitri B. Chklovskii
NYU Langone Medical Center
Flatiron Institute
dchklovskii@flatironinstitute.org

Abstract

Many neurons in the brain, such as place cells in the rodent hippocampus, have localized receptive fields, i.e., they respond to a small neighborhood of stimulus space. What is the functional significance of such representations and how can they arise? Here, we propose that localized receptive fields emerge in similarity-preserving networks of rectifying neurons that learn low-dimensional manifolds populated by sensory inputs. Numerical simulations of such networks on standard datasets yield manifold-tiling localized receptive fields. More generally, we show analytically that, for data lying on symmetric manifolds, optimal solutions of objectives, from which similarity-preserving networks are derived, have localized receptive fields. Therefore, nonnegative similarity-preserving mapping (NSM) implemented by neural networks can model representations of continuous manifolds in the brain.

1 Introduction

A salient and unexplained feature of many neurons is that their receptive fields are localized in the parameter space they represent. For example, a hippocampus place cell is active in a particular spatial location [1], the response of a V1 neuron is localized in visual space and orientation [2], and the response of an auditory neuron is localized in the sound frequency space [3]. In all these examples, receptive fields of neurons from the same brain area tile (with overlap) low-dimensional manifolds.

Localized receptive fields are shaped by neural activity as evidenced by experimental manipulations in developing and adult animals [4, 5, 6, 7]. Activity influences receptive fields via modification, or learning, of synaptic weights which gate the activity of upstream neurons channeling sensory inputs. To be biologically plausible, synaptic learning rules must be physically local, i.e., the weight of a synapse depends on the activity of only the two neurons it connects, pre- and post-synaptic.

In this paper, we demonstrate that biologically plausible neural networks can learn manifold-tiling localized receptive fields from the upstream activity in an unsupervised fashion. Because analyzing

*M. Tepper and C. Pehlevan contributed equally to this work.

Preprint. Work in progress.

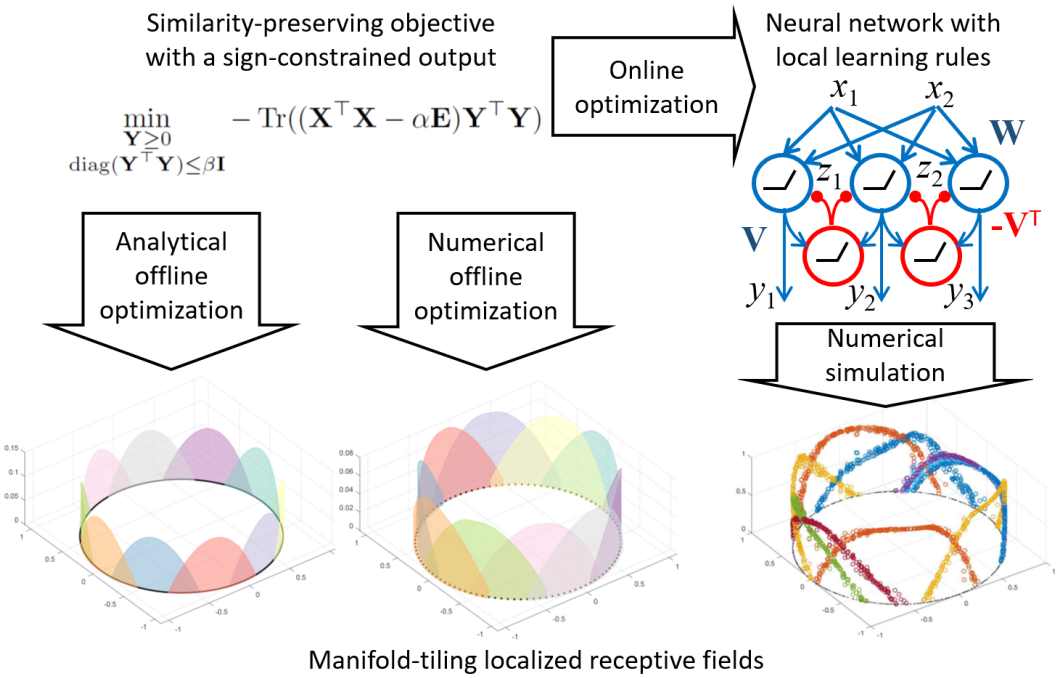

Figure 1: A schematic illustration of our normative approach.

the outcome of learning in arbitrary neural networks is often difficult, we take a normative approach, Fig. 1. First, we formulate an optimization problem by postulating an objective function and constraints, Fig. 1. Second, for inputs lying on a manifold, we derive an optimal offline solution and demonstrate analytically and numerically that the receptive fields are localized and tile the manifold, Fig. 1. Third, from the same objective, we derive an online optimization algorithm which can be implemented by a biologically plausible neural network, Fig. 1. We expect this network to learn localized receptive fields, the conjecture we confirm by simulating the network numerically, Fig. 1.

Optimization functions considered here belong to the family of similarity-preserving objectives which dictate that similar inputs to the network elicit similar outputs [8, 9, 10, 11, 12]. In the absence of sign constraints, such objectives are provably optimized by projecting inputs onto the principal subspace [13, 14, 15], which can be done online by networks of linear neurons [8, 9, 10]. Constraining the sign of the output leads to networks of rectifying neurons [11] which have been simulated numerically in the context of clustering and feature learning [11, 12, 16, 17], and analyzed in the context of blind source extraction [18]. In the context of manifold learning, sign-constrained optimal solutions of similarity-preserving objectives have been missing because optimization of existing similarity-preserving objectives is challenging.

Our main contributions are:

- Analytical sign-constrained optimization of similarity-preserving objectives for input originating from symmetric manifolds.
- Derivation of biologically plausible similarity-preserving, manifold-learning neural networks.
- Offline and online algorithms for manifold learning of arbitrary manifolds.

The paper is organized as follows. In Sec. 2, we derive a simplified version of a similarity-preserving objective with sign-constrained output. Much of our following analysis can be carried over to other similarity-preserving objectives but with additional technical considerations. In Sec. 3, we derive a necessary condition for the optimal solution. In Sec. 4, we consider solutions for the case of symmetric manifolds. In Sec. 5, we derive online optimization algorithm and a neural network. In Sec. 6, we present the results of numerical experiments.

2 A Simplified Similarity-preserving Objective Function

To introduce similarity-preserving objectives, let us define our notation. The input to the network is a set of vectors, $\mathbf{x}_t \in \mathbb{R}^n, t = 1, \ldots, T$, with components represented by the activity of n upstream neurons at time, t. In response, the network outputs an activity vector, $\mathbf{y}_t \in \mathbb{R}^m, t = 1, \ldots, T$, m being the number of output neurons.

Similarity preservation postulates that similar input pairs, $\mathbf{x}_t$ and $\mathbf{x}_{t'}$, evoke similar output pairs, $\mathbf{y}_t$ and $\mathbf{y}_{t'}$. If similarity of a pair of vectors is quantified by their scalar product and the distance metric is Euclidean, we have

$$\min_{\forall t \in \{1,\ldots,T\}:\, \mathbf{y}_t \in \mathbb{R}^m} \frac{1}{2} \sum_{t,t'=1}^{T} (\mathbf{x}_t \cdot \mathbf{x}_{t'} - \mathbf{y}_t \cdot \mathbf{y}_{t'})^2 = \min_{\mathbf{Y} \in \mathbb{R}^{m \times T}} \frac{1}{2} \|\mathbf{X}^\top \mathbf{X} - \mathbf{Y}^\top \mathbf{Y}\|_F^2, \tag{1}$$

where we introduced a matrix notation $\mathbf{X} \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{n \times T}$ and $\mathbf{Y} \equiv [\mathbf{y}_1, \ldots, \mathbf{y}_T] \in \mathbb{R}^{m \times T}$ and $m < n$. Such optimization problem is solved offline by projecting the input data to the principal subspace [13, 14, 15]. The same problem can be solved online by a biologically plausible neural network performing global linear dimensionality reduction [8, 10].

We will see below that nonlinear manifolds can be learned by constraining the sign of the output and introducing a similarity threshold, α:

$$\min_{\mathbf{Y} \geq 0} \frac{1}{2} \|\mathbf{X}^\top \mathbf{X} - \alpha \mathbf{E} - \mathbf{Y}^\top \mathbf{Y}\|_F^2 = \min_{\forall t:\, \mathbf{y}_t \geq 0} \frac{1}{2} \sum_{t,t'} (\mathbf{x}_t \cdot \mathbf{x}_{t'} - \alpha - \mathbf{y}_t \cdot \mathbf{y}_{t'})^2, \tag{2}$$

where $\mathbf{E}$ is a matrix of all ones. In the special case, $\alpha = 0$, Eq. (2) reduces to the objective in [11, 19, 18].

Intuitively, Eq. (2) attempts to preserve similarity for similar pairs of input samples but orthogonalizes the outputs corresponding to dissimilar input pairs. Indeed, if the input similarity of a pair of samples t, t' is above a specified threshold, $\mathbf{x}_t \cdot \mathbf{x}_{t'} > \alpha$, output vectors $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ would prefer to have $\mathbf{y}_t \cdot \mathbf{y}_{t'} \approx \mathbf{x}_t \cdot \mathbf{x}_{t'} - \alpha$, i.e., it would be similar. If, however, $\mathbf{x}_t \cdot \mathbf{x}_{t'} < \alpha$, the lowest value $\mathbf{y}_t \cdot \mathbf{y}_{t'}$ for $\mathbf{y}_t, \mathbf{y}_{t'} \geq 0$ is zero meaning that that they would tend to be orthogonal, $\mathbf{y}_t \cdot \mathbf{y}_{t'} = 0$. As $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ are nonnegative, to achieve orthogonality, the output activity patterns for dissimilar patterns would have non-overlapping sets of active neurons. In the context of manifold representation, Eq. (2) strives to preserve in the $\mathbf{y}$-representation local geometry of the input data cloud in $\mathbf{x}$-space and let the global geometry emerge out of the nonlinear optimization process.

As the difficulty in analyzing Eq. (2) is due to the quartic in $\mathbf{Y}$ term, we go on to derive a simpler quadratic in $\mathbf{Y}$ objective function that produces very similar outcomes. To this end, we, first, introduce an additional power constraint: $\operatorname{Tr} \mathbf{Y}^\top \mathbf{Y} \leq k$ as in [9, 11]. We will call the input-output mapping obtained by this procedure NSM-0, with NSM standing for **N**onnegative **S**imilarity-preserving **M**apping:

$$\operatorname*{argmin}_{\substack{\mathbf{Y} \geq 0 \\ \operatorname{Tr} \mathbf{Y}^\top \mathbf{Y} \leq k}} \frac{1}{2} \|\mathbf{X}^\top \mathbf{X} - \alpha \mathbf{E} - \mathbf{Y}^\top \mathbf{Y}\|_F^2 = \operatorname*{argmin}_{\substack{\mathbf{Y} \geq 0 \\ \operatorname{Tr} \mathbf{Y}^\top \mathbf{Y} \leq k}} -\operatorname{Tr}((\mathbf{X}^\top \mathbf{X} - \alpha \mathbf{E}) \mathbf{Y}^\top \mathbf{Y}) + \|\mathbf{Y}^\top \mathbf{Y}\|_F^2, \tag{NSM-0}$$

where we expanded the square and kept only the $\mathbf{Y}$-dependent terms.

We can redefine the variables and drop the last term in a certain limit (see the Supplementary Material, Sec. A.1, for details) leading to the optimization problem we call NSM-1:

$$\min_{\substack{\mathbf{Y} \geq 0 \\ \operatorname{diag}(\mathbf{Y}^\top \mathbf{Y}) \leq \beta \mathbf{I}}} -\operatorname{Tr}((\mathbf{X}^\top \mathbf{X} - \alpha \mathbf{E}) \mathbf{Y}^\top \mathbf{Y}) = \min_{\substack{\forall t \in \{1,\ldots,T\}: \\ \mathbf{y}_t \geq 0,\ \|\mathbf{y}_t\|_2^2 \leq \beta}} -\sum_{t,t'} (\mathbf{x}_t \cdot \mathbf{x}_{t'} - \alpha) \mathbf{y}_t \cdot \mathbf{y}_{t'}. \tag{NSM-1}$$

Intuitively, just like Eq. (2), NSM-1 preserves similarity of nearby input data samples while orthogonalizing output vectors of dissimilar input pairs. Conceptually, this type of objective has proven successful for manifold learning [20]. Indeed, a pair of samples t, t' with $\mathbf{x}_t \cdot \mathbf{x}_{t'} > \alpha$, would tend to have $\mathbf{y}_t \cdot \mathbf{y}_{t'}$ as large as possible, albeit with the norm of the vectors controlled by the constraint $\|\mathbf{y}_t\|^2 \leq \beta$. Therefore, when the input similarity for the pair is above a specified threshold, the vectors $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ would prefer to be aligned in the same direction. For dissimilar inputs with $\mathbf{x}_t \cdot \mathbf{x}_{t'} < \alpha$, the corresponding output vectors $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ would tend to be orthogonal, meaning that responses to these dissimilar inputs would activate mostly nonoverlapping sets of neurons.

3 A Necessary Optimality Condition for NSM-1

In this section, we derive the necessary optimality condition for Problem (NSM-1). For notational convenience, we introduce the Gramian $\mathbf{D} \equiv \mathbf{X}^\top \mathbf{X}$ and use $[\mathbf{z}]_+$, where $\mathbf{z} \in \mathbb{R}^T$, for the component-wise ReLU function, $([\mathbf{z}]_+)_t \equiv \max(z_t, 0)$.

Proposition 1. *The optimal solution of Problem (NSM-1) satisfies*

$$[(\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}]_+ = \mathbf{\Lambda}\mathbf{y}^{(a)}, \tag{3}$$

where $\mathbf{y}^{(a)}$ designates a column vector which is the transpose of the a-th row of $\mathbf{Y}$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_T)$ is a nonnegative diagonal matrix.

The proof of Proposition 1 (Supplementary Material, Sec. A.2) proceeds by introducing Lagrange multipliers $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_T) \geq 0$ for the constraint $\mathrm{diag}(\mathbf{Y}^\top\mathbf{Y}) \leq \beta\mathbf{I}$, and writing down the KKT conditions. Then, by separately considering the cases $\lambda_t y_{at} = 0$ and $\lambda_t y_{at} > 0$ we get Eq. (3).

To gain some insight into the nature of the solutions of Eq. (3), let us assume $\lambda_t > 0$ for all t and rewrite it as

$$y_{at} = \left[\frac{1}{\lambda_t}\sum_{t'}(D_{tt'} - \alpha)y_{at'}\right]_+. \tag{4}$$

Eq. (4) suggests that the interaction within each pair of $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ has a different sign, i.e., it is excitatory or inhibitory, depending on their similarity. If $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ are similar, $D_{tt'} > \alpha$, then $y_{at'}$ has excitatory influence on y_{at}. Otherwise, if $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ are farther apart, the influence is inhibitory. Such models often give rise to localized bump solutions [21]. Since, in our case, the variable y_{at} gives the activity of the a-th neuron as the t-th input vector is presented to the network, such a bump would define a localized receptive field of neuron, a, in the space of inputs. Below, we will derive such solutions with local receptive fields for inputs arising from symmetric manifolds.

4 Solution for Symmetric Manifolds via a Convex Formulation

So far, we set the dimensionality of $\mathbf{y}$, i.e., the number of output neurons, m, a priori. However, as this number depends on the dataset, we would like to allow for flexibility of choosing the output dimensionality adaptively. To this end, we introduce the Gramian, $\mathbf{Q} \equiv \mathbf{Y}^\top\mathbf{Y}$, and do not constrain its rank. Minimization of our objective functions requires that the output similarity expressed by Gramian, $\mathbf{Q}$, captures some of the input similarity structure encoded in the input Gramian, $\mathbf{D}$.

Redefining the variables makes the domain of the optimization problem convex. Matrices like $\mathbf{D}$ and $\mathbf{Q}$ which could be expressed as Gramians are symmetric and positive semidefinite. In addition, any matrix, $\mathbf{Q}$, such that $\mathbf{Q} \equiv \mathbf{Y}^\top\mathbf{Y}$ with $\mathbf{Y} \geq 0$ is called *completely positive*. The set of completely positive $T \times T$ matrices is denoted by $\mathcal{CP}^T$ and forms a closed convex cone [22].

Then, NSM-1, without the rank constraint, can be restated as a convex optimization problem with respect to $\mathbf{Q}$ belonging to the convex cone $\mathcal{CP}^T$:

$$\min_{\substack{\mathbf{Q}\in\mathcal{CP}^T \\ \mathrm{diag}(\mathbf{Q})\leq\beta\mathbf{I}}} -\mathrm{Tr}((\mathbf{D} - \alpha\mathbf{E})\mathbf{Q}). \tag{NSM-1a}$$

Despite the convexity, for arbitrary datasets, optimization problems in $\mathcal{CP}^T$ are often intractable for large T [22]. Yet, for $\mathbf{D}$ with a high degree of symmetry, below, we will find the optimal $\mathbf{Q}$.

Imagine now that there is a group $G \subseteq S_T$, S_T being the permutation group of the set $\{1, 2, \ldots, T\}$, so that $D_{g(t)g(t')} = D_{tt'}$ for all $g \in G$. The matrix with elements $M_{g(t)g(t')}$ is denoted as $g\mathbf{M}$, representing group action on $\mathbf{M}$. We will represent action of g on a vector $\mathbf{w} \in R^T$ as $g\mathbf{w}$, with $(g\mathbf{w})_t = w_{g(t)}$.

Theorem 1. *If the action of the group G is transitive, that is, for any pair $t, t' \in \{1, 2, \ldots, T\}$ there is a $g \in G$ so that $t' = g(t)$, then there is at least one optimal solution of Problem (NSM-1a) with $\mathbf{Q} = \mathbf{Y}^\top\mathbf{Y}, \mathbf{Y} \in \mathbb{R}^{m\times T}$ such that*

(i) for each a, the transpose of the a-th row of $\mathbf{Y}$, termed $\mathbf{y}^{(a)}$, satisfies

$$[(\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}]_+ = \lambda\mathbf{y}^{(a)}, \forall a \in \{1, 2, \ldots, m\}, \tag{5}$$

(ii) $\mathbf{Y}$ *could be written as*

$$\mathbf{Y}^\top = \frac{1}{\sqrt{|G/H|}}[g_1\mathbf{y}^{(1)} g_2\mathbf{y}^{(1)} \dots g_{|G/H|}\mathbf{y}^{(1)}], \tag{6}$$

where g_i *are members of distinct left cosets in* G/H*,* H *being the stabilizer subgroup of* $\mathbf{y}^{(1)}$*, namely* $H = \{h \in G | h\mathbf{y}^{(1)} = \mathbf{y}^{(1)}\}$*.*

In other words, when the symmetry group action is transitive, all the Lagrange multipliers are the same. Also the different rows of the $\mathbf{Y}$ matrix could be generated from a single row by the action of the group. A sketch of the proof is as follows (see Supplementary Material, Sec. A.3, for further details). For part (i), we argue that a convex minimization problem with a symmetry always has a solution which respects the symmetry. Thus our search could be limited to the G-invariant elements of the convex cone, $\mathcal{CP}^G = \{\mathbf{Q} \in \mathcal{CP}^T \mid \mathbf{Q} = g\mathbf{Q}, \forall g \in G\}$, which happens to be a convex cone itself. We then introduce the Lagrange multipliers and define the Lagrangian for the problem on the invariant convex cone and show that it is enough to search over $\mathbf{\Lambda} = \lambda\mathbf{I}$. Part (ii) follows from optimality of $\mathbf{Q} = \mathbf{Y}^\top\mathbf{Y}$ implying optimality of $\bar{\mathbf{Q}} = \frac{1}{|G|}\sum_g g\mathbf{Q}$.

Eq. (5) is a non-linear eigenvalue equation that can have more than one solution. In simple cases, those solutions are related to each other by symmetry. We will find such explicit solutions in the following subsections.

4.1 Solution for Inputs on the Ring with Cosine Similarity in the Continuum Limit

In this subsection, we consider the case where inputs, $\mathbf{x}_t$, lie on a one-dimensional manifold shaped as a ring centered on the origin:

$$\mathbf{x}_t = \left[\cos(\tfrac{2\pi t}{T}), \sin(\tfrac{2\pi t}{T})\right]^\top,$$

where $t \in \{1, 2, \dots, T\}$. Then, we have $D_{tt'} = \cos\left[\frac{2\pi(t-t')}{T}\right]$.

In the limit of large T, we can replace a discrete variable, t, by a continuous variable, θ: $\frac{2\pi t}{T} \to \theta$, $D_{tt'} = \cos\left[\frac{2\pi(t-t')}{T}\right] \to \cos(\theta - \theta')$, $y_t^a \to Cu_\phi(\theta)$ and $\lambda \to T\mu$ leading to

$$\left[\frac{1}{2\pi}\int_0^{2\pi}\left\{\cos(\theta-\theta') - \alpha\right\}u_\phi(\theta')d\theta'\right]_+ = \mu u_\phi(\theta), \tag{7}$$

with C adjusted so that $\int u_\phi(\theta)^2 dm(\phi) = 1$ for some measure m in the space of ϕ, which is a continuous variable labeling the output neurons. We will see that ϕ could naturally be chosen as an angle and the constraint becomes $\int_0^{2\pi} u_\phi(\theta)^2 d\phi = 1$.

Eq. (7) has appeared previously in the context of the ring attractor [21]. While our variables have a completely different neuroscience interpretation, we can still use their solution:

$$u_\phi(\theta) = A\left[\cos(\theta - \phi) - \cos(\psi)\right]_+ \tag{8}$$

whose support is the interval $[\phi - \psi, \phi + \psi]$.

Eq. (8) gives the receptive fields of a neuron, ϕ, in terms of the azimuthal coordinate, θ, shown in the bottom left panel of Fig. 1. The dependence of μ and ψ on α is given parametrically by two self-consistency conditions (Supplementary Material, Sec. A.4). So far, we have only shown that Eq. (8) satisfies the necessary optimality condition in the continuous limit of Eq. (7). In Sec. 6, we confirm numerically that the optimal solution for a finite number of neurons approximates Eq. (8), Fig. 2.

While we do not have a closed-form solution for NSM-0 on a ring, we show that the optimal solution also has localized receptive fields (see Supplementary Material, Sec. A.5).

4.2 Solution for Inputs on Higher-dimensional Compact Homogeneous Manifolds

Here, we consider two special cases of higher dimensional manifolds. The first example is the 2-sphere, $S^2 = SO(3)/SO(1)$. The second example is the rotation group, $SO(3)$, which is a

three-dimensional manifold. It is possible to generalize this method to other compact homogeneous spaces for particular kernels.

We can think of a 2-sphere via its 3-dimensional embedding: $S^2 \equiv \{\mathbf{x} \in \mathbb{R}^3 | \|\mathbf{x}\| = 1\}$. For two points Ω, Ω' on the 2-sphere let $\mathbf{D}(\Omega, \Omega') = \mathbf{x}(\Omega) \cdot \mathbf{x}(\Omega')$, where $\mathbf{x}(\Omega), \mathbf{x}(\Omega')$ are the corresponding unit vectors in the 3-dimensional embedding.

Remarkably, we can show that solutions satisfying the optimality conditions are of the form

$$u_{\Omega_0}(\Omega) = A\big[\mathbf{x}(\Omega_0) \cdot \mathbf{x}(\Omega) - \cos\psi\big]_+. \tag{9}$$

This means that the center of a receptive field on the sphere is at Ω_0. The neuron is active while the angle between $\mathbf{x}(\Omega)$ and $\mathbf{x}(\Omega_0)$ is less than ψ. For the self-consistency conditions, determining ψ, μ in terms of α, see Supplementary Material, Sec. A.6.

In the case of the rotation group, for $g, g' \in SO(3)$ we adopt the 3×3 matrix representations $\mathbf{R}(g), \mathbf{R}(g')$ and consider $\frac{1}{3}\operatorname{Tr}\left(\mathbf{R}(g)\mathbf{R}(g')^\top\right)$ to be the similarity kernel. Once more, we index a receptive field solution by the rotation group element, g_0, where the response is maximum:

$$u_{g_0}(\Omega) = \tfrac{A}{2}\left[\operatorname{Tr}\left(\mathbf{R}(g_0)^\top \mathbf{R}(g)\right) - 2\cos\psi - 1\right]_+ \tag{10}$$

with ψ, μ being determined by α through self-consistency equations. This solution has support over $g \in SO(3)$, such that the rotation gg_0^{-1} has a rotation angle less than ψ.

5 Online Optimization and Neural Networks

Here, we derive a biologically plausible neural network that optimizes NSM-1. To this end, we transform NSM-1 by, first, rewriting it in the Lagrangian form:

$$\min_{\forall t:\, \mathbf{y}_t \geq 0} \max_{\forall t:\, \mathbf{z}_t \geq 0} -\frac{1}{T}\sum_{t,t'}(\mathbf{x}_t \cdot \mathbf{x}_{t'} - \alpha)\mathbf{y}_t \cdot \mathbf{y}_{t'} + \sum_t \mathbf{z}_t \cdot \mathbf{z}_t(\beta - \mathbf{y}_t \cdot \mathbf{y}_t). \tag{11}$$

Here, unconventionally, the nonnegative Lagrange multipliers are factorized into inner products of two nonnegative vectors ($\mathbf{z}_t \cdot \mathbf{z}_t$). Second, we introduce auxiliary variables, $\mathbf{W}, \mathbf{b}, \mathbf{V}_t$ [10]:

$$\begin{aligned}\min_{\forall t:\, \mathbf{y}_t \geq 0} \max_{\forall t:\, \mathbf{z}_t \geq 0} \min_{\mathbf{W}} \max_{\mathbf{b}} \min_{\forall t:\, \mathbf{V}_t \geq 0} \; & T\operatorname{Tr}(\mathbf{W}^\top\mathbf{W}) - T\|\mathbf{b}\|_2^2 + \\ + \sum_t \Big(-2\mathbf{x}_t\mathbf{W}^\top\mathbf{y}_t + 2\sqrt{\alpha}\mathbf{y}_t \cdot \mathbf{b} + \beta\|\mathbf{z}_t\|_2^2 - 2\mathbf{z}_t\mathbf{V}_t\mathbf{y}_t + \operatorname{Tr}(\mathbf{V}_t^\top\mathbf{V}_t)\Big).\end{aligned} \tag{12}$$

This form suggests a two-step online algorithm. For each input $\mathbf{x}_t$, in the first step, one solves for $\mathbf{y}_t$, $\mathbf{z}_t$ and $\mathbf{V}_t$, by projected gradient descent-ascent-descent,

$$\begin{bmatrix}\mathbf{y}_t \\ \mathbf{z}_t \\ \mathbf{V}_t\end{bmatrix} \longleftarrow \begin{bmatrix}\mathbf{y}_t + \gamma_y\left(\mathbf{W}\mathbf{x}_t - \mathbf{V}_t^\top\mathbf{z}_t - \sqrt{\alpha}\mathbf{b}\right) \\ \mathbf{z}_t + \gamma_z\left(-\beta\mathbf{z}_t + \mathbf{V}_t\mathbf{y}_t\right) \\ \mathbf{V}_t + \gamma_V\left(\mathbf{z}_t\mathbf{y}_t^\top - \mathbf{V}_t\right)\end{bmatrix}_+, \tag{13}$$

where $\gamma_{y,z,V}$ are step sizes. This iteration can be interpreted as the dynamics of a neural circuit (Fig.1, Top right panel), where $\mathbf{y}_t$ are activity variables of an excitatory neuron population, $\mathbf{b}$ is a bias term, $\mathbf{z}_t$ are activity variables of an inhibitory neuron population, $\mathbf{W}$ is the feedforward connectivity matrix, and $\mathbf{V}_t$ is the synaptic weight matrix from excitatory to inhibitory neurons, which goes through a fast time-scale anti-Hebbian plasticity. In the second step, $\mathbf{W}$ and $\mathbf{b}$ are updated by a gradient descent-ascent step:

$$\mathbf{W} \longleftarrow \mathbf{W} + \eta\left(\mathbf{y}_t\mathbf{x}_t^\top - \mathbf{W}\right), \qquad \mathbf{b} \longleftarrow \mathbf{b} + \eta\left(\sqrt{\alpha}\mathbf{y}_t - \mathbf{b}\right), \tag{14}$$

where $\mathbf{W}$ is going through a slow time-scale Hebbian plasticity and $\mathbf{b}$ through homeostatic plasticity. η is a learning rate. Application of this algorithm to symmetric datasets is shown in Fig. 2 and Fig. 3.

6 Experimental Results

In this section, we solve both offline and online optimization problems numerically. Our results confirm the theoretical predictions in Sec. 4. Moreover, our algorithms yield manifold-tiling localized receptive fields on real-world data.

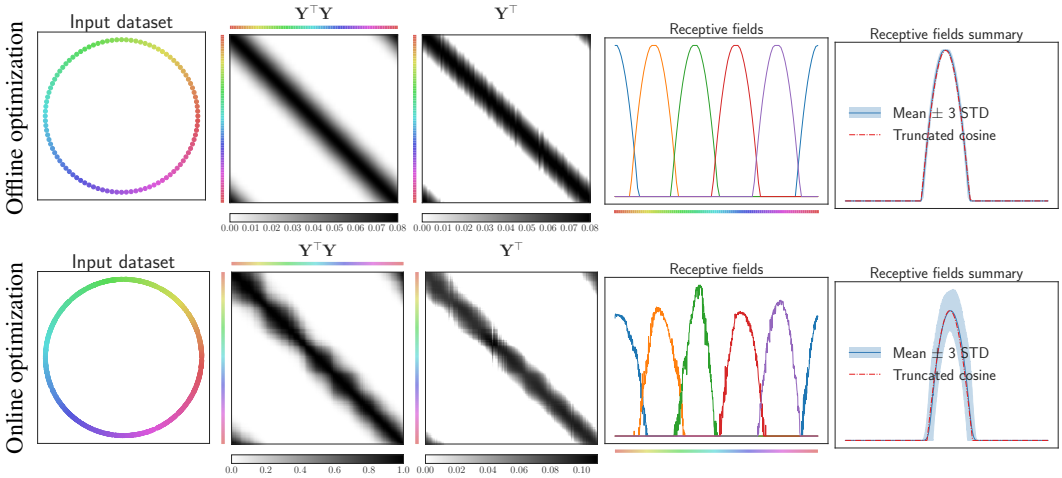

Figure 2: Action of NSM-1 on a 2d ring. We show, from left to right, the input dataset $\mathbf{X}$, the output similarity, the output neural activity matrix $\mathbf{Y}$, and a few localized receptive fields as 1d curves. On the right, we align all receptive fields and plot their mean plus/minus three standard deviations, showing that they are essentially the same object translated along the ring. Fitting a truncated cosine function provides an accurate description of the receptive fields.

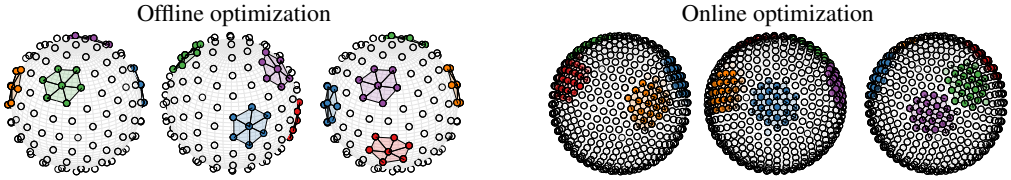

Figure 3: NSM tiles the sphere with overlapping localized receptive fields (soft-clusters), providing an accurate and useful data representation. We show a few receptive fields in different colors over three different views of the sphere. An advantage of the online optimization is that it can handle a larger number of points (in fact, arbitrarily large).

According to the theoretical results in Sec. 4, for the input data lying on a ring, optimization without a rank constraint yields truncated cosine solutions, see Eq. (8). Here, we show numerically that fixed-rank optimization yields the same solutions, Fig. 2: the computed matrix $\mathbf{Y}^\top\mathbf{Y}$ is indeed circulant, all receptive fields are equivalent to each other, are well approximated by truncated cosine and tile the manifold with overlap. Similarly, for the input lying on a 2-sphere, we find numerically that localized solutions tile the manifold, Fig. 3.

For the offline optimization we used a Burer-Monteiro augmented Lagrangian method [23, 24]. Traditionally, the number of rows m of $\mathbf{Y}$ is chosen to be βT (observe that $\operatorname{diag}(\mathbf{Y}^\top\mathbf{Y}) \leq \beta\mathbf{I}$ implies that $\operatorname{Tr}(\mathbf{Y}^\top\mathbf{Y}) \leq \beta T$, making βT an upper bound of the rank). We use the non-standard setting $m \gg \beta T$, as a small m might create degeneracies (i.e., hard-clustering solutions).

Dealing with real-world data. For normalized input data with every diagonal element $D_{tt} = \|\mathbf{x}_t\|_2^2$ above the threshold α, the term $\alpha \operatorname{Tr}(\mathbf{EQ}) = \alpha \sum_{tt'} \mathbf{y}_t \cdot \mathbf{y}_{t'}$ in NSM-1 behaves as described in Sec. 2. For unnormalized inputs, it is preferable to control the sum of each row of $\mathbf{Q}$, i.e. $\sum_{t'} \mathbf{y}_t \cdot \mathbf{y}_{t'}$, with an individual α_t, instead of the total sum.

Additionally, enforcing $\sum_t \|\mathbf{y}_t\|_2^2 \leq \beta T$ is in many cases empirically equivalent to enforcing $\|\mathbf{y}_t\|_2^2 \leq \beta$ but makes the optimization easier. We thus obtain the objective function

$$\min_{\substack{\forall t:\, \mathbf{y}_t \geq 0, \\ \sum_t \|\mathbf{y}_t\|_2^2 \leq \beta T}} -\sum_{t,t'} (\mathbf{x}_t \cdot \mathbf{x}_{t'} - \alpha_t) \mathbf{y}_t \cdot \mathbf{y}_{t'}, \tag{15}$$

which, for some choice of α_t, is equivalent to

$$\min_{\mathbf{Y} \geq 0} -\operatorname{Tr}(\mathbf{X}^\top\mathbf{X}\mathbf{Y}^\top\mathbf{Y}) \quad \text{s.t.} \quad \mathbf{Y}^\top\mathbf{Y}\mathbf{1} = \mathbf{1},\ \operatorname{Tr}(\mathbf{Y}^\top\mathbf{Y}) \leq \beta T, \tag{NSM-2}$$

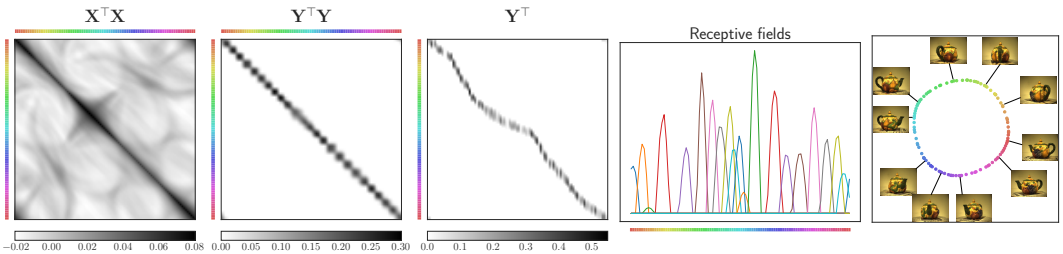

Figure 4: Manifold representation with NSM-2. The dataset is composed of 100 images obtained by viewing a teapot from different angles in a plane. The input size n is 23028 (76×101 pixels, 3 color channels). We build a 2d linear embedding (PCA) from the solution to NSM. The obtained 1d manifold uncovers the change in orientation (better seen with zoom), thanks to the overlapping localized receptive fields that tile it.

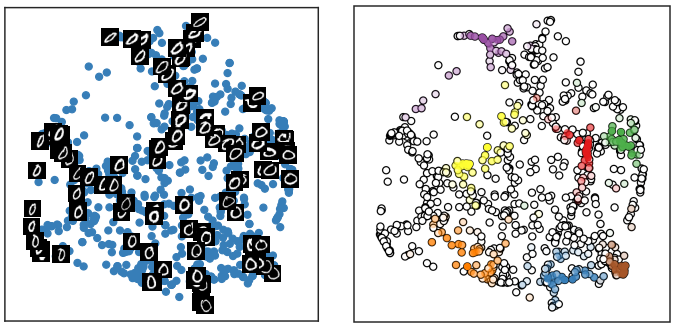

Figure 5: NSM-2 tiles the dataset formed by the digit 0 in MNIST (with input size, $n = 28 \times 28 = 784$) with localized receptive fields, providing an accurate and useful data representation. **Left:** Two-dimensional linear embedding (PCA) of $\mathbf{Y}$. The data gets organized according to different visual characteristics of the hand-written digit (e.g., orientation and elongation). **Right:** We show a few receptive fields in different colors over the low-dimensional embedding.

where $\mathbf{1} \in \mathbb{R}^T$ is a column vector of ones.

For highly symmetric datasets without constraints on rank, NSM-2 has the same solutions as NSM-1 (see Supplementary Material, Sec. A.7). Relaxations of this optimization problem have been the subject of extensive research to solve clustering and manifold learning problems [25, 26, 27, 28]. A biologically plausible neural network solving this problem was proposed in [12]. For the optimization of NSM-2 we use an augmented Lagrangian method [23, 24, 28, 29].

We have extensively applied NSM-2 to datasets previously analyzed in the context of manifold learning [28, 30] (see Supplementary Material, Sec. B). Here, we include just two representative examples, figs. 4 and 5, showing the emergence of localized receptive fields in a high-dimensional space. Despite the lack of symmetry and ensuing loss of regularity, we obtain neurons whose receptive fields, taken together, tile the entire data cloud. Such tiling solutions indicate robustness of the method to imperfections in the dataset and further corroborate the theoretical results derived in this paper.

7 Discussion

In this work, we show that objective functions approximately preserving similarity, along with nonnegativity constraint on the outputs, provide neurally plausible representation of manifolds. Neural networks implementing such algorithms rely only on local (Hebbian or anti-Hebbian) plasticity of synaptic strengths.

Our algorithms, starting from a linear kernel, $\mathbf{D}$, generate an output kernel, $\mathbf{Q}$, restricted to the sample space. Whereas the associations between kernels and neural networks was known [31], previously proposed networks used random synaptic weights with no learning. In our algorithms, the weights are

learned from the input data to optimize the objective. Therefore, our algorithms learn data-dependent kernels adaptively.

In addition to modeling biological neural computation, our algorithms may also serve as general-purpose mechanisms for generating representations of manifolds adaptively. Unlike most existing manifold learning algorithms [32, 33, 34, 35, 36, 37], ours can operate naturally in the online setting. Also, unlike most existing algorithms, ours do not output low-dimensional vectors of embedding variables but rather high-dimensional vectors of assignment indices to centroids tiling the manifold, similar to radial basis function networks [38]. The advantage of such high-dimensional representation becomes obvious if the output representation is used not for visualization but for further computation, e.g., linear classification [39].

Acknowledgments

We are grateful to Johannes Friedrich, Victor Minden, Eftychios Pnevmatikakis, and the other members of the Flatiron Neuroscience group for discussion and comments on an earlier version of this manuscript. We thank Sanjeev Arora, Moses Charikar, Jeff Cheeger, Surya Ganguli, Dustin Mixon, Afonso Bandeira, Marc'Aurelio Ranzato, and Soledad Villar for helpful discussions.

References

[1] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.

[2] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[3] Eric I Knudsen and Masakazu Konishi. Center-surround organization of auditory receptive fields in the owl. *Science*, 202(4369):778–780, 1978.

[4] Michael P Kilgard and Michael M Merzenich. Cortical map reorganization enabled by nucleus basalis activity. *Science*, 279(5357):1714–1718, 1998.

[5] Daniel E Feldman and Michael Brecht. Map plasticity in somatosensory cortex. *Science*, 310(5749):810–815, 2005.

[6] Takao K Hensch. Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience*, 6(11):877, 2005.

[7] Valentin Dragoi, Jitendra Sharma, and Mriganka Sur. Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, 28(1):287–298, 2000.

[8] Cengiz Pehlevan, Tao Hu, and Dmitri B Chklovskii. A hebbian/anti-hebbian neural network for linear subspace learning: A derivation from multidimensional scaling of streaming data. *Neural computation*, 27(7):1461–1495, 2015.

[9] Cengiz Pehlevan and Dmitri Chklovskii. A normative theory of adaptive dimensionality reduction in neural networks. In *Advances in neural information processing systems*, pages 2269–2277, 2015.

[10] Cengiz Pehlevan, Anirvan M Sengupta, and Dmitri B Chklovskii. Why do similarity matching objectives lead to hebbian/anti-hebbian networks? *Neural computation*, 30(1):84–124, 2018.

[11] Cengiz Pehlevan and Dmitri B Chklovskii. A hebbian/anti-hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features. In *Signals, Systems and Computers, 2014 48th Asilomar Conference on*, pages 769–775. IEEE, 2014.

[12] Cengiz Pehlevan, Alex Genkin, and Dmitri B Chklovskii. A clustering neural network model of insect olfaction. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 593–600, Oct 2017.

[13] Christopher KI Williams. On a connection between kernel pca and metric multidimensional scaling. In *Advances in neural information processing systems*, pages 675–681, 2001.

[14] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. CRC press, 2000.

[15] JM Bibby, JT Kent, and KV Mardia. Multivariate analysis, 1979.

[16] H Sebastian Seung and Jonathan Zung. A correlation game for unsupervised learning yields computational interpretations of hebbian excitation, anti-hebbian inhibition, and synapse elimination. *arXiv preprint arXiv:1704.00646*, 2017.

[17] Yanis Bahroun and Andrea Soltoggio. Online representation learning with single and multi-layer hebbian networks for image classification. In *International Conference on Artificial Neural Networks*, pages 354–363. Springer, 2017.

[18] Cengiz Pehlevan, Sreyas Mohan, and Dmitri B Chklovskii. Blind nonnegative source separation using biological neural networks. *Neural computation*, 29(11):2925–2954, 2017.

[19] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.

[20] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[21] R Ben-Yishai, R Lev Bar-Or, and H Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[22] Abraham Berman and Naomi Shaked-Monderer. *Completely positive matrices*. World Scientific, 2003.

[23] Samuel Burer, Kurt M. Anstreicher, and Mirjam Dür. The difference between $5\times$ doubly nonnegative and completely positive matrices. *Linear Algebra and its Applications*, 431(9):1539–1552, 2009.

[24] Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[25] Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *arXiv preprint arXiv:1406.5647*, 2014.

[26] Pranjal Awasthi, Afonso S Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *ITCS*, 2015.

[27] Jiming Peng and Yu Wei. Approximating K-means-type Clustering via Semidefinite Programming. *SIAM Journal on Optimization*, 18(1):186–205, jan 2007.

[28] Mariano Tepper, Anirvan M Sengupta, and Dmitri Chklovskii. Clustering is semidefinitely not that hard: Nonnegative SDP for manifold disentangling. *arXiv preprint arXiv:1706.06028*, 2017.

[29] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *NIPS*, 2016.

[30] Killan Q. Weinberger and Lawrence K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1683–1686, 2006.

[31] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *NIPS*, 2009.

[32] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500), 2000.

[33] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2000.

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

[35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[36] Kilian Q Weinberger and Lawrence K Saul. Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision*, 70(1):77–90, 2006.

[37] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[38] David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[39] Sanjeev Arora and Andrej Risteski. Provable benefits of representation learning. *arXiv preprint arXiv:1706.04601*, 2017.

[40] Christine Bachoc, Dion C Gijswijt, Alexander Schrijver, and Frank Vallentin. Invariant semidefinite programs. In *Handbook on semidefinite, conic and polynomial optimization*, pages 219–269. Springer, 2012.

Manifold-tiling Localized Receptive Fields are Optimal in Similarity-preserving Neural Networks

Supplementary Material

A Details of Derivations

A.1 Relation between NSM-0 and NSM-1

We will consider the optimization under an additional power constraint: $\operatorname{Tr}\mathbf{Y}^\top\mathbf{Y} \leq k$ as in [9, 11]. To arrive at a simpler objective function, let us first scale the current objective function by an overall of factor of $\frac{\beta}{kT}$, a procedure that does not affect the optimal $\mathbf{Y}$:

$$\underset{\substack{\mathbf{Y}\geq 0\\ \operatorname{Tr}\mathbf{Y}^\top\mathbf{Y}\leq k}}{\operatorname{argmin}} \frac{\beta}{2kT}\|\mathbf{X}^\top\mathbf{X}-\alpha\mathbf{E}-\mathbf{Y}^\top\mathbf{Y}\|_F^2 = \underset{\substack{\mathbf{Y}\geq 0\\ \operatorname{Tr}\mathbf{Y}^\top\mathbf{Y}\leq k}}{\operatorname{argmin}} -\frac{\beta}{kT}\operatorname{Tr}((\mathbf{X}^\top\mathbf{X}-\alpha\mathbf{E})\mathbf{Y}^\top\mathbf{Y})+\frac{\beta}{2kT}\|\mathbf{Y}^\top\mathbf{Y}\|_F^2. \tag{16}$$

Now we rescale the variables: $\tilde{\mathbf{Y}} = \sqrt{\frac{\beta T}{k}}\mathbf{Y}$. If we hold β fixed and let $k \to 0$, the optimal $\tilde{\mathbf{Y}}$ is given by

$$\lim_{k\to 0} \underset{\substack{\tilde{\mathbf{Y}}\geq 0\\ \operatorname{Tr}\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}\leq \beta T}}{\operatorname{argmin}} -\operatorname{Tr}((\mathbf{X}^\top\mathbf{X}-\alpha\mathbf{E})\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}})+\frac{kT}{2\beta}\|\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}\|_F^2 = \underset{\substack{\tilde{\mathbf{Y}}\geq 0\\ \operatorname{Tr}\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}\leq \beta T}}{\operatorname{argmin}} -\operatorname{Tr}((\mathbf{X}^\top\mathbf{X}-\alpha\mathbf{E})\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}, \tag{17}$$

since, in that limit, we can ignore the second term and just minimize the quadratic function of $\tilde{\mathbf{Y}}$.

We will work with the quadratic objective function discussed above. Switching back to $\mathbf{Y}$ from $\tilde{\mathbf{Y}}$ and applying a constraint on each diagonal element of $\mathbf{Y}^\top\mathbf{Y}$ rather than to $\operatorname{Tr}\mathbf{Y}^\top\mathbf{Y}$, we have:

$$\min_{\substack{\mathbf{Y}\geq 0\\ \operatorname{diag}(\mathbf{Y}^\top\mathbf{Y})\leq\beta\mathbf{I}}} -\operatorname{Tr}((\mathbf{D}-\alpha\mathbf{E})\mathbf{Y}^\top\mathbf{Y}) = \min_{\substack{\forall t\in\{1,2,\ldots,T\},\,\mathbf{y}_t\geq 0\\ \|\mathbf{y}_t\|_2^2\leq\beta}} -\sum_{t,t'}(\mathbf{x}_t\cdot\mathbf{x}_{t'}-\alpha)\mathbf{y}_t\cdot\mathbf{y}_{t'}. \tag{18}$$

A.2 Proof of Proposition 1

Proposition 1. *The optimal solution of Problem (NSM-1) satisfies*

$$[(\mathbf{D}-\alpha\mathbf{E})\mathbf{y}^{(a)}]_+ = \mathbf{\Lambda}\mathbf{y}^{(a)}, \tag{19}$$

where $\mathbf{y}^{(a)}$ refers to the transpose of the a-th row of $\mathbf{Y}$ as a column vector and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1,\ldots,\lambda_T)$ is a nonnegative diagonal matrix.

Proof. We introduce Lagrange multipliers $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1,\ldots,\lambda_T)$ for the constraint, $\operatorname{diag}(\mathbf{Y}^\top\mathbf{Y}) \leq \beta\mathbf{I}$:

$$\min_{\substack{\mathbf{Y}\geq 0\\ \operatorname{diag}(\mathbf{Y}^\top\mathbf{Y})\leq\beta\mathbf{I}}} -\operatorname{Tr}((\mathbf{D}-\alpha\mathbf{E})\mathbf{Y}^\top\mathbf{Y}) = \min_{\mathbf{Y}\geq 0}\max_{\mathbf{\Lambda}\geq 0}\operatorname{Tr}\left(-(\mathbf{D}-\alpha\mathbf{E})\mathbf{Y}^\top\mathbf{Y})+\mathbf{\Lambda}(\mathbf{Y}^\top\mathbf{Y}-\beta\mathbf{I})\right). \tag{20}$$

By varying $\mathbf{Y}\in\mathbb{R}^{m\times T}$ we derive a necessary optimality condition:

$$[(\mathbf{D}-\alpha\mathbf{E}-\mathbf{\Lambda})\mathbf{Y}^\top]_+ = 0. \tag{21}$$

The Lagrange multipliers $\lambda_1,\ldots,\lambda_T$ must be adjusted so that $\sum_a y_{at}^2 \leq \beta$, for each t, with $\lambda_t = 0$, when the inequality is strict, i.e., $\sum_a y_{at}^2 < 1$.

To simplify Eq. (21), we designate the transpose of the a-th row of $\mathbf{Y}$ as a column vector, $\mathbf{y}^{(a)}$. Each such $\mathbf{y}^{(a)}$ must satisfy the same equation:

$$[(\mathbf{D}-\alpha\mathbf{E}-\mathbf{\Lambda})\mathbf{y}^{(a)}]_+ = 0. \tag{22}$$

Note that $(\mathbf{\Lambda}\mathbf{y}^{(a)})_t = \lambda_t y_{at} \geq 0$. Now consider the two possibilities:

- If $\lambda_t y_{at} > 0$, then $y_{at} > 0$. Optimality requires that the partial derivative of Eq. (20) with respect to y_{at} is zero:
$$\big((\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}\big)_t = \lambda_t y_{at}. \tag{23}$$
Since the RHS has to be positive, the LHS has to be positive. We can, therefore, replace it by $[\big((\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}\big)_t]_+$, giving us
$$[\big((\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}\big)_t]_+ = \lambda_t y_{at}. \tag{24}$$
- If $\lambda_t y_{at} = 0$, we have
$$[\big((\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}\big)_t]_+ = 0. \tag{25}$$
In this case, we trivially get
$$[\big((\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}\big)_t]_+ = \lambda_t y_{at}. \tag{26}$$
Hence,
$$[(\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}]_+ = \mathbf{\Lambda}\mathbf{y}^{(a)}. \tag{27}$$

□

A.3 Proof of Theorem 1

Invariance under group action could be represented concisely with the help of matrices $\{\mathbf{R}(g) \,|\, g \in G\}$, which form the natural representation of the group G: $\mathbf{R}(g)_{tt'} = \delta_{g(t)t'}$, δ_{mn} being the Kronecker delta. The condition of symmetry for the problem is then stated very simply: $\mathbf{D} = \mathbf{R}(g)\mathbf{D}\mathbf{R}(g)^\top$.

Lemma 1. *To optimize Problem (NSM-1a), it is possible to restrict the search to the G-invariant convex cone $\mathcal{CP}^G = \{\mathbf{Q} \in \mathcal{CP}^T \,|\, \mathbf{Q} = \mathbf{R}(g)\mathbf{Q}\mathbf{R}(g)^\top, \forall g \in G\}$.*

Thanks to convexity of Problem (NSM-1a), there exists one optimal solution $\mathbf{Q}_0$ which is invariant under G, namely $\mathbf{R}(g)\mathbf{Q}_0\mathbf{R}(g)^\top = \mathbf{Q}_0$ for all $g \in G$ [40]. The proof relies on showing that if $\mathbf{Q}$ is an optimal solution then the group orbit average $\mathbf{Q}_0 = \frac{1}{|G|}\sum_{g\in G}\mathbf{R}(g)\mathbf{Q}\mathbf{R}(g)^\top$, an explicitly G-invariant matrix, has equal value of the objective function. Hence, it is enough to search for solutions that are invariant under the group G. Since we search for solutions that are invariant, we restrict ourselves to the convex cone of G-invariant $T \times T$ completely positive matrices: $\mathcal{CP}^G = \{\mathbf{Q} \in \mathcal{CP}^T \,|\, \mathbf{Q} = \mathbf{R}(g)\mathbf{Q}\mathbf{R}(g)^\top, \forall g \in G\}$.

Theorem 1. *If the action of the group G is transitive, that is, for any pair $t, t' \in \{1, 2, \ldots, T\}$ there is a $g \in G$ so that $t' = g(t)$, then there is at least one optimal solution of Problem (NSM-1a) with $\mathbf{Q} = \mathbf{Y}^\top\mathbf{Y}, \mathbf{Y} \in \mathbb{R}^{m\times T}$ such that*

(i) for each a, the transpose of the a-th row of $\mathbf{Y}$, termed $\mathbf{y}^{(a)}$, satisfies
$$[(\mathbf{D} - \alpha\mathbf{E})\mathbf{y}^{(a)}]_+ = \lambda\mathbf{y}^{(a)}, \forall a \in \{1, 2, \ldots, m\}, \tag{28}$$
(ii) $\mathbf{Y}$ could be written as
$$\mathbf{Y}^\top = \frac{1}{\sqrt{|G/H|}}[g_1\mathbf{y}^{(1)} g_2\mathbf{y}^{(1)} \ldots g_{|G/H|}\mathbf{y}^{(1)}], \tag{29}$$
where g_i are members of distinct left cosets in G/H, H being the stabilizer subgroup of $\mathbf{y}^{(1)}$, namely $H = \{h \in G | h\mathbf{y}^{(1)} = \mathbf{y}^{(1)}\}$.

Proof. Part (i): Let us go back the Lagrangian for Problem (NSM-1a):
$$\underset{\substack{\mathbf{Q}\in\mathcal{CP}^G \\ \mathrm{diag}(\mathbf{Q})\le\beta\mathbf{I}}}{\mathrm{argmin}} \; -\mathrm{Tr}((\mathbf{D} - \alpha\mathbf{E})\mathbf{Q}) = \underset{\mathbf{Q}\in\mathcal{CP}^G}{\mathrm{argmin}} \; \underset{\substack{\mathbf{\Lambda}=\mathrm{diag}(\lambda_1,\ldots,\lambda_T) \\ \mathbf{\Lambda}\ge 0}}{\mathrm{argmax}} \; -\mathrm{Tr}((\mathbf{D} - \alpha\mathbf{E})\mathbf{Q}) + \mathrm{Tr}(\mathbf{\Lambda}(\mathbf{Q} - \beta\mathbf{I})). \tag{30}$$
As we are only considering $\mathbf{Q}$ belonging to the G-invariant convex cone $\mathcal{CP}^G$, we can replace $\mathbf{Q}$ by its orbit average $\frac{1}{|G|}\sum_{g\in G}\mathbf{R}(g)\mathbf{Q}\mathbf{R}(g)^\top$ in the expression in Eq. (30). Looking at the $\mathrm{Tr}(\mathbf{\Lambda}(\mathbf{Q} - \beta\mathbf{I})$ term with this replacement, it is clear that in the Lagrangian, we could replace $\mathbf{\Lambda}$ by $\frac{1}{|G|}\sum_{g\in G}\mathbf{R}(g)\mathbf{\Lambda}\mathbf{R}(g)^\top$. So we can restrict ourselves to Lagrange multipliers that are G-invariant. In other words, we can assume $\mathbf{\Lambda} = \mathbf{R}(g)\mathbf{\Lambda}\mathbf{R}(g)^\top$.

Since the group action is transitive, the invariance implies that the diagonal elements of $\mathbf{\Lambda}$ are the same, giving us $\mathbf{\Lambda} = \lambda \mathbf{I}$. Now, the optimality condition becomes

$$[(\mathbf{D} - \alpha \mathbf{E})\mathbf{y}^{(a)}]_+ = \lambda \mathbf{y}^{(a)}. \tag{31}$$

$\boxed{\text{Part (ii):}}$ This part is straightforward if tedious. Imagine we have a G-invariant $\mathbf{Q} = \mathbf{Y}_0^T \mathbf{Y}_0 = \sum_{c=1}^{m_0} \mathbf{y}_0^{(c)} \mathbf{y}_0^{(c)\top}$ which is an optimal solution. The nonnegative vectors $\mathbf{y}_0^{(c)\top}$ form the rows of the matrix $\mathbf{Y}_0 \in R^{m_0 \times T}$, and could be assumed to be nonzero vectors without loss of generality.

Since $\mathbf{Q}$ is G-invariant we have,

$$\mathbf{Q} = \tfrac{1}{|G|} \sum_{g \in G} \mathbf{R}(g) \mathbf{Q} \mathbf{R}(g)^\top = \tfrac{1}{|G|} \sum_{g \in G} \mathbf{R}(g) \mathbf{Y}_0^T \mathbf{Y}_0 \mathbf{R}(g)^\top. \tag{32}$$

Rewrite the last matrix $\frac{1}{|G|} \sum_{g \in G} \mathbf{R}(g) \mathbf{Y}_0^T \mathbf{Y}_0 \mathbf{R}(g)^\top$ as $\sum_{c=1}^{m_0} \frac{1}{|G|} \sum_{g \in G} \mathbf{R}(g) \mathbf{y}_0^{(c)} (\mathbf{R}(g) \mathbf{y}_0^{(c)})^\top$. Let $\tilde{\mathbf{Q}}_c \equiv \frac{\mathrm{Tr}(\mathbf{Q})}{|G| \|\mathbf{y}_0^{(c)}\|_2^2} \sum_{g \in G} \mathbf{R}(g) \mathbf{y}_0^{(c)} (\mathbf{R}(g) \mathbf{y}_0^{(c)})^\top$.

Then $\mathrm{Tr}(\mathbf{Q}_c) = \mathrm{Tr}(\mathbf{Q})$ indicating the diagonal elements of $\mathbf{Q}_c$ are the same as that of $\mathbf{Q}$. Also $\mathbf{Q}$ is a convex combination of $\mathbf{Q}_c$'s.

$$\mathbf{Q} = \sum_{c=1}^{m_0} \tfrac{1}{|G|} \sum_{g \in G} \mathbf{R}(g) \mathbf{y}_0^{(c)} (\mathbf{R}(g) \mathbf{y}_0^{(c)})^\top = \sum_{c=1}^{m_0} \rho_c \mathbf{Q}_c, \tag{33}$$

where $\rho_c = \frac{\|\mathbf{y}_0^{(c)}\|_2^2}{\mathrm{Tr}(\mathbf{Q})} = \frac{\|\mathbf{y}_0^{(c)}\|_2^2}{\sum_{c'=1}^{m_0} \|\mathbf{y}_0^{(c')}\|_2^2}$. Note that $\sum_c \rho_c = 1$. We will show that one of the $\mathbf{Q}_c$ is as good a solution as $\mathbf{Q}$.

Let $c_0 = \mathrm{argmax}_c \, \mathrm{Tr}((\mathbf{D} - \alpha \mathbf{E})\mathbf{Q}_c)$. Since each $\mathbf{Q}_c$ satisfies the same constraints as $\mathbf{Q}$ and $-\mathrm{Tr}((\mathbf{D} - \alpha \mathbf{E})\mathbf{Q}) = -\sum_c \rho_c \, \mathrm{Tr}((\mathbf{D} - \alpha \mathbf{E})\mathbf{Q}_c) \geq -\mathrm{Tr}((\mathbf{D} - \alpha \mathbf{E})\mathbf{Q}_{c_0})$, $\mathbf{Q}_{c_0}$ is an optimal solution.

Now let $\mathbf{y}^{(1)} \equiv \frac{\sqrt{\mathrm{Tr}(\mathbf{Q})}}{\|\mathbf{y}_0^{(c_0)}\|_2} \mathbf{y}_0^{(c_0)}$. Construct $\mathbf{Y}$ such that

$$\mathbf{Y}^\top = \tfrac{1}{\sqrt{|G/H|}} [g_1 \mathbf{y}^{(1)} g_2 \mathbf{y}^{(1)} \dots g_{|G/H|} \mathbf{y}^{(1)}]$$

with g_i belonging to distinct cosets in G/H, where H is the stabilizer subgroup of $\mathbf{y}_1$.

With that, we have

$$\mathbf{Y}^\top \mathbf{Y} = \tfrac{1}{|G/H|} \sum_{g \in \{g_1, g_2, \dots, g_{|G/H|}\}} \mathbf{R}(g) \mathbf{y}^{(1)} (\mathbf{R}(g) \mathbf{y}^{(1)})^\top = \tfrac{1}{|G|} \sum_{g \in G} \mathbf{R}(g) \mathbf{y}^{(1)} (\mathbf{R}(g) \mathbf{y}^{(1)})^\top = \mathbf{Q}_{c_0}, \tag{34}$$

which is an optimal solution. □

A.4 Details of Self-consistency Condition in the Ring Solution

In this section we provide additional details about the derivations presented in Sec. 4.1. Let us focus on the solution for $\phi = 0$

$$\left[\tfrac{1}{2\pi} \int_0^{2\pi} \big\{ \cos(\theta - \theta') - \alpha \big\} u_0(\theta') d\theta' \right]_+ = \mu u_0(\theta). \tag{35}$$

Using $\cos(\theta - \theta') = \cos\theta \cos\theta' + \sin\theta \sin\theta'$ and our ansatz for $u_0(\theta')$ being even in θ', we get

$$\left[\tfrac{\cos\theta}{2\pi} \int_0^{2\pi} u_0(\theta') \cos\theta' d\theta' - \tfrac{\alpha}{2\pi} \int_0^{2\pi} u_0(\theta') d\theta' \right]_+ = \mu u_0(\theta). \tag{36}$$

The LHS is already in the form $A[\cos\theta - \cos\psi]_+$. Hence,

$$\cos\psi = \frac{\alpha \int_0^{2\pi} u_0(\theta') d\theta'}{\int_0^{2\pi} u_0(\theta') \cos\theta' d\theta'} = \frac{\alpha \int_0^{\psi} (\cos\theta' - \cos\psi) d\theta'}{\int_0^{\psi} (\cos\theta' - \cos\psi) \cos\theta' d\theta'}. \tag{37}$$

We now compute μ in terms of α, ψ by setting $\theta = 0$ in Eq. (36). Remembering that $u_0(\theta')$ is proportional to $[\cos\theta' - \cos\psi]_+$

$$\frac{1}{\pi}\int_0^{\psi}(\cos\theta' - \cos\psi)\cos\theta' d\theta' - \frac{\alpha}{\pi}\int_0^{\psi}(\cos\theta' - \cos\psi)d\theta' = \mu(1-\cos\psi). \tag{38}$$

Using Eq. (37), this can be simplified to

$$\mu = \tfrac{1}{\pi}\int_0^{\psi}(\cos\theta' - \cos\psi)\cos\theta' d\theta'. \tag{39}$$

Doing the integrals, we get the dependence of μ and ψ on α, given parametrically by

$$\alpha = \frac{\cos\psi(2\psi - \sin 2\psi)}{4(\sin\psi - \psi\cos\psi)}, \tag{40}$$

$$\mu = \tfrac{1}{4\pi}(2\psi - \sin 2\psi). \tag{41}$$

A.5 Optimality for NSM-0 and the Solution on the Ring

In this section we provide additional details about the derivations presented in Sec. 4.1. If we go back to a nonnegative version of similarity-preserving, we could rewrite the problem as:

$$\underset{\substack{\mathbf{Q}\in\mathcal{CP}\\ \mathrm{Tr}(\mathbf{Q})\le k}}{\mathrm{argmin}}\ \tfrac{1}{2}\,\mathrm{Tr}\left((\mathbf{D}-\alpha\mathbf{E}-\mathbf{Q})^2\right). \tag{NSM-0a}$$

We have chosen to use an inequality $\mathrm{Tr}(\mathbf{Q}) = \mathrm{Tr}(\mathbf{Y}^\top\mathbf{Y}) \le k$ rather than an equality to simplify the discussion. Note that these optimization problems minimize a convex function, in fact a quadratic one in $\mathbf{Q}$, in a convex region. Many of the arguments we have made so far about invariant problems apply to these problems as well. If we derive the optimality condition following our earlier arguments, we get

$$[(\mathbf{D}-\alpha\mathbf{E}-\mathbf{Q})\mathbf{y}^{(a)}]_+ = \lambda\mathbf{y}^{(a)}. \tag{42}$$

The choice of λ decides the constraint on $\mathrm{Tr}(\mathbf{Y}^\top\mathbf{Y})$. If we set λ to be near zero, we get $\mathbf{Q}$ as close to $\mathbf{D}-\alpha\mathbf{E}$ as possible for the elements where $D-\alpha\mathbf{E}$ is nonnegative. As we let λ grow bigger, $\mathbf{Y}$ becomes smaller. As λ gets closer to its maximum possible value, we have $\mathbf{D}-\alpha\mathbf{E} \gg \mathbf{Y}^\top\mathbf{Y} = \mathbf{Q}$. In that limit,

$$[(\mathbf{D}-\alpha\mathbf{E})\mathbf{y}^{(a)}]_+ = \lambda_m\mathbf{y}^{(a)}, \tag{43}$$

which is the nonnegative eigenvector problem we saw for NSM-1 when it has a symmetry group acting transitively. The Lagrange multiplier λ thus takes values between zero and λ_m.

In NSM-0, $\alpha\mathbf{E}+\mathbf{Q}$ provides the inhibitory interaction to make the localized bump solutions. The non-triviality of the analysis comes from having to solve $\mathbf{y}^{(a)}$'s to determine $\mathbf{Q}$ and *vice versa*. That self-consistency problem is non-trivial, in general.

For the ring model, thanks to the symmetries, we have $\mathbf{Q}$ to be a circulant matrix. In this case,we can solve the self-consistency problem numerically quite easily. As shown in Fig. 6, increasing λ reduces the size of the response but also alters the tuning curve. When $\lambda \approx \lambda_m$, $\mathbf{y}^{(a)}$ is approximately given by truncated cosines.

A.6 Details of NSM-1 Solutions for S^2 and for $SO(3)$

In this section we provide additional details about the derivations presented in Sec. 4.2. For S^2, we start from the equation

$$\frac{1}{4\pi}\left[\int_{S^2}\left\{\mathbf{x}(\Omega)\cdot\mathbf{x}(\Omega')-\alpha\right\}u_\sigma(\Omega')d^2\Omega'\right]_+ = \mu u_\sigma(\Omega). \tag{44}$$

This equation invariant under 3-dimensional rotations around the origin of the sphere. For $g \in SO(3)$, let the action of g take Ω to $g\Omega$, with $\mathbf{x}(g\Omega) = \mathbf{R}(g)\mathbf{x}(\Omega)$. $\mathbf{R}(g)$ is the 3×3 orthogonal matrix representation of the rotation. If $u_\sigma(\Omega)$ is a solution, so is $u_\sigma(g\Omega)$. We will use this feature to create a family of solutions by action of rotation group on a single solution.

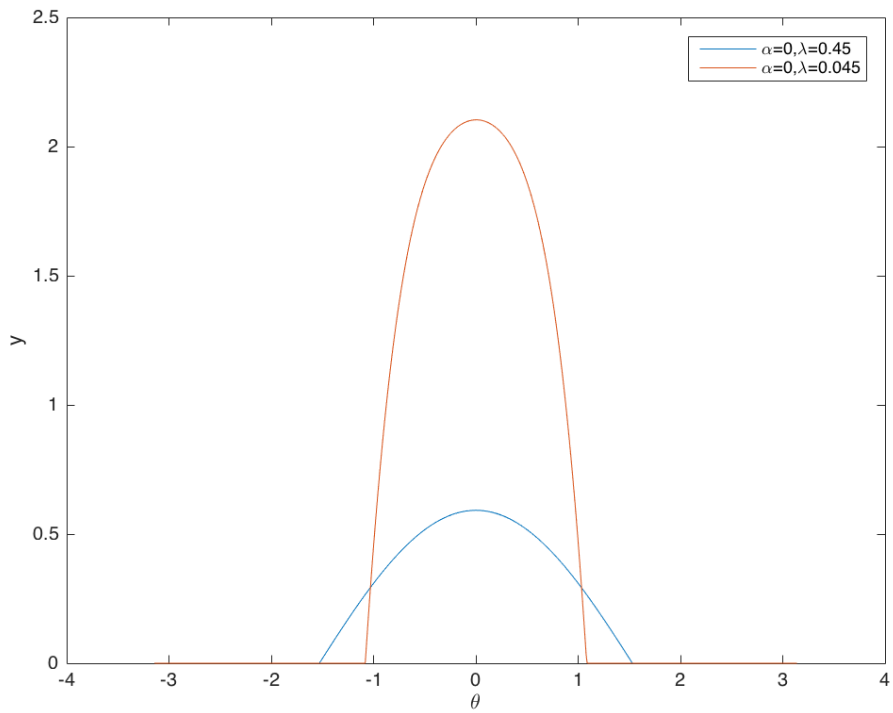

Figure 6: The receptive fields from NSM-0 for two different values of λ. We use $T = 628$, but plot the result against $\theta = 2\pi t/T - \pi$.

Since $\int_{S^2} \mathbf{x}(\Omega) \cdot \mathbf{x}(\Omega') u_\sigma(\Omega') d^2\Omega' = \mathbf{x}(\Omega) \cdot \left[\int_{S^2} \mathbf{x}(\Omega') u_\sigma(\Omega') d^2\Omega' \right]$, we will compute the vector $\int_{S^2} \mathbf{x}(\Omega') u_\sigma(\Omega') d^2\Omega'$. One way to describe the point Ω on the sphere is by using polar coordinates (θ, ϕ) , with $\mathbf{x}(\Omega) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$.

Imagine $u_\sigma(\Omega) = U(\theta)$, namely, it is independent of ϕ. In that case,

$$\int_{S^2} \mathbf{x}(\Omega') u_\sigma(\Omega') d^2\Omega' = \left(0, 0, 2\pi \int_0^\pi d\theta' \sin\theta' \cos\theta' U(\theta')\right), \tag{45}$$

implying that $\int_{S^2} \mathbf{x}(\Omega) \cdot \mathbf{x}(\Omega') u_\sigma(\Omega') d^2\Omega' \propto x_3(\Omega) = \cos(\theta)$. Thus,

$$\int_{S^2} \left\{ \mathbf{x}(\Omega) \cdot \mathbf{x}(\Omega') - \alpha) \right\} u_\sigma(\Omega') d^2\Omega' = a\cos\theta - b, \tag{46}$$

with $a = 2\pi \int_0^\pi d\theta' \sin\theta' \cos\theta' U(\theta')$ and $b = 2\alpha\pi \int_0^\pi d\theta' \sin\theta' U(\theta')$. Defining $A = a/\mu, B = b/\mu$, and using Eq. (44), we have

$$U(\theta) = \left[A\cos\theta - B \right]_+. \tag{47}$$

When $A > B$, $U(\theta) = A\left[\cos\theta - \cos\psi\right]_+$, which is non-zero only for $0 \le \theta \le \psi$. To get to the self-consistency conditions for the parameters, we go through procedures very similar to what we did for the ring. We ultimately get

$$\cos\psi = \frac{\alpha \int_0^\psi (\cos\theta' - \cos\psi) \sin\theta' d\theta'}{\int_0^\psi (\cos\theta' - \cos\psi) \cos\theta' \sin\theta' d\theta'} \tag{48}$$

and

$$\mu = \tfrac{1}{2} \int_0^\psi (\cos\theta' - \cos\psi) \cos\theta' \sin\theta' d\theta'. \tag{49}$$

Working out the integrals gives the parametric equations

$$\alpha = \tfrac{1}{3}\cos\psi(2+\cos\psi), \tag{50}$$

$$\mu = \tfrac{1}{3}\sin^4\frac{\psi}{2}(2+\cos\psi). \tag{51}$$

Moving over to the case of the rotation group, for $g, g' \in SO(3)$ we take the 3×3 matrix representations $\mathbf{R}(g), \mathbf{R}(g')$ and consider $\frac{1}{3}\operatorname{Tr}\left(\mathbf{R}(g)\mathbf{R}(g')^\top\right)$ to be the similarity kernel.

$$\left[\int_{SO(3)} \left\{\tfrac{1}{3}\operatorname{Tr}\left(\mathbf{R}(g)\mathbf{R}(g')^\top\right) - \alpha)\right\} u_\gamma(g')d^3g'\right]_+ = \mu u_\gamma(g). \tag{52}$$

The Haar measure normalized so that $\int_{SO(3)} dg' = 1$. Once more, we can find a solution with γ corresponding rotation group element g_0 where the response is maximum, i.e.,

$$u_{g_0}(\Omega) = \tfrac{A}{2}\left[\operatorname{Tr}\left(\mathbf{R}(g_0)^\top\mathbf{R}(g)\right) - 2\cos\psi - 1\right]_+, \tag{53}$$

with ψ, μ being determined by α through self-consistency equation. Since $\mathbf{R}(g_0)^\top\mathbf{R}(g) = \mathbf{R}(g_0^{-1}g)$ is a rotation, its trace is determined by the angle of rotation θ. Let us work out the self-consistency for the special case where g_0 is the identity element. $u_\gamma(g) = U(\theta)$, where θ is associated with rotation angle of g: $\mathbf{R}(g)$ has the eigenvalues $1, e^{i\theta}, e^{-i\theta}$, corresponding to eigenvectors $\hat{\mathbf{e}}_1, \frac{1}{\sqrt{2}}(\hat{\mathbf{e}}_2 + i\hat{\mathbf{e}}_3), \frac{1}{\sqrt{2}}(\hat{\mathbf{e}}_2 - i\hat{\mathbf{e}}_3)$, respectively. Integrating over g for the same θ corresponds to averaging over a random orthonormal basis $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$. Using that fact, we could argue that $\int_{SO(3)} \mathbf{R}(g')^\top u_\gamma(g')d^3g'$ is proportional to the identity matrix.

Hence,

$$\begin{aligned}
\mu U(\theta) =& \int_{SO(3)} \left\{\tfrac{1}{3}\operatorname{Tr}\left(\mathbf{R}(g)\mathbf{R}(g')^\top\right) - \alpha)\right\} u_\gamma(g')d^3g' \\
=& \tfrac{1}{3}\sum_{ij}\mathbf{R}(g)_{ij}\int_{SO(3)}\mathbf{R}(g')_{ij}u_\gamma(g')d^3g' - \alpha\int_{SO(3)}u_\gamma(g')d^3g' \\
=& \tfrac{1}{3}\sum_{i}\mathbf{R}(g)_{ii}\int_{SO(3)}\mathbf{R}(g')_{ii}u_\gamma(g')d^3g' - \alpha\int_{SO(3)}u_\gamma(g')d^3g' \\
=& \tfrac{1}{9}\operatorname{Tr}(\mathbf{R}(g))\int_{SO(3)}\operatorname{Tr}(\mathbf{R}(g'))u_\gamma(g')d^3g' - \alpha\int_{SO(3)}u_\gamma(g')d^3g' \\
=& \frac{1+2\cos\theta}{3}\int_0^\pi \frac{1+2\cos\theta'}{3}U(\theta')(1-\cos\theta')\frac{d\theta'}{\pi} - \alpha\int_0^\pi U(\theta')(1-\cos\theta')\frac{d\theta'}{\pi}.
\end{aligned} \tag{54}$$

The last step involves changing variables to θ' and getting the appropriate Jacobian $(1-\cos\theta')$, before integrating out the variable related to the rotation axis. The factor $\frac{1}{\pi}$ is needed to maintain that $\int_{SO(3)} dg' = \int_0^\pi (1-\cos\theta')\frac{d\theta'}{\pi} = 1$. Going through similar steps as before, and noting that $U(\theta') \propto [\cos\theta' - \cos\psi]_+$, we get the self-consistency conditions to be

$$1 + 2\cos\psi = \frac{9\alpha\int_0^\psi(\cos\theta'-\cos\psi)(1-\cos\theta')d\theta'}{\int_0^\psi(\cos\theta'-\cos\psi)(1+2\cos\theta')(1-\cos\theta')d\theta'}, \tag{55}$$

$$\mu = \frac{2}{9\pi}\int_0^\psi(\cos\theta'-\cos\psi)(1+2\cos\theta')(1-\cos\theta')d\theta'. \tag{56}$$

Doing these integrals gives us the parametric equations for the $SO(3)$ problem to be

$$\alpha = \frac{(1+2\cos\psi)(6\psi - 3\sin\psi - 3\sin 2\psi + \sin 3\psi)}{27(-2\psi + 4\sin\psi + \sin 2\psi - 4\psi\cos\psi)}, \tag{57}$$

$$\mu = \frac{1}{54\pi}(6\psi - 3\sin\psi - 3\sin 2\psi + \sin 3\psi). \tag{58}$$

A.7 Relation between NSM-1 and NSM-2 in Presence of Symmetry

Written in terms of $\mathbf{Q}$, without rank constraints, and with an inequality for $\operatorname{Tr}\mathbf{Q}$, NSM-2 is stated as

$$\min_{\mathbf{Q}\in\mathcal{CP}^T} -\operatorname{Tr}(\mathbf{DQ}) \quad \text{s.t.} \quad \mathbf{Q1}=\mathbf{1},\ \operatorname{Tr}(\mathbf{Q})\le\beta T. \tag{QNSM-2}$$

Introducing Lagrange multipliers, the Lagrangian version of QNSM-2 becomes

$$\min_{\mathbf{Q}\in\mathcal{CP}^T}\max_{\boldsymbol{\alpha}\in R^T}\max_{\lambda\ge 0} -\operatorname{Tr}(\mathbf{DQ})+\boldsymbol{\alpha}^\top(\mathbf{Q1}-\mathbf{1})+\lambda(\operatorname{Tr}(\mathbf{Q})-\beta T)). \tag{59}$$

When $\mathbf{D}$ is G-invariant, one can restrict the $\mathbf{Q}$ search to the G-invariant cone:

$$\min_{\mathbf{Q}\in\mathcal{CP}^G}\max_{\boldsymbol{\alpha}\in R^T}\max_{\lambda\ge 0} -\operatorname{Tr}(\mathbf{DQ})+\boldsymbol{\alpha}^\top(\mathbf{Q1}-\mathbf{1})+\lambda(\operatorname{Tr}(\mathbf{Q})-\beta T)). \tag{60}$$

The $\boldsymbol{\alpha}$ dependence comes in through the term $\boldsymbol{\alpha}^\top\mathbf{Q1}$. Since $\mathbf{Q}\in\mathcal{CP}^G$, we can play the trick of replacing $\mathbf{Q}$ by $\frac{1}{|G|}\sum_{g\in G}\mathbf{R}(g)\mathbf{QR}(g)^\top$.

$$\boldsymbol{\alpha}^\top\mathbf{Q1}=\tfrac{1}{|G|}\sum_{g\in G}\boldsymbol{\alpha}^\top\mathbf{R}(g)\mathbf{QR}(g)^\top\mathbf{1}=\left(\tfrac{1}{|G|}\sum_{g\in G}\boldsymbol{\alpha}^\top\mathbf{R}(g)\right)\mathbf{Q1}. \tag{61}$$

So we could replace $\boldsymbol{\alpha}$ by $\bar{\boldsymbol{\alpha}}=\frac{1}{|G|}\sum_{g\in G}\mathbf{R}(g^{-1})\boldsymbol{\alpha}$. When G acts transitively, $\bar{\boldsymbol{\alpha}}$ is a constant vector, namely $\alpha\mathbf{1}$. In that case, the optimization problem becomes

$$\min_{\mathbf{Q}\in\mathcal{CP}^G}\max_{\alpha\in R}\max_{\lambda\ge 0} -\operatorname{Tr}(\mathbf{DQ})+\alpha(\operatorname{Tr}(\mathbf{EQ})-T)+\lambda(\operatorname{Tr}(\mathbf{Q})-\beta T)), \tag{62}$$

The necessary condition for $\mathbf{Y}$, where $\mathbf{Q}=\mathbf{Y}^\top\mathbf{Y}$, is now

$$[(\mathbf{D}-\alpha\mathbf{E})\mathbf{Y}^\top]_+=\lambda\mathbf{Y}^\top, \tag{63}$$

which is the same equation we analyzed for G-invariant NSM-1.

B Additional NSM-2 Results

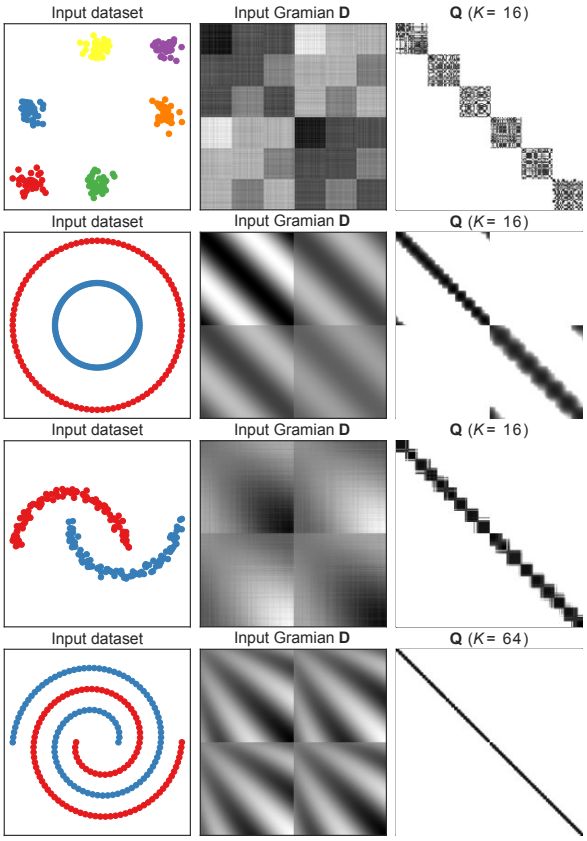

Figure 7: NSM-2 tiles manifolds: additional examples with synthetic datasets.

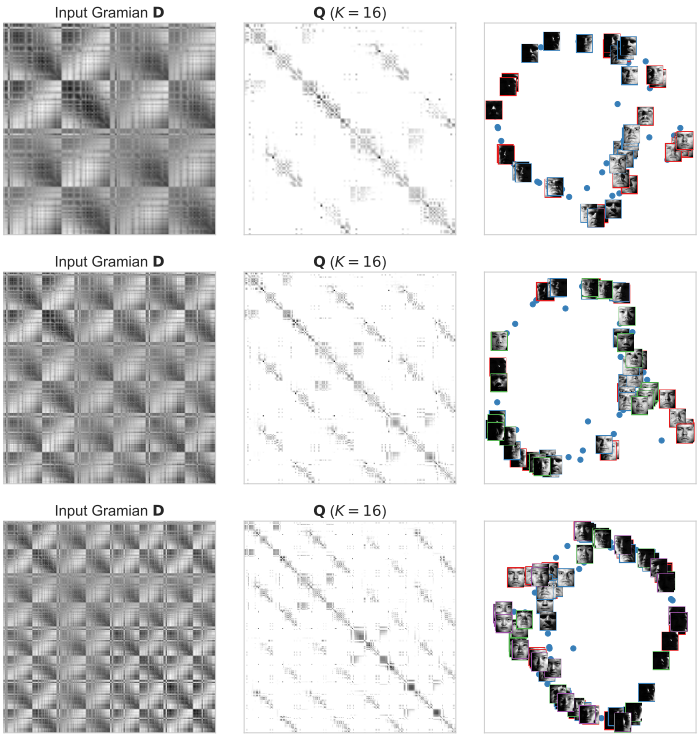

Figure 8: NSM-2 tiles manifolds: additional examples with the Yale Faces dataset. From top to bottom, using 2, 3, and 4 faces, respectively.

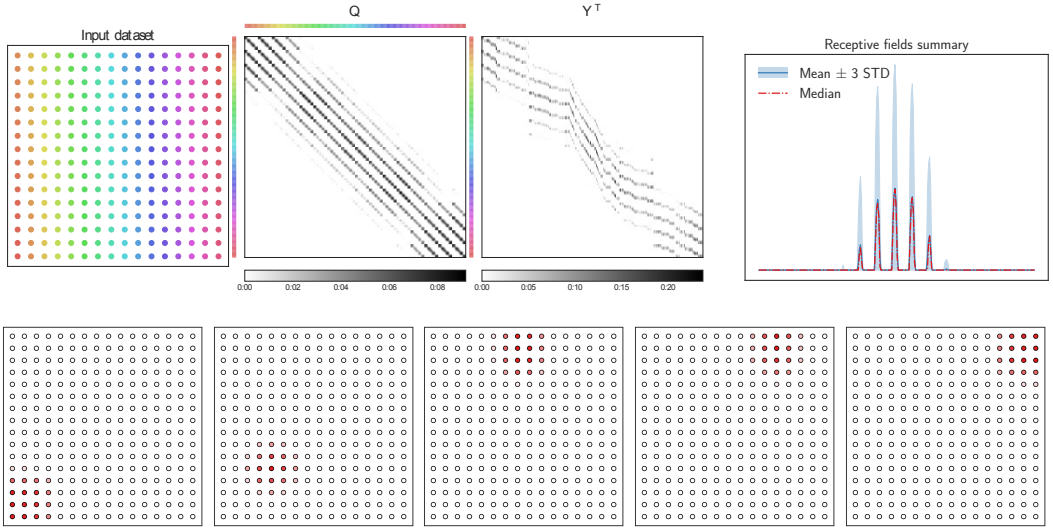

Figure 9: NSM-2 tiles manifolds: additional example with a synthetic 2d grid. **Bottom row:** A few selected receptive fields.