Equivariant Representation in Recurrent Networks with a Continuous Manifold of Attractors

Anonymous Author(s)
Affiliation
Address
email

Abstract

Equivariant representation is necessary for the brain and artificial perceptual systems to faithfully represent the stimulus under some (Lie) group transformations. However, it remains unknown how recurrent neural circuits in the brain represent the stimulus equivariantly, nor the neural representation of abstract group operators. The present study uses a one-dimensional (1D) translation group as an example to explore the general recurrent neural circuit mechanism of the equivariant stimulus representation. We found that a continuous attractor network (CAN), a canonical neural circuit model, self-consistently generates a continuous family of stationary population responses (attractors) that represents the stimulus equivariantly. Inspired by the Drosophila's compass circuit, we found that the 1D translation operators can be represented by extra speed neurons besides the CAN, where speed neurons' responses represent the moving speed (1D translation group parameter), and their feedback connections to the CAN represent the translation generator (Lie algebra). We demonstrated that the network responses are consistent with experimental data. Our model for the first time demonstrates how recurrent neural circuitry in the brain achieves equivariant stimulus representation.

1 Introduction

Neuronal responses in the brain internally represent the stimulus s in the world. When the stimulus changes due to some transformations, the brain updates its internal representation accordingly to faithfully reflects the stimulus change. For example, V1 neurons encode the visual orientation, and their population responses track the changing orientation which enables subsequent brain areas to decode the instant orientation. Transformations acting on a continuous stimulus can be described by the Lie group theory [1], and a neural representation reflecting the stimulus transformation is called group *equivariant* representation [2–4]. Mathematically, supposing a Lie group G acting on a continuous stimulus s, the evoked neural population response $\bar{u}(s)$ (a continuous function of s, and $\bar{u}$ denotes the mean response) is equivariant with the action of group G (group *homomorphism*) if,

$$\hat{T}_g \cdot \bar{u}(s) = \bar{u}(T_g \cdot s), \quad \forall g \in G, \forall s. \tag{1}$$

$\hat{T}_g$ and T_g are the induced actions (operators) with group element $g \in G$ which respectively act on neuronal responses and the stimulus, and $\cdot$ denotes the action of group operators. Intuitively, the equivariant neural representation (Eq. 1) implies the neuronal responses evoked from a transformed stimulus (e.g., rotating the orientation of Gabor images), i.e., $\bar{u}(T_g \cdot s)$ (Fig. 1A, blue path) would be the same as transforming the neuronal responses evoked by the original stimulus along the stimulus manifold, i.e., $\hat{T}_g \cdot \bar{u}(s)$ (Fig. 1A, green path).

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

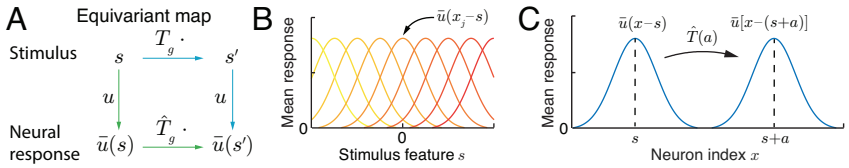

Figure 1: (A) An equivariant map between the stimulus and the neural representation. (B) Tuning curves (mean responses) of a population of homogeneous neurons in response to a one-dimensional stimulus feature s, where each neuron j has its unique value of preferred stimulus x_j. The tunings of neurons have the same profile and only differ in the preferred stimulus x_j. (C) A one-dimensional translation operator $\hat{T}(a)$ shifts the population response $\bar{u}(x-s)$ by amount a.

The equivariant representation is indispensable for the brain to represent the stimulus under transformations. An equivariant neural circuit is not only required to represent the stimulus equivariantly ($\bar{u}(s)$, Eq. 1), but it also needs to *explicitly* represents the Lie group operators ($\hat{T}_g$, Eq. 1). Explicitly representing group operators enables the brain to actively synthesize a transformed stimulus and predict the changing stimulus under transformations. For example, when tracking a moving object, the brain not only represents the object's instantaneous position s, but also represents its moving speed (the translation group parameter) in order to predict the stimulus position in the next time step. Yet it is an open question as to how recurrent neural circuits in the brain equivariantly represent the stimulus, and how abstract group operators are represented in the brain.

Neuroscientists utilized the Lie group to investigate the perception and representation in the brain (e.g., [5–13]), whereas nearly none of previous studies linked the abstract group theory to concrete biological recurrent circuit dynamics. Meanwhile, although equivariant neural networks have been extensively studied recently in machine learning (ML) research to capture transformations of stimuli (input data) defined on graphs (e.g., [14]) or defined on smooth manifolds (e.g., [3, 4, 15, 16]), nearly all these studies focused on developing feedforward weights which act as group convolution kernels in order to yield equivariant outputs under group transformations. There has not been much ML research studying equivariant recurrent networks, nor explicit network representations of group operators.

To provide theoretical insight into this fundamental question, we used a 1D translation group acting on a 1D stimulus feature as an example to investigate the general recurrent circuit mechanism of equivariant representation. We found the equivariant stimulus representation can be naturally realized in continuous attractor networks (CANs) which self-consistently generate a continuous family of stationary population responses (attractors) to represent the stimulus equivariantly. Moreover, the translation group operators are represented by separate neuronal populations whose responses represent the translation speed (translation group parameter). The feedback connections from speed neurons to CANs are shifted on the stimulus manifold, which encode the translation *generator* (Lie algebra). We compare the proposed recurrent circuit model with Drosophila's internal compass circuit to show its biological plausibility. The present study is one of the first studies linking the equivariant representation to a biologically plausible neural circuit model with rigorous mathematical analysis. It will shed light on the neural representations in the brain.

2 Translation group and homogeneous neural population code

We start from a concrete Lie group and neural representation to investigate the general circuit mechanism of equivariant representation. As an example, we consider a 1D translation Lie group $\mathbb{T}$ acting on a 1D stimulus s which can be regarded to be heading direction, orientation, moving direction, spatial location, etc. A group element $T(a) \in \mathbb{T}$ translates the stimulus s to $s+a$, i.e., $T(a) \cdot s = s + a$, and therefore an equivariant neural representation $\bar{u}(s)$ should satisfy (Fig. 1C),

$$\bar{u}\big[T(a) \cdot s\big] = \bar{u}(s+a) = \hat{T}(a) \cdot \bar{u}(s), \tag{2}$$

where $\hat{T}(a)$ is the 1D translation operator inducing the translation of neural representation (Fig. 1C). Several properties of the $\hat{T}(a)$ can be directly derived from the definition (Eq. 2) [17],

$$\hat{T}(0) = 1, \tag{3a}$$

$$\hat{T}(a)\hat{T}(b) = \hat{T}(a+b) = \hat{T}(b)\hat{T}(a), \tag{3b}$$

$$\hat{T}(a)^{-1} = \hat{T}(-a). \tag{3c}$$

First, a translation operator with zero translation should leave the neural representation of stimulus to be the same as no transformation (Eq. 3a). Second, the translation operator can be composed together and the composition is not dependent on the order of operators (Eq. 3b), which implies the translation operator is *commutative* or *abelian* [17]. Third, the inverse of a translation operator corresponds to an operator translating to the opposite direction with the same distance (Eq. 3c). Since the translation is continuous, the amount of translation can be made infinitesimally small. Consider an infinitesimal translation $\delta a \to 0$ on the stimulus s, then the corresponding change of neural representation is,

$$\hat{T}(\delta a) \cdot \bar{u}(s) = \bar{u}(s + \delta a) \approx \bar{u}(s) + \delta a \partial_s \bar{u}(s) = (1 + \delta a \partial_s)\bar{u}(s), \tag{4}$$

where $\partial_s = \partial/\partial s$ denotes the derivative over s, and a first order Taylor expansion was used in above derivation. $\hat{p} \equiv \partial_s$ is called the translation *generator* characterizing the tangential direction of translation in the group space, and forms a basis of the Lie *algebra*. In physics, $\hat{p}$ is usually regarded as the momentum operator [17]. Using the infinitesimal translation, the differential form of a translation operator can be derived as (see details of Lie group in Supplementary Information (SI)),

$$\frac{d\hat{T}(a)}{da} = \lim_{\delta a \to 0} \frac{\hat{T}(a + \delta a) - \hat{T}(a)}{\delta a} = \left(\lim_{\delta a \to 0} \frac{\hat{T}(\delta a) - 1}{\delta a} \right) \hat{T}(a) = \hat{p} \cdot \hat{T}(a). \tag{5}$$

By exponentiating the two sides in Eq. (5), the translation operator can be solved as $\hat{T}(a) = \exp(a\hat{p})$, which is an exponential map of the translation generator $\hat{p}$.

We next define a concrete neural representation based on experimental observations. The neural representation $\bar{u}(s) = \{\bar{u}_j(s)\}_{j=1}^N$ is regarded as the mean responses of a population of N neurons. We are interested in the limit of a large neuronal population ($N \to \infty$), and then $\bar{u}(s)$ effectively converges to a continuous function (a vector of infinite dimension). Inspired by the bell-shape stimulus tuning curves widely observed in the cortex, the mean response of the j-th neuron is modeled as a Gaussian function of the stimulus s (Fig. 1B),

$$\bar{u}_j(s) \equiv \bar{u}(x_j - s) = \mathsf{U}_j \exp[-(x_j - s)^2/2\sigma_j^2], \tag{6}$$

where x_j denotes the preferred stimulus of a neuron, indicating the neuronal response reaches its maximum when the stimulus matches its preferred value, i.e., $s = x_j$. U_j and σ_j denote respectively the peak firing rate and the tuning width of the j-th neuron. A necessary condition to construct an equivariant neural population code is *homogeneous* neurons in the population, which means the tuning curves of all neurons have the same profile, i.e., $\mathsf{U}_j \equiv \mathsf{U}$ and $\sigma_j \equiv \sigma$ for any j, and the preferred stimulus of all neurons, $\{x_j\}$, is uniformly distributed on the stimulus space. In this way, the mean responses of all neurons in response to a stimulus s are (Fig. 1C)

$$\bar{u}(s) \equiv \bar{u}(x - s) = \mathsf{U} \exp[-(x - s)^2/2\sigma^2], \tag{7}$$

which can be verified to satisfy the translation equivariant representation (Eq. 2).

3 Towards an equivariant recurrent neural circuit model

We next investigate how the equivariant homogeneous neural code, i.e., $\bar{u}(x - s)$ in Eq. (7), can be self-consistently generated by a concrete recurrent neural circuit dynamics. Specifically, we explore the recurrent circuit mechanism to generate a stable stimulus representation $\bar{u}(x - s)$ without translation (corresponding to a zero translation operator, $\hat{T}(0)$). Afterward, we investigate the neural representation of translation operators and how it induces the translation of stimulus representation.

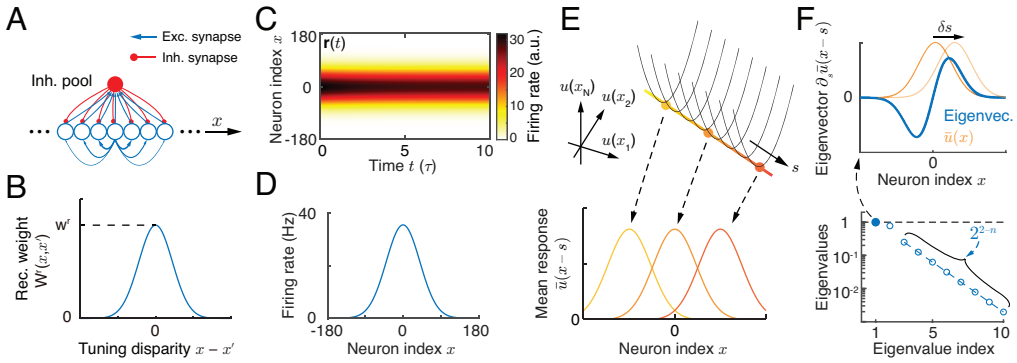

Figure 2: A continuous attractor network (CAN) self-consistently maintains an equivariant stimulus representation. (A) The CAN structure. Excitatory (E) neurons are aligned based on their preferred stimulus x, and a pool of inhibitory neurons provides negative feedback to ensure the network stability. (B) The recurrent connections between E neurons are translation-invariant with respect to the preferred stimulus x. (C-D) An example of spatiotemporal responses of all E neurons in the network model (C) and their temporal average (D). Note that the non-zero response is self-sustained by the network without external input. (E) The state space of CAN's population responses contains a one-dimensional continuous manifold, on which each location denotes a population response $\bar{u}(x-s)$ representing the value of stimulus feature s. Top: a schematic diagram of the energy landscape of the CAN. Bottom: three example population responses which correspond to three states on the continuous manifold. (F) Eigenanalysis of the perturbed CAN dynamics. Top: the eigenvector (blue) with the largest eigenvalue ($\lambda_1 = 1$) corresponds to an infinitesimal change of population responses (orange) along the stimulus feature manifold. Bottom: the eigenvalues of the perturbed CAN dynamics where the largest eigenvalue is 1, indicating neutral stability along the stimulus manifold.

3.1 The autonomous dynamics of continuous attractor networks

Since the recurrent neural circuit dynamics needs to generate non-zero mean responses $\bar{u}(x-s)$ for any value s (Eq. 6), the neural dynamics should be stable at any s. That is, there is a manifold of fixed points (attractors) in the state space of network dynamics. The continuous manifold of attractors can emerge from the autonomous dynamics (no external input drive) of continuous attractor networks (CANs, Fig. 2A), which is a canonical network model explaining the processing of continuous stimuli in neural circuits [18, 19], including heading-direction [20], orientation [21], spatial location [22] and so on. In the present study, we consider a CAN dynamics as follows [23, 24],

$$\tau \frac{\partial u(x,t)}{\partial t} = -u(x,t) + \rho \int \mathsf{W}_r(x,x') \cdot r(x',t) dx', \tag{8a}$$

$$r(x,t) = G[u(x,t)] = \frac{[u(x,t)]_+^2}{1 + k\rho \int [u(x',t)]_+^2 dx'}. \tag{8b}$$

where $u(x,t)$ and $r(x,t)$ denote respectively the synaptic input and firing rate of the excitatory neuron preferring stimulus $s = x$ at time t, and τ is the time constant of synaptic input. $G(\cdot)$ is an activation function that is modeled as divisive normalization where $[\cdot]_+$ is nonnegative rectification and k determines the global inhibition strength. The divisive normalization is a canonical operation widely observed in the cortex [25], and is resulted from the parvalbumin (PV) inhibitory neurons [26].

$\mathsf{W}_r(x,x')$ denotes the recurrent weight from the neuron preferring x' to the neuron preferring x, which is modeled as a Gaussian function of the disparity between the preferred stimuli of the pre- and post-synaptic neurons (Fig. 2B), i.e.,

$$\mathsf{W}_r(x,x') = \mathsf{W}_r(x-x') = \mathsf{w}_r(\sqrt{2\pi}\sigma)^{-1} \exp\left[-(x-x')^2/2\sigma^2\right], \tag{9}$$

where w_r is a scalar and determines the peak recurrent weight, and σ the width of connection in the stimulus space. Because we consider a large population of neurons that are uniformly distributed in the stimulus space, the neuronal interactions effectively become an integral in Eq. 8a, with ρ

denoting the neuronal density in the stimulus space. Furthermore, the translation-invariant recurrent connections, which only depend on $x - x'$, imply that the recurrent interactions are effectively conducting *convolution* on the stimulus space, i.e., $\int \mathsf{W_r}(x-x') \cdot r(x',t)dx' \equiv (\mathsf{W_r} * r)(x,t)$, where the symbol $*$ denotes the convolution over x.

3.2 A continuous attractor network equivariantly represents the stimulus

It can be checked that the CAN dynamics (Eqs. 8a-9) can generate a family of non-zero Gaussian-profile stationary responses that are consistent with the proposed responses (Eq. 6) [24, 27],

$$\bar{u}(x-s) = \mathsf{U} \exp\left[-(x-s)^2/4\sigma^2\right], \quad \bar{r}(x-s) = \mathsf{R} \exp\left[-(x-s)^2/2\sigma^2\right], \tag{10}$$

as long as the recurrent weight is stronger than a critical value which is solved as $\mathsf{w_c} = 2\sqrt{2}(2\pi)^{1/4}\sqrt{k\sigma/\rho}$ (see SI. Sec. 2). It is worth noting that s is a free parameter in Eq. (10), indicating that the network has a continuous family of fixed points, i.e., $\bar{u}(x-s)$, which self-consistently emerge from the recurrent neural dynamics without the influence from external inputs (Eqs. 8a-8b). The stationary population responses (Eq. 10) are consistent with the equivariant homogeneous neural population code (Eq. 7), and therefore they represent the stimulus equivariantly.

The stationary population responses in the recurrent network should be stable in order to maintain the equivariant stimulus representation robustly, otherwise, the population responses can be easily disrupted by noises. Therefore we check whether the continuous family of population responses (Eq. 10) are stable (attractors), by performing the stability analysis of the network dynamics. We perturbed the network state around the fixed points by adding a small perturbation $\delta u(x,t)$, i.e., $u(x,t) = \bar{u}(x-s) + \delta u(x,t)$, and then the dynamics of the perturbation is derived as [27],

$$\tau \frac{\partial}{\partial t}\delta u(x,t) = -\delta u(x,t) + \int K(x,x'|s)\delta u(x',t)dx'. \tag{11}$$

The interaction kernel $K(x,x'|s) = \rho \int \mathsf{W_r}(x-x'')\partial\bar{r}(x''-s)/\partial\bar{u}(x'-s)dx''$, whose analytical expression can be seen at SI. Sec. 3. Treating $K(x,x'|s)$ as an operator acting on the perturbation $\delta u(x,t)$, all of its eigenvalues, λ_n, and (unnormalized) eigenfunctions, f_n, can be analytically calculated (listed with a descending order) [27],

$$\lambda_1 = 1, \qquad f_1(x|s) \propto \partial_s \bar{u}(x-s), \tag{12a}$$

$$\lambda_2 = 1 - \sqrt{1 - \mathsf{w_c}^2/\mathsf{w_r}^2}, \quad f_2(x|s) \propto \bar{u}(x-s), \tag{12b}$$

$$\lambda_n = 2^{2-n} \ \ (n \geq 3), \qquad f_n(x|s) = \cdots . \tag{12c}$$

The explicit expression of eigenfunctions for $n \geq 3$ can be found at SI. Sec. 3. We see no eigenvalues are larger than 1 (Fig. 2F) indicating that the continuous family of fixed points ($\bar{u}(x-s)$ in Eq. 10) are stable, i.e., the network has a continuous family of attractors, and each attractor is an equivariant representation of the stimulus s (Fig. 2E).

4 Representing translation operators in CANs

It is noteworthy that the eigenfunction $f_1(x|s)$ of the perturbed network dynamics (Eq. 2F, top) is parallel to the direction of translation generator (Eq. 12a), i.e., $f_1(x|s) \propto \partial_s \bar{u}(x-s) = \hat{p} \cdot \bar{u}(x-s)$, thus $f_1(x|s)$ corresponds to the translation of neural responses along the continuous attractor manifold representing the stimulus s (Eq. 2F, top). Meanwhile, its corresponding eigenvalue $\lambda_1 = 1$ suggests the eigenfunction $f_1(x|s)$ is *neutrally* stable, which is resulted from the translation invariance of recurrent connections in the network (Eq. 9). Nevertheless, although the CAN dynamics maintains the translation operator resulted from the external perturbation $\delta u(x,t)$, it does not self-consistently support non-zero translation generators in that the stationary network states are static without any translation (Eq. 10, Fig. 2C-D), i.e., corresponding to $\hat{T}(0)$. We next explore a self-sustained network mechanism to represent non-zero translation operators, i.e., $\hat{T}(a),\ (a \neq 0)$ (Eq. 2), in order to induce translations of neuronal responses $\bar{u}(x-s)$.

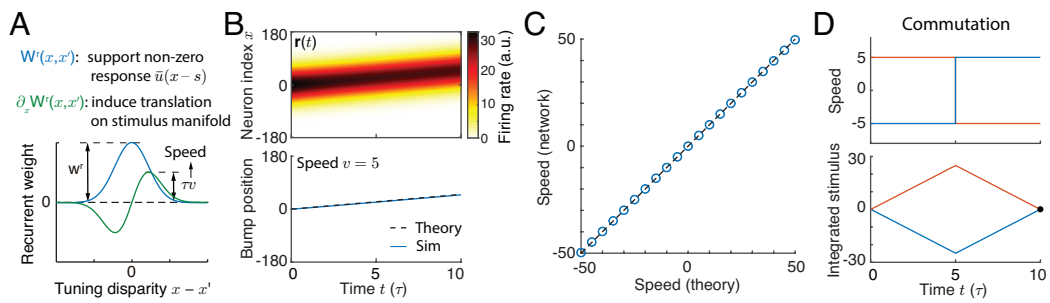

Figure 3: The derivative of recurrent connection, $\partial_x \mathsf{W_r}$, induces translation of CAN's population responses along the stimulus feature manifold, and represents the 1D translation operator. (A) Illustration of the recurrent connections in the network model. (B) An example of spatiotemporal responses in the network (top) and the internally represented stimulus feature moves over time with a constant speed. (C) The actual moving speed of population responses along the stimulus manifold is consistent with theoretical prediction by using the magnitude of $\partial_x \mathsf{W_r}$. (D) Commutative path integration in the CAN.

4.1 Incorporating translation operators in a CAN

To embed translation operators in the CAN dynamics, we derive a differential equation of Eq. (2) and compare it with the CAN dynamics. We consider the translation has instantaneous speed $v(t)$, which determines the translation amount in a time window dt as $da = v(t)dt$. Taking the derivative of Eq. (2) over time t, and utilizing the differential form of translation operators (Eq. 5), we derive that (to simplify notation we set $s = 0$ without loss of generality),

$$\frac{d\bar{u}(x-a)}{dt} = \frac{d\hat{T}(a)\cdot\bar{u}(x)}{da}\frac{da}{dt} = v\left[\frac{d\hat{T}(a)}{da}\bar{u}(x)\right] = v\hat{p}\cdot\hat{T}(a)\cdot\bar{u}(x) = v\hat{p}\cdot\bar{u}(x-a).$$

Replacing the variable $x - a$ by x, we reach the differential equation governing the translation of neuronal responses with speed $v(t)$,

$$\frac{d\bar{u}(x)}{dt} = v(t)\hat{p}\cdot\bar{u}(x). \tag{13}$$

In comparison, the CAN dynamics in the stationary state is $d\bar{u}(x)/dt = 0$ (Eq. 8a), i.e., the network state is static. To incorporate the translation operator with speed $v(t)$ (Eq. 13), we utilize that the stationary CAN state satisfying $\bar{u}(x-s) = \rho\mathsf{W_r} * \bar{r}(x-s)$, and substitute it to the right hand side of Eq. (13),

$$v\hat{p}\cdot\bar{u}(x) = v\partial_s[\mathsf{W_r} * \bar{r}(x)] = -v(\partial_x\mathsf{W_r}) * \bar{r}(x), \tag{14}$$

where we have used $\partial_s = -\partial_x$ (Eq. 7), and swapped the convolution and partial derivative. The above equation provides an insight that the translation generator can be converted into the derivative of recurrent connections in the network, i.e., $\partial_x \mathsf{W_r}$. Therefore, it motivates us to consider a modified CAN dynamics whose recurrent connections have a new component ($\partial_x \mathsf{W_r}$, Fig. 3A) [20, 28],

$$\tau\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \rho\left(\mathsf{W_r} - \tau v\partial_x\mathsf{W_r}\right) * r(x,t). \tag{15}$$

Next we verify whether the modified CAN dynamics (Eq. 15) embeds translation operators, i.e., a spontaneously moving neural response along the stimulus manifold with speed v. We substitute the mean population response ($\bar{u}(x-s)$ and $\bar{r}(x-s)$, Eq. 10) into the above network dynamics, repeatedly use the stationary response that satisfies $\bar{u}(x-s) = \rho\mathsf{W_r} * \bar{r}(x-s)$ to cancel terms,

$$\begin{aligned}\frac{\partial\bar{u}(x-s)}{\partial t} &= \tau^{-1}[-\bar{u}(x-s) + \rho\mathsf{W_r} * \bar{r}(x-s)] - v\rho(\partial_x\mathsf{W_r}) * \bar{r}(x-s),\\ &= -v\partial_x[\rho\mathsf{W_r} * \bar{r}(x-s)] = v\hat{p}\cdot\bar{u}(x-s).\end{aligned} \tag{16}$$

Comparing Eqs. (16) and (13), we see the stationary state in the modified CAN spontaneously translates along the stimulus manifold in a way consistent with the action of translation operators (Eq. 13). Therefore the recurrent connection component $\partial_x \mathsf{W}_\mathsf{r}$ represents the translation generator, and its magnitude determines the moving speed v. Numerical simulations of the network model also confirm our theoretical derivations (Fig. 3B-C), where the actual moving speed is well predicted by the theory using the magnitude of recurrent connection $\partial_x \mathsf{W}_\mathsf{r}$.

When the network response deviates from the stereotypical Gaussian profile (Eq. 10), the recurrent connection component W_r will remove distortions because the eigenvalues of all perturbations perpendicular to the continuous stimulus manifold are smaller than 1 (Eq. 12a - 12c). Once the population response settles into a smooth Gaussian profile, the network response translates along the continuous manifold with the desired speed (see details in SI. Sec. 4).

4.2 Path integration in the modified CAN dynamics

The brain accumulates the effects of a series of translation operators, and the combined translation effects are not dependent on their action orders (Eq. 3b), i.e., translations are commutative. The translation accumulation that is independent of action orders corresponds to the path integration in the neuroscience research [22, 28, 29], where the animals' navigation system integrates their walking path to update their walking distance. The path integration can be achieved in the modified CAN dynamics (Eq. 15). To illustrate this, we compute the translation distance in the modified CAN dynamics (Eq. 16) by using the chain rule, $\frac{\partial \bar{u}(x-s)}{\partial t} = \frac{\partial \bar{u}(x-s)}{\partial s}\frac{\partial s}{\partial t} = -\frac{\partial \bar{u}(x-s)}{\partial x} v(t)$, and substitute it back to Eq. (16). Finally, the translation distance is derived as

$$ds = v(t)dt, \tag{17}$$

where ds is the translation distance along the continuous stimulus manifold in the CAN in a time window lasting dt. As an example, we consider a constant moving speed v_1 lasts for time duration t_1, followed by another constant speed v_2 lasting for duration t_2, and then we swap the order of two speeds (Fig. 3D, top). The moving distance in the two cases can be calculated as the same,

$$|s| = \int_0^{t_1} v_1 dt + \int_{t_1}^{t_1+t_2} v_2 dt = \int_0^{t_2} v_2 dt + \int_{t_2}^{t_2+t_1} v_1 dt = v_1 t_1 + v_2 t_2, \tag{18}$$

which shows the translation in the modified CAN is commutative, i.e., the network achieves path integration. This is also confirmed by numerical simulations (Fig. 3D, bottom).

5 A biologically plausible equivariant recurrent circuit

Although in the modified CAN dynamics the recurrent connection component $\partial_x \mathsf{W}_\mathsf{r}$ (Eq. 15) mathematically represents the translation generator and translates network responses, the connection $\partial_x \mathsf{W}_\mathsf{r}$ is not biologically plausible due to two reasons. First, the magnitude of the connection component $\partial_x \mathsf{W}_\mathsf{r}$ needs to be multiplicatively modulated by the speed v (Eq. 15), but multiplicative modulation of synaptic weights is challenging in neural systems. Second, with Gaussian-profile connections (Eq. 9), the derivative $\partial_x \mathsf{W}_\mathsf{r}$ means that a neuron has both excitatory and inhibitory synapses, which disobeys Dale's law that a neuron can be either excitatory or inhibitory but cannot be both. Therefore, we propose a new neural circuit model to represent the translation operator that is consistent with the biological neural circuitry. Mathematically, we will replace the term of multiplicative synaptic weight modulation (Eq. 15, last term) by a biologically plausible input term that is originated from separate neuronal populations whose responses represent moving speed v.

The new circuit model is inspired by the Drosophila's compass system which consists of three neuronal populations (Fig. 4B, e.g., [30, 31]). There is a ring of heading-direction (stimulus s) neurons whose dynamics is governed by a CAN (Fig. 4B, E-PG neurons in blue disk, [32]). Besides, there are two populations of speed neurons (Fig. 4B, left and right P-EN neurons) whose firing rates encode the speed (translation distance a in a unit time). Moreover, the feedback projections from speed neurons to heading-direction neurons are either clockwise or anti-clockwise *shifted* and they induce translations of heading-direction neurons' responses. Therefore, our new recurrent circuit

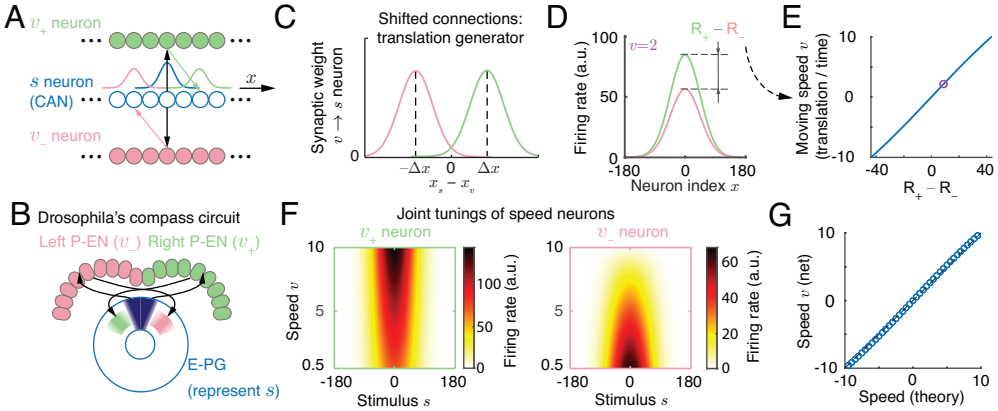

Figure 4: The network model with disentangled representations of stimulus s and speed v (amount of translation). (A-B) The proposed network structure (A) is similar to the Drosophila's internal compass (B). The three bumps shown in the middle represent the recurrent inputs from corresponding color coded neural populations. (C) The connection strength from speed neurons to stimulus neurons are shifted by Δx which represents the translation generator. (D-E) The difference of firing rate between two populations of speed neurons (D) determines the moving speed of population activities of stimulus neurons (E). (F) The joint tuning curves over stimulus and speed of two example neurons. (G) The proposed network implements the path integration with consistent speed as theory.

model has three populations of neurons (Fig. 4A): the stimulus s neurons which is modeled as a CAN as before (Eqs. 8a-8b), and two populations of speed neurons which are supposed to encode the positive and negative speed respectively. The network dynamics is governed by [30, 33–35],

$$\tau \frac{\partial u(x,t)}{\partial t} = -u(x,t) + \rho \mathsf{W}_\mathsf{r} * r(x,t) + \rho \sum\nolimits_{m=\pm} \mathsf{W}_m * r_m(x,t), \tag{19a}$$

$$\tau \frac{\partial u_\pm(x,t)}{\partial t} = -u_\pm(x,t) + \mathsf{w}_{vs} r(x,t), \quad r_\pm(x,t) = [(g_v \pm v) u_\pm(x,t)]_+. \tag{19b}$$

Eq. (19a) is the CAN dynamics of stimulus neurons. $u_\pm(x,t)$ ($r_\pm(x,t)$) are the synaptic inputs (firing rate) of speed neurons which represent positive and negative speeds respectively and are driven by the stimulus neuron $r(x,t)$ preferring stimulus x. Note that stimulus neurons receive feedback inputs from speed neurons (Eq. 19a, last term), which replaces recurrent inputs from biologically implausible recurrent weight $\partial_x \mathsf{W}_\mathsf{r}$ (Eq. 15, last term). Consistent with the Drosophila's compass [30, 31, 33], recurrent connections $\mathsf{W}_\pm(x,x')$ from the speed neuron x' to the stimulus neuron x is shifted by $\mp\Delta x$ towards the purported moving direction encoded by speed neurons (Fig. 4C),

$$\mathsf{W}_\pm(x,x') = \mathsf{w}_{sv}(\sqrt{2\pi}\sigma)^{-1} \exp\left[-(x - x' \mp \Delta x)^2/2\sigma^2\right]. \tag{20}$$

Moreover, speed neurons' firing rates $r_\pm(x,t)$ are multiplicatively modulated by the speed v (Eq. 19b). This operation is necessary to reproduce gain modulation of speed neurons' responses (speed invariant stimulus tuning width) in present rate-based model [30], and it achieves the same computation with multiplicative modulation of synaptic weights (Eq. 15, last term). g_v is a constant determining the baseline activity of speed neurons. To emphasize the main computational mechanism, we simplify the circuit model (Eqs. 19a-19b) compared with the Drosophila's circuit. First, we ignored the interactions between speed neurons. Second, we assumed that a speed neuron is only driven by a stimulus neuron with the same preferred stimulus, which is exhibited by the scalar w_{vs} (Eq. 19b).

5.1 Neural representation of translation operators

We investigate how the abstract translation operator $\hat{T}(a)$ is represented in the circuit model inspired by the Drosophila's compass. Similar to the theoretical analysis performed in Eqs. (15-16), we focus on the latent network dynamics along the dominant continuous stimulus manifold by ignoring the

fluctuations along other directions. Then Eq. (19a) can be simplified into (see more details in SI. 5),

$$\frac{\partial}{\partial t}\bar{u}(x-s) \approx \frac{\rho \mathsf{w}_{sv}}{\sqrt{2}\tau \mathsf{U}}(\mathsf{R}_+ - \mathsf{R}_-)\left[\Delta x\, \hat{p} \cdot \bar{u}(x-s)\right], \tag{21}$$

where R_+ and R_- are the peak firing rates of positive and negative speed neurons, and U is the peak synaptic input received by stimulus neurons (Eq. 10). The shifted connections $\mathsf{W}_\pm(x, x')$ from speed neurons to stimulus neurons introduce the translation generator, in that the magnitude of translation generator is proportional to the connection shift Δx (Eq. 16). The excitatory $\mathsf{W}_\pm(x, x')$ pulls the network response towards shifted directions (implementing the excitatory part of the translation generator), while the inhibitory part of the translation generator is achieved through the network effect of global inhibition (Eq. 8b). Moreover, the firing rate difference between two populations of speed neuron, $\mathsf{R}_+ - \mathsf{R}_-$, determines the translation speed (1D translation operator parameter). Overall, the fixed translation generator $\hat{p}$ is stored in the shifted connections from speed to stimulus neurons, and the time-varying translation speed is represented by the firing rate difference between two speed neuron populations which can change in a short time scale.

Meanwhile, the speed neurons' dynamics (Eq. 19b) suggests $\mathsf{R}_+ - \mathsf{R}_- = [(g_v + v) - (g_v - v)] = 2v\mathsf{w}_{vs}\mathsf{R}$. To translate network responses with the desired speed, the coefficient on the right hand side of Eq. (21) should equal to the speed determining speed neurons' response (Eq. 19b), which requires,

$$\sqrt{2}\rho \mathsf{w}_{sv}\mathsf{w}_{vs}\mathsf{R}\Delta x = \tau \mathsf{U}. \tag{22}$$

We performed numerical simulation of the circuit model (Eqs. 19a-19b), whose parameters are set to satisfy Eq. (22). We firstly verified our theoretical prediction, i.e., the moving speed is proportional to the firing rate difference between two populations of speed neurons, i.e., $\mathsf{R}_+ - \mathsf{R}_-$ (Fig. 4D-E). Indeed, the actual moving speed of stimulus neurons' population responses is consistent with the required speed that determines the firing rate of speed neurons (Fig. 4G), which supports the validity of our theoretical analysis and indicates the accurate path integration in the proposed circuit model. The joint tuning curves of two example speed neurons with respect to the stimulus s and moving speed v are plotted in Fig. 4F, which are multiplicatively modulated by the speed, i.e., the speed invariant stimulus tuning width, and this is consistent with experimental observations [30].

6 Conclusion and Discussion

Equivariant representation is necessary for the brain and artificial perceptual systems to reliably represent the stimulus under transformations [2–6, 36]. The present study investigates how recurrent neural circuits implement equivariant stimulus representation and Lie group operator representation. Our theoretical analysis shows that the stimulus can be equivariantly represented in a homogeneous neural population code that collectively emerges from a CAN dynamics. The Lie translation group operators are represented by extra neuronal populations whose activities represent the translation speed (translation group parameter) and whose feedback connections to the CAN store the translation generator. Our study for the first time links the Lie group equivariance to a concrete, biologically plausible recurrent neural circuit. It gives us insight into understanding the neural coding principle and may inspire a new building block for equivariant representation in ML tasks.

The CAN is a canonical neural circuit model widely used in neuroscience [19, 21, 24, 37, 38]. Recurrent circuit models similar to the present model were proposed before (e.g., [28, 30, 33–35]), whereas no previous studies theoretically linked the network dynamics to the equivariant stimulus representation, nor the Lie group operator representation. Besides, our theory also derives the condition that network responses translate with the desired speed (Eq. 22), which is testable by analyzing experimental data.

To elucidate the concepts clearly, the present study only investigates the 1D translation group while the proposed framework can be generalized to more complicated group transformations. For example, we may develop a recurrent circuit with a 2D translation group equivariant representation to explain the spatial location coding in place and grid cells (e.g., [29, 39, 40]), and (non-commutative) 3D rotation group equivariant recurrent circuits to explain the 3D rotation encoding in the brain. These studies form our future research.

References

[1] Anthony Zee. *Group theory in a nutshell for physicists*, volume 17. Princeton University Press, 2016.

[2] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

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

[4] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pages 2747–2755. PMLR, 2018.

[5] William C Hoffman. The lie algebra of visual perception. *Journal of mathematical Psychology*, 3(1):65–98, 1966.

[6] Peter C Dodwell. The lie transformation group model of visual perception. *Perception & Psychophysics*, 34(1):1–16, 1983.

[7] Jack L Gallant, Jochen Braun, and David C Van Essen. Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259(5091):100–103, 1993.

[8] RPN Rao and DL Ruderman. Learning lie transformation groups for invariant visual perception. *M. Kearns, SA Solla, & D. Cohn, Ad6ances in neural information processing systems*, 11, 1999.

[9] Xu Miao and Rajesh PN Rao. Learning the lie groups of visual invariance. *Neural computation*, 19(10):2665–2693, 2007.

[10] Joel Z Leibo, Qianli Liao, Fabio Anselmi, Winrich A Freiwald, and Tomaso Poggio. View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation. *Current Biology*, 27(1):62–67, 2017.

[11] Andrea Tacchetti, Leyla Isik, and Tomaso A Poggio. Invariant recognition shapes neural representations of visual input. *Annual review of vision science*, 4:403–422, 2018.

[12] Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):1–14, 2021.

[13] Yanis Bahroun, Dmitri Chklovskii, and Anirvan Sengupta. A similarity-preserving network trained on transformed images recapitulates salient features of the fly motion detection circuit. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[15] Taco Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, pages 1755–1763. PMLR, 2014.

[16] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

[17] David J Griffiths and Darrell F Schroeter. *Introduction to quantum mechanics*. Cambridge university press, 2018.

[18] Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, 27(2):77–87, 1977.

[19] Si Wu, KY Michael Wong, CC Alan Fung, Yuanyuan Mi, and Wenhao Zhang. Continuous attractor neural networks: candidate of a canonical model for neural information representation. *F1000Research*, 5, 2016.

[20] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *The Journal of Neuroscience*, 16(6):2112–2126, 1996.

[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] Bruce L McNaughton, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. Path integration and the neural basis of the'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006.

[23] Sophie Deneve, Peter E Latham, and Alexandre Pouget. Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745, 1999.

[24] Si Wu, Kosuke Hamaguchi, and Shun-ichi Amari. Dynamics and computation of continuous attractors. *Neural Computation*, 20(4):994–1025, 2008.

[25] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[26] Cristopher M Niell. Cell types, circuits, and receptive fields in the mouse visual cortex. *Annual review of neuroscience*, 38:413–431, 2015.

[27] C. C Alan Fung, K. Y. Michael Wong, and Si Wu. A moving bump in a continuous manifold: A comprehensive study of the tracking dynamics of continuous attractor neural networks. *Neural Computation*, 22(3):752–792, 2010.

[28] John B Issa and Kechen Zhang. Universal conditions for exact path integration in neural systems. *Proceedings of the National Academy of Sciences*, 109(17):6716–6720, 2012.

[29] Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2):e1000291, 2009.

[30] Daniel Turner-Evans, Stephanie Wegener, Herve Rouault, Romain Franconville, Tanya Wolff, Johannes D Seelig, Shaul Druckmann, and Vivek Jayaraman. Angular velocity integration in a fly heading circuit. *Elife*, 6:e23496, 2017.

[31] Jonathan Green, Atsuko Adachi, Kunal K Shah, Jonathan D Hirokawa, Pablo S Magani, and Gaby Maimon. A neural circuit architecture for angular integration in drosophila. *Nature*, 546(7656):101–106, 2017.

[32] Sung Soo Kim, Hervé Rouault, Shaul Druckmann, and Vivek Jayaraman. Ring attractor dynamics in the drosophila central brain. *Science*, 356(6340):849–853, 2017.

[33] Anna Kutschireiter, Melanie A Basnak, Rachel I Wilson, and Jan Drugowitsch. A bayesian perspective on the ring attractor for heading-direction tracking in the drosophila central complex. *bioRxiv*, 2021.

[34] William Skaggs, James Knierim, Hemant Kudrimoti, and Bruce McNaughton. A model of the neural basis of the rat's sense of direction. *Advances in neural information processing systems*, 7, 1994.

[35] Xiaohui Xie, Richard HR Hahnloser, and H Sebastian Seung. Double-ring network model of the head-direction system. *Physical Review E*, 66(4):041902, 2002.

[36] Mario Livio. Why symmetry matters. *Nature*, 490(7421):472–473, 2012.

[37] H Sebastian Seung and Haim Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90(22):10749–10753, 1993.

[38] James J Knierim and Kechen Zhang. Attractor dynamics of spatially correlated neural activity in the limbic system. *Annual review of neuroscience*, 35:267–285, 2012.

[39] Ben Sorscher, Gabriel Mel, Surya Ganguli, and Samuel Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in neural information processing systems*, 32, 2019.

[40] Ruiqi Gao, Jianwen Xie, Xue-Xin Wei, Song-Chun Zhu, and Ying Nian Wu. On path integration of grid cells: Group representation and isotropic scaling. *Advances in Neural Information Processing Systems*, 34, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sections 5 and 6.
   (c) Did you discuss any potential negative societal impacts of your work? [No] The present paper is a theoretical study investigating the computation of neural circuits in the brain, and thus it won't have negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We clearly state the assumptions in theoretical analysis, and simplifications of the model compared with real neural circuits in the brain across the manuscript.
   (b) Did you include complete proofs of all theoretical results? [Yes] The main text only lists the idea of proof and theoretical results. Detailed proofs and mathematical derivations are presented in the Supplementary Information.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The simulation details are presented in Supplementary Information and the code is also uploaded.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] They are presented in simulation details in Supplementary Information.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]