Associative memory networks for graph-based abstraction

Tatsuya Haga and Tomoki Fukai

Okinawa Institute of Science and Technology, Onna-son, Okinawa, Japan

correspondence: tatsuya.haga3@oist.jp (TH), tfukai@oist.jp (TF)

Abstract

Our cognition relies on the ability of the brain to segment hierarchically structured events on multiple scales. Recent evidence suggests that the brain performs this event segmentation based on the structure of state-transition graphs behind sequential experiences. However, the underlying circuit mechanisms are only poorly understood. In this paper, we propose an extended attractor network model for the graph-based hierarchical computation, called as Laplacian associative memory. This model generates multiscale representations for communities (clusters) of associative links between memory items, and the scale is regulated by heterogenous modulation of inhibitory circuits. We analytically and numerically show that these representations correspond to graph Laplacian eigenvectors, a popular method for graph segmentation and dimensionality reduction. Finally, we demonstrate that our model with asymmetry exhibits chunking resembling to hippocampal theta sequences. Our model connects graph theory and the attractor dynamics to provide a biologically plausible mechanism for abstraction in the brain.

Introduction

The brain builds a hierarchical knowledge structure through the abstraction of groups and segments. This ability of the brain is essential for various cognitive functions such as chunking of items which increases the number of items kept in a limited capacity of working memory¹, segmentation of words which is essential for learning and comprehension of language²⁻⁴, and temporal abstraction of repeated sequential actions which accelerates reinforcement learning⁵.

Experimental evidence suggest that the brain performs segmentation based on the graph structures behind experiences. When the brain experiences a sequence of events, it learns temporal associations between successive events and eventually captures the structure of the state-transition graph behind the experience. It has been shown that the event segmentation performed by human subjects behaviorally reflects community structures (or clusters) of such state-transition graphs and, neurobiologically, sensory events within a same community are

represented by more similar activity patterns than those belonging to different communities^{6,7}. Computationally, such graph segmentation of events is considered to benefit the temporal abstraction of actions in reinforcement learning^{8,9}. Furthermore, graph-based representations can explain many characteristics of place cells and entorhinal grid cells¹⁰.

Despite their behavioral and representational evidence, the biological mechanism that creates graph-based representations remains unknown. Conventionally, the circuit-level mechanisms in hippocampal and cortical processing have been modeled as attractor-based associative memory networks^{11–13}. Experiments have revealed some hallmarks of associative memory networks such as Hebbian learning (as spike-timing dependent plasticity)^{14,15}, pattern completion and attractor states^{16–19} in the brain. In the context of associative memory, temporal associations between successive events can be modeled as hetero-associations between successively activated cell assemblies through Hebbian learning^{20–23}. This learning scheme creates correlated attractors from originally uncorrelated stimuli. Correlations depends on the temporal distance between memorized events along the event sequence, which quantitatively agrees with neural recordings from the monkey brain^{24,25}. Such correlated attractors, and hence the class of associative memory models, can be potentially extended to offer a biologically plausible representational basis for more general graphical structures. However, this hypothesis has not been examined.

In this study, we propose a generalized class of associative memory networks^{20–23} that performs graph-based segmentation and abstraction. We make two major extensions here. First, we generalize the one-dimensional sequential structure of temporal associations in the conventional model^{20–23} to arbitrary symmetric graphs. Second, we allow the model to have negative associative weights which can be interpreted as assembly-specific inhibition²⁶. We found that this network generates mixed representations that are shared by multiple memory items within same communities in the graph, which fits with human experiments^{6,7}. We mathematically revealed that attractors in our model are related to graph Laplacian eigenvectors, a popular mathematical method to perform graph segmentation^{27,28} and nonlinear dimensionality reduction²⁹. Because of this property, we call our model as Laplacian associative memory (LAM), and demonstrate that LAM is applicable to problems related to graph Laplacian eigenvectors such as subgoal finding in reinforcement learning^{8,9}. The scale of representations (the size of represented communities) is modulated by the relative strength of local and global inhibitory circuits, which indicates an active role of target-specific inhibition²⁶ and inhomogeneous neuromodulation of inhibitory circuits^{30,31}. Finally, we show that LAM with asymmetric weight bias generates chunked sequential activities observed in

hippocampus^{32,33}. To our knowledge, our model describes the first theoretical results that connect associative memory networks and graph theory, providing a biologically plausible dynamical mechanism for the hierarchical abstraction in the brain.

Results

Laplacian associative memory model

Laplacian associative memory (LAM) is a novel class of Hopfield-type recurrent network model^{11–13,20,23}. Let us define a network of N binary units $x_i(t)$ ($i = 1, \dots, N$; $0 \leq x_i(t) \leq 1$) with the dynamics:

$$\dot{x}_i = -x_i + \Theta\left(\sum_{j=1}^N w_{ij}x_j\right), \quad (1)$$

where w_{ij} are synaptic weights and $\Theta(x)$ is a step function ($\Theta(x) = 1$ if $x > 0$, otherwise $\Theta(x) = 0$). We assume that each memory item (e.g. sensory stimuli, places or events) is represented by a 0-1 binary random memory pattern ξ_i^μ ($i = 1, \dots, N$; $\mu = 1, \dots, P$) with sparsity p ($P[\xi_i^\mu = 1] = p$). We set synaptic weights from these memory patterns as

$$w_{ij} = \frac{1}{NV} \sum_{\mu=1}^P \sum_{\nu=1}^P (\alpha \delta_{\mu\nu} + H_{\mu\nu}) \tilde{\xi}_i^\mu \tilde{\xi}_j^\nu - \frac{1}{N} (\alpha + 1) \gamma, \quad (2)$$

where $\tilde{\xi}_i^\mu = \xi_i^\mu - P^{-1} \sum_{\mu} \xi_i^\mu$ and $V = p(1 - p)$. The term $\alpha \delta_{\mu\nu}$ represents auto-association within each item, where $\delta_{\mu\nu}$ is Kronecker delta and α is a modifiable parameter that determines the strength of auto-association. On the other hand, $H_{\mu\nu}$ is hetero-associative weights between memory items μ and ν ($H_{\mu\mu} = 0$). The parameter $\gamma \geq 0$ gives an additional global inhibitory effect¹³. In short, this network stores multiple cell assemblies (P memory patterns) through auto-associative Hebbian learning and further links them through hetero-associative Hebbian learning (Figure 1, left). Assuming Hebbian learning between successively activated cell assemblies^{22,25}, we construct hetero-associative weights from a normalized adjacency matrix of a state-transition graph, or generally, other graphs such as semantic relationships (see Methods for the detail). Thus, the structure of hetero-associative links is a graph reflecting the statistical structure behind experiences in which we may find some communities.

LAM can be regarded as the generalization of previous associative memory models. When $\alpha > 0$ and all $H_{\mu\nu}$ are zero, LAM is analogous to conventional Hopfield-type model storing biased memory patterns^{11–13}. If only adjacent items are associated ($H_{\mu,\mu+1} = H_{\mu+1,\mu} > 0$, and

all other $H_{\mu\nu}$ are zero) so that associative links form a one-dimensional chain, the model coincides with an associative memory model for a temporal sequence^{20,23}. However, unlike those previous models, LAM can also take other arbitrary hetero-associative link structures. Furthermore, we do not restrict the parameter α to be positive, allowing inhibitory auto-association. We found unique behaviors of LAM mostly in the regime of negative auto-association, which has not been extensively investigated previously.

We clarify biological interpretation of the model by decomposition of excitatory and inhibitory components. As in a previous work²³, we can decompose the weights as

$$w_{ij} = w_{ij}^E - (\alpha_{\max} - \alpha)w_{ij}^L - (1 + \alpha)w_{ij}^G. \quad (3)$$

In the range $-1 < \alpha < \alpha_{\max}$, decomposed weights w_{ij}^E , w_{ij}^L and w_{ij}^G (always non-negative values) represent excitatory connections, pattern-specific local inhibition and non-selective global inhibition, respectively (see Methods for the detailed description). Therefore, LAM can be regarded as a circuit with local and global inhibition, in which the parameter α determines the ratio between strengths of two types of inhibitory circuits (Figure 1, right). Biologically, the difference of α may correspond to anatomical inhomogeneity of interneurons. Otherwise, the balance of inhibition can be changed through inhomogeneous modulation of interneurons by acetylcholine^{30,31}.

Multiscale representation of community structures in LAM

To demonstrate representations in LAM, we tested three representative graph structures. The first one is the graph used previously to study how humans segment temporal sequences obeying probabilistic state-transition rules^{6,7} (Figure 2a). The second is karate club network³⁴ which is a popular dataset for testing community detection methods in graph theory (Figure 2f). The third is a graph representing the structure of compartmentalized rooms (Figure 2k) which is often used as the state-transition graph for reinforcement learning^{5,8-10}. For each graph, we assigned a random binary pattern to each node and constructed a LAM network by setting hetero-associative weights $H_{\mu\nu}$ from adjacency matrices of the graph. We initialized the activity pattern of LAM with one of memory patterns (trigger stimulus corresponding to each node), and simulated dynamics of the network for sufficiently long time to converge to an attractor state. We regard that attractor pattern as a neural representation of the node.

For each attractor state, we calculated an index called as pattern overlaps to evaluate the degree of retrieval of each memory pattern:

$$m^\mu = \frac{1}{NV} \sum_{i=1}^N \tilde{\xi}_i^\mu x_i. \quad (4)$$

Large positive values mean the significant activation of the memory pattern. Furthermore, we calculated pattern correlation between attractor patterns obtained from different trigger nodes.

LAM converges to various attractor patterns depending on trigger nodes and the value of auto-associative weight α . Generally, memory recall (large positive maximum pattern overlap) was observed in the parameter region $\alpha > -1$ (Figure 2d,i,n). When α takes large positive values, attractor patterns locally represent one or a few nodes in the graph (Figure 2b,g,right) and patterns for different nodes are uncorrelated to each other (Figure 2c,h,right). These attractors correspond to those obtained by conventional Hopfield-type models. In contrast, when α is negative ($-1 < \alpha < 0$), multiple memory patterns are active in attractor states. Quantitatively, the average number of active patterns is maximum at $\alpha \approx -1$ and decreases as α increases (Figure 2e,j,o; see Methods for the definition of active patterns). Especially in $\alpha \approx -1$, distributions of pattern overlaps represent large communities in graphs (Figure 2b,g,left), and accordingly, pattern correlation between attractor patterns are high within each community (Figure 2c,h,left). When α took an intermediate value, LAM exhibited a representation for a mesoscale community in the four-room graph (Figure 2l, right). This result demonstrates that LAM generates mixed representations for communities in the hetero-associative links by partially recalling multiple memory patterns simultaneously in attractor states. Accordingly, representations for nodes within a community are highly correlated, which agrees with experiments^{6,7}.

Theoretical relationship between LAM and graph Laplacian

Next, we analyze the mathematical mechanism behind representations of LAM. We found that the representations are related to graph Laplacian (GL). GL is a matrix defined for a graphical structure and its eigenvectors are used for various applications. One popular application is graph segmentation (also called as community detection) because it has been shown that signs of elements in GL eigenvectors indicates optimal two-fold segmentation of a graph^{27,28} (examples are shown in Figure 3a-c). GL eigenvectors give segmentation in various levels depending on their eigenvalues (small eigenvalue corresponds to the coarse resolution with large communities), thus combinations of multiple eigenvectors enables multi-level segmentation. This property is utilized for image segmentation^{27,28} for example. In another aspect, GL eigenvectors is also used for nonlinear dimensionality reduction²⁹, which gives low-dimensional representations for nodes (data points) in which the structure is represented

through similarity. As for the connection to neural representations, GL eigenvectors become grid-like code in the homogenous space, and their distortion brought by inhomogeneity fits with experimental observation of grid cells, and predictive spatial representations in hippocampus can be eigendecomposed into GL eigenvectors¹⁰. See Methods for the definition of GL and a brief review of mathematical properties.

We performed a formal theoretical analysis to show the relationship between LAM and GL. With symmetric normalization of hetero-associative weights (see Methods for the detail) which gives quantitatively same results as results shown above (Supplementary Figure 1), we can define an energy function of the model as

$$E = -\frac{1}{NV} \sum_{i=1}^N \sum_{j=1}^N w_{ij} x_i x_j. \quad (5)$$

As in the conventional Hopfield model¹¹, dynamics of LAM monotonically decreases this energy (Supplementary Figure 2). We consider a vector of pattern overlaps $\mathbf{m} = (m^1, \dots, m^P)^T$ and a vector rescaled by degrees of the graph $\tilde{\mathbf{m}}$, then we can rewrite the energy function as

$$E = \tilde{\mathbf{m}}^T \mathbf{L} \tilde{\mathbf{m}} + (\alpha + 1) [-\tilde{\mathbf{m}}^T \mathbf{D} \tilde{\mathbf{m}} + \gamma V^{-1} (m^0)^2], \quad (6)$$

where $\mathbf{L}$ is GL for the hetero-associative link structure (the state-transition graph) and $\mathbf{D}$ is a degree matrix, m^0 is the mean activity level in the network. Here we find the minimization of $\tilde{\mathbf{m}}^T \mathbf{L} \tilde{\mathbf{m}}$ under the constraint of $\tilde{\mathbf{m}}^T \mathbf{D} \tilde{\mathbf{m}}$, which is the same objective with graph segmentation²⁷ and graph-based dimensionality reduction²⁹, for which GL eigenvectors give optimal solutions. Therefore, we can expect GL eigenvectors appear in the rescaled pattern overlap vector $\tilde{\mathbf{m}}$ through the energy minimization of LAM. Furthermore, we derived that a GL eigenvector with an eigenvalue λ_k is activated in the pattern overlap vector in the condition $\lambda_k < \alpha + 1$ (see Methods). Noting that the minimum eigenvalue of GL is always zero and smaller eigenvalue corresponds to coarser graph segmentation, this result indicates that representations of the largest community (the eigenvector with the second smallest eigenvalue, which is called as Fiedler vector) appears in LAM when α is slightly higher than -1. As α increases, eigenvectors with higher eigenvalues are also activated, thus it is expected that represented communities get smaller. This analysis fits with the results shown in the previous section, especially the similarity between pattern overlaps in $\alpha \approx 1$ and Fiedler vectors (Figure 2b,g,l and Figure 3a-c). Although this analysis of the energy function depends on symmetricity of synaptic weights, we also derived same activation thresholds of GL eigenvectors through Turing instability analysis for complex networks³⁵ which does not require such constraint on connectivity. Therefore, we can expect similar transient dynamical properties also for LAM with asymmetric connections although the existence of attractors is

not guaranteed.

Alternatively, we can also interpret the energy minimization above as the combination of two conflicting optimization. First, the minimization of $\tilde{\mathbf{m}}^T \mathbf{L} \tilde{\mathbf{m}}$ is equivalent to the minimization of differences between pattern overlaps m^μ for two strongly connected cell assemblies²⁹. This results in smoothing (or diffusion) on the graph, which leads to non-sparse solutions, observed as mixed representations of multiple cell assemblies. Second, the term $-\tilde{\mathbf{m}}^T \mathbf{D} \tilde{\mathbf{m}}$ is same as the conventional Hopfield model, which basically leads to the activation of a single memory pattern. Minimization of the mean activity m^0 also helps to create sparse activity patterns. Therefore, the latter part of the energy function acts for the sparsification, that is, reduction of the number of activated memory patterns. In sum, the energy function is composed of two components for smoothness and sparsity, respectively, and the value $\alpha + 1$ determines the trade-off. If $\alpha < -1$, the effect of sparsification vanishes, thus no pattern is preferentially active. The number of activate patterns is maximized when α is slightly higher than -1 because of the strong smoothing effect. As α increases in the region $\alpha > -1$, the number of activated patterns gradually decreases and the model gets close to the conventional Hopfield model. This intuitive interpretation also fits with the actual behavior of the model.

To quantitatively validate the relationship between representations in LAM and GL eigenvectors, we performed linear regression of pattern overlaps of LAM (shown in Figure 2) by various numbers of GL eigenvectors, and calculated variance explained. Eigenvectors were chosen from those with small eigenvalues to those with large eigenvalues. The result shows that pattern overlaps in LAM with low auto-associative weights α were mostly explained by small numbers of GL eigenvectors with small eigenvalues, and eigenvectors with large eigenvalues were gradually recruited as α increased (Figure 3d-f). This result fits with our theoretical analyses and demonstrates the mathematical mechanism behind representations in LAM.

Based on the relationship with GL, we tested graph-based image segmentation by LAM, which is one of the well-established applications of GL²⁷. We assigned a random binary pattern to each pixel and defined hetero-associative links between pixels based on spatial proximity and similarity of RGB values similarly to the previous study²⁷. LAM successfully extracted large segments corresponding to a GL eigenvector when the auto-associative weight α was close to -1 (Supplementary Figure 3). When α was increased, LAM extracted relatively small segments (Supplementary Figure 3). These results show that LAM is

also applicable to non-ideal graphs constructed from real-world data.

Finding subgoals by graph-based representations and novelty detection

One of important applications of GL eigenvectors is to find appropriate subgoals for hierarchical reinforcement learning^{8,9}. In this framework, a set of primitive actions (called as options) are optimized through learning to reach subgoals. Desirable subgoals are “bottlenecks” shared by many trajectories on the state-transition graph. GL eigenvectors have been used to find bottlenecks through graph segmentation. We tested whether representations in LAM can also be used for subgoal finding by comparing results between LAM and GL.

To identify bottlenecks, we calculated “novelty index” of each node, which measures the expected changes of representations caused by the movements from a node to surrounding nodes (see Methods for the mathematical definition). In hierarchical reinforcement learning, subgoals are treated as pseudo-reward for agents. It is biologically natural to treat novelty as pseudo-reward because dopamine cells are activated by not only reward but also novelty³⁶. With GL, we constructed low-dimensional representations of nodes from GL eigenvectors with low eigenvalues (Laplacian eigenmap)²⁹. On the other hand, with LAM, activity patterns in attractor states are directly used as representations.

With GL, the novelty index successfully detected nodes located at the bottlenecks (doorways) between distinct compartments (Figure 4a). Furthermore, the sensitivity of bottleneck detection was controlled by the dimension of representation vectors: the use of a single eigenvector with low eigenvalue (Fiedler vector) extracted the narrowest bottlenecks, and higher dimensional representations enabled the detection of other bottlenecks. We obtained equally good performance by LAM (Figure 4b). In LAM, auto-associative weight α regulates the number of active GL eigenvectors in representations, hence it changes the sensitivity of bottleneck detection. This result demonstrates that novelty detection in LAM enables multi-resolution subgoal finding comparable to GL eigenvectors. The idea of using novelty has been suggested in the literature of hierarchical reinforcement learning³⁷ but LAM provides a more biologically plausible mechanism based on a neural network.

Chunked sequential activities in asymmetric LAM

So far, we have analyzed attractor patterns in LAM with symmetric links. Next, we show dynamic properties of asymmetric LAM. We constructed asymmetric LAM with a ring-shape graph in which link weights were slightly stronger in one direction than in the opposite

direction (Figure 5a). We simulated the neural activity while continuously changing the value of the auto-associative weight α (Figure 5b, top). The network showed a sequential activity in which embedded memory patterns are consecutively retrieved at a variable speed (Figure 5b, bottom). Rapid state transitions selectively occurred when the value of α became negative and close to -1 (Figure 5d), at which the distribution of pattern overlaps was maximally expanded (Figure 5c). These results indicate that negative auto-associative weights in asymmetric LAM not only generate macroscopic representations for large communities but also increases the sensitivity to asymmetry in synaptic weights and facilitates sequential transitions across memories.

Motivated by this dynamic property and the relationship between LAM and GL, we examined whether the sequential activities in asymmetric LAM are chunked according to the communities in hetero-associative links. For simulations, we specifically focused on hippocampal theta sequences in which chunking was experimentally observed³². We assumed a virtual animal running on a ring-shape track, and we modeled the hippocampus of the animal by asymmetric LAM with a ring-shape hetero-associative link structure (Figure 6a). We simulated neural activities of LAM with the fixed parameter $\alpha = -0.9$ while regularly stimulating a cell assembly encoding the current location of the animal. In the simulation, we obtained rhythmic sequential activities along the ring (Figure 6b), which are a simplified model of theta sequences.

In this model, we tested three hetero-associative link structures (see Supplementary Figure 4 for the detail of structures). In a uniform ring structure without chunks (Figure 6c) sequential activities propagated homogenously (Figure 6f, i). When the structure has local bottlenecks (Figure 6d), sequential propagation was constrained at the bottlenecks (Figure 6g,j). This result is analogous to that of symmetric LAM with the four-room graph (Figure 2k). Finally, when the structure is chunked by local over-representations (Figure 6e) which were implemented as densely connected nodes, sequential propagation was also chunked at over-represented locations and representations strongly correlated within chunks (Figure 6h,k). Although both the bottleneck model and the over-representation model exhibited chunking effects, the over-representation model is particularly consistent with the experimental observation because theta sequences are segmented at salient landmarks and rewards³² which are over-represented by hippocampal place cells^{38,39}. These results demonstrates that LAM provides a unified mechanism for graph-based representations^{6,7} and chunking of sequential activities³².

Discussion

In this paper, we proposed Laplacian associative memory (LAM), an extension of the Hopfield-type network models to compute community structures in hetero-associative links. While structural segmentation has been attempted by hierarchical networks with different time constants⁴⁰, our model provides a novel framework for multiscale information processing in a single network and accounts for experimentally observed graph-based representations^{6,7}. Furthermore, we showed that LAM with asymmetry can generate chunked sequential activities which reproduced experimentally observed chunking of theta sequences³². Notably, a model parameter crucial for segmentation can be interpreted as the strength ratio of local (assembly-specific) and global (non-specific) inhibitory feedback. This interpretation offers a novel insight into the computational roles of inhomogeneous neuromodulations of interneuron circuits, which is a focus of recent experimental studies^{30,31}.

We interpreted the model parameter α that controls the scale of the representation as the balance between global and local inhibition. In the light of recent experimental evidence, we further speculate the specific circuit mechanisms to regulate the parameter α . In visual cortex, somatostatin-expressing (SOM) interneurons and parvalbumin-expressing (PV) interneurons are considered to serve for anatomically global and local inhibitory feedback, respectively⁴¹. Furthermore, the cholinergic activation of vasoactive intestinal peptide-expressing (VIP) interneurons selectively inhibits SOM interneurons, which changes the balance between SOM and PV interneurons^{30,31}. Then, it seems reasonable to hypothesize that acetylcholine elicits a local-inhibition-dominant (PV-dominant) state in cortical circuits and hence generates the macroscopic representations corresponding to large communities. This hypothesis is consistent with the computational model of cortical inference in which acetylcholine signals uncertainty⁴² because acetylcholine creates mixed and ambiguous representations over many states. However, our model offers a novel prediction that uncertainty of the inference (expansion of probabilistic distributions) is constrained by communities in the graph structure behind experience. Our hypothesis also predicts that the malfunctioning of PV interneurons results in the deficit of processing uncertainty and macroscopic information, and difficulty of transition between states. This may be consistent with the symptom and the likely cause of schizophrenia^{43–45}. These possibilities should be pursued by more specific and detailed modeling and fitting to experimental data.

We used graph-based representations in LAM for subgoal finding in HRL^{8,9}. Another way to perform reinforcement learning with graph-based representations is to use successor

representation⁴⁶. Successor representation gives prediction of near-future state occupancy from current states that is useful for value estimation, and has been shown to be consistent with many experimental findings in hippocampal information representations¹⁰ including representational similarity within community structures shown in this study. Our model shares similar properties with successor representation as its eigendecomposition yields GL eigenvectors¹⁰ and modifying timescale of representation results in changing eigenvalues. Therefore, we expect some theoretical connection between successor representation and our model. However, we note that our model directly exhibits the Fiedler vector in $\alpha \approx -1$, but successor representation cannot give such representation in any parameter setting (see Methods). Therefore, in the current form, there is a little quantitative difference between our model and successor representation. Which model better explains experimental observations and whether our model (with some modification) exhibits successor representation are interesting open problems.

In asymmetric LAM, we found that negative auto-associative weights facilitate sequential transitions across memory patterns. Previously, we found that negative auto-association significantly increases the sensitivity of correlated attractors to external perturbation²³. We speculate that changes in the propagation speed presented here depends on a similar mechanism. If the auto-association is strongly positive, attractors are stable and invulnerable to directional biases in link weights. However, as α gets close to -1, the attractors are gradually instabilized and become sensitive to weight biases and external perturbations. This property suggests that macroscopic representations are dynamic in the brain, and are unlikely to serve for robust working memory in the brain as conventional attractor networks^{20,24,25}.

We found that both local bottlenecks and over-representations induce the chunking of sequential activities in asymmetric LAM. The over-representation model is particularly interesting because it accounts for the role of salient landmarks and rewards which are over-represented by place cells^{38,39}. We may be able to apply a ring-shape structure with two over-representations for modeling typical experiments in which animals run back and forth on a 1-D track to get rewards at both ends, considering that many place cells are direction-selective in such experimental setting⁴⁷. In contrast, to our knowledge, how bottlenecks affect hippocampal sequential activities has not been tested experimentally. An adequate design of bottlenecks does not seem to be trivial in spatial navigation tasks because animals may recognize spatial bottlenecks as salient landmarks which would be over-represented in the brain. A proper design of the task structure requires a careful control of saliency of each state.

The simple model with the asymmetric LAM produced sequential activities similar to chunked hippocampal theta sequences³² (Figure 6). However, the hippocampal circuits generate more complex oscillatory dynamics, which is also likely to contribute to segmentation. For instance, in hippocampal replays of spatial trajectories, a boundary of chunks (a bifurcating point) in the spatial structure is locked to troughs of LFP power in concatenated sharp wave ripples³³. Furthermore, hippocampal circuits repeat convergence to and divergence from discrete attractors every gamma cycle during sharp wave ripples¹⁶. Our simplified model cannot address the relationship between such complex oscillatory dynamics and segmentation. A detailed network model involving realistic spiking neurons and inhibitory circuits is necessary for studying such the relationship.

Previously, processing hierarchical knowledge in associative memory models was implemented by embedding artificially correlated memory patterns⁴⁸. This model successfully reproduced the dynamics of hierarchical information processing in the temporal visual cortex^{49,50}. The relationship between our model with hetero-associative links and the previous model with correlated memory patterns is currently unclear, and worth exploring. If similar graphical computation is possible with correlated memory patterns, the brain may perform graphical computation based on not only temporal association (hetero-associative links in our model) but semantic similarity between items (correlation between memory patterns). However, we emphasize that our finding that associative memory networks can autonomously compute mathematically well-defined communities in complex graphs is previously not known, because previous models have tested only simple structures and negative auto-association has not been considered.

The mechanism proposed in this paper will give a novel method to solve an arbitrary eigenvalue problem by using associative memory models. In the present model, we constructed hetero-associative weights from normalized adjacency matrices of graphs. However, the proposed dynamical mechanism to solve eigenvalue problem is generic and does not depend on this specific condition. For example, if we employ covariance matrix between encoded variables for a hetero-associative weight matrix, a network is theoretically expected to perform principal component analysis. Because eigenvalue problems ubiquitously appear in applied mathematics and machine learning, other computational methods may also be mapped to brain functions through the similar mechanism. Our model suggests much more powerful computing ability of associative memory models than previously thought and may provide a bridge across artificial intelligence and brain science.

Acknowledgements

We are grateful to H. Shiwaku for helpful discussion. We also thank OIST Scientific Computation and Data Analysis section for the technical support for scientific computing. This work was partially supported by Kakenhi nos. 18H05213 and 19H04994 from the MEXT (Ministry of Education, Culture, Sports, Science and Technology, Japan).

Author contributions

T. H. and T. F. conceived the project and wrote the manuscript. T. H. mathematically designed the model, and performed simulations and analyses.

Competing financial interests

The authors declare no competing interests.

Methods

Definition and mathematical properties of graph Laplacian

Let us assume a symmetric graph which has an adjacency matrix $\mathbf{A}$ whose element A_{ij} denotes the existence of an edge with 0 and 1 (unweighted graphs) or the weight of the edge (weighted graphs) between the node i and the node j . We also define a degree matrix $\mathbf{D}$, in which diagonal elements are degrees (the number of edges connected to each node) $d_i = \sum_j A_{ij}$ and other elements are zero. Graph Laplacian is a matrix defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. There are two ways of normalization: a symmetric one $\mathbf{L}^{\text{sym}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$ and an asymmetric one $\mathbf{L}^{\text{asym}} = \mathbf{D}^{-1} \mathbf{L} = \mathbf{I} - \mathbf{D}^{-1} \mathbf{A}$ ($\mathbf{I}$ is an identity matrix). These two matrices have qualitatively same properties²⁸.

An important characteristic of graph Laplacian matrix is that its eigenvectors give optimal graph segmentation. Here optimality is defined by the min-cut criterion that prefers a two-fold division of a graph obtained by cutting the minimum number of edges. It has been proven that min-cut graph segmentation can be performed by solving the generalized eigenvalue problem $\mathbf{L}\mathbf{y} = \lambda\mathbf{D}\mathbf{y}$, or equivalently, eigenvectors of normalized graph Laplacian $\mathbf{L}^{\text{sym}}$ and $\mathbf{L}^{\text{asym}}$ ^{27,28}. A sign of each element in the eigenvector $\mathbf{y}$ indicate a segment that each node should be assigned, and multiple eigenvectors correspond to the two-fold segmentation in various levels, depending on their eigenvalues. The eigenvector with the second smallest eigenvalue (Fiedler vector) is regarded as the best non-trivial solution which corresponds to the largest community structure (which achieves minimum cut) in the graph. Eigenvectors with larger eigenvalues are suboptimal solutions perpendicular to other eigenvectors, and tend to give subdivision of large communities into subclusters.

Another useful interpretation of graph Laplacian eigenvectors is low-dimensional representations for nodes in the graph which is called as Laplacian eigenmap²⁹. The generalized eigenvalue problem $\mathbf{Ly} = \lambda \mathbf{Dy}$ gives perpendicular solutions for $\min_{\mathbf{y}} \mathbf{y}^T \mathbf{Ly}$ subject to $\mathbf{y}^T \mathbf{Dy} = 1$ and the eigenvalue indicates the minimized value. Because of the relationship

$$\mathbf{y}^T \mathbf{Ly} = \frac{1}{2} \sum_{i,j} A_{ij} (y_i - y_j)^2, \quad (7)$$

minimization of $\mathbf{y}^T \mathbf{Ly}$ can be regarded as assigning values y_i to nodes such that strongly connected nodes are represented by close values. In this sense, low-dimensional representations constructed from graph Laplacian eigenvectors with low eigenvalues captures the graph structure through their similarity, which is the appropriate property for nonlinear dimensionality reduction.

Construction of hetero-associative weights

In this study, we hypothesize that hetero-associative weight matrix $\mathbf{H} = (H_{\mu\nu})_{1 \leq \mu \leq P, 1 \leq \nu \leq P}$ is constructed as $\mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$ (symmetric normalization) or $\mathbf{D}^{-1} \mathbf{A}$ (asymmetric normalization) where $\mathbf{A}$ and $\mathbf{D}$ are the adjacency matrix and the degree matrix of the graph, respectively. As in graph Laplacian, two normalization yields qualitatively same results. However, the symmetric normalization model enables us formal theoretical analyses and the asymmetric normalization model gives biologically plausible interpretation of the model.

Asymmetrically normalized weights directly correspond to the transition probability matrix for random walk on the graph²⁸, thus they can be naturally learned in sequential experiences through Hebbian learning at the state transition^{22,25}. With some additional assumptions, we can specify the learning rule for weights of asymmetric normalization model. We assume that an agent experiences a sequence of discrete states $s[t]$, and the neural activity $r_i[t]$ exhibits a binary pattern $\xi_i^{s[t]}$ corresponding to the current state. We consider an activity-dependent learning rule:

$$w_{ij}[t+1] = w_{ij}[t] + \eta (ar_i[t]r_j[t] + br_i[t]r_j[t-1] - r_i[t]w_{ij}[t]). \quad (8)$$

In a stable state, $E[w_{ij}[t+1] - w_{ij}[t]] = 0$ and $E[w_{ij}[t]]$ can be regarded as a constant w_{ij} . Then,

$$w_{ij} = \frac{aE[r_i[t]r_j[t]] + bE[r_i[t]r_j[t-1]]}{E[r_i[t]]}. \quad (9)$$

Using $r_i[t] = \xi_i^{s[t]}$,

$$w_{ij} = \frac{a \sum_{\mu} P(s[t] = \mu) \xi_i^{\mu} \xi_j^{\mu} + b \sum_{\mu, \nu} P(s[t] = \mu, s[t-1] = \nu) \xi_i^{\mu} \xi_j^{\nu}}{\sum_{\mu} P(s[t] = \mu) \xi_i^{\mu}}. \quad (10)$$

If patterns ξ_i^{μ} are sparse and overlaps between patterns are negligible (each neuron participates a single pattern at most),

$$w_{ij} = a\delta_{\mu\nu} + bP(s[t-1] = \nu | s[t] = \mu) \text{ for } i, j \text{ subject to } \xi_i^{\mu} = 1, \xi_j^{\nu} = 1,$$

and $w_{ij} = 0$ if neuron i or j does not belong to any pattern (assuming initial weights are zero). Finally, with $a = \frac{\alpha_{\max}}{NV}$, $b = \frac{1}{NV}$ and $P(s[t-1] = \nu | s[t] = \mu) = H_{\mu\nu}$, we obtain

$$w_{ij} = \frac{1}{NV} \sum_{\mu, \nu} (\alpha_{\max} \delta_{\mu\nu} + H_{\mu\nu}) \xi_i^{\mu} \xi_j^{\nu}. \quad (11)$$

In a random walk on a state-transition graph, $H_{\mu\nu} = P(s[t-1] = \nu | s[t] = \mu)$ corresponds to an element in the normalized adjacency matrix $\mathbf{D}^{-1}\mathbf{A}$.

Decomposition of excitatory and inhibitory synaptic weights

With the asymmetric normalization model, synaptic weights can be decomposed as

$$w_{ij} = w_{ij}^E - (\alpha_{\max} - \alpha)w_{ij}^L - (1 + \alpha)w_{ij}^G, \quad (12)$$

$$w_{ij}^E = \frac{1}{NV} \sum_{\mu=1}^P \sum_{\nu=1}^P (\alpha_{\max} \delta_{\mu\nu} + H_{\mu\nu}) \xi_i^{\mu} \xi_j^{\nu}, \quad (13)$$

$$w_{ij}^L = \frac{1}{NV} \sum_{\mu=1}^P \xi_i^{\mu} \xi_j^{\mu} \quad (14)$$

$$w_{ij}^G = \frac{P}{NV} \left(\frac{1}{P} \sum_{\mu=1}^P \xi_i^{\mu} \right) \left(\frac{1}{P} \sum_{\nu=1}^P \xi_j^{\nu} \right) + \frac{1}{N} \gamma. \quad (15)$$

Here, we used the constraint $\sum_{\nu=1}^P H_{\mu\nu} = 1$ and approximated $\sum_{\nu=1}^P H_{\mu\nu} \approx 1$. All decomposed weights are always non-negative. The component w_{ij}^E , w_{ij}^L , and w_{ij}^G are excitatory connections (which reflects the structure of cell assemblies), assembly-specific local inhibition, and non-selective global inhibition, respectively.

Analysis of the energy function of LAM

Here we consider symmetric normalization model in which hetero-associative links are constructed as $\mathbf{H} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. As in the main text, the energy function is

$$E = -\frac{1}{NV} \sum_{i=1}^N \sum_{j=1}^N w_{ij} x_i x_j, \quad (16)$$

We define a pattern overlap

$$m^\mu = \frac{1}{NV} \sum_{i=1}^N \tilde{\xi}_i^\mu x_i, \quad (17)$$

a pattern overlap vector $\mathbf{m} = (m^1, \dots, m^P)^T$, and the mean activity $m^0 = N^{-1} \sum_{i=1}^N x_i$. Then, we can rewrite the energy function as

$$E = \mathbf{m}^T \mathbf{L}^{\text{sym}} \mathbf{m} + (\alpha + 1) [-\mathbf{m}^T \mathbf{m} + \gamma V^{-1} (m^0)^2]. \quad (18)$$

The matrix $\mathbf{L}^{\text{sym}} = \mathbf{I} - \mathbf{H}$ is symmetric normalized graph Laplacian²⁷ if we regard the hetero-associative weight matrix as the normalized adjacency matrix. By rescaling $\mathbf{m}$ by the degree matrix of the hetero-associative links $\mathbf{D}$ as $\tilde{\mathbf{m}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{m}$, we further obtain

$$E = \tilde{\mathbf{m}}^T \mathbf{L} \tilde{\mathbf{m}} + (\alpha + 1) [-\tilde{\mathbf{m}}^T \mathbf{D} \tilde{\mathbf{m}} + \gamma V^{-1} (m^0)^2], \quad (19)$$

where $\mathbf{L}$ is unnormalized graph Laplacian.

To see the relationship with graph Laplacian eigenvectors more quantitatively, we expand the overlap vector $\mathbf{m}$ by a linear combination of eigenvectors of symmetric normalized graph Laplacian $\boldsymbol{\phi}_k$ (with corresponding eigenvalues λ_k) as

$$\mathbf{m} = \sum_{k=1}^P c_k \boldsymbol{\phi}_k. \quad (20)$$

Then, the energy function can be written as

$$E = -\sum_{k=1}^P (c_k)^2 [\alpha + 1 - \lambda_k] + \gamma V^{-1} (\alpha + 1) (m^0)^2. \quad (21)$$

If $\gamma = 0$, the minimization of this energy requires $c_k \neq 0$ if $\lambda_k < \alpha + 1$, which gives the approximate threshold of the activation of an eigenvector $\boldsymbol{\phi}_k$ in the representation (note that the actual threshold can be shifted because of $\gamma > 0$).

Turing instability analysis of LAM

For the analysis, we first replace the step function $\Theta(x)$ in Eq. (1) by a differentiable monotonically increasing function $f(\epsilon x)$ that converges to $\Theta(x)$ in the limit of $\epsilon \rightarrow \infty$ (e.g. a logistic function $f(\epsilon x) = (1 + \exp(-\epsilon x))^{-1}$). As in the main text, we define pattern overlaps as

$$m^\mu = \frac{1}{NV} \sum_{i=1}^N \tilde{\xi}_i^\mu x_i. \quad (22)$$

From Eq. (1)(2), dynamics of overlaps can be obtained as

$$\dot{m}^\rho = -m^\rho + \frac{1}{NV} \sum_{i=1}^N \xi_i^\rho f \left(\epsilon V^{-1} \sum_{\mu=1}^P \sum_{v=1}^P (\alpha \delta_{\mu v} + H_{\mu v}) \xi_i^\mu m^v \right) \quad (23)$$

Next, with vectors $\mathbf{m} = (m^1, \dots, m^P)^T$, $\tilde{\xi}_i = (\xi_i^1, \dots, \xi_i^P)^T$, and the hetero-associative weight matrix $\mathbf{H} = (H_{\mu v})_{1 \leq \mu \leq P, 1 \leq v \leq P}$, we obtain a vector representation as

$$\dot{\mathbf{m}} = -\mathbf{m} + \frac{1}{NV} \sum_{i=1}^N \tilde{\xi}_i f(\epsilon V^{-1} \tilde{\xi}_i^T (\alpha \mathbf{I} + \mathbf{H}) \mathbf{m}) \quad (24)$$

where $\mathbf{I}$ is an identity matrix. When N and P are sufficiently large and memory patterns are random, $\mathbf{m} = \mathbf{0}$ is an equilibrium point for this dynamical equation. Furthermore, in that condition, the matrix $\frac{1}{NV} \sum_{i=1}^N \tilde{\xi}_i \tilde{\xi}_i^T$ becomes a correlation matrix for random memory patterns, which can be approximated by an identity matrix. Thus, we obtain the following equation by linearizing $f(x)$ around $\mathbf{m} = \mathbf{0}$:

$$\dot{\mathbf{m}} = -\mathbf{m} + \epsilon V^{-1} f'(0) [(\alpha + 1) \mathbf{I} - \mathbf{L}] \mathbf{m} \quad (25)$$

Here, we defined $\mathbf{L} = \mathbf{I} - \mathbf{H}$ (this is either symmetric or asymmetric normalized graph Laplacian). Finally, we expand $\mathbf{m}$ with eigenvectors $\boldsymbol{\phi}_n$ ($n = 1, \dots, P$) of the matrix $\mathbf{L}$ as

$$\mathbf{m} = \sum_{n=1}^P d_n \exp(\beta_n t) \boldsymbol{\phi}_n \quad (26)$$

Substituting this into the linearized equation yields

$$\sum_{n=1}^P [(\alpha + 1 - \lambda_n) \epsilon V^{-1} f'(0) - 1 - \beta_n] d_n \exp(\beta_n t) \boldsymbol{\phi}_n = \mathbf{0} \quad (27)$$

where λ_n is an eigenvalue for $\boldsymbol{\phi}_n$. This equation has non-trivial solutions ($d_n \neq 0$) only if $\beta_n = (\alpha + 1 - \lambda_n) \epsilon V^{-1} f'(0) - 1$, which gives exponential growth rates along each eigenvector around $\mathbf{m} = \mathbf{0}$. If there exists a positive growth rate, the network becomes instable along the corresponding eigenvectors; otherwise the network is stabilized at $\mathbf{m} = \mathbf{0}$. In the limit of $\epsilon \rightarrow \infty$ ($f(\epsilon x) \rightarrow \Theta(x)$), the sign of β_n is solely determined by the sign of $\alpha + 1 - \lambda_n$. This result suggests that the overlap vector $\mathbf{m}$ is activated (instabilized) along the k -th eigenvector of the graph Laplacian matrix $\mathbf{L}$ if $\alpha > \lambda_k - 1$ (λ_k is the eigenvalue for the k -th eigenvector).

Simulations of the network model

In numerical simulations, we used decomposed asymmetric normalization model unless specified otherwise. We first initialized activities by one of memory patterns ($x_i[0] = \xi_i^\mu$) and updated activities by a discretized version of Eq. (1):

$$x_i[t + 1] = x_i[t] + \eta \left[-x_i[t] + \Theta \left(\sum_{j=1}^N w_{ij} x_j[t] + e_i[t] \right) \right], \quad (28)$$

where $e_i[t]$ is an external input applied in simulations in Figure 6. We used $\eta = 0.01$ for simulations with symmetric graphs (Figure 2, 3, 4 and Supplementary Figure 1, 2, 3) and $\eta = 0.1$ for simulations of sequential activities (Figure 5 and 6). The number of neurons were $N = 30000$ for image segmentation tasks, and $N = 10000$ for all other simulations. The additional inhibition parameter was $\gamma = 0.6$ for image segmentation and simulations of sequential activities, and $\gamma = 0.3$ for all other simulations. Sparsity p was 0.1 (approximately 10% of neurons are active in each pattern) throughout the study.

Attractor patterns of the network model with symmetric graphs were obtained by simulations of 3,000 time-steps. Simulations of sequential activities in Figure 5 was performed for 10,000 time-steps. Simulations of sequential activities in Figure 6 were performed for 30,000 time-steps and repeated three times using different random seeds for each setting. We averaged mean pattern overlaps at each location and correlations between mean pattern overlaps over those three trials. We truncated negative mean pattern overlaps to zero in this calculation.

We counted the number of active patterns by counting the number of μ that satisfies $m^\mu > 0.05$ and $m^\mu > \frac{1}{2} \max_{\mu} m^\mu$.

Settings for image segmentation

For image segmentation tasks, we took images from pixabay (<https://pixabay.com/>). We trimmed and down-sampled images so that it contains 1000 - 1500 pixels (e.g. $P \approx 1000$). We note that images shown in figures are ones before down-sampling. We constructed link weights by the same way with Shi & Marik (2000)²⁷:

$$A_{ij} = \begin{cases} \exp \left(-\frac{\|\mathbf{F}_i - \mathbf{F}_j\|_2^2}{\sigma_1} - \frac{\|\mathbf{X}_i - \mathbf{X}_j\|_2^2}{\sigma_X} \right), & \text{if } \|\mathbf{X}_i - \mathbf{X}_j\|_2 < r \\ 0, & \text{otherwise} \end{cases}, \quad (29)$$

where vectors $\mathbf{F}_i$ and $\mathbf{X}_i$ denotes the RGB value (normalized between 0 and 1) and the spatial location of pixel i , respectively. Parameters were $\sigma_1 = 0.1, \sigma_X = 4, r = 5$. After setting values, we performed asymmetric normalization of weights to get hetero-associative weights.

Definition of the novelty index for subgoal finding

First, we calculate a representation of each node by either constructing low-dimensional vectors from graph Laplacian eigenvectors (Laplacian eigenmap)²⁹ or simulating an attractor pattern of LAM triggered by a memory pattern of each node (ξ_i^μ). We define similarity between representations of two nodes as $s(\mu, \nu)$ by either cosine similarity between two representations of Laplacian eigenmap, or correlation between attractor patterns of LAM. The novelty index of a node μ is defined as

$$NI(\mu) = \frac{1}{2} \sum_{\nu} T_{\mu \rightarrow \nu} (1 - s(\mu, \nu)), \quad (30)$$

where $T_{\mu \rightarrow \nu}$ denotes the transition probability from μ to ν in random walk on the graph (which is equivalent with elements in $\mathbf{D}^{-1}\mathbf{A}$). The novelty index $NI(\mu)$ spans from 0 to 1 and indicates the expected change of information representations that an agent experience in the transition from the node μ .

Asymmetric Laplacian associative memory and the model of a virtual animal

To construct asymmetric hetero-associative weights, we converted symmetric graphs into mutually connected asymmetric weighted graphs. We set the weight of asymmetric links in the biased direction (including diagonal connections) to 110, and the weight for opposite directions to 90. All links horizontal to the biased direction (radial connections) was 100. After constructing adjacency matrices, we performed asymmetric normalization as in symmetric graphs.

In the simulation in Figure 6, we represented the current location of the virtual animal on the track by a continuous value $z[t]$ ranging from 0 to 90 which corresponds to 90 nodes in the uniform ring-shape graph (uniform model). The velocity $z[t+1] - z[t]$ was constant but we performed resampling of the velocity from a range [0.02, 0.04] at the timings determined by Poisson process (the mean interval was 1000 time-steps). We determined the index of stimulated pattern by truncating $z[t]$ to an integer. We applied stimulation to the network every 150 time-steps (the uniform model and the over-representation model) or 200 time-steps (the bottleneck model). The amplitude and the length of stimulation was 0.3 and 50 time-steps, respectively.

In the bottleneck model and the over-representation model, we connected additional nodes at the side of the uniform ring-shape graph (as shown in Supplementary Figure 4). We did not stimulate patterns corresponding to additional nodes. We calculated the pattern overlap for each location by averaging over nodes in the central ring and additional nodes at the same location.

Eigenvalues of successor representation

Successor representation is defined for a pair of states s and s' as

$$M(s, s') = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}(s_t = s') | s_0 = s \right], \quad (31)$$

where γ indicates a discount factor. We consider a matrix of successor representation $\mathbf{M}$, a transition probability matrix $\mathbf{T}$, and $\mathbf{L}^{\text{asym}} = \mathbf{I} - \mathbf{T}$ is an asymmetric normalized graph Laplacian. Eigenvectors and eigenvalues of $\mathbf{L}^{\text{asym}}$ are defined as $\boldsymbol{\phi}_i$ and λ_i^{L} , respectively. Then, they satisfy

$$\mathbf{L}^{\text{asym}} \boldsymbol{\phi}_i = \lambda_i^{\text{L}} \boldsymbol{\phi}_i. \quad (32)$$

Using the relationship $\mathbf{M} = (\mathbf{I} - \gamma \mathbf{T})^{-1} \mathbf{T}$, we can rewrite this equation with a successor representation matrix:

$$\mathbf{M} \boldsymbol{\phi}_i = \left(1 + \gamma(\lambda_i^{\text{L}} - 1) \right)^{-1} \boldsymbol{\phi}_i. \quad (33)$$

Therefore, eigenvectors of the successor representation matrix are equivalent to those of graph Laplacian, and eigenvalues are $\lambda_i^{\text{M}} = \left(1 + \gamma(\lambda_i^{\text{L}} - 1) \right)^{-1}$. This relationship becomes $\lambda_i^{\text{M}} = (\lambda_i^{\text{L}})^{-1}$ in the limit of $\gamma \rightarrow 1$, and $\lambda_i^{\text{M}} = 1$ in the limit of $\gamma \rightarrow 0$. Thus, although the contribution of Fiedler vector increases as γ goes to 1, it is impossible to have $\lambda_i^{\text{M}} > 0$ for only Fiedler vector if the graph has a sufficiently complex structure and λ_i^{L} is continuously distributed.

References

1. Thalmann, M., Souza, A. S. & Oberauer, K. How does chunking help working memory? *J. Exp. Psychol. Learn. Mem. Cogn.* **45**, 37–55 (2019).
2. Saffran, J. R., Aslin, R. N. & Newport, E. L. Statistical learning by 8-month-old infants. *Science (80-.)*. **274**, 1926–1928 (1996).
3. Perruchet, P. & Vinter, A. PARSER: A Model for Word Segmentation. *J. Mem. Lang.* **39**, 246–263 (1998).
4. Perruchet, P., Vinter, A., Pacteau, C. & Gallego, J. The formation of structurally relevant units in artificial grammar learning. *Q. J. Exp. Psychol. Sect. A Hum. Exp. Psychol.* **55**, 485–503 (2002).
5. Botvinick, M. M., Niv, Y. & Barto, A. C. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition* **113**, 262–280 (2009).
6. Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B. & Botvinick, M. M. Neural representations of events arise from temporal community structure. *Nat. Neurosci.* **16**, 486–492 (2013).
7. Schapiro, A. C., Turk-Browne, N. B., Norman, K. A. & Botvinick, M. M. Statistical learning of temporal community structure in the hippocampus. *Hippocampus* **26**, 3–8 (2016).
8. Şimşek, Ö., Wolfe, A. P. & Barto, A. G. Identifying useful subgoals in reinforcement learning by local graph partitioning. *ICML 2005 - Proc. 22nd Int. Conf. Mach. Learn.* 817–824 (2005). doi:10.1145/1102351.1102454
9. Machado, M. C., Bellemare, M. G. & Bowling, M. A laplacian framework for option discovery in reinforcement learning. *34th Int. Conf. Mach. Learn. ICML 2017* **5**, 3567–3582 (2017).
10. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
11. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* **79**, 2554–2558 (1982).
12. Tsodyks, M. & Feigel'man, M. V. The Enhanced Storage Capacity in Neural Networks with Low Activity Level. *Europhys. Lett.* **6**, 101 (1988).
13. Buhmann, J., Divko, R. & Schulten, K. Associative memory with high information content. *Phys. Rev. A* **39**, 2689–2692 (1989).
14. Bi, G. Q. & Poo, M. M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18**, 10464–10472 (1998).

15. Mishra, R. K., Kim, S., Guzman, S. J. & Jonas, P. Symmetric spike timing-dependent plasticity at CA3-CA3 synapses optimizes storage and recall in autoassociative networks. *Nat. Commun.* **7**, 1–11 (2016).
16. Pfeiffer, B. E. & Foster, D. J. Autoassociative dynamics in the generation of sequences of hippocampal place cells. *Science (80-.)*. **349**, 180–183 (2015).
17. Carrillo-Reid, L., Yang, W., Bando, Y., Peterka, D. S. & Yuste, R. Imprinting Cortical Ensembles. *Science (80-.)*. **353**, 691–694 (2016).
18. Carrillo-Reid, L., Han, S., Yang, W., Akrouh, A. & Yuste, R. Controlling Visually Guided Behavior by Holographic Recalling of Cortical Ensembles. *Cell* **178**, 447–457.e5 (2019).
19. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).
20. Griniasty, M., Tsodyks, M. V. & Amit, D. J. Conversion of Temporal Correlations Between Stimuli To Spatial Correlations Between Attractors. *Neural Comput.* **5**, 1–17 (1993).
21. Amit, D. J., Brunel, N. & Tsodyks, M. V. Correlations of cortical Hebbian reverberations: theory versus experiment. *J. Neurosci.* **14**, 6435–6445 (1994).
22. Brunel, N. Hebbian learning of context in recurrent neural networks. *Neural Comput* **8**, 1677–1710 (1996).
23. Haga, T. & Fukai, T. Extended Temporal Association Memory by Modulations of Inhibitory Circuits. *Phys. Rev. Lett.* **123**, 78101 (2019).
24. Miyashita, Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* **335**, 817–820 (1988).
25. Yakovlev, V., Fusi, S., Berman, E. & Zohary, E. Inter-trial neuronal activity in inferior temporal cortex: a putative vehicle to generate long-term visual associations. *Nat. Neurosci.* **1**, 310–317 (1998).
26. Barron, H. C., Vogels, T. P., Behrens, T. E. & Ramaswami, M. Inhibitory engrams in perception and memory. *Proc. Natl. Acad. Sci.* **114**, 201701812 (2017).
27. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
28. Meila, M. & Shi, J. A random walks view of spectral segmentation. *AI Stat.* **57**, 5287 (2001).
29. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373–1396 (2003).
30. Pi, H. J. *et al.* Cortical interneurons that specialize in disinhibitory control. *Nature* **503**, 521–524 (2013).

31. Fu, Y. *et al.* A cortical circuit for gain control by behavioral state. *Cell* **156**, 1139–1152 (2014).
32. Gupta, A. S., Van Der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. Segmentation of spatial experience by hippocampal theta sequences. *Nat. Neurosci.* **15**, 1032–1039 (2012).
33. Wu, X. & Foster, D. J. Hippocampal replay captures the unique topological structure of a novel environment. *J. Neurosci.* **34**, 6459–6469 (2014).
34. Zachary, W. W. An Information Flow Model for Conflict and Fission in Small Groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
35. Nakao, H. & Mikhailov, A. S. Turing patterns in network-organized activator-inhibitor systems. *Nat. Phys.* **6**, 544–550 (2010).
36. Kakade, S. & Dayan, P. Dopamine: Generalization and bonuses. *Neural Networks* **15**, 549–559 (2002).
37. Şimşek, Ö. & Barto, A. G. Using relative novelty to identify useful temporal abstractions in reinforcement learning. *Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004* 751–758 (2004). doi:10.1145/1015330.1015353
38. Hollup, S. A., Molden, S., Donnett, J. G., Moser, M. B. & Moser, E. I. Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *J. Neurosci.* **21**, 1635–1644 (2001).
39. Sato, M. *et al.* Distinct Mechanisms of Over-Representation of Landmarks and Rewards in the Hippocampus. *Cell Rep.* **32**, 107864 (2020).
40. Yamashita, Y. & Tani, J. Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLoS Comput. Biol.* **4**, (2008).
41. Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z. J. & Scanziani, M. A neural circuit for spatial summation in visual cortex. *Nature* **490**, 226–230 (2012).
42. Yu, A. J. & Dayan, P. Acetylcholine in cortical inference. *Neural Networks* **15**, 719–730 (2002).
43. Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science (80-.).* **334**, 1569–1573 (2011).
44. Belforte, J. E. *et al.* Postnatal NMDA receptor ablation in corticolimbic interneurons confers schizophrenia-like phenotypes. *Nat. Neurosci.* **13**, 76–83 (2010).
45. Gonzalez-Burgos, G., Hashimoto, T. & Lewis, D. A. Alterations of cortical GABA neurons and network oscillations in schizophrenia. *Curr. Psychiatry Rep.* **12**, 335–344 (2010).
46. Dayan, P. Improving Generalization for Temporal Difference Learning: The Successor

- Representation. *Neural Comput.* **5**, 613–624 (1993).
47. Buzsáki, G. Theta rhythm of navigation: Link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus* **15**, 827–840 (2005).
 48. Matsumoto, N., Okada, M., Sugase-Miyamoto, Y. & Yamane, S. Neuronal mechanisms encoding global-to-fine information in inferior-temporal cortex. *J. Comput. Neurosci.* **18**, 85–103 (2005).
 49. Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. Global and fine information coded by single neurons in the temporal visual cortex. *Nature* **400**, 869–873 (1999).
 50. Matsumoto, N., Okada, M., Sugase-Miyamoto, Y., Yamane, S. & Kawano, K. Population dynamics of face-responsive neurons in the inferior temporal cortex. *Cereb. Cortex* **15**, 1103–1112 (2005).

Figure 1 Schematics of Laplacian associative memory (LAM) model. Left: associative memory network model with auto-association and hetero-association. The parameter α indicates the auto-associative strength. Right: Equivalent biological neural network model which contains local (assembly-specific) and global (non-specific) inhibition. α indicates the ratio between local and global inhibition.

Figure 2 LAM generates multi-scale representations for community structures. (a) Graph used in Schapiro et al. (2013)⁶. (b) Pattern overlaps of example attractor patterns. (c) Correlation matrices between attractor patterns triggered by different initial patterns obtained. (d) Maximum pattern overlaps obtained by various values of α . (e) Numbers of active patterns obtained by various values of α . In d and e, values from all attractor patterns triggered by different initial patterns are averaged for each α . (f) Karate club network³⁴. (g-j) Results for Karate-club network. (k) A compartmentalized room structure used in the context of reinforcement learning^{5,9,10,37} (four-room graph). (l-o) Results for four-room graph.

Figure 3 The relationship between Graph Laplacian eigenvectors and LAM. (a-c) GL eigenvectors for example graphs. Fiedler vector is an eigenvector with a second smallest eigenvalue. (a) Fiedler vector for the graph in Schapiro et al. (2013). (b) Fiedler vector for karate-club network. (c) Fiedler vector and the eigenvector with the third smallest eigenvalue for the four-room graph. (d-f) The ratio of variance of pattern overlap vectors explained by various number of graph Laplacian eigenvectors. The color indicates the value of α . Values from all attractor pattern triggered by different initial patterns are averaged for each α . (d) Results from the graph in Schapiro et al. (2013). (e) Results for karate-club network. (f) Results for the four-room graph.

Figure 4 Subgoal finding by novelty detection with representations in LAM. (a) Subgoal finding using low-dimensional (1, 2, or 3) representations constructed from graph Laplacian eigenvectors. The color indicates the novelty index for each node. (b) Subgoal finding using representations in LAM obtained by different values of α .

Figure 5 Parameter-dependent sequential activities in asymmetric LAM. (a) A ring structure of hetero-associative links for the simulation of asymmetric LAM. Hetero-associative weights were biased towards one direction. (b) The time course of α (top) and pattern overlaps (bottom) in the simulation of asymmetric LAM. Negative pattern overlaps were truncated to zero. (c) Peak-aligned mean pattern overlap distributions in different ranges of α . (d) Mean speed of the peak shift of the pattern overlap distribution in different ranges of α .

Figure 6. Chunked sequential activities in asymmetric LAM. (a) Schematics for the simulation setting. We assumed a virtual animal running around a ring track, and regularly stimulated patterns in asymmetric LAM corresponding to the current location. (b) Pattern overlaps of simulated neural activities in the uniform model. The red dotted line indicates the actual location of the virtual animal. (c-e) Schematics of the hetero-associative link structure. A uniform ring (c), a ring chunked with local bottlenecks (d), and a ring chunked with local over-representations (e). Details are shown in supplementary figure 4. (f-h) Mean pattern overlaps (the strength of representations of locations) at each actual location of a virtual animal in the simulation of a uniform model (f), a bottleneck model (g), and over-representation model (h). We truncated negative pattern overlaps to zero in these figures. (i-k) Correlations between mean pattern overlaps at different locations in the simulation of a uniform model (i), a

bottleneck model (j), and over-representation model (k). White dotted lines in g, h, j, k indicate locations of chunk borders (bottlenecks or over-representations).